The University Of Sheffield.

How Do We Hate? Defining and Detecting Hate Speech (on Twitter).

Zeerak Talat
Univeristy of Sheffield

Comp. Ling. Colloquium
Potsdam
8/05/2017

Why Should We Care?

Psych Effects
Stigmatisation

Social Cohesion
Democratic inclusion
Integration
Political Depolarisation
Preventing Stigmatisation

Hate Crime

How do we understand “hate speech”

Common Understanding

- Slurs
- Threats
- Violence (sometimes)
- Vandalism (sometimes)
- Stereotyping (sometimes)
- Disparaging speech (sometimes)

Legal Understanding

- Disparaging speech
- Violence (sometimes)
- Threats (sometimes)
- Vandalism

Some more understandings

Academic Understanding

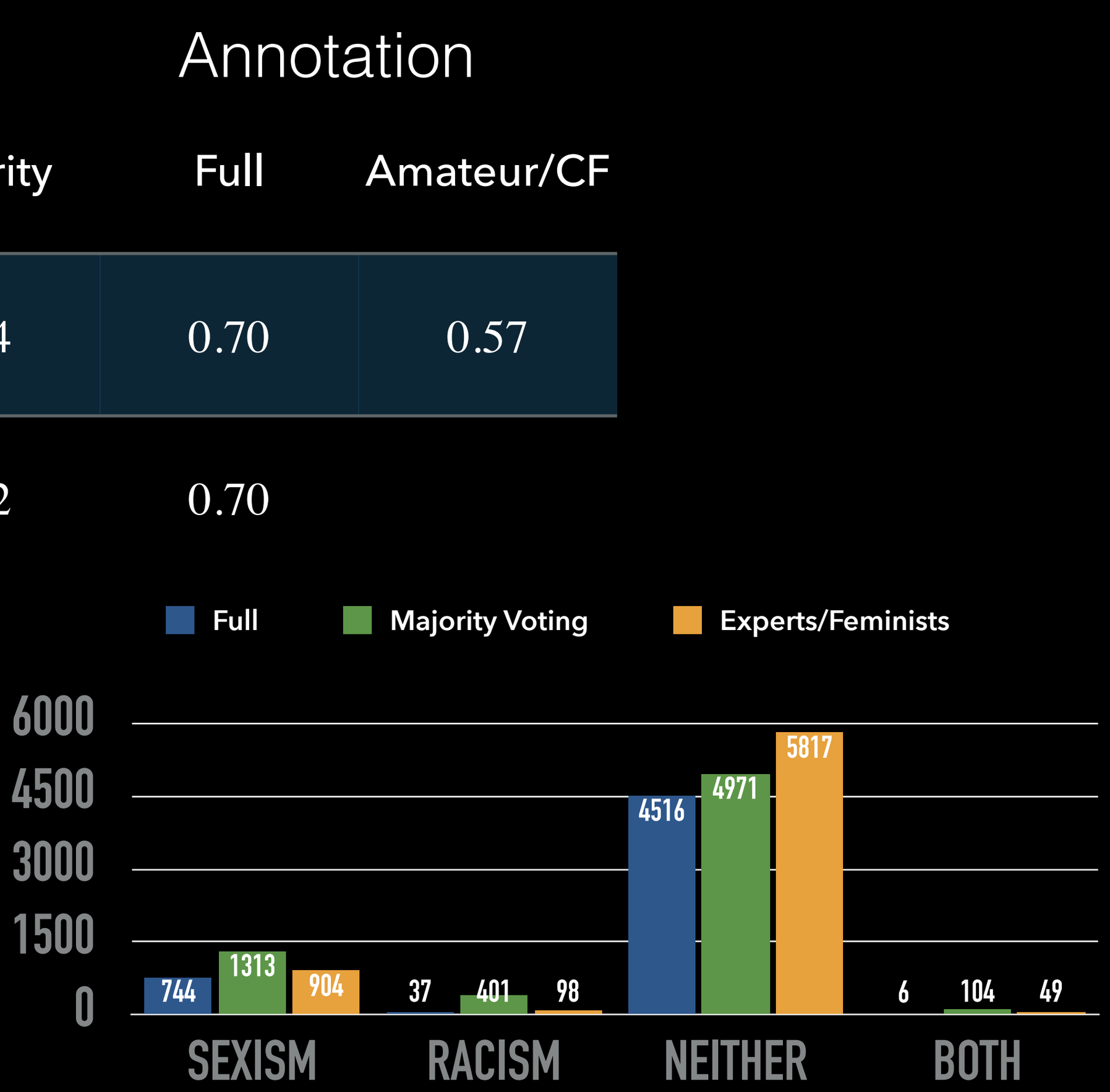
- Slurs
- Stereotyping
- Disparaging speech
- Ridicule
- Minimising
- Undermining
- Lack of representation
- Violence
- Threats
- Vandalism

Social Media Co's Understanding

- Slurs
- Disparaging speech
- Threats (sometimes)
- Violence (sometimes)

Annotation

	Majority	Full	Amateur/CF
Cohen's kappa	0.34	0.70	0.57
Krippendorf's alpha	0.32	0.70	



Classification

F1-Scores

	Amateur	Expert
Character n-*gram*	86.41	91.24
Token n-*gram*	86.37	91.55
Token unigram	86.46	91.15
3-Skip-grams	86.27	91.53
Length	83.16	86.43
Binary Gender	76.64	77.77
GenderProbability	86.37	81.30
Brown Clusters	84.50	87.74
AHST	71.71	55.40

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.”

Some suggested features

	Explicit	Implicit
Directed	Mentions, bad-words dictionary, POS, NER	bad-words dictionary, Semi-supervised methods for finding euphemisms, mentions, word-embeddings
Generalized	Lexical features, named, demographies, bad-words dictionary	Semi-supervised methods for building euphemisms, lexical features, named demographies, word-embeddings

Conclusions and challenges