| **Analysis in Food Emoji Prediction** |
| --- |
|  |
| **Word Wizards: Danial Syed, Alex Iliarski,**  **Jiayi Huang, Sai Tarun Inaganti** |
|  |
|  |
|  |
|  |

**Abstract**

Emojis have become increasingly prevalent in online communication. As large language models are beginning to have increasing popularity and use across many fields and industries, it begs the question of how well these models can process and understand emojis, given how they can be used in so many different ways depending on the context. We analyze a narrow slice of LLM emoji understanding, testing whether OpenAI’s ChatGPT can correctly label a food-related tweet with a corresponding food emoji.

1. **Introduction**

Emojis have become an increasingly prevalent part of digital communication. It is important that we understand the capabilities of modern NLP models to understand and accurately utilize emojis, and also their abilities on sentiment analysis and expression. As such, we strive to analyze ChatGPT’s performance for a narrow and fairly unambiguous situation: using the appropriate food emoji to describe a food-related tweet. By analyzing how well ChatGPT performs in comparison to human labellers, we gain a further understanding of ChatGPT’s abilities to understand both text and emoji usage.

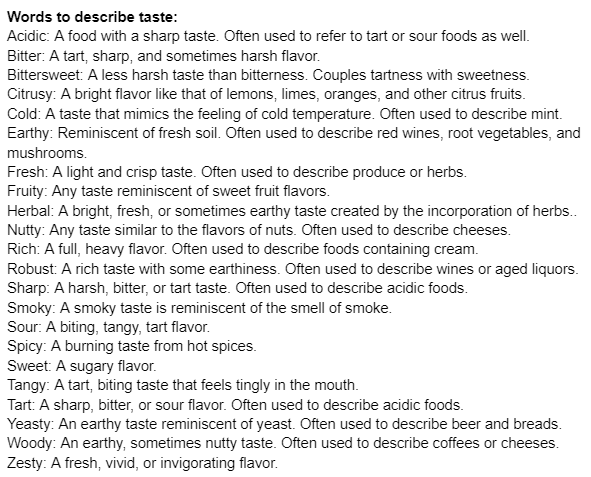
[*How is it done today by other researchers? What are the limitations and challenges of current practice?]*

A possible application of this research is for corporate marketing teams. A great amount of modern marketing is conducted online, on platforms such as Twitter. It is commonplace for companies and other users to include emojis in their tweets. If companies can use NLP models to automatically generate marketing materials that include emojis, this can save marketing teams time while still producing high-quality marketing materials. Additionally, the NLP can analyze responses which include emojis to these marketing campaigns to determine market sentiment. As such, it is important that NLP models can develop capabilities to understand emojis accurately. By conducting our research, we can determine the current state of NLP models’ emoji analysis abilities and recommend whether it is a viable alternative marketing tool. There are also many other viable use cases where NLP models' understanding of emojis is vital to its performance: text messages, social media posts and comments, etc.

1. **Approach**
   1. **Tweet Scraping**

First, we scraped food-related tweets from Twitter. This was done by using the snscrape API to search tweets with keywords. For all the tweets we scraped, we used a timeframe between 07-05-2021 to 07-06-2022 in order to incorporate a diverse set of tweets spanning across a year to sample from. This was done intentionally as some emojis may be more common than others depending on the time of the year. For example, the candy emoji ‘[🍬’,](https://emojipedia.org/candy/) is definitely more common during the month of October as many are preparing and celebrating Halloween.

From there, we used several different keywords to search for tweets containing them. Using a common description of food critiques (<https://www.webstaurantstore.com/article/53/how-to-write-a-menu.html>), a list of 73 common food adjectives were compiled. Due to time constraints, we limited our keyword search to only 23 food adjectives that specifically refer to taste. Some examples of the words we used were ‘spicy’ and bitter’. The rest of the words, such as ‘crunchy’ and ‘flaky’ more so referred to the physical texture of the food. We decided to just hone in on taste adjectives as it allowed us to just analyze the sentiment of the taste, rather than incorporating other senses such as look and feel which would make our analysis more convoluted.



