Sexually Explicit Deepfakes

Recent advancements in generative AI have significantly lowered the costs of and barriers to producing hyperrealistic but false images and videos. While some deepfakes are humorous, others carry malicious intent, harming the reputations of high-profile individuals, including politicians. Involving pornographic content, non-consensual intimate imagery (NCII) is an exceptionally damaging form of deepfakes. AI-generated NCII disproportionately targets women and has already impacted figures such as US Representative Alexandria Ocasio-Cortez. We seek to combine an observational study of data sourced from Instagram and X with an online experiment that uses generative AI to create sexually explicit images of fictitious politicians. Our study examines how these sexually explicit deepfakes affect the public's perception of politicians' credibility and competence. Our project underscores the serious threat AI-generated NCII poses, especially to women, and aims to develop protective strategies to uphold public trust in the democratic process.
(see *pre-analysis plan*)

Data

- Comments approx. 2 months before/after the surfacing of the AI-generated sexually explicit images
- Cases
  - Cara Hunter (May 2022): `Post_carahuntermla.json`
  - AOC (Feb 2024): still in development

Tasks

1. Bring JSON data into tabular format, i.e., "flatten" data at the level of comments
2. Compute features for each comment (see pre-analysis plan or suggested classifiers. Please let me know if you have other/better suggestions)
   1. Sentiment
   2. Abusive language
   3. Misogyny
3. Create a sample of computed features we can validate.

Codebook

This is an extract of the full JSON data you get when querying the Meta Graph API. For many fields, I can also just "guestimate" what they are.

- id (int): Unique internal ID of post in MongoDB database.
- code (str): Unique public meta code of post. Part of the public URL of a post (https://www.instagram.com/p/CODE)
- owner_id (int): Meta ID of account. Should be the same like in "platform_id" in the other dataset.
- created_at (datetime): Time of creation of post (or caption, to me more explicit).
- taken_at (datetime): Time media was recorded. Can be different form created_at
- device_timestamp (datetime): Time on device when post was created. Can be used to infer timezone from which post was created.

Replies are just nested comments with the same structure (save the "replies" array).