#mood: The Emoji Context Project

Caroline Locke, Ryan Flint, Chris Muncey, Bryceton Bible, and Manny Bhidya

***Abstract*— This data analysis project uses Twitter posts to extract mood and sentiment of emojis based on the context in which they are used. By scraping Twitter data and comparing it against a list of compiled mood words, we were able to rank emojis by overall "sentiment" and mood as well as group together emojis that have similar mood and meaning.**

I. Objective

#Mood: The Emoji Project seeks to find the underlying meanings of emojis based on the words that typically accompany them. Most emojis have a face value that can be interpreted by just viewing the image — for example, the smiley face represents a smile or happiness; however, language mutates rapidly on social media and so the underlying meanings and usage of these emojis changes with current trends and slang. Sometimes, one must be engrossed in an internet subculture or social platform to be able to fully understand the true meaning of a message and the emoji used because of the layers of subtext that have been created in these bubbles. This project seeks to learn if the underlying meanings of emojis can be extracted purely from analyzing the surrounding text without being in the social media bubble in which the meaning is being created. By collecting words that commonly appear with an emoji and assigning a mood word or connotation to it, it may be possible to interpret these underlying lying meanings just from the textual evidence. This project will determine if this is a viable way to understand the ways in which users enhance their communications with emojis.

II. Motivation

In today's world, a large proportion of communication takes place electronically — via social media, text messages, email, forums, websites, etc. Although emojis aren't typically used in professional or academic electronic communications, they are nearly ubiquitous in casual settings. From the dawn of the internet, users have created ASCII-based faces and caricatures meant to communicate a feeling or idea i.e. the smiley [:)], the surprise face [:O], etc. Overtime, these ASCII figures have evolved into "emojis" — small pictures representing everything from face emotions, to animals, to blood types. They have become integrated into the way we communicate with our peers and are available for use in most texting and communication apps. Text based communication is rife with misinterpretations and miscommunications due to lack of tone or context, but emojis help add back the emotion that is lost in translation. Because of the vast amount of emojis (with more being added all the time), users are finding new and interesting ways to use these pictures to emphasize their words, clarify meanings and enhance their words. Because of social media, emojis can take on brand new meanings with elaborate layers of context that requires the reader to be "in the know" in specific ways to fully understand the meaning of both the emoji and the contextual words. Many emojis often have multiple contextual meanings alongside their face value meaning. This project attempts to unearth these hidden meanings through textual analysis of the words accompanying these emojis to try to understand the mood or connotation being imbued by the users.

III. Methods

This section discusses the methods in which data was obtained and how it was analyzed.

A. Data Collection

The data analyzed for this project was collected via Twitter. Twitter is one of the major social media sites and a hub of emoji usage. Many viral memes, videos, quotes, and jokes come from this site. With over 300 million active users, this source provides a large sampling population. Twitter also has an easy to use API for scraping the data. Before scraping, we narrowed down the thousands of emojis currently on devices down to the top 50. We referenced the website "Emoji Frequency", which displays all emojis in circulation in order of frequency in which they appear on various social media sites. From there, we took the top 50 to plug into our collection script. To collect the data, we created a Python script that searches Twitter using the API connection and collects English language tweets that contains these target emojis. We ran two instances of the script each looking for 25 emojis each. The scripts ran for 8 hours each collecting a total of 16 hours of data. We collected nearly 5 million tweets overall. The tweet data was formatted and placed in a file to be parsed using Python analysis tools.