For each keyword, we asked our model to scrape 20 tweets with the given keyword and append the word “food” onto it so we only receive food related tweets with the adjectives. Furthermore, for some keywords, we decided to expand our search further. For example, instead of just searching ‘bitter food’, we also searched ‘bitter snack’ or ‘sharp flavor’. This allowed us to receive tweets on a wider scale that allow for a little more context. Some keyword adjectives came up with less than 20 tweets. For example, one of our adjectives we searched with were “bittersweet food” which produced less than 20 tweets.

Once our algorithm was set in place to search our keywords, we would extract the text of the tweet as well as the user and date of the date. From there, our data was compiled into a large data table containing the three previously mentioned metrics. Bearing all this in mind, we finished our scraping with 300 tweets.

* 1. **Emoji Bank**

We developed an emoji bank with 20 food emojis. Our main priority was to find a range of foods that are most common to the average person. We wanted to get a wide variety of food ranging from snacks to full meals. These are also foods that are fairly common throughout the world and not just one culture or nationality.



With the advice of our teaching assistant, Risako Owan, we decided it was best to limit our emoji set to only 20 emojis in total. We made this decision knowing that we were on a severe time constraint and manually labeling a few hundred tweets with a more appropriate set of 60 emojis would be far more time consuming.

Also, we wanted to limit the number of emojis that are very similar to each other. For example, pancakes and waffles are fairly similar. They are both breakfast foods that are typically eaten with syrup and butter. With this in mind, we can simply remove waffles as an option as it would be difficult to deduce a tweet such as “what a sticky and sweet breakfast!” as either pancakes or waffles.

Another reason for choosing the emojis we did was also to diversify the typical taste sentiment that these foods would have. We wanted a wide variety of tastes such as spicy, sweet, cold, or hot. Our goal was to analyze the accuracy of ChatGPT, so limiting the amount of redundant foods that have similar taste was important. Removing this redundancy would allow ChatGPT and ourselves to not just randomly break a tie between two emojis, and rather come to one absolute option.

