Abusive Language

Ethics in NLP: Including Society in Discourse and Design

March 7th, 2019

Ryan Georgi

A Few Updates

- OpenAI did *not* release training code.
  - Though they used a transformer network, which Google has available in T2T

Dealing With Abusive Language

- **Defining It**
  - How to define the language we're looking at?
- **Detecting It**
  - How much is encoded in the language itself?
  - How much is encoded in pragmatic and non-linguistic information?
- **Acting Upon It**
  - What to do if we detect such language?

Abusive Language: Defining It

- Hate speech — Waseem & Hovy (2016)
- Aggression & Bullying — TRAC-2018 Workshop
  - (Overtly Aggressive, Covertly Aggressive, Non-Aggressive)
- Toxic Comment Classification Challenge — Kaggle
  - (Toxic, Severely Toxic, Obscene, Threat, Insult, Identity Hate)

Abusive Language: Defining It

- Example of defining hate speech, from Waseem & Hovy (2016)
  - uses a sexist or racial slur.
  - attacks a minority.
  - seeks to silence a minority.
  - criticizes a minority (without a well founded argument).
  - promotes, but does not directly use, hate speech or violent crime.
  - criticizes a minority and uses a straw man argument.
  - blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims.
  - shows support of problematic hash tags. E.g. "#BanIslam", "#whoriental", "#whitegenocide"
  - negatively stereotypes a minority.
  - defends xenophobia or sexism.
  - contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria.

Abusive Language: Defining It

- Many of these are pragmatic cues on ***stance*** detection
  - How is the topic at hand situated by the writer?

Abusive Language: Defining It

- One of the key things to look for in such papers:
  - Inter-annotator agreement (IAA)
    - Krippendorff's α
    - Cohen's κ
- Does a high IAA mean a good definition? Or a narrow one?

Abusive Language: Detecting It

- Some problems identified in van Aken et. al, (2018):
  - Words are often OOV, sometimes deliberately so
  - Long-range dependencies
  - Multi-word Expressions
    - (Expressions where words are not toxic in isolation, but take on a different meaning when combined as an MWE)
  - Quoting others to respond to their speech
  - Pragmatics of what is an insult?
    - ("she looks like a horse" → Requires cultural/world knowledge)
    - Sarcasm

Abusive Language: Dealing With It

- Deplatforming?
  - Is it effective?
  - Research by (Chandrasekharan et. al, 2017) by users of banned subreddits in 2015:
    - Many users from banned subreddits left the site, instead of migrating
    - Many that "migrated" did not take their behavior with them to new subreddits
      - "invading" users of other communities continued doing so
      - But existing users did not change their speech
    - Some did find alternative subreddits (e.g. r/CoonTown → r/The_Donald)
    - Users that stayed exhibited less hate speech in their posts
      - (due to largely posting in different places)

Abusive Language: Dealing With It

- Interventions?
  - Munger (2016)
  - Created a twitter bot to @reply users that used racial slurs
  - Found a statistically significant dropoff of this language when "treated" by the white male avatar with high follower count
  - Ethics of this programmatic intervention?

Failure Cases

- Rosenberg, Eli (7/5/2018). Facebook censored a post for 'hate speech.' It was the Declaration of Independence. *The Washington Post*.
  - *"[King George III] has excited domestic insurrections amongst us, and has endeavoured to bring on the inhabitants of our frontiers,* ***the merciless Indian Savages****, whose known rule of warfare, is an undistinguished destruction of all ages, sexes and conditions" [link]*
  - But also, is there overlap between posts containing hate speech and excerpts of this document?

Failure Cases

- Angwin, Julia et. al (6/28/2017). Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children. *ProPublica*.
- Gibbs, Samuel. (12/5/2019). Facebook bans women for posting 'men are scum' after harassment scandals. *The Guardian*.
- Lux, Dottie and "Lil' Miss Hot Mess" (8/14/2017). Facebook's Hate Speech Policies Censor Marginalized Users. *Wired*.
- Riley, John. (1/16/2019). Brian Sims was called a "faggot" on Facebook. They banned him for sharing it. *Metro Weekly*.
- Sunderland, Mitchell (5/25/2014). Facebook Blocked Me Because I Said 'Faggot,' Even Though I'm Gay. *Vice*.

Failure Cases

- Rosenberg, Eli (7/5/2018). Facebook censored a post for 'hate speech.' It was the Declaration of Independence. *The Washington Post*.
  - *"[King George III] has excited domestic insurrections amongst us, and has endeavoured to bring on the inhabitants of our frontiers,* ***the merciless Indian Savages****, whose known rule of warfare, is an undistinguished destruction of all ages, sexes and conditions" [link]*
  - But also, is there overlap between posts containing hate speech and excerpts of this document?

Inspecting Failure Cases

- From Angwin et. al (2017):
  - Unlike American law, which permits preferences such as affirmative action for racial minorities and women for the sake of diversity or redressing discrimination, Facebook's algorithm is designed to defend all races and genders equally.

    "Sadly," the rules are "incorporating this color-blindness idea which is not in the spirit of why we have equal protection," said Danielle Citron, a law professor and expert on information privacy at the University of Maryland.
- A definition of abuse that ignores power structures is incomplete.
  -

Resources

- First Workshop on Trolling, Aggression, and Cyberbullying 2018 (TRAC-2018)