M3-1 Hypothesis

Group Name: SEM
Group Leader: Emily Feng (ejf9kwf)
Group Members: Mei Gilhousen (mlg8jc), Samarth Saxena (ss9nfb)
DS 4002, 11/2/2022

Narrative Paragraphs:
In "Differences in Emoji Sentiment Perception between Readers and Writers", an experiment was conducted that looked at the emoji sentiment of authors versus that of writers. Berengueres and Castro found that readers and writers broadly agree on the sentiments of emojis. The largest discrepancy between readers and writers occurs for negative emojis, such as the sad face emoji. We would like to see how emoji sentiment perception differs from country to country [1].
Even though emojis are widely used to provide a reaction through digital communication, it is still better to keep in mind that some meanings of emojis are not universally accepted. In China, the regular smiley face  or  is reserved for bosses. But in other countries, it usually conveys a wide range of positive sentiments. Moreover,  represents "bye" in most of the countries. However, in China, it means "I don't want to see you anymore." Based on the annual report of the country's most popular social app, it is found that emoji use is not only different by the users' generation, but also by geography and culture. [2] Inspired by two articles mentioned above, the team would like to analyze the sentiment of the top 50 most used emojis in the world.

Hypothesis:
In the top 50 most used emojis in the world, at least 50% of them denote more positive feelings than negative feelings, with a sentiment score above 0 on a scale from -1 to 1. Additionally, the correlation between the popularity of emojis and their sentiment scores are greater than 0.5.

Modeling Approach:
The initial type of model we are going to use is a decision tree model for feature selection. We decided to use this model because the model is easily applicable to our situation and dataset, as the decision tree model returns a feature ranking [3]. We would like to make a decision tree model based on the sentiment of each emoji, and let our model tell us which emojis have the highest sentiment difference [4].

References:
[1] Berengueres, Jose."Differences in Emoji Sentiment Perception between Readers and Writers," *arxiv.org*. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1710/1710.00888.pdf. [Accessed: 02-Nov-2022].

[2] Borak, Masha. "In China, the smiley face emoji does not mean what you think it means," *Tech in Asia*, 11-Jan-2019. [Online]. Available: https://www.techinasia.com/smiley-face-emoji-meaning-china. [Accessed: 02-Nov-2022].

[3] H. Patel , "What is feature engineering — importance, tools and techniques for ...," *Towards Data Science*, 30-Aug-2021. [Online]. Available: https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10. [Accessed: 02-Nov-2022].

[4] Akhil. "Feature selection using Decision Tree – Busigence." *Busigence*, 21-Oct-2016. [Online]. http://busigence.com/blog/feature-selection-using-decision-tree. [Accessed Nov. 02, 2022].