a.

i. Possible evidence suggesting discrimination can be deduced by the regions that the model highlights as obscene content. For example if an image that displays women with a variety of skin colors that all somewhat have the same clothing. If the model tends to highlight regions of womens with darker skin as obscure content would be significant evidence for bias in the model. A similar situation would be when women with various skin colors all have images that have the same background. If the model tends to highlight images with darker skinned women, then it would also be significant evidence of bias.

ii. Given the same scenarios if the model highlights regions that are explicitly obscure content regardless of the skin color of the women, would be good evidence that the model is being objective.

b. The trade-off that I would consider would be between continuing using the model to filter content that has obscure content and even though the initial LIME evidence does not suggest any bias, there is no guarantee that the model has no bias. I would explain this situation clearly to the social media platform. Highlighting the potential issue about LIME when trying to generalize local interpretations.