Project Proposal

Project title: Cross Cultural Social Media Prompting for Japanese and American Audiences

Authors: Jordan Otsuji and Chris Stevenson

The scope of projects must be notably different from homeworks. E.g. just gradient-based attribution over text even if task is not sentiment classification is not sufficient.

What are you trying to do? (150 words max)
[*This serves as an overview of the proposal. A high-level description of the problem and the gap in the literature you are aiming to solve (without mentioning prior works, you'll write that next), and a high-level description of your solution.*]

Explanations tailored to US and Japanese cultural groups may differ. We wish to explore the difference between explanations of US Tweets produced by prompts tailored to US and Japanese audiences. These Tweets require an understanding of American culture to understand. Therefore, if the model's performance suffers when prompted to provide explanations to a different cultural group, then the tool is less useful or potentially harmful to that group.

How is it done today, and what are the limits of current practice? If someone (you? 😺) addresses identified limits, what difference will that make? (150 words max)
[*The most critical themes/works related to the problem you aim to solve. Do not write a comprehensive literature review. From here it should be clear that there is a notable gap in the literature that must be addressed.*]

Automated translation between languages is often inaccurate, and never provides cultural background that might be necessary for people of other languages and cultures to understand. Manual translation will likely be accurate and may include background information, but takes a long time and requires a human translator.

What is new in your approach and why do you think it will be successful? What motivated your choice to incorporate explainability into your approach for this problem? In other words, what limitations or drawbacks are associated with using methods that lack explainability? (150 words max)
[*Here you can provide more details of your approach and illustrate why this is a viable path to solving the gap above.*]

Our approach of using LLMs to explain the cultural background of different media will theoretically provide the accuracy and thoroughness of a manual translation, with the speed and accessibility of machine translation. If this method can be thoroughly explored and the reliable

use cases can be defined, LLMs can be a powerful tool for other cultures and underrepresented communities to comprehensively understand media from other cultures.

What are the midterm (after 2 weeks) and final exams (after a month) to check for success? If a user study does not fit as evaluation for your problem, explain why. (150 words max)
[*What is the concrete evaluation after half of the project period (on Nov 2)? What are all possible outcomes of the mid-term evaluation? What is the concrete evaluation for the final check (on Nov 22)?*
If you are introducing explainability because you argue it will help people achieve something better, you must conduct a user study on Nov 22. If you are a team of undergraduates alone you need to prepare a user study templates in the class session dedicated for this, but you do not need to undertake the study if it turns out you do not get the needed model outputs on time.]

By November 2, we expect to have infrastructure in place to facilitate experimentation with the prompts of interest.

By the final evaluation, we expect to have analyzed roughly 50 examples in both English and Japanese. We expect to conduct analysis on the results to provide quantitative information regarding the problem of context-dependent translation of Tweets.

A plan of how the workload will by divided among teammates. (150 words max)
[*You will used a shared github repo. I will check commits later to see how much everyone contributed to the codebase. It is not possible that one person implements, while the other prepares reports and the poster.]*

The bulk of the work will be compiling the dataset of tweets and other media and reviewing each generated explanation using a rubric that we will define, and these tasks can be split evenly between both team members. Jordan will score (and translate into English if needed for peer reviews) the Japanese explanations, and Chris will score the English ones.

