The Role of Context and Demographics in Emotional Online Interpretation

Anika Ahluwalia

May 12, 2023

Abstract

This paper investigates the ease and accuracy of the interpretation of emotional online communication through the use of emoji-only posts and accompanying demographic information. An anonymous online survey was used to collect data from participants, who were asked to interpret four emoji-only posts displayed with accompanying demographic information (either accurate or inaccurate). The study examines how different demographics of the poster and interpreter impact accuracy and confidence in interpretation, as well as how the inclusion of demographic information or the perceived similarity between the poster and the interpreter impacts interpretation. The findings show that gender itself does not impact accuracy or confidence, but differences in interpretation are seen between genders when interpreting each other, specifically that women were approximately 10% more skilled at interpreting men than men themselves were, and that men are about 21% easier to interpret than women overall. Age, student status, and familiarity with emojis impact confidence in interpretation, but not accuracy. Younger participants are approximately 20% more confident in their interpretations than older participants, but were not more accurate. Additionally, the inclusion of demographic information and perceived demographic similarity do not impact interpretation, however, actual demographic similarity between the poster and interpreter impacts confidence of interpretation. Overall, this study provides insights into how context and demographics play a role in the interpretation of emotional online communication.
1 Introduction

Communication can be tricky. The emergence of social media and online communication in recent decades have greatly changed the modes and means of communication and social interaction. However, as online communication has become more prevalent, so has the potential for misunderstandings to arise. A significant challenge in online communication is accurately interpreting the emotions of others and the true meaning of what they are communicating. When communicating face-to-face, nonverbal cues such as facial expressions or body language can help convey emotion or hidden meanings, however, in online communication, the message must be interpreted based on content alone.

Given the prevalence and reliance on online communication and social media, understanding how individuals interpret and express emotions online is becoming more important. The COVID-19 pandemic, in particular, has accelerated the use of online communication platforms, making them an even more critical tool for socializing, learning, and working. Understanding how people interpret emotions in online settings is essential for promoting effective communication.

Moreover, research has shown that different demographics may communicate differently online [Prada et al.(2018), Tossell et al.(2012), Rodrigues et al.(2018)]. For example, younger generations are generally more comfortable expressing their emotions through digital communication, while older generations rely more on traditional face-to-face communication [Drouin et al.(2018)]. Cultural differences may also play a role in emotional expression and interpretation online [Sadiq et al.(2019), Sugiyama(2015)]. Understanding these demographic differences in communication patterns can help us design better communication tools and develop more effective strategies for resolving conflicts and promoting positive social interactions online.

Emojis have become a mainstay in online communication, with research demonstrating that their use has increased dramatically over the past decade [Evans(2017)]. In fact, over 90% of worldwide smartphone users incorporate emojis in their text messages [McCulloch(2020)]. Furthermore, many online messages are either enhanced through the use of emojis or can be comprised of solely emojis. Generally, emojis can be used to communicate complex meanings and emotions that are difficult to express through text alone. They can also supplement messages by conveying emotions and cues that could be missed in text-only communication. With relation to the interpretation of emojis online, research has shown that individuals use emojis in different ways depending on their demographic characteristics, cultural background, and personality traits [Prada et al.(2018), Sadiq et al.(2019), Sugiyama(2015)]. Ultimately, emojis play an important role in digital communication, serving as a crucial means of expressing emotions and attitudes.

Overall, this paper aims to investigate the following research questions:

1. How do different demographics interpret emotional posts online?
2. Does actual demographic similarity between a poster and an interpreter impact interpretation accuracy or confidence?
3. Does knowing demographic information about the poster impact interpretation of the post (in either confidence or accuracy)?
4. Does perceived demographic similarity between a poster and an interpreter impact interpretation accuracy or confidence?

5. Does the positivity or emotion of a post impact ability to interpret?

2 Related Work

This section explores previous work on the usage of emojis in online communication, the role and limits of context, and how different demographic communicate emotionally and online.

2.1 Emoji Usage and Benefits

Studying emojis is crucial because they play an important role in digital communication. They allow individuals to express emotions within the text-only environment of online communication. Understanding how different people use and interpret emojis can help us better understand how online communication works, the role of emotions in online communication, and potential ways to improve it. Moreover, the addition of emojis in mobile text messages has reduced misinterpretation and added context to these messages [Derks et al. (2008)], and their usage leads to both user enjoyment and perceived usefulness [Willoughby and Liu (2018)].

One study, “Emoji as Emotion Tag for Tweets,” [Wood and Ruder (2016)] removed emojis from several randomly selected tweets and asked annotators to identify key emotions in the text. The emojis of the tweet were then compared with the identified emotions found in the text. Ultimately, the study concluded that emojis could be helpful as emotion labels or tags for models of emotion in text. In this study, emojis were also compared to other indicators of the context of the tweet, such as hashtags. When compared, the authors argued that “emoji are found to be used far more extensively than hasht tags and we argue that they present a more faithful representation of a user’s emotional state” [Wood and Ruder (2016)].

Emoji-only communication is also growing in prevalence. Many researchers are investigating the nuances of this form of communication and what factors may impact interpretability. In “Opico: A Study of Emoji-first Communication in a Mobile Social App,” [Khandekar et al. (2019)] researchers build the “Opico” application in which users could create emoji posts (of one to five emojis) that were tied to a location. These posts are shared on a social feed in which users are allowed to like and comment upon posts. This study included a follow up in which these emoji only posts were interpreted and translated by other users of the application and measured for accuracy. Ultimately, the study found that “emoji usage and interpretation varies across age, gender, geographic location, and social status,” reinforcing the hypotheses for this study [Khandekar et al. (2019)]. This study simply measured the accuracy of each interpretation, but did not explore which factors impacted accuracy.

It is important to understand the nuances behind emoji communication and interpretation not only because of the prevalence of emojis but also because of their potential benefits in aiding in online communication. As previously mentioned, emojis can use helpful as a means of supplementing textual communication to include better indications of emotions.
and nonverbal cues. In fact, a recent study in which participants were asked to use a mobile mental health app centered around emoji posting found that the use of emojis in the context of mobile health demonstrates great potential for youth seeking mental health treatment online because participants became more self aware and more motivated to change their behavioral patterns [Van Dam et al.(2019)]. Within our study, we examine the accuracy and confidence of different emotional posts online. We also examine demographic differences in interpretation and differences in the interpretation of various emotions which could have potential ramifications in the online mental health space.

2.2 Demographic Differences in Emoji Usage

Because emojis themselves are symbols and icons, their interpretation requires context and a shared understanding of the meanings behind different emojis. Often, these shared understandings come from context and surrounding culture, which can be impacted by both location and demographics [Sadiq et al.(2019), Sugiyama(2015), Butterworth et al.(2019)].

Many papers reinforce the shared understanding that emoji interpretation is impacted by culture. For example, previous studies have demonstrated that emoji usage is heavily impacted by social context [Derks et al.(2008), Park et al.(2013)]. Cultural background and language environment both also influence emoji use. In particular, some specific uses of emoji are closely related to cultural background, such as Finnish, Indian, and Pakistani users, who use specific emoticons according to their own culture [Sadiq et al.(2019)]. Additionally, another study explored how Japanese teenagers utilize emojis in new ways to express themselves in a subculturally specific way in which they prioritize the aesthetics of the emojis over their meaning when choosing to include them in a [Sugiyama(2015)]. Finally, another study found that nearly 60% of emojis have different meanings and emotional associations between China and the U.S. [Chui(2020)].

Additionally, language environment significantly influences the use and interpretation of emojis. Research suggests that emojis exhibit high context sensitivity in cross-language communication, making them highly dependent on their linguistic and textual environment [Vandergriff(2013)]. As a result, the meaning and connotations of emojis can vary across different cultures and languages, leading to misinterpretations and misunderstandings.

2.2.1 Gender and Emojis

We recognize that culture is not just limited to nationality. This study questions not just the impact of regionally specific culture on interpretation, but the impact of the interplay between demographic cultures on emoji interpretation. In particular, the impacts of age and gender on both emoji usage and interpretation will be examined. While some papers find that men and women understand the functionality of emoji similarly [Herring and Dainas(2018)], others find that the ways in which each gender uses emoticons are quite different. For example, females use emojis more often and positively [Prada et al.(2018)] while males use a wider variety of emoticons [Tossell et al.(2012)]. Research also demonstrates that the emotions evoked in the interpreter differ when different genders use the same emoji [Butterworth et al.(2019)]. The study found that women who use affectionate icons are viewed as more appropriate and attractive, while men are considered appropriate and attractive when they use less affectionate
but friendly emojis. These nuances in demographic usage of emojis can also impact their understanding and interpretability. For example, a study demonstrated that females perceive emoticons as more familiar, clear and meaningful than males did [Rodrigues et al.(2018)]. Our study continues exploration of this area while integrating ideas of perceived and actual differences of demographics and examining the impacts of display on interpretation.

