ZEERAK TALAT | UNIVERSITY OF SHEFFIELD

CA' FOSCARI | 18.10.2018

PROGRESS IN HATE SPEECH DETECTION AND THE THINGS WE LOST IN THE FIRE

STATING OUR AIMS

- High level:
  - Protection of people
    - Without coinciding with societal discrimination
    - and allowing for dissent
- Low Level
  - Removal of content
  - Find means of dealing with hate speech as it occurs on online platforms.

STATING OUR AIMS

- Which means…

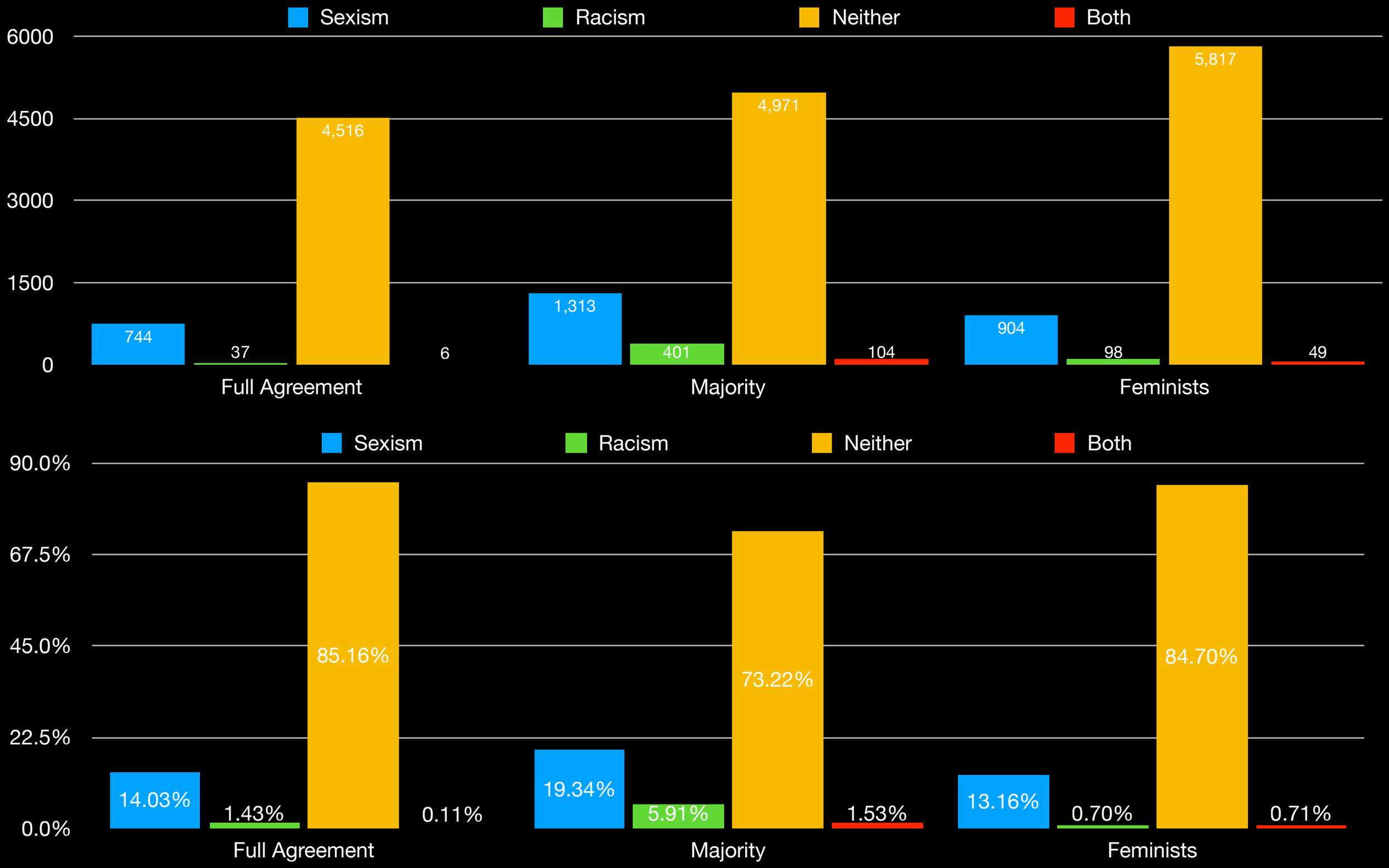
  A. Remove candidates of abuse; or

  B. Understanding societal contexts to allow for dealing with abuse and hate.

ANNOTATION

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

	Majority	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

Waseem et al. (2017)	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*.

Training Objective		Feats	F_1-Scores of Predictions on Test Sets							
Primary	Aux		W/W+H				Davidson			
			R	S	N	Avg	H	O	N	Avg
W/W+H	-	BoW	0.70	0.65	0.88	0.82	0.00	0.64	0.42	0.57
W/W+H	-	Emb	0.30	0.42	0.85	0.71	0.01	0.04	0.29	0.08
W/W+H	-	B+E	0.00	0.00	0.82	0.57	0.00	0.00	0.29	0.05
Davidson	-	BoW	0.22	0.29	0.69	0.56	0.32	0.94	0.84	0.89
Davidson	-	Emb	0.00	0.32	0.60	0.48	0.19	0.92	0.69	0.84
Davidson	-	B+E	0.25	0.33	0.70	0.58	0.39	0.82	0.94	0.89
Both	-	BoW	0.21	0.54	0.81	0.70	0.20	0.92	0.77	0.86
Both	-	Emb	0.21	0.45	0.76	0.64	0.05	0.90	0.64	0.80
Both	-	B+E	0.17	0.53	0.81	0.69	0.31	0.92	0.77	0.86
