ZEERAK TALAT | UNIVERSITY OF SHEFFIELD | 03.09.2019

“IT AIN’T ALL GOOD”

MARGINALISATION IN THE NAME OF PROTECTION

“THANGS AIN'T GOIN' LIKE YOU THINK THEY SHOULD, IT'S ALL ON YOU”

“All Good” - De La Soul featuring Chaka Khan

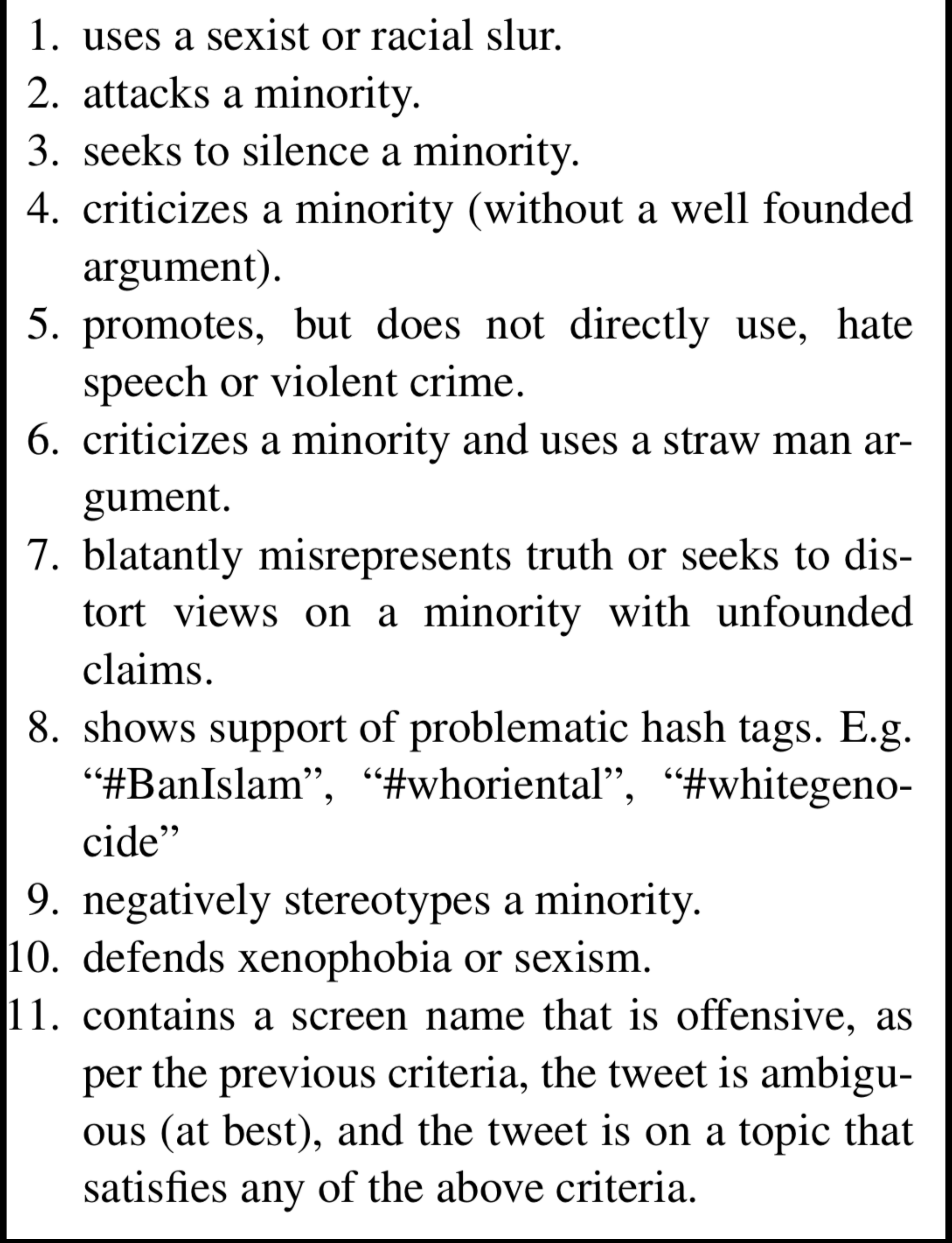
the best time to delete this tweet was immediately after sending it, the second best time is now