* 1. **Tweet Labeling**

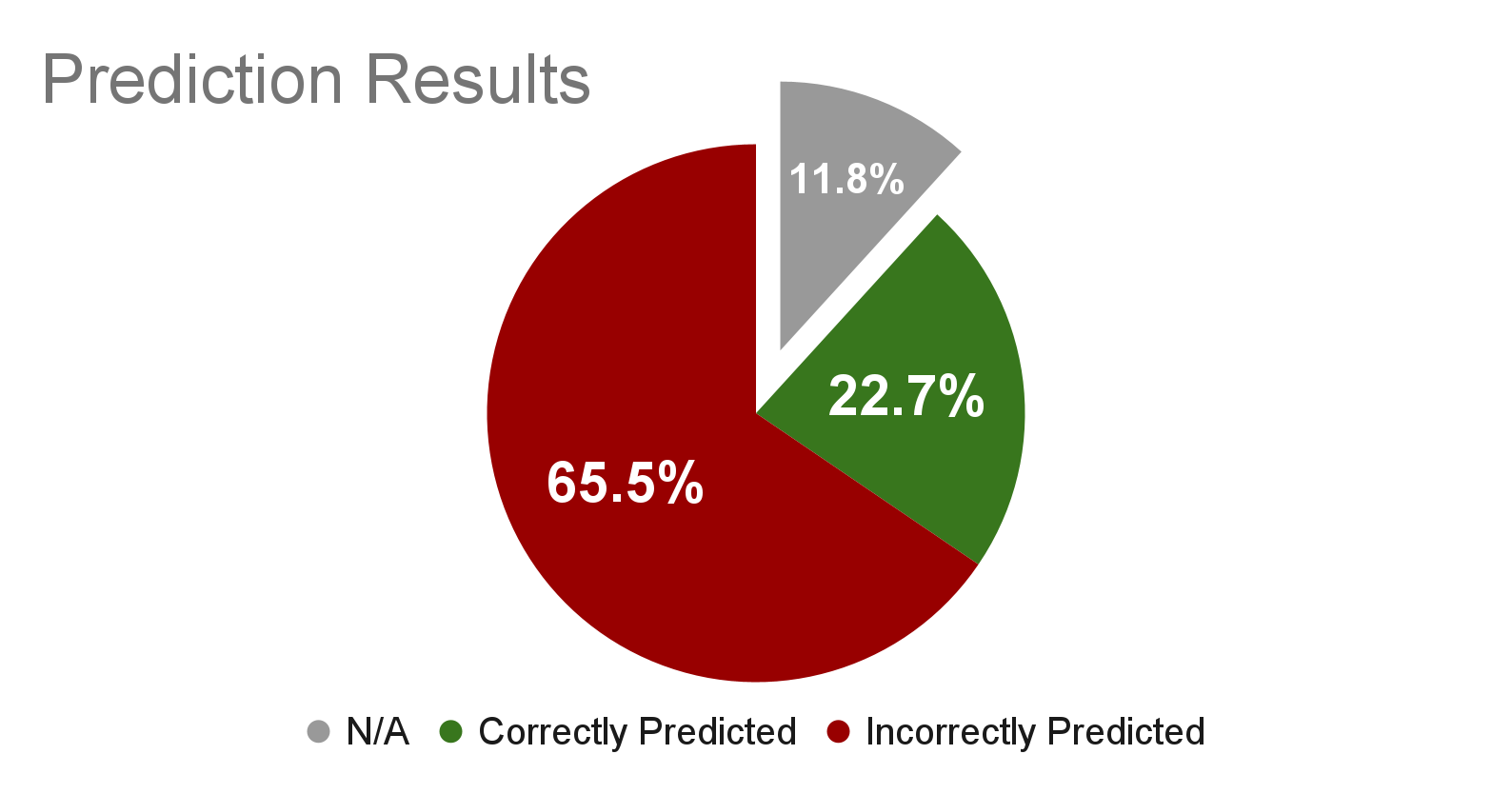
Next, after pulling a dataset of 300 tweets from 2021 to 2022, we began the human labeling process. For each tweet, three group members would come to a consensus for which emoji from our emoji bank best applied as a label for the tweet. If the tweet was entirely unrelated to food, we labeled the tweet as ‘N/A’ and did not consider it further.

We anticipated that this process would be challenging and ambiguous. To resolve this, we used a team of three human labelers to resolve disputes. Indeed, it was extremely difficult to accurately label each tweet. This is due to the limitations set by only having 20 emojis to choose from in the emoji bank. When a tweet did not directly correspond to an available emoji, labeling was mostly a subjective determination of which emoji most closely corresponded with the tweet. There was often much variability in the three human’s initial choices, before coming to a consensus on a single emoji.

Since there is so much subjectivity, even from the human end, it would be somewhat unreasonable to expect ChatGPT’s emoji choices to correlate near-perfectly with the predicted emojis. As such, we hypothesized a low correlation between ChatGPT’s predictions and the human labels.

1. **Experiments, Results, and Error Analysis** 
   1. **One Emoji Response**

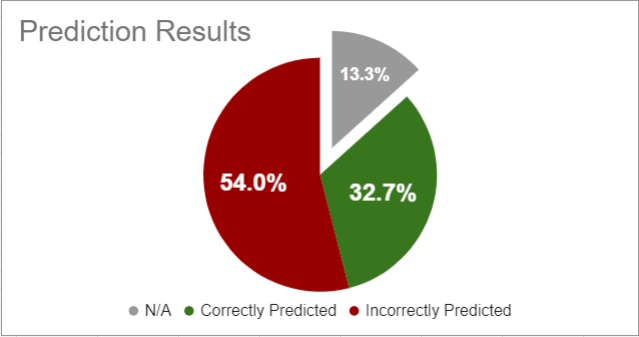
Initially, we asked ChatGPT to suggest a single emoji for a tweet. The prompt was as follows: “Suggest a food-related emoji for the following tweet … *[tweet text here]* … You must suggest exactly one emoji that is defined in the following emoji bank … *[emoji bank here]*”. ChatGPT was considered to be successful if the emoji it suggests is the same emoji that the human labelers attributed to that tweet. The results of this experiment are seen below, with “N/A” attributed to tweets that were not food-related, and thus disregarded from this experiment.