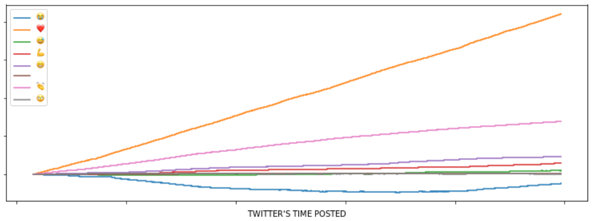
B. Data Processing

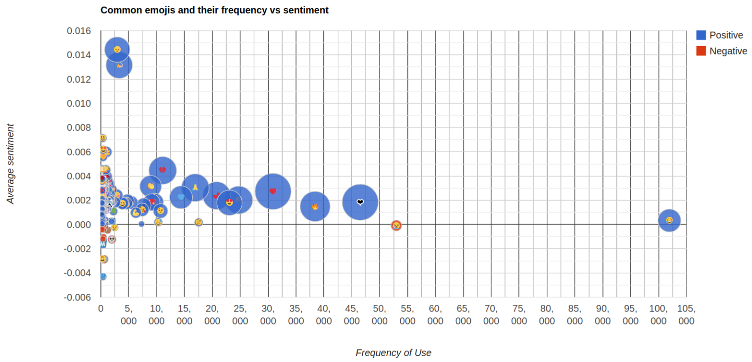
Once the data collection was complete, we began to parse and analyze the data. If multiple emojis appeared in a tweet, the words in the tweet were included for each emoji. We did not analyze any stop words (like 'of', 'but', 'and', and any slurs). We first processed the tweets by identifying emojis and calculating a sentiment score. To calculate the sentiment of a tweet, we utilised a positive and negative opinion lexicon generated by Minqing Hu and Bing Liu's "Mining and Summarizing Customer Reviews." and Bing Liu, Minqing Hu and Junsheng Cheng's "Opinion Observer: Analyzing and Comparing Opinions on the Web." We then use sci-kit-learn's TF-IDF, Term Frequency Inverse Document Frequency, Vectorizer algorithm to transform the tweet corpus and positive and negative lexicons into a matrix of TF-IDF features. The sentiment for a tweet is finally derived by taking the difference of the tweet's cosine similarity to the positive and negative lexicon vectors. After calculating sentiment scores for tweets in a corpus sample of half of a million, we can determine the emoji features including counts, average sentiment, maximum and minimum sentiment, the standard deviation of sentiment, and the 25 ,50, and 75 quantile values. The emoji features are then processed using sci-kit-learn's K-Means clustering algorithm to group emojis into seven distinct groups. Sci-kit-learn's Principal Component Analysis also processes the emoji features for Dimensionality Reduction, the decomposed features and cluster labels are then plotted onto a two-dimensional graph for analysis. The final processing step was to generate a data frame of chronologically ordered tweets with emojis and a cumulative sentiment score. With these preprocessing steps complete, we generate the following visualisations for analysis: (see Appendix for corresponding item numbers and the images associated with those descriptions)

1) An Emoji cluster chart based on the calculated features and sentiments.
2) Emojis grouped into seven distinct groups.
3) Bar charts for all Emojis visualising count, minimum sentiment, maximum sentiment, average sentiment, sentiment deviation, and 25, 50, and 75 quantile values
4) Bar charts for the top 25 Emojis visualising count, minimum sentiment, maximum sentiment, average sentiment, sentiment deviation, and 25, 50, and 75 quantile values
5) Bar charts for the bottom 25 Emojis visualising count, minimum sentiment, maximum sentiment, average sentiment, sentiment deviation, and 25, 50, and 75 quantile values
6) Pie charts for the top 25 Emojis visualising count, minimum sentiment, maximum sentiment, average sentiment, sentiment deviation, and 25, 50, and 75 quantile values