W/W+H	Davidson	BoW	0.64	0.63	0.87	0.80	0.39	0.94	0.84	0.89
W/W+H	Davidson	Emb	0.32	0.50	0.84	0.72	0.10	0.91	0.64	0.82
W/W+H	Davidson	B+E	0.51	0.53	0.86	0.75	0.16	0.93	0.78	0.86
Davidson	W/W+H	BoW	0.66	0.62	0.86	0.79	0.37	0.94	0.83	0.89
Davidson	W/W+H	Emb	0.39	0.49	0.84	0.73	0.09	0.91	0.62	0.81
Davidson	W/W+H	B+E	0.60	0.57	0.85	0.77	0.14	0.93	0.78	0.86

Waseem et al. (2018)

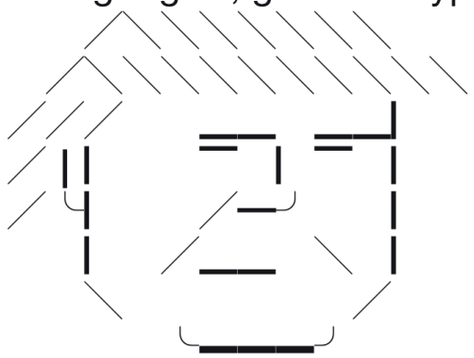
CURRENT TRENDS

- Reducing gender bias (Park et al. EMNLP 2018).
- Expanding on crowd-sourced annotation (Founta et al., ArXiv 2018).
- Deep Learning Models (Badjatiya et al., WWW 2017).

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

- Waseem et al. (2017): ***Understanding abuse: A typology of abusive language detection subtasks***; *Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber; ALW 2017.*
- *Waseem et al. (2018):* ***Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection***; *Zeerak Waseem, James Thorne, Joachim Bingel*; Online Harassment 2018.
- Park et al. (2018): **Reducing Gender Bias in Abusive Language Detection**; *Ji Ho Park, Jamin Shin, Pascale Fung*; EMNLP 2018.
- Founta et al. (2018): **Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior**; *Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, Nicolas Kourtellis*; ArXiv 2018.
- Badjatiya et al. (2017)**: Deep learning for hate speech detection in tweets**; *Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma;* WWW 2017