2.3 Demographic Differences in Online Behavior & Interpretation

The variance in online behaviors exhibited by various groups can be attributed to the differences in their usage and comprehension of emoji icons. One of the most consistent findings is that young people tend to use social media platforms more frequently than older generations (e.g. [Drouin et al.(2018), Lehtinen et al.(2009), Weiß et al.(2020)]). Moreover, social media usage has been positively associated with higher levels of social capital, such as the ability to access and mobilize resources and social support networks [Lu and Hampton(2017)]. However, usage of social media has also been linked to higher rates of depression, anxiety, and loneliness in college students [Drouin et al.(2018)], impacting how they interact online and the positivity of those interactions. Those with high levels of depression and anxiety tend to have different social media usage patterns and preferences, and are more likely to share emotions online.

Another important demographic difference in online behavior is gender. Research has consistently found that women tend to value emotional online communication more than men [Procio and Procopio(2007)]. In particular, women are more likely to engage in expressive communication, such as sharing personal experiences and emotions, while men tend to engage in instrumental communication, such as sharing information and problem-solving. This gender difference in communication style has been linked to social and cultural norms around gender roles and expression.

Finally, educational differences have been found to play a large role in determining predictions of tie strength on social media platforms [Gilbert and Karahalios(2009)]. Gilbert et al. examined existing social networks and found that a variety of demographic factors are reliable and accurate predictors of tie strength, showing that online bonds and communication is generally stronger and more interpretable when those engaged in the conversation share demographic similarities. Interestingly, people with higher levels of education tend to have stronger and more diverse social networks online, while those with lower levels of education tend to have weaker and more homogeneous networks [Gilbert and Karahalios(2009)]. Other factors, such as age differences, political affiliation, and frequency of usage of social media, also impact tie strength and the nature of online interactions.

2.4 Demographic Knowledge and Online Communication

In this study, we also hope to explore how the addition and the accuracy of demographic knowledge could impact online communication and interpretation. One such factor that impacts online communication is anonymity. Research has shown that anonymity increases the aggression displayed in online posts [Nakano et al.(2016)]. However, people find it easier to speak about their emotions online when it is anonymous because anonymity provides a sense of safety and comfort in discussing sensitive topics [Drouin et al.(2018)]. Additionally, a study
about self disclosure in a computer-mediated environment revealed that participants were more likely to disclose personal information when they were visually anonymous compared to when they were not [Joinson(2001)]. This type of self disclosure in online communication is important to study because it often leads to long lasting bonds and is a crucial part of relationships in today’s digital era [McKenna et al.(2002)].

It is also interesting to see that the addition of demographic information to an online post can boost feelings of social support because generally people prefer to talk to others with the same problems as them, as they are more likely to understand their situation and provide support [Drouin et al.(2018), Wright et al.(2003)]. Additionally, some research explores the impacts of deception on online relationships, such as gender deception or other online identity manipulation, and find that legitimate online relationships are still able to be initially formed under these conditions [Parks and Floyd(1996)].

3 Methods

In order to better understand the impacts of context on emoji-based online communication, an online survey was distributed. The survey asked participants to interpret several emoji posts created on the Chime application.

Researchers determined that an online survey would be most effective in recruiting a wide variety of participants. The only constraints on participation were that survey takers were over the age of 18, residing in the U.S., and had not participated in the previous Chime study held in 2022. Generally, it was expected that participants would be vaguely familiar with emojis and their usage, hence, online advertising and distribution seemed appropriate. Emoji usage and familiarity was also asked in the survey.

In this short, 15 minute survey, participants were asked to briefly provide some demographic information and then were presented with four online emoji-only posts. Participants were asked to interpret each post and report their confidence on their interpretation.

3.1 Participants and Recruiting

Participants were recruited from a variety of online sources (notably Reddit, Twitter, and Mastodon), and locally through placing flyers and posters in the community. The platform from which participants found the study (Reddit, physical media, etc) was noted and stored.

Participant recruitment was done primarily through online sources in order to reach a demographically diverse population. In accordance, the survey was advertised in general survey taking subreddits in order to avoid limiting the participant pool to people specifically interested in emojis, including r/SurveyCircle, r/PaidStudies, etc. However, as active online users (especially on Reddit) are more likely to be tech savvy than the average person, the survey was also advertised using physical posters in order to get a diverse population in terms of online experience. Finally, researchers utilized Twitter and Mastodon for the final advertising methods.

For this study, all participants were required to be over the age of 18, reside in the United States, and not have participated in the previous Chime study held in 2022 as these same posts were visible to participants of that study. Research demonstrated that
emoji interpretation between different national cultures is very varied [Sadiq et al.(2019), Sugiyama(2015)], however this study focused on other demographic markers (i.e. gender, age, and student status). Thus participants were limited to those who currently reside in the United States to control for cultural variance.

A total of 108 valid responses from eligible participants were collected (55.5% male, 36.1% female, 8.3% non-binary). 54.6% of participants were students. The age of participants was also collected (Mean = 27.9, Median = 26). For the purposes of this survey, age was separated into either young adult (ranging from 18 to 25 years old) or adult (26 years or older). Between these two categories, 49.1% of participants were young adults, and 50.9% were adults.

### 3.2 Pilots and Study Creation

This study aimed to explore the various aspects of the impact of demographics on emoji understanding. Thus, researchers needed to limit for external factors and ensure that the survey was as easy to interpret as possible. In order to do so, a series of seven pilot sessions of the initial survey were conducted.

All pilot sessions were conducted virtually through a video call. Upon entering the call, participants were sent a link to the online survey. They were not given context or allowed to ask questions until they had completed the survey. As they were taking the survey, they were asked to share their screen and use the think aloud method. Researchers would remain silent in the call until the survey was completed. Various participants were asked to take the survey on different screen sizes (laptop, desktop, mobile, etc), in order to ensure that the survey was accessible from a variety of devices.

At the end of each pilot session, researchers asked the participant several questions related to their understanding of the survey questions, the length of the survey, and their overall impression and thoughts. After each of these sessions, small adjustments to the survey were made based on the observations of researchers. Thus, the survey was continually improved upon following each session. Gradual improvements included: adjusting the location in which the demographic information of the poster was displayed, adding a progress bar, adjusting the wording of open-ended interpretation questions, and replacing a numerical scale with a Likert scale.

After conducting seven pilot sessions, researchers felt as though the survey was an appropriate length, accessible, and each to understand and interpret. The questions, format, and wording were finalized.

This study was approved by our institution’s IRB as a human subjects survey.

### 3.3 Study Procedure

Within the survey, participants were presented four emoji posts from the Chime application. When creating a post in the Chime application, users are first asked “How are you feeling?” and allowed to select one main emoji from a set of nine “expression” emojis. Following the selection of a main emoji, users are asked “Want to share why?” and are allowed to select up to five additional emojis. The survey displayed the posts with this context, and with the main emoji presented first and at double the size of the additional story emojis.
Each participant was shown the same four emoji posts in a randomized order. Posts were chosen from the previous Chime study for which “ground truth labels” were collected. This was crucial in ensuring that researchers could measure accuracy of interpretation. These labels were provided in a follow-up survey of the initial Chime study in which participants were shown a post that they had created in the Chime application and asked to explain the post. This provided explanation is considered to be the “ground truth label” of the post.

Additionally, posts were selected that had different perceptions and sentiments of the main emoji in order to understand the nuances of the interpretations of different emotions. Ultimately, the survey included one post with a main emoji that was clearly positive, one with a main emoji that was clearly negative, one in which the main emoji’s positive or negative sentiment was ambiguous, and another in which the main emoji’s meaning altogether was ambiguous. Meanings of an emojis were determined through analysis of perceptions in a survey in the initial Chime study. Before beginning usage of the Chime application, participants took a survey in which they were shown each of the nine possible main emojis and asked to write their perception of that emoji. These responses were analyzed using a sentiment analysis model (VADER) and chose our four main emojis based on the responded and the previously defined criteria.

Once the four main emojis were selected, a stratified sample of all posts with each main emoji was used to chose the four Chime posts that would be displayed in the survey. Posts that did not have at least three unique emojis were removed from this sampling in order to reduce ambiguity. The four selected posts can be seen in Figures 1a, 1b, 1c, and 1d.

