Case Study Title: Emoji prediction from social media

Background:

Emojis are extensively used and are the evolution of character-based emoticons (Pavalanathan and Eisenstein, 2015). They are powerful for expressing ideas of topics such as food, mood, place, etc. Suppose that your goal is to have emoji suggestions when a user is posting a tweet.

Current solution:

Currently, emoji can be recommended based on a single word meaning for iMessage. For example, if the user types “pizza”, keyboard will suggest a pizza emoji. According to Emojipedia, the Face with Tears of Joy is the most popular emoji used in 2018. However, this emoji can be interpreted into different meanings such as amusement or embarrassment based on the context.

Challenges:

Many emojis have constantly evolved across cultures over time. There is little to no computational model that can understand emoji with whole content (Barbier et al., 2018). Other challenges include extracting, processing, and analyzing the data from social media channels such as Twitter. Scrapping real time data is usually resulting in a highly imbalanced data set. Finally, the raw data from twitter is usually noisy and contains many unnecessary information.

Needs:

Researching the connection may help sentiment analysis and NLP tasks such as information retrieval (Novak et al., 2015). Moreover, using emoji as a query for searching content that doesn’t contain emoji could be very powerful.

Additional information: data cleaning, filtering and preparing for training/testing

There are a few steps to clean the data. First, filter out tweet that are not English (or in the target language). This may lead to reduction of 50% or more. Then, you should drop duplicate tweet and rows that have missing values. Moreover, make sure to extract emoji and map to an index number as this will be your ‘response’ column for training/testing data. Some tweets have more than one emoji. Initially, you can select only tweets that contain one targeted emoji, which will

result in another 80% reduction of the data. However, after the initial system has been set up, you may want to consider tweets with more than one emoji.

Sample questions:

1) What are some NLP and text analysis techniques that could be used to fulfill the needs highlighted in the case?
2) How would you evaluate these methods? What are some metrics and paradigms that you could draw upon?
3) Imagine that you are an advertiser. Could you potentially use emojis or emoji predictions to better target your potential customers? In what cases might that not be useful? Give specific examples of companies and industries that come to mind for either extreme.
4) Are emojis still relevant in an age of video, such as YouTube and TikTok? Why or why not?
5) What is one reason advertisers might prefer emojis over simple linguistic keyword-based cues? *Hint: are emojis very tied to language or culture, or do you think there is something universal about some emojis?*
6) Design an architecture for solving this problem based on your knowledge of text analytic pipelines and supervised machine learning.
7) Why do you think the suggestion to initially consider tweets with only one emoji is a good one? What complications arise if we consider tweets with many emojis?
8) Considering the above, should we further filter so that we consider tweets with only one emoji *instance* or is it okay to have one *unique* emoji, even if that emoji occurs multiple times in the tweets?

References:

Barbieri, Francesco, Ballesteros, Miguel, Ronzano, Francesco, Saggion, Horacio. Multimodal Emoji Prediction. arXiv.org. April 2018. http://search.proquest.com/docview/2072033947/.

Li X., Yan R., Zhang M. (2017) Joint Emoji Classification and Embedding Learning. In: Chen

L., Jensen C., Shahabi C., Yang X., Lian X. (eds) Web and Big Data. APWeb-WAIM 2017. Lecture Notes in Computer Science, vol 10367. Springer, Cham

Francesco Barbieri, Luis Espinnosa-Anke, Jose Camacho-Collados, Steven Schockaert, adn Horacio Saggion. 2018. Interpretable Emoji Prediction via Label-Wise Attention LSTMs. Association for Computational Linguistics. P.4766—4771

Petra Kralj Novak, Jasmina Smailovic´, Borut Sluban, and Igor Mozetic˘. 2015. Sentiment of emojis. PloS one, 10(12):e0144296.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Emoticons vs. emojis on Twitter: A causal inference approach. arXiv preprint arXiv:1510.08480.

Emoji Sentiment Map. (n.d.). Retrieved October 25, 2019, from http://kt.ijs.si/data/Emoji_sentiment_ranking/emojimap.html.

Emojitracker (n.d.). realtime emoji use on twitter. Retrieved October 21, 2019, from http://www.emojitracker.com/.