THE END IS NIGH!
THE END IS NIGH!
THE END IS NIGH!
END NIGH!
THE IS NIG
THE END IS NIGH!
MATT GROENING

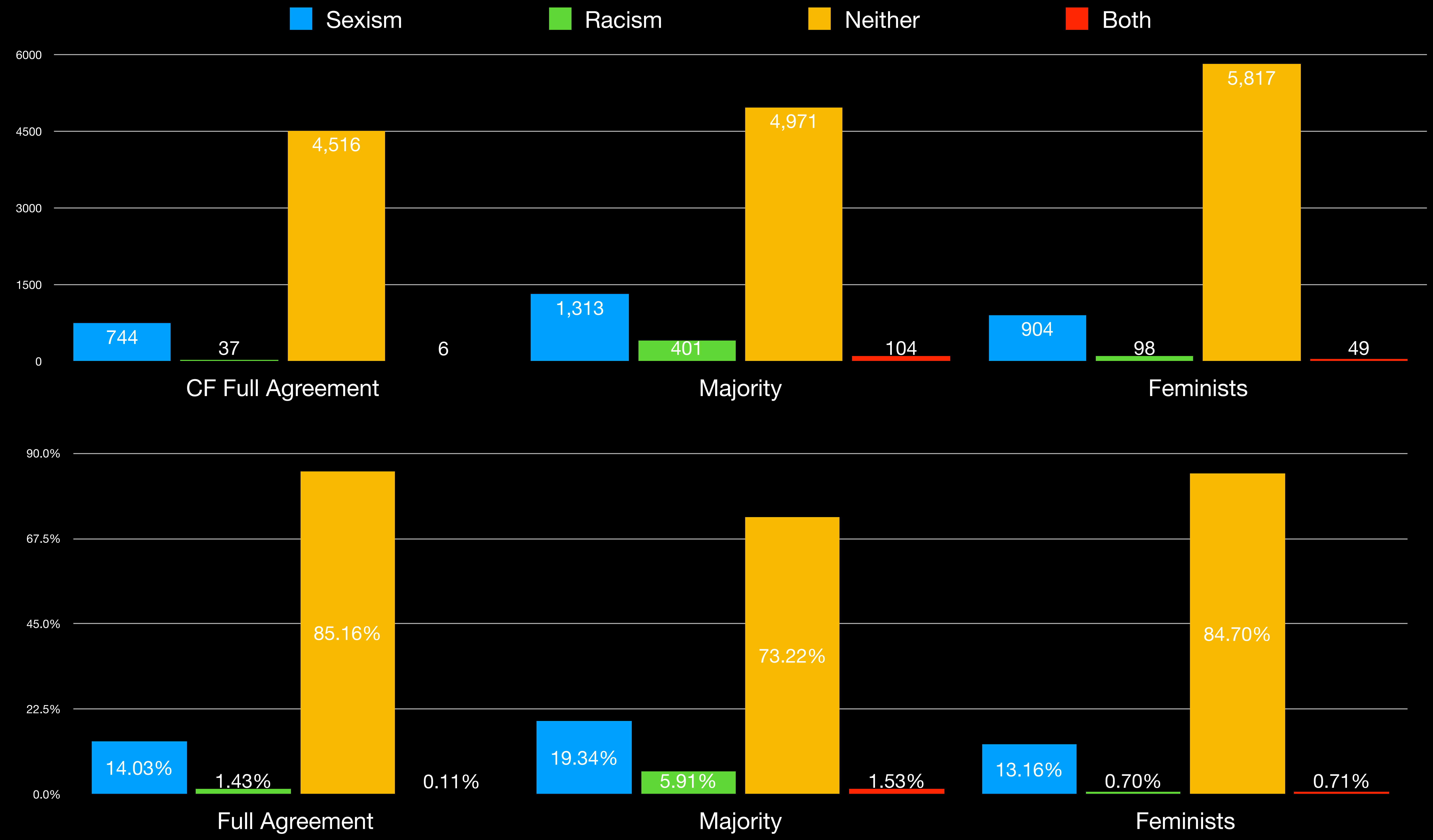
IMAGINATIONS

LAW

ANNOTATION



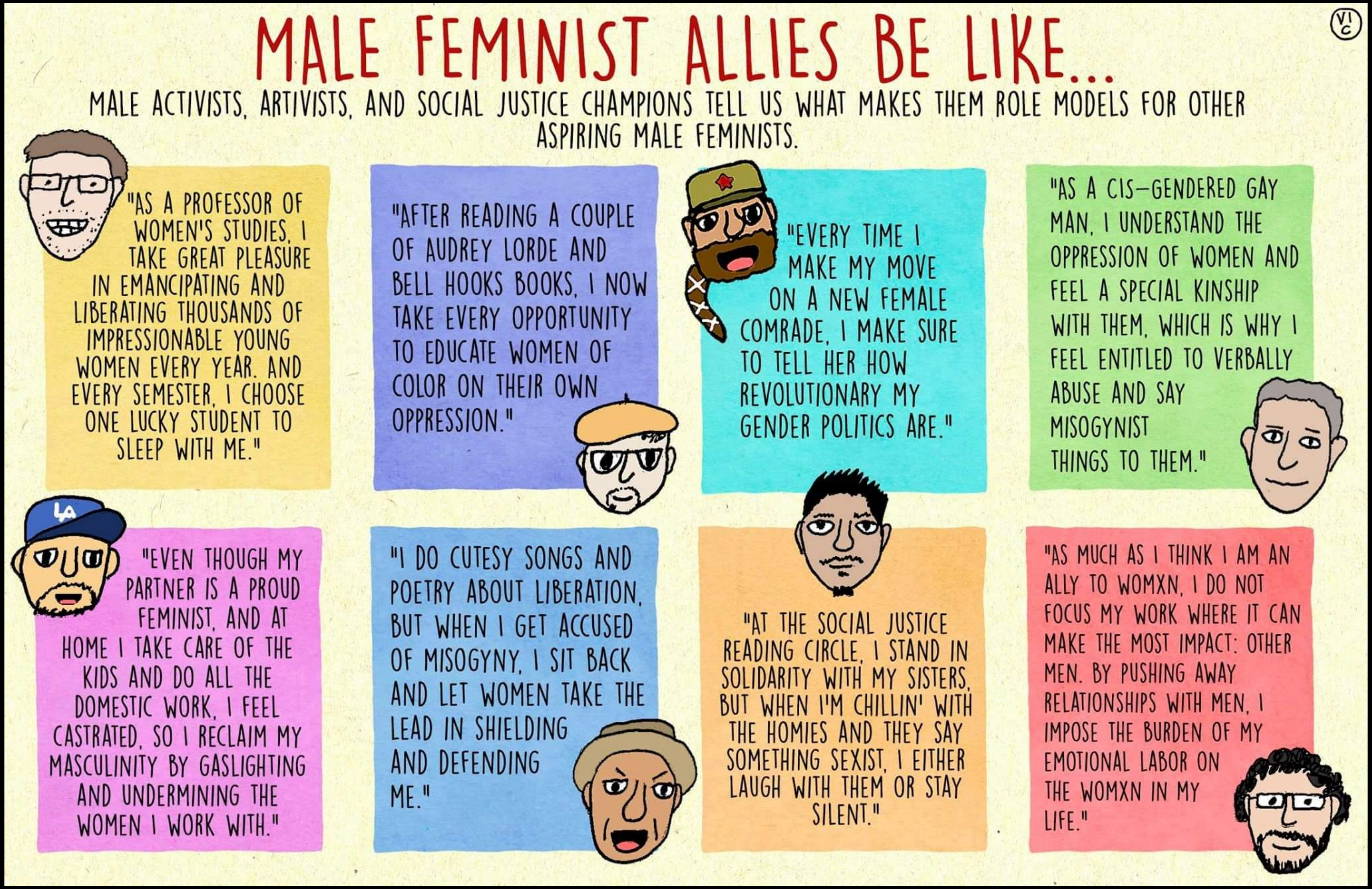
1. uses a sexist or racial slur.
2. attacks a minority.
3. seeks to silence a minority.
4. criticizes a minority (without a well founded argument).
5. promotes, but does not directly use, hate speech or violent crime.
6. criticizes a minority and uses a straw man argument.
7. blatantly misrepresents 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.