For each of these posts, a ground truth label was provided by the original poster and used for accuracy scoring. One example of a ground truth label is the following explanation provided for the “tired” post:

“They were tired so they got coffee; then they grabbed their backpack and took the bus to work. They did work then saw their boyfriend/girlfriend later.”

Additionally, the posts were each accompanied by randomly selected demographic information. Three categories of demographic information were provided: gender, age group, and student status. Within these three categories, two options were established that would be “opposites.” For gender, it was male and female, for age group it was young adult and adult, and for student status, it was student or non-student. Please note that for non-binary
participants, the “opposite” value was randomly selected between male and female. These options were chosen in order to simplify the way in which demographic information was provided with each post and easily track the options that could be displayed.

With these three categories, there were five options of the actual information that would be provided with the post:

1. **The Ground Truth Demographic Information.** The post would be accompanied by all three categories as they actually were for the original poster.

2. **The Opposite of the Ground Truth Demographic Information.** The post would be accompanied by the exact opposite of the truth for each category of demographic information. For example, if the post was actually made by a male young adult who is a student, the post would be shown as being made by a female adult who is not a student.

3. **A Reflection of the Participant’s Demographic Information.** The post would be displayed with whatever the survey taker’s demographic information is. This information is collected in the initial part of the survey, and then is coded to be stored and later displayed in this portion. The demographic information of the poster has no impact on this option.

4. **The Opposite of the Reflection of the Participant’s Demographic Information.** The post would be displayed with the exact opposite demographic information of the survey taker. For example, if the survey taker was a female adult who is a student, the post would be displayed as being made by a male young adult who is not a student. The demographic information of the poster has no impact on this option.

5. **No Demographic Information.** The post would show no demographic information at all (for any of the three categories).

In addition to the post interpretation, participants were asked to rate the confidence of their interpretation on a scale of one to ten.

After each post was displayed, participants were taken to the next screen in which they were shown “another person’s interpretation of the post” which would actually be the ground truth label of the post. The ground truth labels displayed in this portion of the survey were perturbed to remove any personal information and sound as though they were coming from an external perspective. After being shown this other interpretation, participants were asked if their confidence in their initial interpretation had changed (on a Likert scale). They were also asked to explain the change (or lack of change) in short answer text.

The survey was distributed over the course of 3 and a half weeks.

All participants were required to sign a consent form in order to complete this survey. Following the consent form, it was confirmed that the participants met all eligibility criteria. If a participant was ineligible, they were not shown the rest of the survey. Eligible participants were asked for demographic information (including their age, gender, student status, and employment status). They were also asked about their familiarity with and usage of emojis, and about their current emotional state.

The full survey can be found in the appendix.
3.4 Compensation

Participants who completed the survey were able to enter their email addresses through a separate form to join a raffle. 10% (one in ten) of participants that entered their email addresses in this raffle were awarded a $25 virtual gift card, sent to the provided email.

3.5 Data Cleaning and Analysis

After the data was collected, it was first hand cleaned. Hand cleaning consisted of two researchers going through each entry by hand and removing entries that were clear duplicates (an issue that came with posting on Reddit), were clearly unrelated to the prompts (ex. single letter responses), and entries that were not in English.

Once the data was hand cleaned, it was processed and cleaned through a cleaning script. This script began by removing extra columns that were not needed for analysis. It modified the responses to questions with images to store the text that the image represented (rather than the file path of the image), and replaced Likert scale text with numerical values for analysis. It also removed ineligible participants; those who did not finish the survey, those with contradictions in their responses (ex. living the US longer than the provided age), and clear timing outliers.

After the data was hand cleaned, it was reformatted for analysis. Each survey response was broken up into four separate entries, one for each interpretation that was provided. Each interpretation was given its own row in the dataset, tied back to the participant by a stored participant ID. Each data point also included both the displayed and accurate demographic information about the poster whose post was interpreted.

After the data was reformatted, interpretations needed to be scored for accuracy. Two researchers acted as readers and analyzed each interpretation in order to score it. Interpretations were scored by taking the ground truth label of the post, extracting key themes, and generating a code book for each post. For example, for the happy post (1a), the ground truth label of the post was as follows:

“They were feeling good and happy because a lot of good things happened. They went to the gym and saw a very nice sunset. They shared photos with a friend and they talked about it. They also got enough sleep and got some things done during the day.”

From this text, the key themes of “happy,” “exercise,” “sunset,” “social/communication,” “enough sleep,” and “productive/got things done” were extracted and used as accuracy markers for the code book. If an interpretation mentioned a subset of these themes, the percentage that they included would represent the accuracy score for that interpretation.

After 4 iterations of the code book, it was used on the study data by the two readers. A final score was assigned to each post by taking the average of the scores provided by both readers. Inter-rater reliability was 0.92, as measured by the Cohen Kappa score. The full code book can be found in the appendix.
4 Results

This analysis focuses on evaluating our data based on two main scores: the accuracy of each interpretation and the confidence level of the interpretation. Once the data was cleaned, it was formatted so that each interpretation represented its own data point, including information about accuracy, confidence, the participant’s demographics, the poster’s demographics, the displayed demographics, and the post that was shown itself.

4.1 Overarching Statistics

To begin, the average accuracy score across all interpretations was 0.45 (out of 1) and the average confidence level was 5.8 (out of 10). The overall distribution of these two metrics can be found in Figures 2a and 2b.

4.2 Differences Between Demographic Groups

The analysis began by examining differences in accuracy and confidence of interpretation between the two options of each of the three underlying demographics that were identified earlier: gender, age, and student status.

4.2.1 The Lack of Impact of Gender on Overall Interpretation

The data was first separated by gender differences. No significant difference was found between men and women in terms of accuracy or confidence. As the data (240 interpretations made by men, 155 interpretations made by women) was not found to be normally distributed by the Shapiro-Wilkes test, the two measures were compared using a Mann-Whitney Test for accuracy ($U = 16565.0$, $p = 0.066$) and confidence ($U = 20134.5$, $p = 0.167$).
4.2.2 Gender Differences in Interpretation Between Groups

While examining differences in poster and interpreter demographics, researchers hoped to explore how well different demographics interpret themselves and each other. This study chose to focus on gender because age and student status were correlated and slightly skewed. Additionally, all of the original chime posts that were presented for interpretation were made by young adult students, so this study was unable to explore how more nuanced differences between the interpreter and poster impacted interpretation for older demographics.

First, the analysis explored if members of the same gender are more skilled at interpreting themselves or others. When comparing both the accuracy of interpretation and confidence level of interpretations in which both the poster and the interpreter were female (group size = 77), both the poster and the interpreter were male (group size = 120), and when the poster and the interpreter were of different genders (group size = 234), there was a significant difference between the three groups in terms of accuracy ($H(2) = 12.21, p = < 0.01$), but an insignificant difference for confidence level ($H(2) = 5.1, p = 0.08$). This analysis was done using a Kruskal Wallace test.

The each category was compared to each other in order to find where interpretations differed. Ultimately, nearly all of the comparisons demonstrated significant differences in accuracy and none were significant in terms of confidence level. The only comparison that did not show significant differences in terms of accuracy levels was the comparison of women predicting women to men predicting women. The results of these comparisons can be found in Table 1. For a more detailed look at the relative accuracy of each group, refer to Table 2.

<table>
<thead>
<tr>
<th>Gender Interpretation Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
</tr>
<tr>
<td>Men Interpreting Men</td>
</tr>
<tr>
<td>Women Interpreting Women</td>
</tr>
<tr>
<td>Men Interpreting Men</td>
</tr>
<tr>
<td>Women Interpreting Men</td>
</tr>
<tr>
<td>Men Interpreting Men</td>
</tr>
<tr>
<td>Women Interpreting Women</td>
</tr>
</tbody>
</table>

Table 1: Statistics and P Values when comparing the accuracy and confidence when different genders interpret each other. Nearly all comparisons are significantly different in accuracy but none are significant for confidence. Significant p-values are bolded in the above table.
Gender Interpretation Accuracy and Confidence

<table>
<thead>
<tr>
<th>Group</th>
<th>Group Size</th>
<th>Mean of Accuracy Scores</th>
<th>Mean of Confidence Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men Interpreting Men</td>
<td>120</td>
<td>0.51</td>
<td>6.15</td>
</tr>
<tr>
<td>Men Interpreting Women</td>
<td>120</td>
<td>0.35</td>
<td>5.87</td>
</tr>
<tr>
<td>Women Interpreting Women</td>
<td>77</td>
<td>0.35</td>
<td>5.61</td>
</tr>
<tr>
<td>Women Interpreting Men</td>
<td>78</td>
<td>0.61</td>
<td>5.92</td>
</tr>
</tbody>
</table>

Table 2: Average accuracy and confidence scores of each gender interpreting another. Posts made by women are more difficult to interpret.

