Adi Srikanth

Bad Data Science in the Wild

---

a) Please provide a link to some online resource (video, website, blog, social media post, etc.) where you saw "bad data science".

Washington Post Article:
https://www.washingtonpost.com/opinions/2021/12/03/biden-media-coverage-worse-trump-favorable/

The Data in Question:
https://www.dropbox.com/s/tag7liqrbqzye67/milbank-wapo-biden-media-sentiment-2021-v3.xlsx?dl=0

b) Tell us specifically why you think it is "bad". For instance, it could be that they misunderstood something (and are now spreading the misunderstanding), use methods far outside of the scope of their intended use (beyond the limitations of their usefulness), are naive about the assumptions of their methods, use the wrong definition of standard terminology, draw the wrong conclusions from their results, have leakage in their cross-validation, use the wrong model, are overfitting, etc. - unfortunately, all of these things are very common

This article was brought to my attention by Nate Silver, during an episode of the FiveThirtyEight politics podcast. The main issue with this article is that it leans on a conclusion that was made based off of a major misunderstanding of the data.

Essentially, there was data published that leveraged a sentiment analysis model in order to categorize media coverage as either positive or negative with regards to President Biden. In theory, this makes sense. However, the model itself does not seem to be very strong in doing what it claims to be doing. Specifically, many articles that are categorized as positive or negative do not actually seem to make sense in their categorization. For example, the article tagged as “most positive” for Biden is headlined *Biden tax plans: Higher taxes for more spending*. To the model, the words “higher” and “more” (and maybe even “spending”) likely triggered a positive response. However, reading the article it is clear that higher taxes are not painted in a positive light.

There are many more examples of bizarre categorizations of articles. In short, it seems that the sentiment analysis model is simply not very accurate. Or, more specifically, the model does not do a good job of properly contextualizing the articles beyond simply the tone of the words. So, when the Washington Post article claims that Biden has been treated more harshly than Trump by the media, that claim (while it could still be true) rests on the results of a dubious model. As such, this is Bad Data Science in the Wild.