ZEERAK TALAT | UNIVERSITY OF SHEFFIELD

TURING INSTITUTE | 09/05/2018

HATE SPEECH, ABUSIVE LANGUAGE, AND THINGS WE LOSE IN THE FIRE

WHY WOULD ANYONE DO THIS?

- Gamergate
- The Task
  - Detecting abuse against women
  - … and people of colour
- The Data
  - Twitter

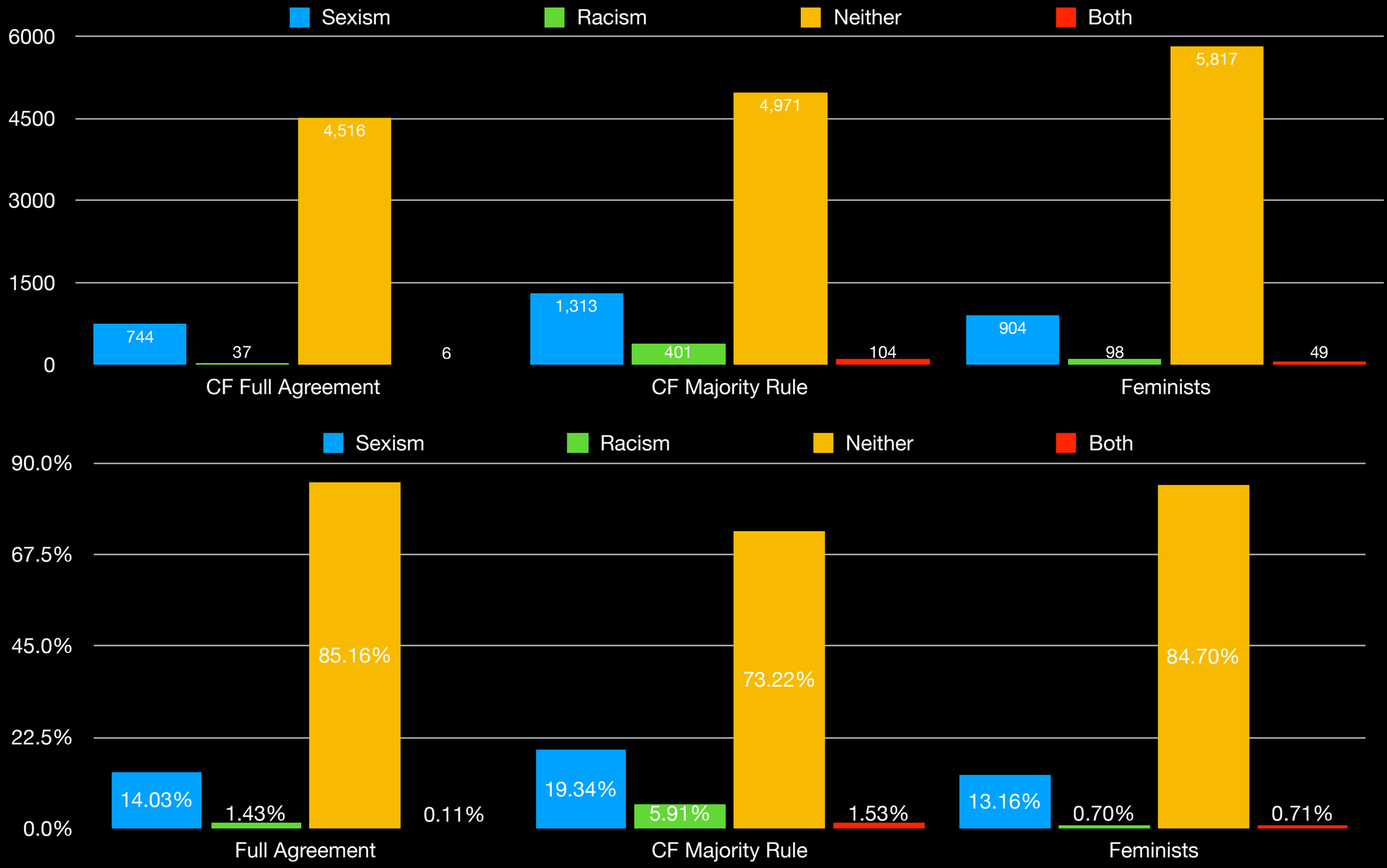
WHAT KIND OF WORLD ARE WE LOOKING TO FOSTER?

- Anti-fascist
- Democratic removal
- Alternatives

ANNOTATING

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.

ISSUES WITH ANNOTATION



HOW WELL DO OUR ANNOTATORS AGREE

	Majority v. Feminists	Full Agreement v. Feminists	All CF Annotators
Cohen's kappa	0.34	0.70	0.57
Krippendorf's alpha	0.32	0.70	

WHAT DOES OUR CLASSIFIER THINK

F1-Scores

	Amateur (Majority Vote)	Expert/Feminists
Character n-*gram*	86.41	91.24
Token n-*gram*	86.37	91.55
Binary Gender	76.64	77.77
GenderProbability	86.37	81.30
Brown Clusters	84.50	87.74
AHST	71.71	55.40

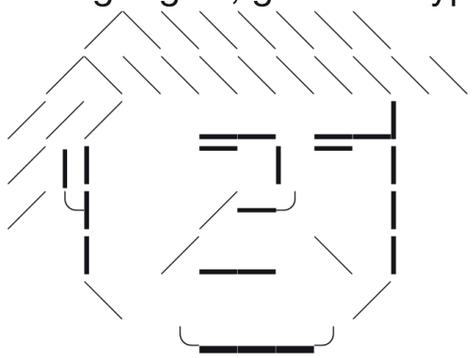
	Explicit	Implicit
Directed	Unambiguous in its potential to be abusive, i.e. use of slurs directed at an individual/entity.	Not immediately clearly abusive. Often obscured by ambiguous terms, sarcasm, lack of profanity, etc. Directed at an entity/ individual.
Generalized	Unambiguous in its potential to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

WHAT DO WE SEE?

погром и созидание
@proudrus

Follow

kill the googles, gas the skypes



7:35 AM - 31 Jan 2017

1 Like

1

AdamSmith
@AdamSmithFan

Replying to @kinsellawarren

#ClimateBarbie fit her perfectly.

It's not sexist. It's brainist.

9:50 PM - 7 Nov 2017

2 Likes

2

Government Shutdown Anime Girl
@RacistAnimeGirl

Follow

"Not all heroes wear capes."
Then explain this photo.

2:31 PM - 29 Oct 2017

11 Retweets 35 Likes

1 11 35

SHAMELESS PROMOTION

- 2nd Workshop on Abusive Language Online @ EMNLP
- Looking for reviewers outside of NLP
- … and suggestions for speakers/panelists working with abuse/hate speech from a policy perspective
- … and ideas on how we can make people from fields outside of NLP submit

in Brussels Oct 31st/Nov 1st

	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.”