IV. Results

From our sentiment analysis, we found that the "heart" emoji by far had the most positive aggregated sentiment (i.e., accounting for emoji use frequency). At a far second, we had the "hand clapping" emoji. The most negative aggregate sentiment was associated with the crying emoji. Interestingly, the crying emoji was not as negative as expected. Over all tweets collected, it trended back towards the other emojis. This shows that the emoji may have other uses besides displaying sadness, but further analysis would be needed to parse the exact usage. When analysing charts that measure average sentiment (through the interactive chart or in the "clustering" set of charts) a binary sentiment is less clear. The frequency of use of an emoji seems to correlate to a more "neutral" sentiment. Emojis used less frequently seem to have a more extreme sentiment. Additionally, with the exception of the "crying" emoji, all emojis classified as "negative" (sentiment of less than 0 on average.) Excepting the "crying" emoji, the most frequently used negative emoji is used more than 40 times less often than the most popular positive emoji.Another interesting finding was the neutrality of the "crying laughing" emoji. Although it's meant to convey an emotion of something being so funny, the reader is pushed to tears, it is ranked as only slightly "positive." We hypothesize that this emoji is frequently used to set the tone of a message as being friendly and used as "punctuation." It is one of the most frequently used emojis for this reason. Anecdotally, the team members have all used this emoji or seen it used at the end of sentences without actually meaning "crying with laughter." More research would need to be done to verify this hypothesis.





V. Primary Issues

The first issue we encountered was deciding the length of time used to collect data. After our first trial run, we collected hundreds of tweets in just a 20 minute time span. Our run of 16 hours yielded more than enough tweets to analyze; however, this also limits our ability to extrapolate meaning since the mood or sentiment could change over time. More information about emoji meanings could be extracted if we ran the collection over a longer period of time. Another issue was the problem with finding a font that supported emojis. Surprisingly, Python had no trouble accepting emojis as search parameters for twitter and yielded the correct result. The issue came when attempting to generate charts and graphs using Python libraries. After implementing a few workarounds, we were able to display the emojis in our charts and graphs. Finally, we struggled with narrowing down what information to parse and display for our final analysis. The twitter scraping gave us an overwhelming amount of data to work with and there are many, many ways to sort, analyze and display the information we collected. Ultimately, we landed on grouping emojis that are similar and analyzing the overall "mood" or sentiment, since that was closest to our original intention and proposal.

VI. Future Work

This project laid the groundwork for many other interesting future projects. One such project would be collecting tweets in other languages and comparing the results of the emoji sentiment across languages. A more fine-grained data collection would allow us to see the exact contextual meaning of the emojis collected beyond the general mood or sentiment of the tweets. More data collected over a larger time span would aid in this analysis. With enough analysis and data collected, we could begin to assign emojis to certain words based on how closely they are associated with those words in the online vernacular usage. One use of this data would allow users to write sentences and tweets and offer suggested emojis to "complement" the overall message or tone. Another use would allow users to type in a string of emojis and have the site translate it into an English sentence.

VII. Organization Chart

All members wire responsible for project design including selecting emojis to scrape for, final assignment of moods and meanings, and evaluation of resource as well as some coding and parsing of the script. Each member will was also tasked with specific elements of the project

- Caroline Locke: project management (including documentation, timeline management, and planning), final data analysis, written report
- Bryceton Bible: obtaining resources for emojis and mood/connotative word lists, parsing resultant data, graph generation
- Ryan Flint: twitter scraping, API maintenance, scripting
- Chris Muncey: twitter scraping, API maintenance, stop word list creation
- Manny Bhidya: parsing resultant data, displaying the associated words data meaningfully via graphs/visuals