Sexism
Racism
Neither
Both
6000
4500
3000
1500
0
744
37
4,516
6
CF Full Agreement
1,313
401
4,971
104
Majority
904
98
5,817
49
Feminists
90.0%
67.5%
45.0%
22.5%
0.0%
14.03%
1.43%
85.16%
0.11%
Full Agreement
19.34%
5.91%
73.22%
1.53%
Majority
13.16%
0.70%
84.70%
0.71%
Feminists

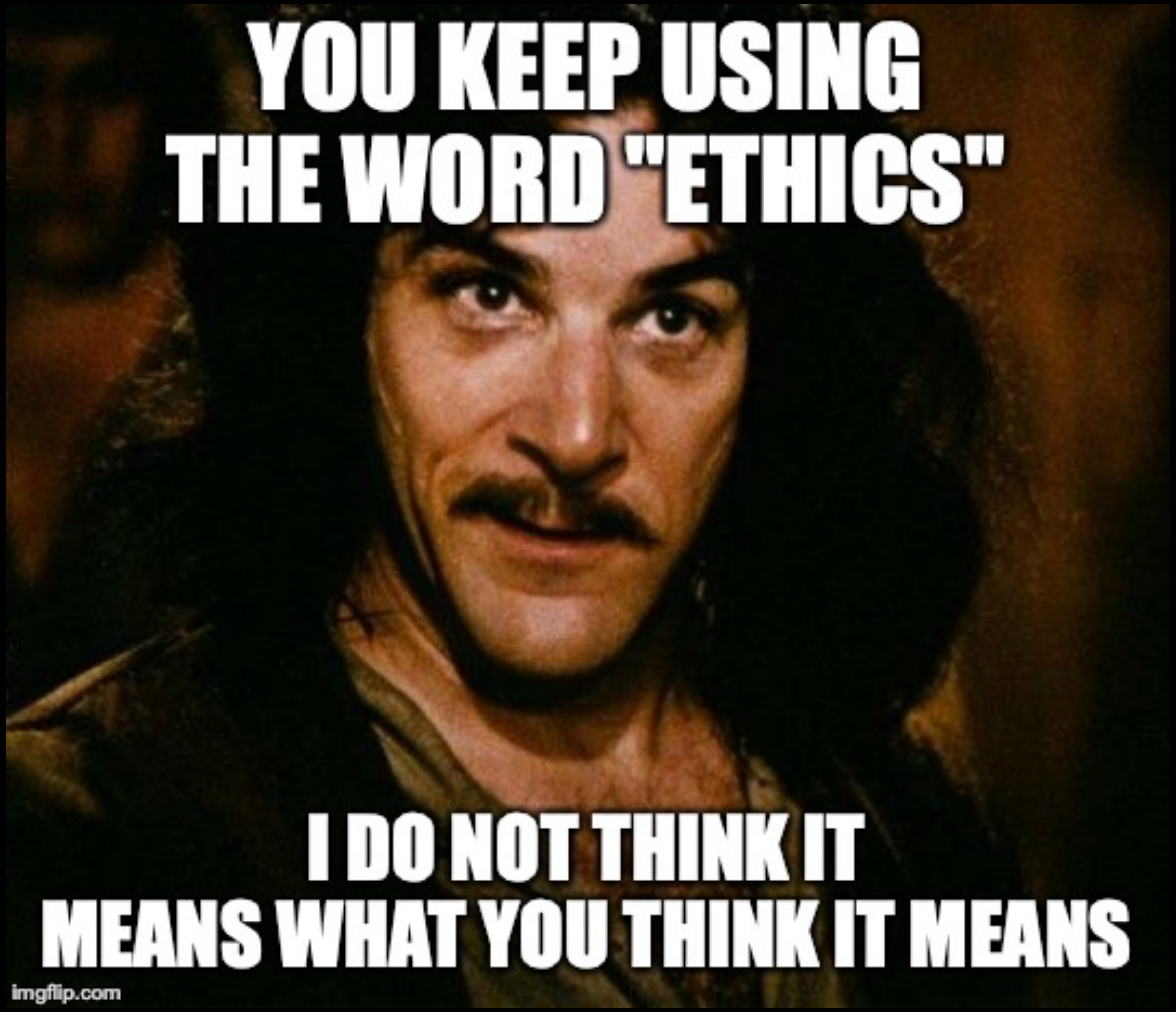
	Full Agreement v. Feminists	Majority v. Feminists	All CF Annotators
Cohen’s kappa	0.70	0.34	0.57
Krippendorf’s alpha	0.70	0.32	

