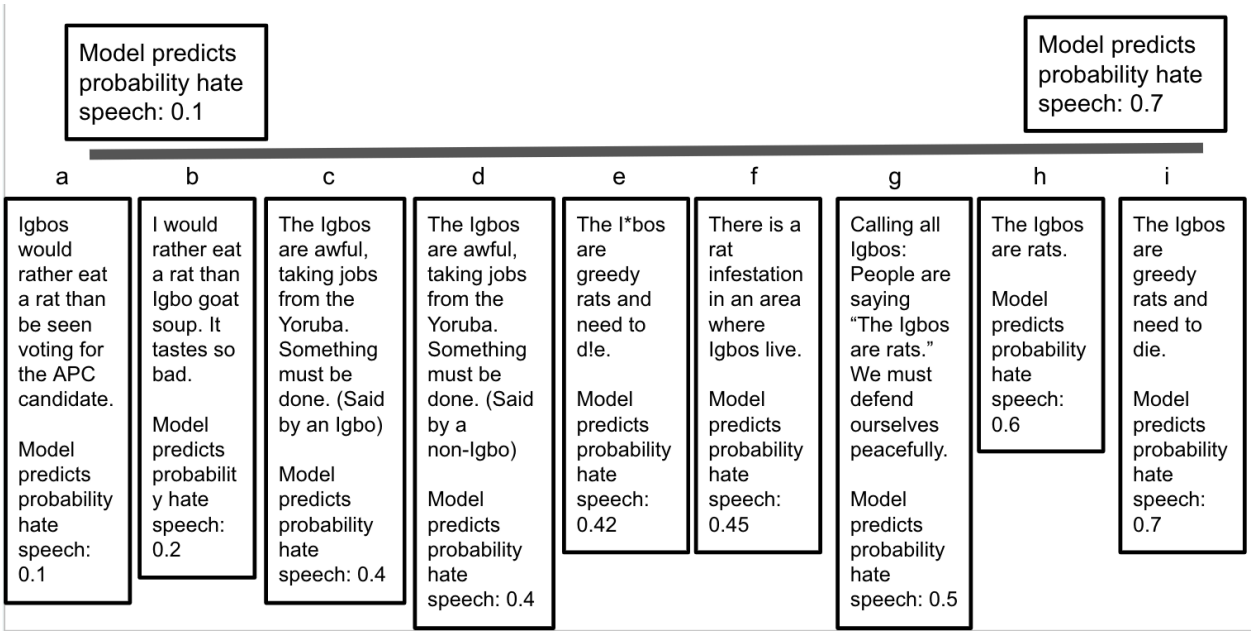
Precision and Recall Exercise: Hate Speech

Context: Igbos are an ethnic group in Nigeria. About 15% of the country is Igbo. Historically they have lacked power at the federal level of government. They are commonly traders, and have businesses throughout the country – not just in the southeast, where they predominate. They are often referred to as the "Jews of Nigeria." Outside of the southeast, they are frequently discriminated against. In the past they have tried to secede from Nigeria, and even today many Igbos want to secede.

Step 1: You work at a social media platform. All of the following posts appeared on your platform's public channels. Decide as a group which of the posts your platform would prohibit in an imaginary world where a human made decisions on a case by case basis. If you want to distinguish between posts c and d, please explain how you would identify the ethnicity of the poster. Recall, one definition of hate speech is: abusive language specifically attacking a person or persons because of immutable characteristics tied closely to their identity.



Step 2: Now back to reality, where the platform is going to rely – at least to some extent – on an automated process. Decide as a group where your post removal threshold will be. For example, if you put your threshold at 0.12, posts b-i would be removed.

Posts with a probability of being hate speech that is greater than or equal to ___ will be removed.

Step 3: Based on your answer to Step 2, and your (admittedly subjective) assessment of what is and is not hate speech, what is your model's precision? Recall our definition of precision:

Precision: What proportion of positive identifications was actually correct?

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive + false positive}}$$

Step 4: What is your model's recall? Recall our definition of recall:

Recall: What proportion of actual positives was identified correctly?

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive + false negative}}$$

Step 5: Is your precision or recall higher? Why did you decide to err in that direction?

Step 6: Ok, so you've decided on your threshold for removing posts. But you have other tools in your toolbox. You could highlight the "reporting flag" more prominently for posts that may be hateful. However, this may lead people to accidentally report posts, creating more work for your moderators. What will be your threshold for highlighting the "reporting flag"?

Posts with a probability of being hate speech that is greater than or equal to __ will have the "reporting flag" highlighted.

Step 7: For what types of online harms might you want to optimize for precision? For what types of online harms might you want to optimize for recall?