A. Timeline

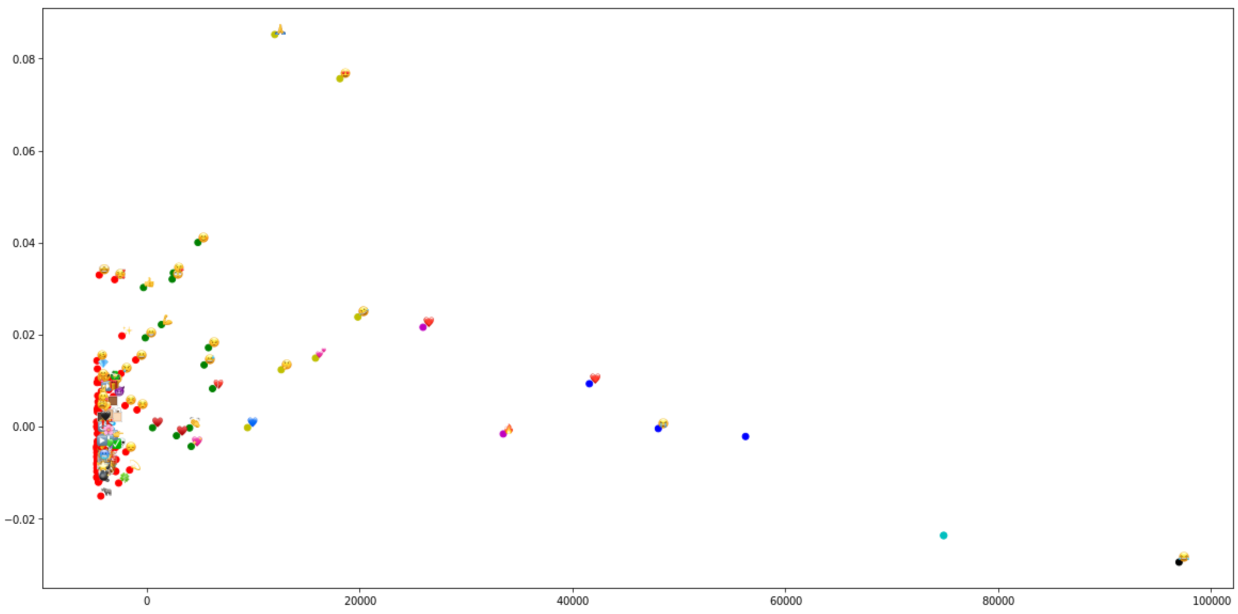
Below is a timeline for the completion of this project:
10/10: Scraping bot completion
10/13: Emojis chosen for bot
10/14 - 10/31: Collect tweets via scraping for all of the emojis.
10/20: Mood/connotation word sources selected
11/1: Began scripting to parse out words for each emoji
11/11: Finished parsing and displaying data via Python
11/15: Began assigning moods/connotations from collected data
11/20: Document findings and complete final report
11/24: Final presentation

VIII. Summary

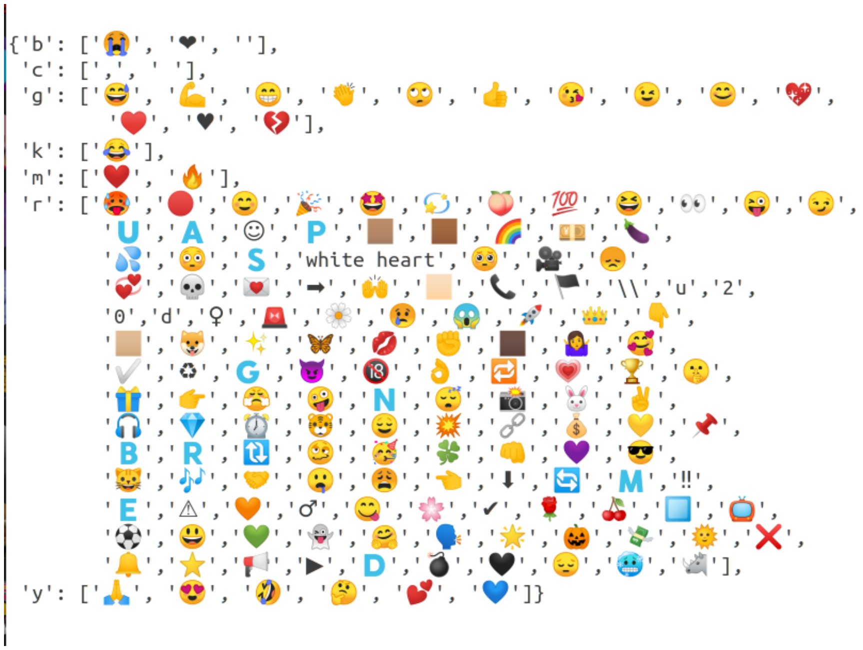
In conclusion, in this project sentiment analysis was used to classify popular emoji as "positive" or "negative" by analysing the contexts in which they appear, to gain insight into the perhaps unexpected ways that emoji are used. By collecting hundreds of thousands of tweets, we can extrapolate our findings to the general uses of these emoji in the English language as a whole. These findings could be leveraged by companies or researchers to gain a further insight into the rapid changes in human communication through online platforms. Overall, though the scope of this project was limited to a number of the most commonly used emoji, it would be trivial to expand these concepts to larger sets of emoji or related concepts such as emoticons or kaomoji. Further, sentiment data used for classification could be updated as trends change or limited retroactively to provide an insightful glimpse into the past. However, the set used encompasses the majority of emoji used in daily life, and thus succeeds in providing a thorough understanding of emoji's uses in and effects on English communication.