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

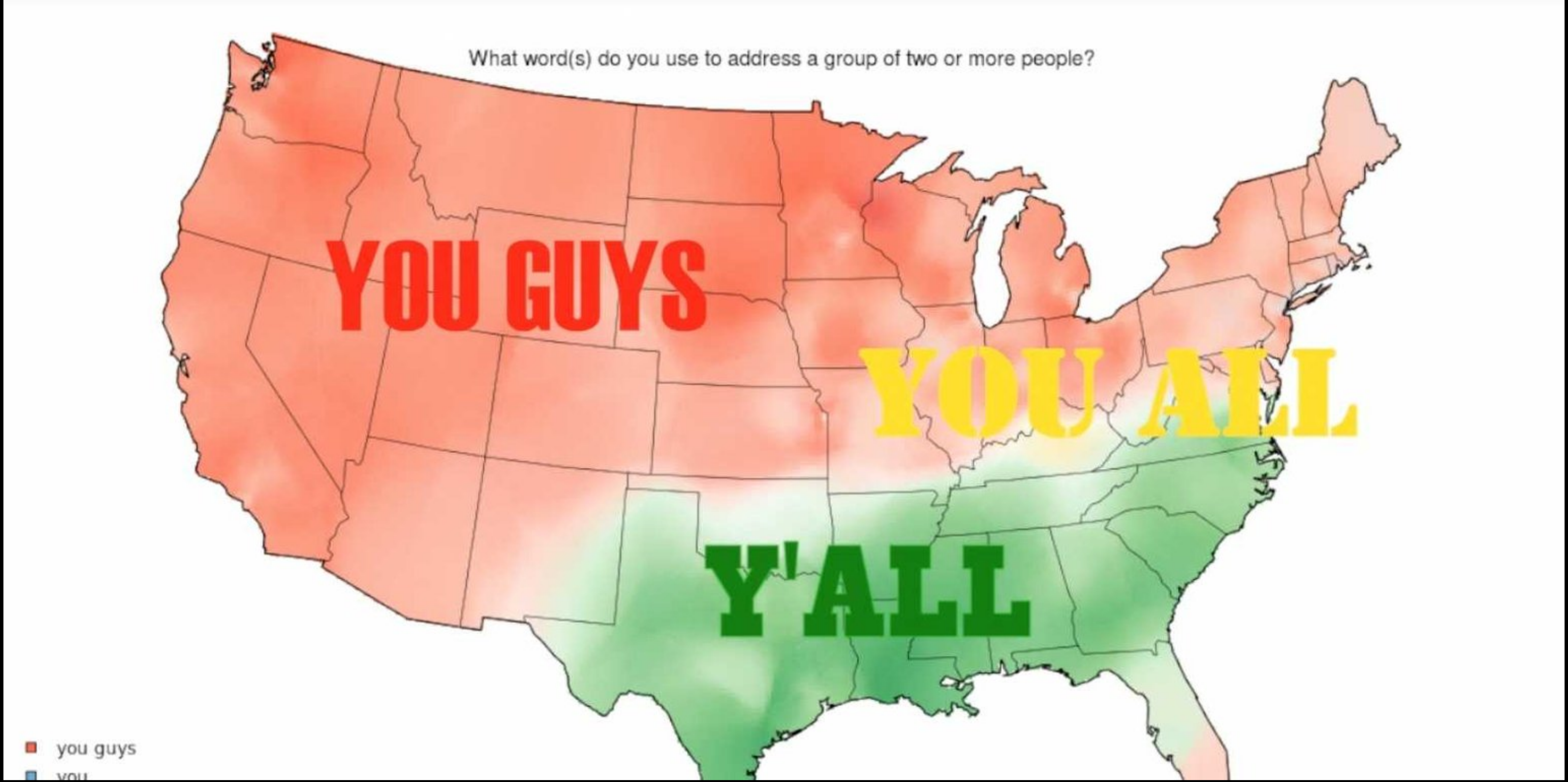

MALE FEMINIST ALLIES BE LIKE...
MALE ACTIVISTS, ARTIVISTS, AND SOCIAL JUSTICE CHAMPIONS TELL US WHAT MAKES THEM ROLE MODELS FOR OTHER ASPIRING MALE FEMINISTS.
"AS A PROFESSOR OF WOMEN'S STUDIES, I TAKE GREAT PLEASURE IN EMANCIPATING AND LIBERATING THOUSANDS OF IMPRESSIONABLE YOUNG WOMEN EVERY YEAR. AND EVERY SEMESTER, I CHOOSE ONE LUCKY STUDENT TO SLEEP WITH ME."
"AFTER READING A COUPLE OF AUDREY LORDE AND BELL HOOKS BOOKS, I NOW TAKE EVERY OPPORTUNITY TO EDUCATE WOMEN OF COLOR ON THEIR OWN OPPRESSION."
"EVERY TIME I MAKE MY MOVE ON A NEW FEMALE COMRADE, I MAKE SURE TO TELL HER HOW REVOLUTIONARY MY GENDER POLITICS ARE."
"AS A CIS-GENDERED GAY MAN, I UNDERSTAND THE OPPRESSION OF WOMEN AND FEEL A SPECIAL KINSHIP WITH THEM, WHICH IS WHY I FEEL ENTITLED TO VERBALLY ABUSE AND SAY MISOGYNIST THINGS TO THEM."