Table 2 demonstrates that posts made by women were consistently more difficult to interpret than posts made by men. It also shows that women were better at interpreting posts made by men than men were at interpreting those same posts.

It is important to consider the fact that the “angry” post was the easiest to interpret overall, and that post was made by a man. When this post is excluded from analysis, the difference between men interpreting men and women interpreting men is no longer statistically significant, however, women do still have a higher average accuracy when interpreting men.

4.2.3 The Impact of Age on Interpretation

Next, differences based on age groups were examined. Participants were classified as either a young adult or adult based on their age (18-25 as young adult, 26 and above as adult). There was not a significant difference between young adults and adults in the accuracy of their interpretations (\(U = 22322, p = 0.49\)), however, there was a significant difference in the confidence of their interpretations (\(U = 30285.5, p =< 0.01\)) in that young adults were more confident than adults (mean of young adult confidence = 6.4, mean of adult confidence = 5.2). Because the data for the scores and confidence levels of these two groups (211 interpretations made by young adults, 220 interpretations made by adults) were not normally distributed so a Mann Whitney test was used.

To dig deeper into this finding, the relationship between the actual numerical age value of each interpreter and the accuracy and confidence of their interpretation was examined. Using a Pearson’s correlation test, a significant correlation between age and confidence in interpretation (\(r = -0.249, p =< 0.01\)) was found. As expected based on the previous result, the relationship between age and accuracy was not significant (\(r = -0.0138, p = 0.775\)).

4.2.4 The Impact of Student Status on Interpretation

The last identified demographic group is students compared to non-students. Using a Mann Whitney test, it was found that there was not a significant difference between the two groups (235 interpretations made by students and 196 interpretations made by non-students) in terms of the accuracy of their interpretations (\(U = 21708.0, p = 0.31\)), but the difference in confidence was significant (\(U = 28685.5, p =< 0.01\)). The data demonstrated that students are more confidence than non-students (mean of student confidence = 6.24, mean of non-student confidence = 5.25). Within the group of students, the interpretations of different
undergraduate years were also compared to each other using a Kruskal Wallace test and the results were insignificant (accuracy: \( p = 0.68 \), confidence: \( p = 0.10 \)).

The average age of a student in this data is 23.1 years old and the average age of non-students is 33.7 years old, demonstrating that age could likely have an impact on student status. The tetrachoric correlation score between the two groups to be 0.88, which is fairly high, indicating that age likely impacted student status and its related results and findings.

### 4.2.5 Impact of Familiarity with Emojis

Alongside the three main demographics, information about the participant’s familiarity with emojis was collected. Using a Spearman’s test, the impacts of familiarity with emojis on both accuracy and confidence of interpretation were explored and it was found that there was not a significant impact on accuracy of interpretation (\( r_s = 0.05, p = 0.3 \)), but there was on confidence (\( r_s = 0.13, p =< 0.01 \)). Namely, as familiarity with emojis increased, so did confidence in interpretation. Participants evaluated their own familiarity with emojis on the Likert scale. In order to be used in analysis, this scale was converted to numerical values.

Because this finding aligned with the previous result related to age impacting confidence, researchers explored how familiarity with emojis and age were related. Using the Spearman’s test again, it was found that the relationship between age and familiarity with emojis was statistically significant (\( r_s = -0.22, p =< 0.01 \)). In particular, young adults are more familiar with emojis than adults are.

### 4.3 Differences in Displayed, Actual, & Participant Demographics

#### 4.3.1 Differences in Poster and Interpreter Demographics

Once differences in confidence and accuracy between the different demographic groups that participants belonged to were examined, it was time to see how demographic similarities between the poster of the post that was shown to survey participants and the participants themselves could potentially impact their interpretation. This analysis was aimed at exploring the hypothesis of if actual demographic similarity between the poster and interpreter would impact confidence and accuracy of interpretation.

For each individual interpretation, the demographic information of the poster and the interpreter were examined on all three demographic categories (gender, age, and student status), and interpretations were sorted into categories based on the level of the match. For example, if all three categories completely matched between the poster and the survey taker, it was classified as a complete match. If only one or two matched, it was classified as a partial match, and if all three did not match, it was classified as no match.

Between these three groups, there was not a statistically significant difference between the levels of demographic matching for accuracy of interpretation (\( H(2) = 1.4, p = 0.5 \)), but there was for confidence of interpretation (\( H(2) = 23.8, p =< 0.01 \)). Both the scores and confidence levels of the interpretations of each group were not normally distributed (using the Shapiro Wilkes test), and the Kruskal Wallace test was used to examine the differences between them. The group sizes and average confidence levels for each level of demographic match can be found in Table 3.
<table>
<thead>
<tr>
<th>Level of Match</th>
<th>Mean Accuracy Score</th>
<th>Mean Confidence Level</th>
<th>Group Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Match</td>
<td>0.42</td>
<td>6.43</td>
<td>87</td>
</tr>
<tr>
<td>Partial Match</td>
<td>0.46</td>
<td>5.94</td>
<td>248</td>
</tr>
<tr>
<td>No Match</td>
<td>0.46</td>
<td>4.83</td>
<td>96</td>
</tr>
</tbody>
</table>

Table 3: Average levels of accuracy and confidence for interpretations categorized by the level of demographic match of the poster and the interpreter. A complete match of demographics results in the highest confidence, followed by a partial match of demographics. No match in demographics results in the lowest confidence score.

Then, each of the match levels were compared to each other to find significant differences. When comparing a complete match to a partial match, there was not a significant difference between confidence levels ($U = 12107.5, p = 0.09$). However, when comparing a complete match to no match, a significant difference in confidence levels was found ($U = 5782.0, p <= 0.01$). Additionally, the difference between a partial match and no match was also significant for confidence ($U = 15173.0, p <= 0.01$). When comparing the match levels to each other, Mann Whitney tests were used because none of the confidence level groups were normally distributed.

### 4.3.2 The Lack of Impact of Displayed Demographics

Next, analysis explored how displayed differences in demographics impacted accuracy and confidence in interpretation. Ultimately, there was neither a significant difference between the displayed demographic option in terms of accuracy ($H(4) = 4.27, p = 0.37$) nor confidence ($H(4) = 4.86, p = 0.30$). For each post, the survey randomly displayed demographic information from 5 predetermined options: Ground Truth (101 interpretations), Opposite of the Ground Truth (74 interpretations), Reflection (103 interpretations), Opposite of Reflection (73 interpretations), and No Displayed Information (80 interpretations). See Methods for more information on these labels. A Kruskal Wallace test was used to see if there were differences in accuracy and confidence of interpretation between the five labels, and it indicated that there were not.

Then, interpretations were classified into similar groups based on how much of the displayed demographic information shown to the interpreter with the post matched with the true demographic information of the actual poster. Posts were categorized by the level of the match, including Complete Match, Partial Match, No Match, and No Displayed Information.

In order to address the initial research question about how knowing demographic information about the poster could impact accuracy and confidence, interpretations that had no demographic information provided (group size = 80) were compared to those who had some information displayed, whether or was accurate or not (group size = 351). Despite previous research that demonstrated that knowledge of a poster’s demographics can impact the way in which an online post is interpreted [Fichman and Sanfilippo(2015)], this analysis demonstrated that there was not a significant difference between either group in accuracy of interpretation ($U = 14375.5, p = 0.74$) or confidence of interpretation ($U = 14007.0,$
The data (scores and confidence levels) for these two groups were not normally distributed so a Mann Whitney test was used to compare them.

To continue to explore how what one knows about the poster could impact interpretation, each of the levels of information (complete match, partial match, etc) were compared to each other. Once again, there was no significant difference between the groups in terms of accuracy ($H(2) = 1.61, p = 0.66$) or confidence level ($H(2) = 3.16, p = 0.37$). For each of these three groups, both the scores and confidence levels of the interpretations were not normally distributed (using the Shapiro Wilkes test), and the Kruskal Wallace test was used to examine the differences between them.

Next, the research question regarding how the level of similarity between the displayed demographic information of the post and the interpreter’s own demographics could impact the accuracy and confidence of their interpretation was investigated. Once again, each post was classified into a group, based on the level of the match of the displayed information and the participant’s information, including Complete Match, Partial Match, No Match, and No Displayed Information.