IX. Appendix

A. Item 1

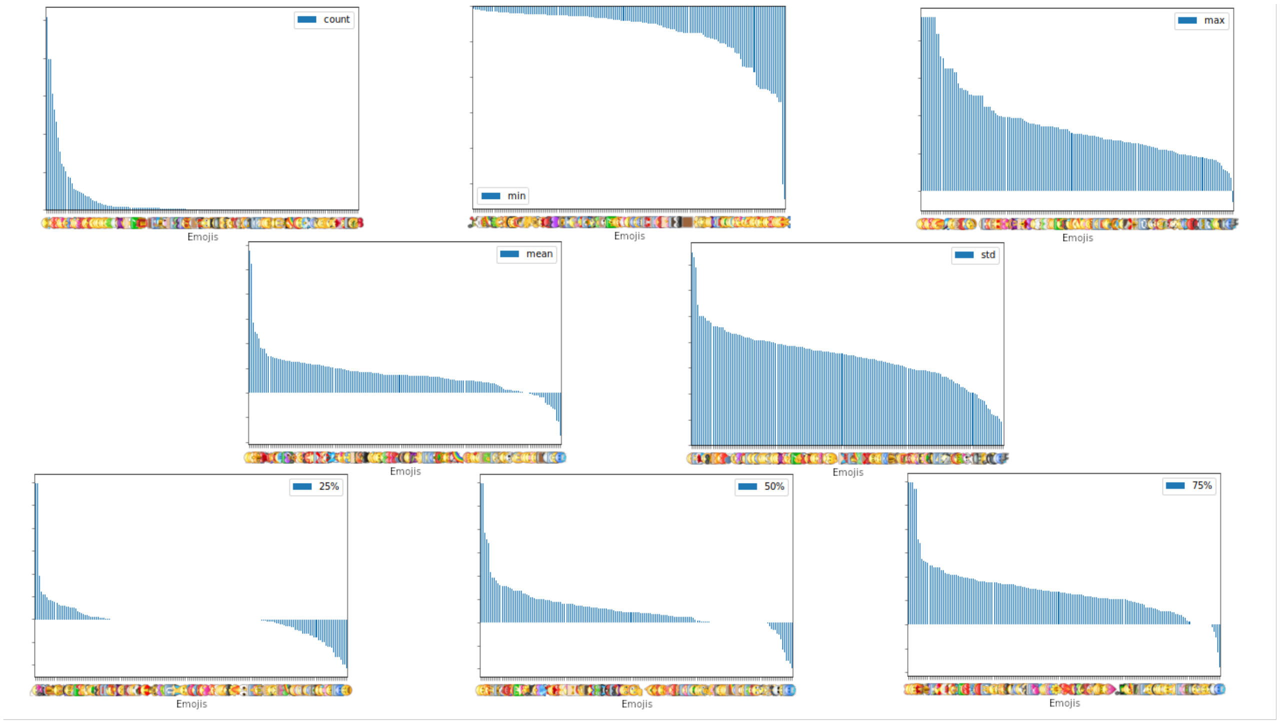


B. Item 2

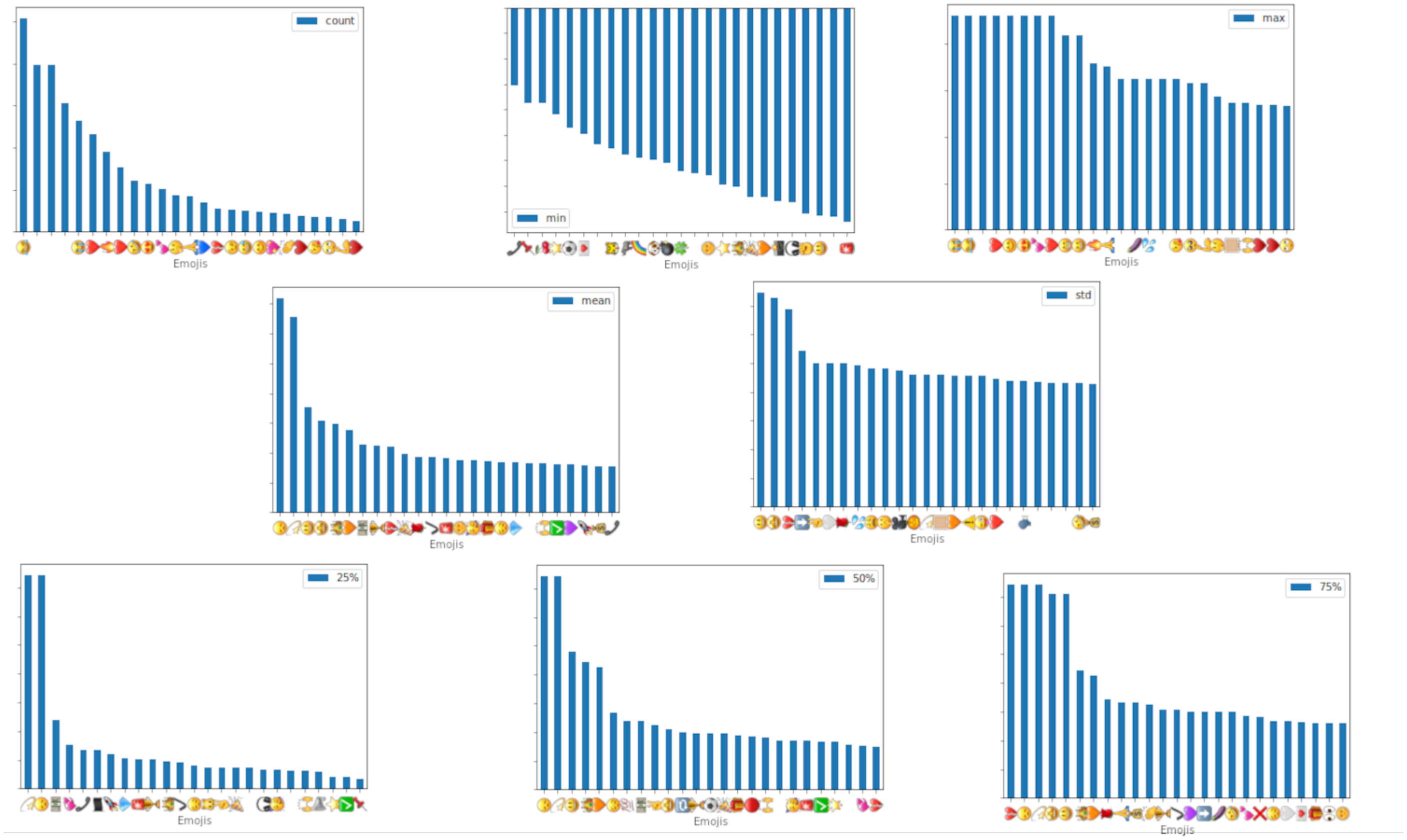


```
{'b': ['😭', '❤', ''],
 'c': [',', ' '],
 'g': ['😅', '💪', '😁', '👏', '🙄', '👍', '😘', '😉', '😊', '💖',
       '❤️', '♥', '💔'],
 'k': ['😂'],
 'm': ['❤️', '🔥'],
 'r': ['🥵','🔴','😊','🎉','🤩','💫','🍑','💯','😆','👀','😜','😏',
       '🇺','🇦','☺','🇵','🏽','🏾','🌈','💴','🍆',
       '💦','😳','🇸','white heart','🥺','🎥','😞',
       '💞','💀','💌','➡','🙌','🏻','📞','🏴','\\','u','2',
       '0','d','♀','🚨','🌼','😢','😱','🚀','👑','👇',
       '🏼','🐶','✨','🦋','💋','✊','🏿','🤷','🥰',
       '✔️','♻','🇬','😈','🔞','👌','🔁','💗','🏆','🤫',
       '🎁','👉','😤','🤪','🇳','😴','📸','🐰','✌',
       '🎧','💎','⏰','🐯','😌','💥','🔗','💰','💛','📌',
       '🇧','🇷','🔃','🥴','🥳','🍀','👊','💜','😎',
       '😺','🎶','🤝','🤤','😩','👈','⬇','🔄','🇲','‼',
       '🇪','⚠','🧡','♂','😋','🌸','✔','🌹','🍒','🟦','📺',
       '⚽','😃','💚','👻','🤗','🗣','🌟','🎃','💸','🌞','❌',
       '🔔','⭐','📢','▶','🇩','💣','🖤','😔','🥶','🦏'],
 'y': ['🙏', '😍', '🤣', '🤔', '💕', '💙']}
```