We observe that ChatGPT is mostly unsuccessful in this task, producing incorrect emojis approximately 3 times more often than correct emojis. This result was anticipated in our hypothesis, due to the subjectivity of the task.

* 1. **Three Emoji Response**

Since the human labelers only attributed one emoji to each tweet due to time constraints there are cases where there could feasibly be multiple correct responses. Thus, we repeated the experiment with ChatGPT suggesting three emojis, instead of just one. This is meant to give ChatGPT the opportunity to list multiple correct responses. For example, take the tweet *“@ThatAldenDiaz I gotta throw Nutty Bars and the Hostess Apple/Cherry pies in the mix. I know it's absolute garbage food but that garbage is delicious lol”*. In this iteration of the experiment, ChatGPT could predict both the candy bar and pie emojis. The results of this experiment are seen below.



ChatGPT performed much better in this iteration of the experiment. This is likely due to some ambiguous tweets being covered by the ability to suggest multiple emojis. However, it may simply be due to the fact that suggesting multiple emojis induces a statistically higher chance of suggesting a correct emoji. Regardless, results were much improved by this modification to the experiment, though still failing more often than not.

* 1. **Error Analysis**

This task sees a very high percentage of failure, even when allowed to suggest multiple responses. This is likely due to the ambiguity in assigning an emoji for many of the tweets scraped from Twitter. For example, take the tweet “*@CHKnyght Someday I'd like to know who figured out olives as a snack food. "Just for giggles, I'll take this super bitter thing from that tree, dunk it in lye, and then chunk it in a bottle of salt water. Yum."*“. The only food mentioned in the tweet is an olive. However, olives are not in the emoji bank. Thus, it was up to the human labelers to determine the closest match. However, if a task is difficult and unclear for a human to complete, it is no surprise that ChatGPT would struggle so immensely. Almost all of the tweets with this sort of ambiguity of not having the correct emoji in the emoji bank ended up incorrectly predicted by ChatGPT. This was the most common failure case for this task.

Perhaps a larger emoji set would be useful to remove ambiguity. However, due to time constraints, this was not feasible. Regardless, there would likely be tweets mentioning foods that do not have an emoji representation at all, such as quinoa, which was mentioned in one of the tweets.

1. **Discussion**
   1. **Replicability**

The replicability of this project is contingent upon several factors, including the specific version of ChatGPT utilized, the prompting techniques employed, and the quality of the dataset. During the experiment, ChatGPT occasionally deviated from the prompts, returning multiple emojis, emojis not found in the emoji bank, or even paragraphs of text. Due to the inherent complexity of ChatGPT, some level of randomness is likely unavoidable without further comprehension of its inner workings. Nonetheless, the results of the experiment were generally consistent across different runs when utilizing the same prompt, with the exception of occasional bugs. We would expect that with the same prompts, same datasets and same version of ChatGPT, others could easily reproduce the same results beside few deviations due to the inherent complexity of ChatGPT.

* 1. **Datasets**

Our tweet table containing 300 tweets had a few mishaps that caused some inefficiencies and limitations in our analysis. One issue that arose was that not all of our tweets were actually food related, despite searching up the keyword “food”. A tweet like, “I could go for some food right now, feeling kind of bitter after that Vikings game” is an example of a tweet that isn’t pointing at a particular food, but fits our search for “bitter food”.

Another issue with our dataset was that some tweets contained several foods that didn’t produce any specific sentiment. For example, the tweet, “I was raw alkaline for about four years. Single ingredients. Fruit, veg, nuts, seeds, whole grains. No vegan garbage. I did it because I understood the damaging effects of acidic food not for the animals or environment etc lol” doesn't produce any foods in specifics other than maybe nuts. Fruit and veg is incredibly vague and doesn’t properly give a single or few food emojis, but rather dozens that would all fit this tweet.