For each of these three groups, there was no significant difference between the groups in terms of accuracy of interpretation ($H(2) = 0.34, p = 0.95$) or confidence level ($H(2) = 0.41, p = 0.94$). Both the scores and confidence levels of the interpretations for each group were not normally distributed (using the Shapiro Wilkes test), and the Kruskal Wallace test was used to examine the differences between them.

### 4.4 Differences in Emotions

Next, the analysis examined how the interpretations varied between each of the four displayed posts, in particular, how it varied from emotion to emotion. Each interpretation was grouped by the emotion of the post to which it was analyzing; happy, angry, satisfied, and tired, in order to explore the hypothesis of if the positivity of a post impacts interpretability. It was found that there was a significant difference between each post in terms of accuracy of interpretation ($H(3) = 104.4, p =< 0.01$), but not in terms of confidence ($H(3) = 5.47, p = 0.14$). Through the Shapiro Wilkes test, it was found that the accuracy and confidence for these groups were all not normally distributed, and differences between them were analyzed using the Kruskal Wallace test. The mean accuracy and confidence scores for each emotion can be found in Table 4, which demonstrates that “angry” had the highest mean accuracy score and “happy” had the lowest.

<table>
<thead>
<tr>
<th>Emotion of Post</th>
<th>No. of Interpretations</th>
<th>Average Accuracy Score</th>
<th>Average Confidence Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>107</td>
<td>0.29</td>
<td>5.36</td>
</tr>
<tr>
<td>Angry</td>
<td>108</td>
<td>0.71</td>
<td>5.82</td>
</tr>
<tr>
<td>Satisfied</td>
<td>108</td>
<td>0.41</td>
<td>5.92</td>
</tr>
<tr>
<td>Tired</td>
<td>108</td>
<td>0.40</td>
<td>6.04</td>
</tr>
</tbody>
</table>

Table 4: Average accuracy and confidence scores for posts, categorized by emotion. Angry posts had the highest accuracy and happy posts had the lowest accuracy of interpretation.
Furthermore, Mann Whitney tests were conducted between each emotion to investigate which differences in accuracy were significant. The results can be found in Table 5 and show that happy and angry were significantly different from all other emotions, while satisfied and tired were similar.

<table>
<thead>
<tr>
<th>Comparison of Accuracy Scores</th>
<th>Between Emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>U</td>
</tr>
<tr>
<td>Happy VS Angry</td>
<td>1755.5</td>
</tr>
<tr>
<td>Happy VS Satisfied</td>
<td>3849.0</td>
</tr>
<tr>
<td>Happy VS Tired</td>
<td>3993.5</td>
</tr>
<tr>
<td>Angry VS Satisfied</td>
<td>9033.0</td>
</tr>
<tr>
<td>Angry VS Tired</td>
<td>9040.5</td>
</tr>
<tr>
<td>Satisfied VS Tired</td>
<td>6031.5</td>
</tr>
</tbody>
</table>

Table 5: Mann Whitney tests between each emotion to compare accuracy scores. Significant p-values are bolded. The only insignificant comparison is tired and satisfied. Happy and angry are significantly different from everything.

4.4.1 Demographic Differences in Emotional Interpretation

In order to fully examine the nuances of the differences between the accuracy of interpretation between each emotion, our analysis also explored how well each demographic interpreted each post. Researchers chose to explore how the accuracy and confidence of each post varied between different genders and age groups, excluding student status because of its correlation with age. Analysis revealed that the only statistically significant differences are in the accuracy of interpretation between men and women for angry posts, and in the confidence level between young adults and adults on angry and tired posts. The detailed results of this analysis can be found in Table 6.

<table>
<thead>
<tr>
<th>Sentiment</th>
<th># Male</th>
<th># Female</th>
<th>Gender Accuracy</th>
<th>Confidence</th>
<th># Young</th>
<th># Adult</th>
<th>Age Accuracy</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>60</td>
<td>38</td>
<td>1169.0 0.84</td>
<td>1227.0 0.53</td>
<td>58</td>
<td>49</td>
<td>1745.0 0.04</td>
<td>1751.0 0.04</td>
</tr>
<tr>
<td>Angry</td>
<td>60</td>
<td>39</td>
<td>846.0 0.02</td>
<td>1336.5 0.24</td>
<td>59</td>
<td>49</td>
<td>1283.0 0.32</td>
<td>1833.0 0.02</td>
</tr>
<tr>
<td>Satisfied</td>
<td>60</td>
<td>39</td>
<td>1082.0 0.53</td>
<td>1273.0 0.46</td>
<td>59</td>
<td>49</td>
<td>1163.5 0.08</td>
<td>1684.5 0.14</td>
</tr>
<tr>
<td>Tired</td>
<td>60</td>
<td>39</td>
<td>1032.5 0.33</td>
<td>1192.5 0.88</td>
<td>59</td>
<td>49</td>
<td>1269.5 0.28</td>
<td>1911.5 &lt;0.01</td>
</tr>
</tbody>
</table>

Table 6: Comparisons of accuracy and confidence between gender and age groups, split up by the emotion of the post. Men and women interpret angry posts with different accuracy and that confidence differs the most between young adults and adults on tired and angry posts. Note that all comparisons were conducted with a Mann Whitney test. Significant p-values are bolded in the above table.
4.4.2 The Impacts of Mood on Interpretation

Next, the impacts of the current mood of the participant on their interpretations were explored. To begin, differences in confidence and accuracy overall based on each mood were explored and it was found that there was a significant difference both in accuracy ($H(7) = 14.82$, $p = 0.04$) and in confidence level ($H(7) = 16.79$, $p = 0.02$) between the different moods. Because more than one of the groups was not normally distributed for score and accuracy, a Kruskal Wallace test was used for analysis. Note that the “Scared” emotion was removed from this analysis because none of the participants selected this as their current mood. The mean scores for both accuracy and confidence of interpretation by each mood (along with group size) can be found in Table 7.

<table>
<thead>
<tr>
<th>Current Mood</th>
<th>Number of Interpretations</th>
<th>Mean Accuracy Score</th>
<th>Mean Confidence Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sad</td>
<td>8</td>
<td>0.5</td>
<td>6.25</td>
</tr>
<tr>
<td>Satisfied</td>
<td>123</td>
<td>0.41</td>
<td>6.25</td>
</tr>
<tr>
<td>Tired</td>
<td>96</td>
<td>0.48</td>
<td>5.01</td>
</tr>
<tr>
<td>Happy</td>
<td>64</td>
<td>0.48</td>
<td>5.89</td>
</tr>
<tr>
<td>Bored</td>
<td>84</td>
<td>0.48</td>
<td>5.83</td>
</tr>
<tr>
<td>Angry</td>
<td>4</td>
<td>0.52</td>
<td>5.5</td>
</tr>
<tr>
<td>Excited</td>
<td>36</td>
<td>0.36</td>
<td>6.11</td>
</tr>
<tr>
<td>Stressed</td>
<td>16</td>
<td>0.56</td>
<td>5.38</td>
</tr>
</tbody>
</table>

Table 7: Presents the groups size, mean accuracy score, and mean confidence level for interpretations, categorized by the current mood of the survey taker.

To investigate which comparisons in both accuracy of interpretation and confidence were significant, a Mann Whitney comparison was conducted between all current moods. Ultimately, the only significant differences between moods were between tired and satisfied for confidence of interpretation (in which tired scored much lower than satisfied), and between excited and stressed for accuracy of interpretation (in which excited had worse accuracy than stressed). The detailed results can be found in Table 8.
## 5 Discussion

### 5.1 General Trends in Accuracy

The average accuracy of interpretation of posts in the study was 0.45. Because accuracy was taken to be the percentage of themes from the ground truth label that were present in an
interpretation, it is observable that approximately half of the content of emoji-only posts is widely interpretable. Furthermore, this finding demonstrates that emojis placed in a certain order still imply meaning, because no additional textual information about the post was given to participants, and yet they were able to glean nearly half of the original meaning of the post.

5.2 Gender Differences in Interpretation

Next, we explored how the demographics of the poster and the interpreter would impact the interpretation of a post. Interestingly enough, the gender of the interpreter did not impact either accuracy or confidence in interpretation. Overall, men and women demonstrated similar scores of accuracy and confidence in their interpretation. This result was surprising because previous work has demonstrated that women are more likely to engage in expressive communication online [Procopio and Procopio(2007)], and therefore, we predicted that they would be better at interpreting emotional posts.

