happy sad sleepy mad

an analysis of the usage of emojis

ksang1 mkimurat mwu27 yyao38

Introduction

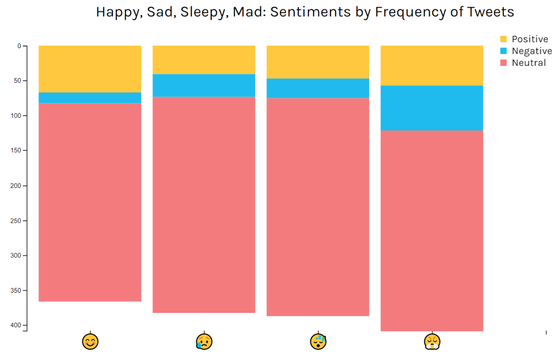
Emojis have become an increasingly popular form of communication that transcends language. Emojis appear to be universally understood, but can still carry a notion of ambiguity. We wanted to explore the nuances in how emojis are used across different languages, as well as the overall sentiments of emojis themselves.

Data

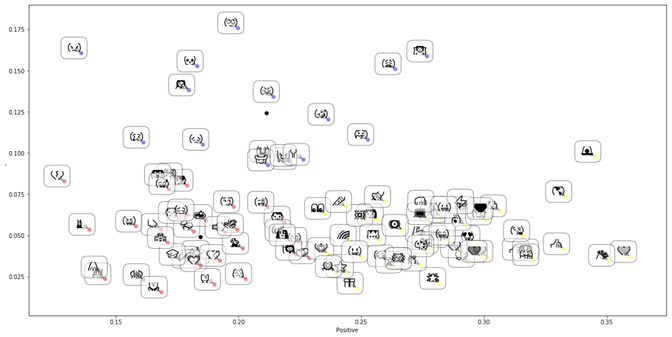
We scraped data about emojis, descriptions, categories, and UTF-8 codes to anticipate how we were going to analyze our data. We also scraped 250,000+ tweets from Twitter over the course of March and April.

Sentiment Analysis

We used TextBlob sentiment analysis to classify tweet sentiments. This became the launching point for our visualizations, where we examined the frequency of the sentiment tags to see if we could find any patterns. Then, we used K-means analysis with three clusters to classify the sentiment of each emoji.

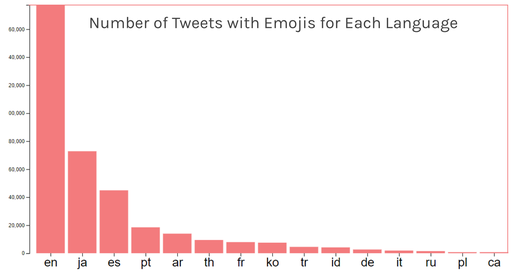


Sentiments of Emojis

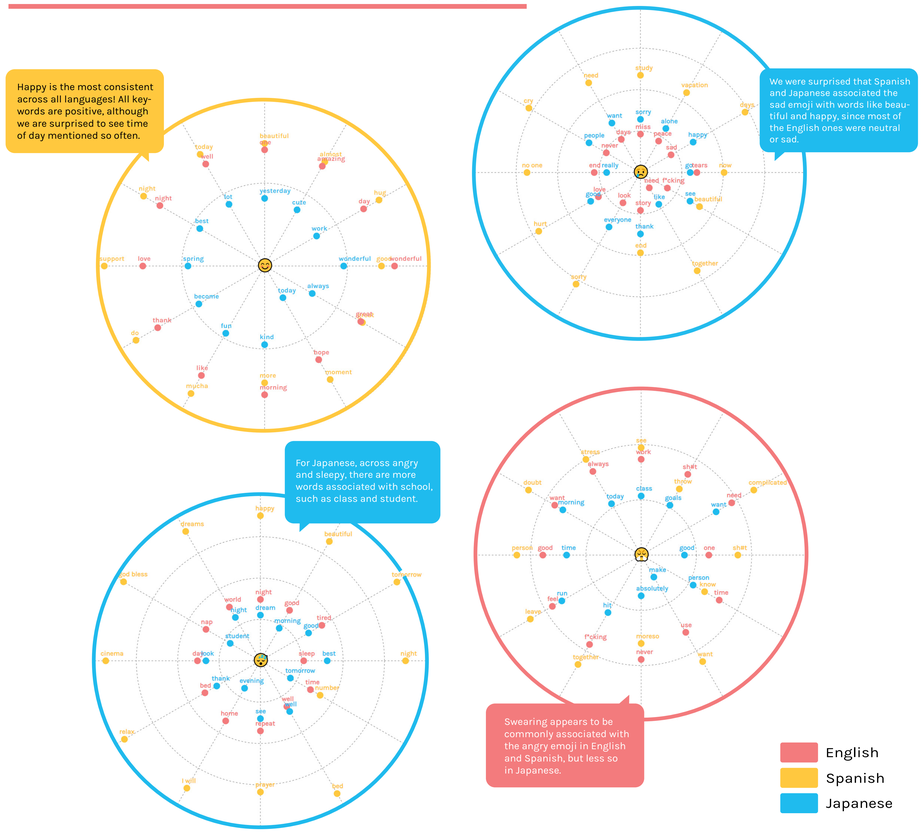


Emojis in Different Languages

We explored word associations with emojis across different languages. First, we determined which languages most commonly used emojis (English, Japanese and Spanish), indicated by the graph below. To analyze the word associations of the emojis, we only kept the target emoji, created word vectors using a context window of 4, and determined which vectors were closely related to each emoji. In our visualizations below, the cosine distance is illustrated by each word's distance from the center.



Common Words Used Across Languages



Challenges

Most of our problems were related to the lack of technological support for emojis.

Data collection and cleaning

- Readability and representation of emojis in database (emojis can have multiple Unicodes)
- Spam tweets and tweets incorrectly tagged as English

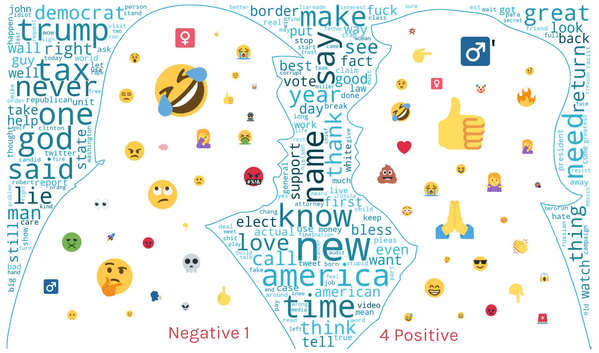
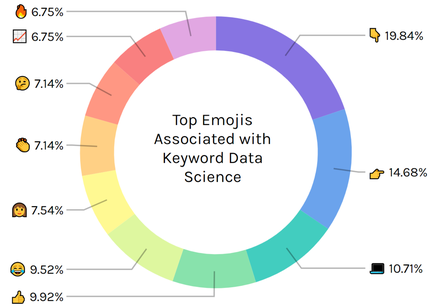
Processing and analysis

- Displaying emojis with Matplotlib
- Regex on emojis for processing tweets
- Translating tweets into English with emojis

Conclusion

In our sentiment analysis, we were limited by the Textblob, which was more inclined to tag tweets as neutral. However, we were still able to discover some interesting findings, such as the 😂🤣 emojis being tagged as negative.

In regards to language, we were able to find nuances in how different languages used the sad emoji. English users mostly used negative words with 😢, while Spanish and Japanese users used more positive keywords.

💩s and giggles

Some fun associations we found!