Gaps in Understanding: Algorithmic Discrimination in Natural Language Processing

Zeerak Talat
Post-X Politics, Leuphana,
23/11/2017

Forgetting the Math

- Jørgensen et al. (No math, yay!)
- Blodgett & O'Conner
    - Probability of correctly classified white-aligned tweets - probability of correctly classified African-American aligned tweets.
- Zhao et al.
    - Aims:
        - Identify bias in data set.
        - Identify bias in predictions (how much does the model overgeneralise?)
        - Over the entire test set, do updates to the model (and predictions) until the model bias is within a margin of the data set bias.

Queering Homophily

- Moving away from collaborative filtering
- Userbase level information versus user information

So let’s talk about bias

A Typology

	Explicit	Implicit
Directed	Unambiguous in its potential to be abusive, i.e. use of slurs directed at an individual/entity.	Not immediately clearly abusive. Often obscured by ambiguous terms, sarcasm, lack of profanity, etc. Directed at an entity/ individual.
Generalized	Unambiguous in its potential to be abusive, i.e. use of slurs directed at a generalised *other*.	Not immediately clearly abusive. Often obscured by ambiguous terms, sarcasm, lack of profanity, etc. Directed at an generalised *other*.

Questions about Machine Learning?

Annotation Guidelines

1. uses a sexist or racial slur.
2. attacks a minority.
3. seeks to silence a minority.
4. criticizes a minority (without a well founded argument).
5. promotes, but does not directly use, hate speech or violent crime.
6. criticizes a minority and uses a straw man argument.
7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims.
8. shows support of problematic hash tags. E.g. “#BanIslam”, “#whoriental”, “#whitegenocide”
9. negatively stereotypes a minority.
10. defends xenophobia or sexism.
11. 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.

Typology

	Explicit	Implicit
Directed	“@User shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga”	“(((@User))) and what is your job? Writing cuck articles and slurping Google balls? #Dumbgoogles”
Generalized	“So an 11 year old n*gger girl killed herself over my tweets? ^ ^ thats another n*gger off the streets!!”	“Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home.”