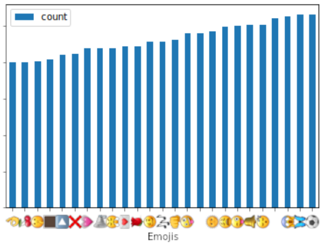
C. *Item 3*

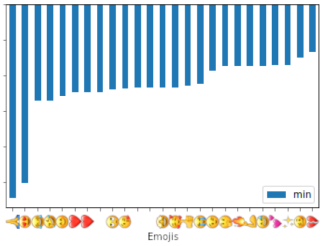


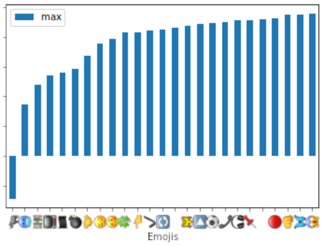
D. *Item 4*

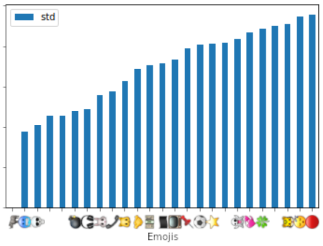


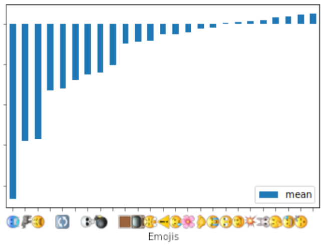
E. Item 5

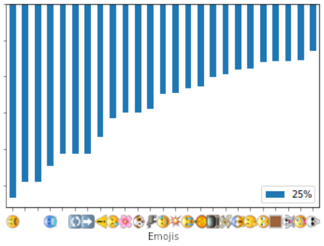


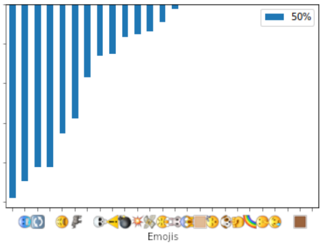


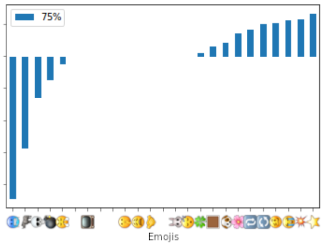












F. Item 6