LA
"EVEN THOUGH MY PARTNER IS A PROUD FEMINIST, AND AT HOME I TAKE CARE OF THE KIDS AND DO ALL THE DOMESTIC WORK, I FEEL CASTRATED, SO I RECLAIM MY MASCULINITY BY GASLIGHTING AND UNDERMINING THE WOMEN I WORK WITH."
"I DO CUTESY SONGS AND POETRY ABOUT LIBERATION, BUT WHEN I GET ACCUSED OF MISOGYNY, I SIT BACK AND LET WOMEN TAKE THE LEAD IN SHIELDING AND DEFENDING ME."
"AT THE SOCIAL JUSTICE READING CIRCLE, I STAND IN SOLIDARITY WITH MY SISTERS, BUT WHEN I'M CHILLIN' WITH THE HOMIES AND THEY SAY SOMETHING SEXIST, I EITHER LAUGH WITH THEM OR STAY SILENT."
"AS MUCH AS I THINK I AM AN ALLY TO WOMXN, I DO NOT FOCUS MY WORK WHERE IT CAN MAKE THE MOST IMPACT: OTHER MEN. BY PUSHING AWAY RELATIONSHIPS WITH MEN, I IMPOSE THE BURDEN OF MY EMOTIONAL LABOR ON THE WOMXN IN MY LIFE."

