Trust and Safety Models

=======================

We decided to open source the training code of the following models:

- pNSFWMedia: Model to detect tweets with NSFW images. This includes adult and porn content.

- pNSFWText: Model to detect tweets with NSFW text, adult/sexual topics.

- pToxicity: Model to detect toxic tweets. Toxicity includes marginal content like insults and certain types of harassment. Toxic content does not violate Twitter's terms of service.

- pAbuse: Model to detect abusive content. This includes violations of Twitter's terms of service, including hate speech, targeted harassment and abusive behavior.

We have several more models and rules that we are not going to open source at this time because of the adversarial nature of this area. The team is considering open sourcing more models going forward and will keep the community posted accordingly.