* 1. **Ethics**

The presented work has been recognized as harmless and posed no risk to society, as it is primarily aimed at evaluating ChatGPT's ability to understand, summarize, and interpret tweets and emojis. However, the results of this study could potentially be utilized by various social media tools designed to generate AI-based tweets to motivate them to include the capacity to tweet with emojis. This raises concerns regarding the possibility of fraud and dissemination of fake information as the use of emojis is an important indicator of the identity of the writer. Addressing this issue would require the involvement of social media platforms and the development of better tools to distinguish between AI-generated text and human-generated text, which fall outside the scope of this project.

* 1. **Limitations**

There are several limitations to our study. The most glaring limitation we encountered was that our bank of only 20 emojis wasn’t sufficient for several tweets. About 12% of the time, ChatGPT would straight up ignore our emoji bank and choose an emoji outside of our list of 20, simply because it made the most sense. This shows our emoji bank did not incorporate a wide enough variety of emojis. This is most likely due to many cultural foods such as ramen and dumplings not being properly represented. For example, a tweet such as “boy, do I need a nice warm Japanese dish to warm me up on this cold night” is clearly referring to a food such as ramen. In a more perfect scenario, we would most likely bump up our list to about at least 50 emojis.

Another limitation we encountered was that we used too few people to manual label our data. Only 3 people is a very small sample size and a much larger sample size would have given a much larger range of sentiment. Our 3 manual labellers are all University students around the age of 20, so there is not much diversity given those two aspects alone.

300 tweets also seems like a fairly small sample size. Snscrape only searches tweets in English and not foreign tweets in other languages. As mentioned previously, there are tons of foods such as ramen and dumplings that are more common in foreign cultures and languages. Moreover, the use of emoji to represent food items can also be limited. Different cultures may have different interpretations of what a particular emoji means or may use a different set of emojis to represent the same food. Therefore, it is important to consider cultural context when interpreting the results of the project.

Lastly, there technically is no singular correct answer since sentiments can always vary. A basic tweet such as “the food was very cold” could relate to many different foods such as a cake, ice cream, or even a sandwich. ChatGPT’s ‘accuracy’ is only dependent on our human based evaluation, which as mentioned, isn’t very diverse to begin with. In a perfect situation, there should be no limit to how many emojis are appropriate as a response.

* 1. **Future Research**

Multiple studies can be conducted based on this presenting work. First, the complexity of labeling a large number of tweets within a short period of time with limited human resources has restricted the scope of the scraped tweets. As an illustration, tweets were labeled with only one food emoji from the available emoji bank to minimize the workload. However, this approach also restricts the performance of ChatGPT since some tweets cannot be accurately described by a single emoji. To address this limitation, future studies could consider extending the food emoji bank and using multiple emojis to summarize a sentence to evaluate ChatGPT's ability to correctly understand and summarize a sentence.

Moreover, we have observed that some tweets retrieved by the current keyword search cannot be described using the available emoji bank or even the existing emoji system. For instance, we didn’t include wine in our emoji bank, but it showed up often in the tweets we scraped. We’ve noticed that ChatGPT can successfully capture the characteristics of wine and use a grape to summarize the sentence. This discovery highlights the importance of evaluating ChatGPT's inferential ability. Future studies could incorporate emojis representing raw materials such as milk, greens, and fruits, while analyzing tweets related to processed foods such as salads, cheese, and jams. This would require ChatGPT to identify the important foods mentioned in the sentence. Then using reasonings and inference to find the corresponding raw material(s).

**Acknowledgments**

This report is the final project for CSCI 5541 during Spring 2023 at the University of Minnesota, Twin Cities. Many thanks to Risako Owen for her constant support and care throughout the project, Professor Dongyeop Kang introducing and teaching us the world of natural language processing, and Shirley Anugrah Hayati for providing previous research within this field that served as a base for our project.

**References**

1. **Appendices**
2. **Supplementary Material**