Interestingly enough, upon closer examination of how different genders interpreted each other, the results were statistically significant. Ultimately, posts made by women scored lower in interpretation accuracy than posts made by men, but were equally difficult for men and women to interpret. This is an interesting finding because studies demonstrate that women use emojis more often and positively than men [Prada et al.(2018)]. However, this result could be related to the fact that research demonstrates that men communicate online in a more instrumental manner than women, which could be more straightforward and easier to understand [Procopio and Procopio(2007)].

This finding raises questions about the extent to which gender differences in communication styles contribute to online misunderstandings and conflicts. Further research could explore these questions by examining different types of online communication (e.g., email, social media) and comparing interpretation accuracy across genders and communication styles.

Moreover, posts made by men were interpreted with more accuracy by women than by men. Previous work reinforces this finding, because it has demonstrated that women perceive emojis as more familiar, clear, and meaningful than men do [Rodrigues et al.(2018)]. Additionally, women are more likely to engage in online expressive communication than men [Procopio and Procopio(2007)], this may imply that they are more skilled at interpreting emotions online.

We also analyzed how well different demographics interpreted each emotion, and found that both comparisons between gender and age yielded similar results in terms of accuracy, with the exception being that women were more skilled at interpreting “angry” posts than men. This finding aligns with the fact that women have historically utilized online communication in a more expressive and emotive way than men [Procopio and Procopio(2007)].

It should be noted that the angry post that was being interpreted by participants was made by a man and was the easiest post to interpret overall, which may have impacted the results. While our results could point to the fact that men communicate in a more straightforward manner than women, it could also be due to the fact that women generally experience more negative emotions than men [Alexander and Wood(2000)], and therefore are more skilled at interpreting them. It could also be confounded by the fact that our angry post had fewer themes than the other emotions. However, when the consistency between
posts is examined, angry was not the most consistent overall, satisfied was.

5.3 Age, Student Status, and Familiarity with Emojis

Next, we continued to explore how demographic differences in participants would impact interpretation by examining differences in age and student status. It is important to note that a high correlation between age group and student status was found in our data. Within our participant pool most students were young adults, and most adults were not students. Additionally, the level of familiarity with emojis of each participant was also related to age. Young adults were more familiar with emojis, while adults tended to be less familiar. The correlation between age and experience with emojis was ultimately found to be significant within our data.

While we cannot conclude if it was one particular factor of these three influencing our results or not, it is interesting to note that all of these factors had a significant impact on confidence but not accuracy of interpretation. Both the correlation between familiarity with emojis and age and their impact on confidence align well with the findings of previous research. Past studies have shown that young people tend to use social media and online resources more than older generations [Drouin et al.(2018), Lehtinen et al.(2009)], and therefore, have more experience interacting with and using emojis [Weiß et al.(2020)] which could boost their confidence when interpreting them. It is intriguing to find that more experience and familiarity with emojis did not appear to give younger generations an advantage in accuracy of interpretation, and that older generations were able to interpret emoji posts with the same level of accuracy, even though they had less experience with emojis.

5.4 Underlying Demographic Similarity and Confidence

To test our hypothesis that actual demographic similarity between the poster and the interpreter would impact accuracy and confidence of interpretation, we examined how the level of demographic match between the poster and interpreter was related to these two metrics. Ultimately, higher levels of demographic similarity were again associated with higher levels of confidence in interpretation, but did not impact the accuracy of the interpretation.

This finding was surprising because we expected that accuracy of interpretation would increase with increased demographic similarity. In fact, previous work demonstrated that communication is generally stronger and more interpretable when those engaged in the conversation share demographic similarities [Drouin et al.(2018)]. This finding is also intriguing because participants may or may not have known the level of demographic similarity when posting, so our results demonstrate that participants’ confidence was boosted by the underlying similarity and not the presented similarity.

This finding may have important implications for the development of online platforms designed to facilitate communication and interaction between individuals who possess similar demographic backgrounds. For example, platforms that emphasize the importance of demographic similarity in connecting users may not necessarily result in greater accuracy or understanding in communication. Instead, other factors that promote effective communication and interpretation should be explored in order to boost the effectiveness of these applications.
5.5 Examining Differences in Displayed Demographics

Next, we explored our initial research question of whether knowing demographic information about the poster would impact confidence or accuracy of interpretation. To begin, posts that had no demographic information provided were compared to those who had some demographic information provided (whether it was accurate or not). Surprisingly, the results demonstrated that the inclusion of demographic information did not impact accuracy or confidence of interpretation when compared to posts with no demographic information.

These findings have important ramifications for the role of anonymity online. Furthermore, it demonstrates that anonymity and a lack of demographic information does not impact interpretability of emotional online posts. This finding challenges the intuitive idea that anonymity is detrimental to communication and interpretation. While anonymity can certainly enable negative behaviors such as trolling and cyberbullying [Nakano et al.(2016)], it does not necessarily hinder accurate interpretation of emotional posts, suggesting that the role of anonymity in online communication presents both potential benefits and risks. Furthermore, because anonymity does not impede a reader’s ability to interpret an emotional online post, this finding could suggest that online communication can be effective even without extensive personal information sharing.

Next, to continue our exploration of whether what the interpreter knew about the poster would impact the interpretation, each interpretation was categorized based on the accuracy of the demographic information displayed with the post and searched for differences in confidence and accuracy of interpretation. Once again, the level of accuracy of the displayed demographic information of the post neither impacted accuracy nor confidence of interpretation.

This finding has interesting implications related to online aliases. Because the accuracy of the information displayed has no impacts on interpretability, we can consider the potential benefits of allowing individuals to use online aliases or pseudonyms while still allowing their content to be equally interpretable. This could be particularly important in contexts in which individuals feel the need to protect their identity or express views that may be unpopular or controversial. However, it could have negative significance when considered in the context of the spread of misinformation online. With no real difference between inaccurate and accurate information in terms of interpretation, it could be easier for individuals to spread false information or engage in malicious activities without being held accountable.

Finally, we investigated our research question of if perceived demographic similarity between a poster and an interpreter could impact the interpretation of a post. Interpretations were categorized by the level of demographic match between what was displayed and the participant themselves. Interestingly enough, the level of perceived demographic similarity did not have an impact on confidence or accuracy of interpretation. This result is interesting in the context of previous research that demonstrates that people generally feel more comfortable communicating online with those who have a similar demographic background as them [Drouin et al.(2018)]. However, this finding could reveal that comfort and interpretability are not related. While perceived demographic similarity may boost feelings of social support [Drouin et al.(2018)], it does not impact the ability of a reader to interpret a post.

When we examine these two findings related to displayed demographics, combined with the previous result demonstrating that underlying actual demographic similarity impacted confidence in interpretation, demonstrate that actual demographic similarity between online
communicators is more impactful than perceived similarity.

5.6 Interpretability of Different Emotions

To investigate our hypothesis related to how the positivity of a post impacts interpretability, each interpretation was compared to each other based on the mood of the main emoji of the post which it was interpreting. Although there was not a significant difference between the confidence level when interpreting each emotion, it was found that each emotion did have a very different average accuracy level. Namely, that “happy” was by far the most difficult emotion to interpret, and that “angry” was clearly the easiest. The high accuracy associated with the “angry” post could be related to the fact that “angry” had only three themes present in our codebook, and therefore was easier to attain a higher accuracy on. Still, this finding is supported by previous work that shows that the “angry” emoji is the most emotionally representative emoji compared to others [Franco and Fugate(2020)]. However, the “happy” post only had six themes present, while “tired” and “satisfied” both had seven. This finding is surprising because previous work demonstrates that the “slightly smiling face” emoji (associated with satisfied) is notoriously also used in a sarcastic manner and therefore would be expected to be more difficult to interpret than the “happy” emoji [Wang([n. d.])].

These findings that different emotions are more difficult to interpret than others could have potential problematic impacts in terms of emotional communication online. For example, the fact that the “happy” emotion is the most difficult to interpret could be potentially problematic in environments in which clear communication is critical, such as online therapy.

It is also interesting to see that the confidence levels did not change throughout the different emotions displayed. Previous work has shown that certain emotions and their respective emojis are widely disagreed upon and found to not be representative of the emotion as a whole, such as the “slightly smiling face emoji” [Wang([n. d.])], while others are recognized as much easier to interpret and associate with an emotion [Franco and Fugate(2020)]. We would therefore expect that satisfied would be the most difficult emotion to interpret because it is represented by the “slightly smiling face” emoji.