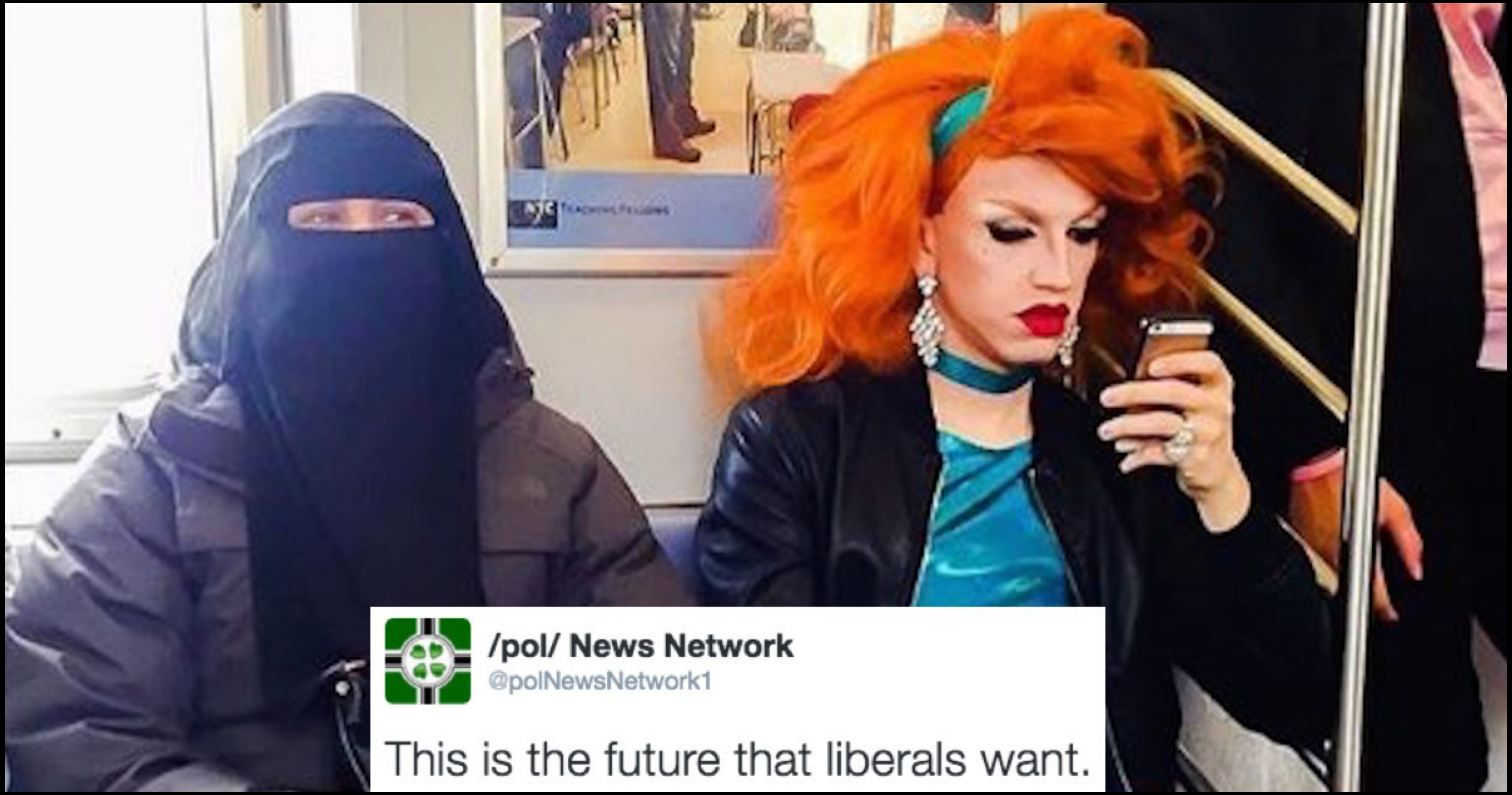
ETHICS AND BIAS



Actually,
IT'S! ABOUT! ETHICS! IN! VIDEO! GAME! JOURNALISM!

TERRIBLY SORRY
EARTH PONIES ONLY BEYOND THIS POINT!
imgflip.com


What word(s) do you use to address a group of two or more people?
YOU GUYS
YOU ALL
Y'ALL
🟥 you guys
🟦 you



/pol/ News Network
@polNewsNetwork1

This is the future that liberals want.

REFERENCES

- The Risk of Racial Bias in Hate Speech Detection. Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi & Noah A Smith. ACL (2019)
- Racial Bias in Hate Speech and Abusive Language Detection Datasets. **Thomas Davidson, Debasmita Bhattacharya and Ingmar Weber**. Third Workshop on Abusive Language Online (2019)
- Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. **Hila Gonen and Yoav Goldberg**. NAACL (2019)
- Hateful symbols or hateful people? predictive features for hate speech detection on twitter. **Zeerak Waseem and Dirk Hovy**. NAACL SRW (2016)
- Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. **Zeerak Waseem**. The First Workshop on NLP and Computational Social Science (2016)
- Understanding Abuse: A Typology of Abusive Language Detection Subtasks. **Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber**. The First Workshop on Abusive Language Online (2017).
- Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection. **Zeerak Waseem, James Thorne, and Joachim Bingel**. In Jennifer Golbeck (editor), Online Harassment (2018)
- A Reductions Approach to Fair Classification. **Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, Hanna Wallach**. ICML (2018)
- What's in a Name? Reducing Bias in Bios Without Access to Protected Attributes. A**lexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky and Adam Kalai**. NAACL (2019)