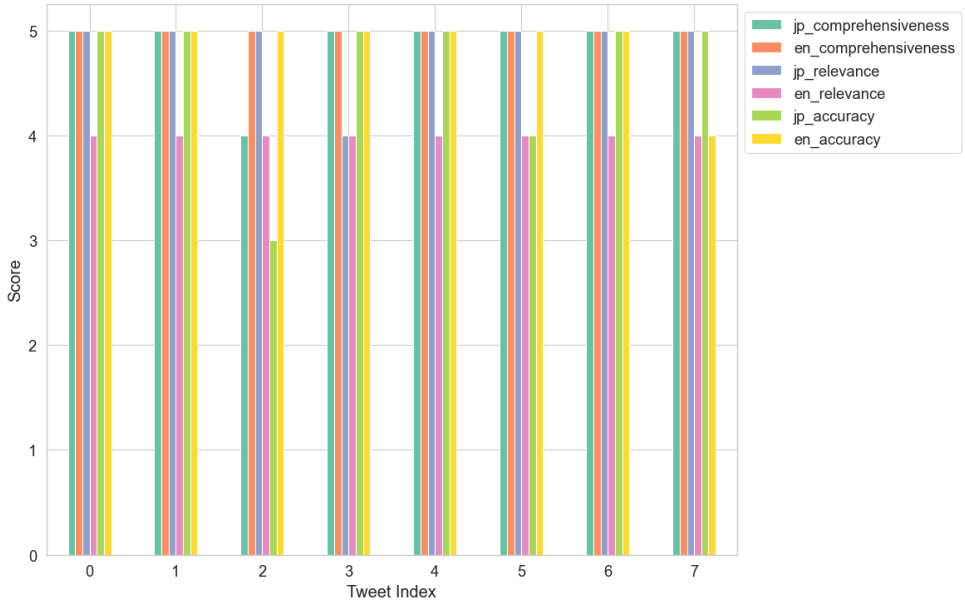
Two-Week Report

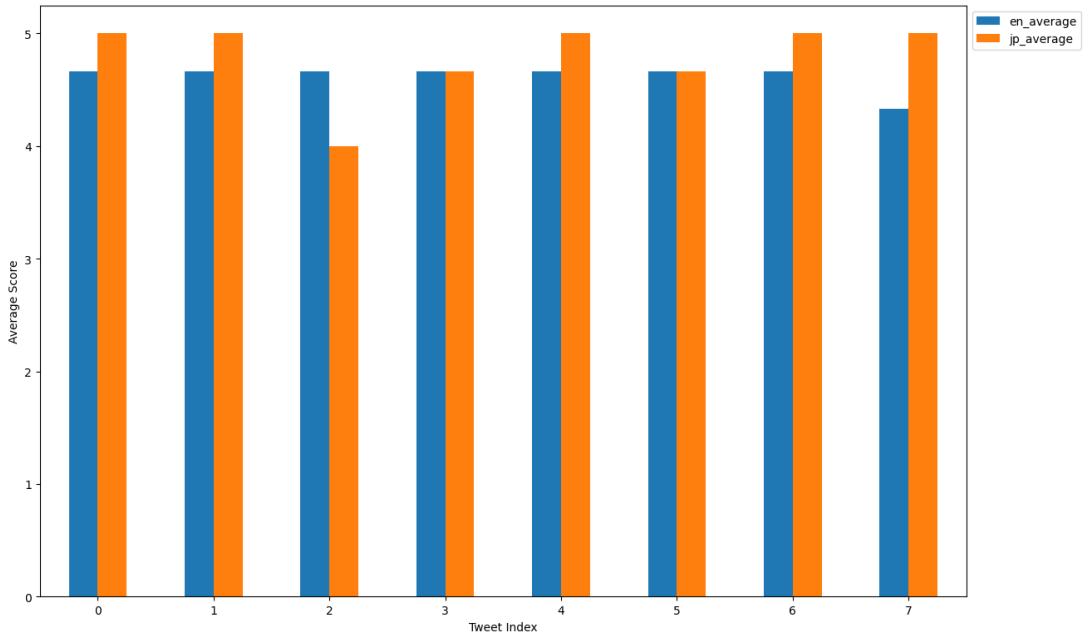
Repeat here your planned evaluation measurement for the two-week report or the one that the instructor has suggested. If you realized that another measurement would be more appropriate, additionally describe it here, while still repeating what you initially planned. (max 150 words)

At two weeks we will have the prompts we will use for our evaluation (requesting explanations for tweets in English and Japanese for their respective audiences) and test them on a small set of manually selected tweets that require some cultural background knowledge to understand. These queries will also be tested on various models, and each assigned a satisfaction score based on comprehensiveness, relevance, and accuracy (whether or not the model interpreted the tweet correctly). We plan on each of these categories to be assigned a score from 1-5

If you managed to conduct that evaluation, share your results (a table or a plot) and comment on what the result tells you about the problem you aim to address so far. If you didn't manage to conduct the evaluation move on to the next question. (max 200 words without tables/plots)

We managed to sample 10 tweets from Elon Musk's twitter account that we decided were appropriate for testing (tweets that require some kind of background information to understand) and we queried Claude (an anthropic LLM model) for explanations in English and Japanese. We assigned each response with a score of 1-5 for overall satisfaction.





As expected, the models generally do a great job at explaining the tweets and providing all of the necessary background information, and only make occasional false assumptions and interpretations. An ANOVA test reveals that the only metric whose mean varies across Japanese and English judgments is “relevance;” however, the English scores included the fact that the model we tested on consistently included irrelevant info about copyright and were

penalized while the Japanese scores did not include that, this discrepancy will be corrected in the future.

Due to changes in Twitter's API, scraping tweets has become very difficult and we're attempting to find solutions through other scraping methods or existing datasets.

If you haven't managed to get concrete outcomes yet, describe what challenges you encountered that prevented you from achieving them. (max 150 words)

N/A

Check whether you need to revise your final evaluation, and if so, write what you aim to quantify in the following two weeks instead. (max 150 words)

We plan on revising the metrics we plan on using to evaluate the degree of satisfaction of each explanation produced by the models. We will also explore different prompts in the future such as prompts including more information about the tweet such as date. Our current prompts seem to function well with the models we've done preliminary testing on, but the prompts are simple and there is room for improvement.