5.7 The Impacts of Mood on Interpretation

Finally, we explored how the current mood of the interpreter could impact the accuracy and confidence of the interpretation. Between the nine options for “current mood” presented to our participants, one option (scared) was never chosen and was therefore excluded from analysis. Ultimately, most moods had a similar accuracy with the exception of “excited” participants having a significantly lower accuracy score than others. Additionally, “tired” participants were the least confident, however, their accuracy was similar to the overall average. These findings highlight the complex relationship between mood, behavior, and interpretation.

It was also intriguing to find that participants whose mood matched that of the post which they were interpreting did not impact their confidence or accuracy of interpretation. Previous work found that people prefer online communication with those in a similar emotional state as them [Drouin et al.(2018)], so it is interesting to find that this preference still did not impact interpretability in any way.
5.8 Future Work

Future work in this area should explore the relationship between gender and ease of interpretation more. Without displaying demographics at all or by only displaying true demographics, it would be interesting to see how men and women interpret each other in a more controlled context. Future work should also explore how men and women interpret each other in emotional online posts that are supplemented by emojis, not only solely comprised of emojis. It would also be intriguing to see which emotions men are more skilled and confident at interpreting and which women are better at interpreting. Our study was limited by the number of emotions and posts that we had to choose from, but future work could explore this through a variety of posts that are equally complex and difficult to interpret.

Additionally, our study was limited by the correlation between age and student status, and by the fact that our posters were all young adults, but future work should investigate how these two demographics interpret each other. It would be interesting to see the accuracy and confidence with which adults interpret young adults and how young adults interpret adults. Future work could also examine the environments in which adults and young adults are more skilled at interpreting emotions online. For example, are adults more skilled at interpretation when emojis are used to supplement text or in posts comprised solely of emojis? Does the length of the emoji string or number of repeated emojis have an impact in interpretability?

Finally, future work should explore the nuances of displayed demographics more in-depth. While we explored if differing displayed demographics could impact accuracy and confidence of interpretation, further research could be done in exploring how displayed demographics impact the themes present in the interpretation. The impacts of displayed demographics on feelings of social support would also be interesting to explore.

6 Limitations

This study has a number of limitations. Firstly, the original Chime posts that were selected for interpretation on our survey were all from the previous Chime study. The participants in this student were limited to college students, specifically first years. Thus, all of the original posts were drawn from a pool in which the posters were all young adults, students, and residing in Providence, RI. This limited the extent to which we could explore and analyze the impacts of the age and student status of the interpreter on the accuracy and confidence of interpretation. While our participant population from the survey was relatively evenly spread across the two, there were no instances in which someone could see an accurate demographic that was also their own demographic if they were not a student or were above the age of 25. Thus, more of our in depth analysis relied on the variance and spread of gender between posters and interpreters, rather than also considering student status and age.

Moreover, a link between student status and age was discovered in our study. Our participant population was skewed in that there was a very high correlation between age and student status. The majority of young adults taking the survey were students, and the majority of adults taking the survey were not students, making it difficult to identify which demographic was impacting the changes in confidence that were examined. This also made it more difficult to run separate analyses on the two, because they confounded each other.
Additionally, both of the positive posts (satisfied and happy) were made by women and both of the negative posts (angry and tired) were made by men. We recognized that the angry post was the easiest to interpret universally, and so when this post was removed from our analysis, no statistically significant difference between the accuracy of male and female interpreters was found.

It is also important to note that the majority of our survey responses came from Reddit, Twitter, and Mastodon. It is widely known that users of these platforms are presumed to be tech savvy, so we can draw the assumption that our participant population is skewed towards those who are technologically advanced and have an online presence. Furthermore, the survey required internet access, so this could also impact the participant pool and exclude some groups.

When advertising, some participants were recruited through non-online sources in order to broaden our participant pool. However, this was done by placing physical posters across Providence, RI, limiting the impact the number and population of non-technologically savvy participants that could be reached.

As discussed earlier, research demonstrates that culture has a great impact on emoji interpretation [Sadiq et al.(2019), Sugiyama(2015)]. However, our survey was in English, which could have not only limited the accessibility of the survey, but also prevented us from exploring the effects of our demographic changes across cultures not studied in this paper.

Finally, responses to our survey questions were all self-reported and not monitored. Though the data was cleaned and filtered for the most obvious instances of this, participants could have exaggerated or randomly filled out their survey responses.

7 Conclusion

This study investigated the accuracy and confidence of interpretation of four emoji-only posts with various demographic information accompanying them to explore how context affects interpretation. Ultimately, accuracy averaged to 45% overall, demonstrating that in emoji-only posts approximately half of the main themes and meaning are generally interpretable. Additionally, the findings revealed that gender differences did not affect accuracy or confidence of interpretation, but posts made by women were more difficult to interpret than those made by men, and women were more skilled at interpreting men. Age, student status, and familiarity with emojis impacted confidence but not accuracy of interpretation. Moreover, demographic similarity between the poster and interpreter only impacted confidence and not accuracy of interpretation. Surprisingly, the inclusion of accurate or inaccurate demographic information did not impact accuracy or confidence of interpretation, demonstrating that actual demographic similarity is more impactful than perceived demographic similarity. Ultimately, this research provides insights into the role of context in emotional interpretation and highlights the importance of actual demographic similarity between online communicators in boosting confidence in interpretation.

8 Acknowledgments

Thank you to my advisors, Talie Massachi and Jeff Huang.
Talie, I could not have completed this thesis without all of your help and guidance. Thank you so much for being such an amazing mentor, supporter, and friend throughout this process. I have learned so much from you and have loved working with you.

Jeff, thank you so much for your advice, guidance, and input throughout this thesis. Your expertise was extremely helpful, and your guidance and support throughout the year were invaluable. It has been a pleasure working with you throughout the year.

Thank you to Andries van Dam who has served as my advisor since freshman year. Andy, I am so grateful for all of your advice and guidance throughout my entire computer science journey at Brown. None of this would have been possible without you, and your constant support means so much to me.

Thank you to Tim Nelson, my reader for this thesis. Tim, thank you so much for your advice on this work, and for always cheering me on throughout my entire four years at Brown. It has been so wonderful to work with you and get to know you.

Finally, and most importantly, thank you to my entire family for their unwavering support. Thank you to my sister, Aneesha, for being my best friend, biggest supporter, and role model for life. Thank you to my parents for always encouraging me, pushing me to try new things, and believing in me no matter what.
References


9 Appendix

9.1 Distributed Survey

The following is a recreation of the survey that was sent to participants. All of the portions in italics are demographic information accompanying posts that could vary randomly.

- Confirmation of Consent
  - By choosing this option, you are confirming that you have read and understood the information on this page, are above the age of 18, and that you agree to volunteer as a research participant for this study.
  - I do not consent.

- Are you at least 18 years of age?
  - Yes
  - No

- Are you currently residing in the United States?
  - Yes
  - No

- Have you ever used the Chime social mood tracking application before OR have you ever participated in a previous study held at Brown University involving the Chime social mood tracking application?
  - Yes
  - No

- How do you identify yourself?
  - Female
  - Fluid/Genderfluid
  - Gender Non-Conforming/Genderqueer
  - Non-Binary
  - Transgender Man / Trans Man
  - Transgender Woman / Trans Woman
  - Not Listed
  - Prefer not to say

- How old are you?
• Which of the following describes your current student status?
  Please select all that apply.
  □ I am currently a part time student.
  □ I am an undergraduate student
  □ I am a graduate student
  □ I am an international student
  □ I am not a student

• Which year of undergraduate study are you in?
  □ First Year
  □ Second Year
  □ Third Year
  □ Fourth Year
  □ Other

• How long have you lived in the US?

• Which of the following describes your current work status?
  □ I am currently employed full time
  □ I am currently employed part time
  □ I am not currently employed

• What area do you work in? (ex. higher education, technology, construction, healthcare)

• How often do you use emojis in your day to day life?
  □ Always
  □ Often
  □ Sometimes
  □ Rarely
  □ Never

• How are you feeling? Choose an emoji to represent your mood.
The following post was made on the Chime app. The first of these emojis is a response to the prompt “How are you feeling?”, and the following emojis are responding to the prompt “Want to share why?”. This post was written by a female young adult who is currently a student.

What is your best guess of what the author of the post meant to convey?

Another user provided the following interpretation of the post: They were feeling good and happy because a lot of good things happened. They went to the gym and saw a very nice sunset. They shared photos with a friend and they talked about it. They also got enough sleep and got some things done during the day.

How confident are you that you correctly interpreted the original poster’s intended message compared to before you saw this other user’s interpretation?

