I have been sitting on a class on Thursday mornings this quarter so I have a standing conflict with tomorrow morning's RP meeting. At present I plan on attending the class so will miss the meeting. If you would like me to be in the meeting I can skip the class without too much hassle so please do let me know if you would prefer that.

In lieu of presenting some updates in person, I've decided to write them down here.

1 On BART

I have run the sample of articles that everyone is assessing by hand through the BART evaluation model. I have attached the results as .html documents to this email. I would not recommend looking at them to avoid an anchoring bias. Some details below:

- The articles are in the same order as in the Google Sheets spreadsheet (link : here)
- Recall that BART returns results as the probability that a text segment is similar to a higher-dimensional representation of the labels that we provide it (in this case "Positive," "Neutral," and "Negative") even though it was never provided these labels/topics when it was trained (by Facebook).
- As users of this model we have to trust the black box that the higher-dimensional representation of "positive" actually resembles something that a human would identify as being semantically similar to "positive."
- We could alternatively provide different labels. For example, I tried "Pro-business," "Neutral" and "Pro-labor" but a cursory inspection of the results looked like random noise. I am convinced this is in part because the way that BART was trained — primarily with affects — gives it no reliable cluster for "pro-business" etc.
- This should introduce some skepticism about the validity/consistency of "Positive," "Neutral," and "Negative," but thinking about the way BART was trained, it intuitively makes sense

these labels map more closely to the affects that the model had labels for, so I'm willing to interpret the results more charitably (more discussion below.)

2 Files & Color Schemes

- The `FIVE_COLORS` file colors paragraphs and sentences with the same color scheme as used in the Google Sheet: Dark green representing "Very positive," light green representing "positive" etc.
- The `THREE_COLORS` file colors paragraphs and sentences with color scheme used in the Google Sheet but collapsing the "Very positive" and "Positive," and "Very negative" and "Negative" labels together.
- The `SMOOTH_COLORS` file colors paragraphs uses a continuous coloring scheme.

Color	Positive	Neutral	Negative
Blue	LOW	HIGH	LOW
Purple	LOW	MED	MED
Mauve	MED	MED	LOW
Brown	MED	MED	MED
Red	LOW	LOW	HIGH
Green	HIGH	LOW	LOW

Table 1: **BART Score to RGB Color Mapping**

3 Brief Discussion of Results

There are a couple of observations that I think are noteworthy at this early stage:

First, BART becomes very uncertain when text segments are of higher resolution (i.e. shorter). That is, sentence-level scores are much more mixed than are article-level scores. This intuitively makes sense as BART is trying to make an assessment with less information. This, however, poses a considerable problem for us. In my reading (and looking at some early evaluations from others), *The Economist* usually makes sentiment-weighing statements about a particular leader in nuanced

and subtle ways that rely on word choice at the sentence level. Rarely are entire articles uniformly positive or negative about a leader. (The Bolsonaro article is the exception that makes the rule). Rather, to my reading, *The Economist* expresses its dis/approval of a leader in select sentences that human readers give much more weight than does BART. That is, BART is picking up too much uninformative noise in sentences not about leaders and misses the informative signal of leader-sentences when it assess the article as a whole.

Additionally, most articles are not uniformly about a single leader. Even if sentence-level scores were not mired by the uncertainty problem described above, articles often mention colleagues, international counterparts, or domestic rivals. (Consider the Rajoy-Sanchez article.) That means that we cannot rely on the article-level sentiment scores as accurate assessments of a leader. (Consider this article on Abbott vs Rudd.)

4 What's Next

Please ask everyone to complete as much scoring as possible by Friday noon. Even if I am skeptical of this methodology, I think that we should see this exercise of comparing human and BART results through to the end. I plan on computing pairwise human-human, pairwise human-BART, and human-group-BART agreement statistics to see if 1) even human readers can agree on what kind of sentiment an article is expressing about a leader; 2) if any one of us is good at scoring articles in a BART-like fashion; and 3) if we are collectively good at scoring in a BART-like fashion.

I think that at least you and I should discuss these results at length on Monday and how we should proceed with this project.

Let me know if you have any questions.

Joshua