The following post was made on the Chime app. The first of these emojis is a response to the prompt “How are you feeling?”, and the following emojis are responding to the prompt “Want to share why?”. This post was written by a male young adult who is currently a student.
What is your best guess of what the author of the post meant to convey?

• Another user provided the following interpretation of the post: They came down with a cold but they went shopping and also got guacamole and chips!
  How confident are you that you correctly interpreted the original poster’s intended message compared to before you saw this other user’s interpretation?

• The following post was made on the Chime app. The first of these emojis is a response to the prompt “How are you feeling?”, and the following emojis are responding to the prompt “Want to share why?”. This post was written by a female young adult who is currently a student.

What is your best guess of what the author of the post meant to convey?

• Another user provided the following interpretation of the post: They are happy because it is sunny today and they got watch the sunset. They spend a lot of the day outside, in nature where the wind was blowing and that made them feel good and refreshed. They ended the day with a yummy slice of pizza for dinner.
  How confident are you that you correctly interpreted the original poster’s intended message compared to before you saw this other user’s interpretation?

• The following post was made on the Chime app. The first of these emojis is a response to the prompt “How are you feeling?”, and the following emojis are responding to the prompt “Want to share why?”. This post was written by a male young adult who is currently a student.

What is your best guess of what the author of the post meant to convey?

• Another user provided the following interpretation of the post: They were tired so they got coffee; then they grabbed their backpack and took the bus to work. They did work then saw their boyfriend/girlfriend later.
  How confident are you that you correctly interpreted the original poster’s intended message compared to before you saw this other user’s interpretation?
9.2 Code Books

The following are the code books used to score accuracy for each interpretation, sorted by the emotion of the post which they are referring to.

Each emotion displays its ground truth label along with the identified themes (in bold). Each theme is weighted equally and includes specific instructions for when it should and should not be counted.
<table>
<thead>
<tr>
<th>Post</th>
<th>Ground Truth</th>
<th>Themes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>They were feeling good and happy because a lot of good things happened.</td>
<td>happy</td>
</tr>
<tr>
<td></td>
<td>They went to the gym and saw a very nice sunset. They shared photos with a</td>
<td>exercise</td>
</tr>
<tr>
<td></td>
<td>friend and they talked about it. They also got enough sleep and got some</td>
<td>sunset</td>
</tr>
<tr>
<td></td>
<td>things done during the day.</td>
<td>social /</td>
</tr>
<tr>
<td></td>
<td></td>
<td>communication</td>
</tr>
<tr>
<td></td>
<td></td>
<td>enough sleep</td>
</tr>
<tr>
<td></td>
<td></td>
<td>productive /</td>
</tr>
<tr>
<td></td>
<td></td>
<td>got things done</td>
</tr>
<tr>
<td></td>
<td>Count when: specifically talking about feeling a certain way and then a</td>
<td>Count when:</td>
</tr>
<tr>
<td></td>
<td>positive thing (good, happy, positive, etc)</td>
<td>Count when:</td>
</tr>
<tr>
<td></td>
<td>Don’t count when: it’s not an emotion, even if it’s positive (so don’t</td>
<td>Count when:</td>
</tr>
<tr>
<td></td>
<td>count “having a good day”) unless it says happy or is specifically an</td>
<td>Count when:</td>
</tr>
<tr>
<td></td>
<td>emotion word.</td>
<td>Count when:</td>
</tr>
<tr>
<td></td>
<td>Don’t count when: Mentions specific types of exercise (went on run, lifted</td>
<td>Don’t count when:</td>
</tr>
<tr>
<td></td>
<td>Don’t count when: mentions of nightime, dusk, sunrise. Only sunsets allowed</td>
<td>Don’t count when:</td>
</tr>
<tr>
<td></td>
<td>Counts when: Talking about general exercise or gym going (ex. went to work</td>
<td>Counts when:</td>
</tr>
<tr>
<td></td>
<td>Count when: Mentions of sunsets specifically</td>
<td>Counts when:</td>
</tr>
<tr>
<td></td>
<td>Counts when: any kind of communication or socializing (hanging out with,</td>
<td>Counts when:</td>
</tr>
<tr>
<td></td>
<td>Don’t count when: when no mention of communication or socializing,</td>
<td>Counts when:</td>
</tr>
<tr>
<td></td>
<td>Don’t count when: any reference to needing rest or sleep, naps, saying they</td>
<td>Counts when:</td>
</tr>
<tr>
<td></td>
<td>Don’t count when: a specific food item that is not guac and chips</td>
<td>Counts when:</td>
</tr>
<tr>
<td>Angry</td>
<td>They came down with a cold but they went shopping and also got guacamole</td>
<td>cold / illness</td>
</tr>
<tr>
<td></td>
<td>and chips!</td>
<td>shopping</td>
</tr>
<tr>
<td></td>
<td>Count when: any symptoms of being sick are mentioned, generally sick or</td>
<td>avocado /</td>
</tr>
<tr>
<td></td>
<td>Count when: anything to do with a purchase (bought medicine), or shopping,</td>
<td>guacamole</td>
</tr>
<tr>
<td></td>
<td>Don’t count when: doctor, or took medicine, or wearing mask, a specific</td>
<td>Counts when:</td>
</tr>
<tr>
<td></td>
<td>Don’t count when: going out, spending money, errands</td>
<td>Counts when:</td>
</tr>
<tr>
<td></td>
<td>Counts when: mention of avocado, or guacamole, &quot;chips&quot;, or generally food</td>
<td>Counts when:</td>
</tr>
<tr>
<td></td>
<td>Counts when: a specific food item that is not guac and chips</td>
<td>Counts when:</td>
</tr>
<tr>
<td>Post</td>
<td>Themes</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Satisfied</td>
<td>They are happy because it is sunny today and they got watch the sunset. They spend a lot of the day outside, in nature where the wind was blowing and that made them feel good and refreshed. They ended the day with a yummy slice of pizza for dinner.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>happy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sunny weather</td>
<td></td>
</tr>
<tr>
<td></td>
<td>watching sunset</td>
<td></td>
</tr>
<tr>
<td></td>
<td>nature</td>
<td></td>
</tr>
<tr>
<td></td>
<td>outdoors</td>
<td></td>
</tr>
<tr>
<td></td>
<td>wind</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pizza</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Count when: specifically talking about feeling a certain way and then a positive thing (good, happy, positive, etc)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>don't count when: it's not an emotion, even if it's positive (so don't count &quot;having a good day&quot;) unless it says happy or is specifically an emotion word.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Count when: has to be a reference to the weather being nice / good / sunny / warm / pleasant</td>
<td></td>
</tr>
<tr>
<td></td>
<td>don't count when: no reference to weather, or bad weather</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Count when: Mentions of sunsets specifically</td>
<td></td>
</tr>
<tr>
<td></td>
<td>don't count when: doesn't use word nature</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Don't count when: backpack, a general bag (stachel, bag)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Count when: bus, schoolbus, public transit</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Count when: work, job, workplace, office</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Count when: accomplished, got work done, productivity, did homework</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Don’t count when: specific food that is not pizza (and no pizza is mentioned)</td>
<td></td>
</tr>
</tbody>
</table>

| Tired | They were tired so they got coffee; then they grabbed their backpack and took the bus to work. They did work then saw their boyfriend / girlfriend later. |
|       | tired                                                                 |
|       | coffee                                                                 |
|       | backpack                                                              |
|       | bus                                                                   |
|       | work (place)                                                          |
|       | doing work                                                            |
|       | significant other                                                     |
|       | Count when: feeling tired, exhausted, sleepy, drowsy, specific mentions to tired feelings |
|       | don’t count when: lazy, unmotivated, reference to sleep that’s not a feeling, “slept enough” |
|       | Count when: mention coffee, caffeine, hot drink, types of coffee      |
|       | don’t count when: specific drink that is not coffee                   |
|       | Count when: backpack, a general bag (stachel, bag)                    |
|       | don’t count when: a specific type of bag that is not a backpack (purse, luggage, tote) |
|       | Count when: bus, schoolbus, public transit                            |
|       | don’t count when: specific transit that is not a bus (ex. car, train, bike, travel) |
|       | Don’t count when: school, mention of work as an action rather than a place |
|       | Don’t count when: work in reference to a place rather than an action, no specific work (writing a letter, a paper) |
|       | Don’t count when: boyfriend / girlfriend / love interest, saw boyfriend / girlfriend, must reference seeing or hanging out with a SO / crush / whatever |
|       | Don’t count when: love letter, confession, asked someone out, if no reference to spending time together or seeing each other |