Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English

Su Lin Blodgett

Brendan O'Connor

Talk at FAT/ML Workshop -- August 2017

College of Information and Computer Sciences

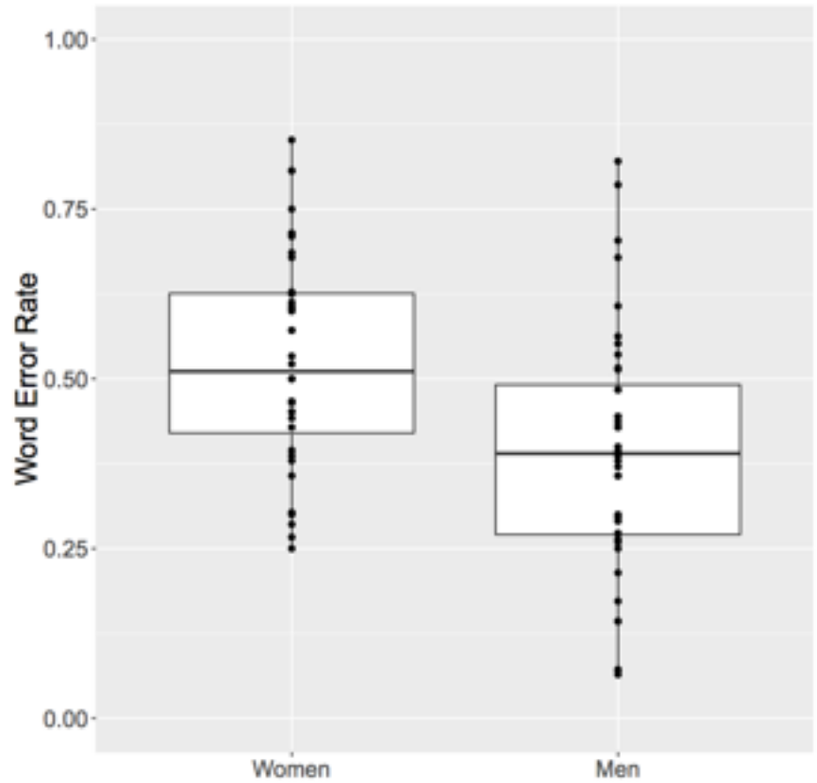
University of Massachusetts Amherst

Paper, etc.: http://brenocon.com

Disparity in language technology

- Language technologies analyze the linguistic behavior of people
- Language is affected by social context and attributes

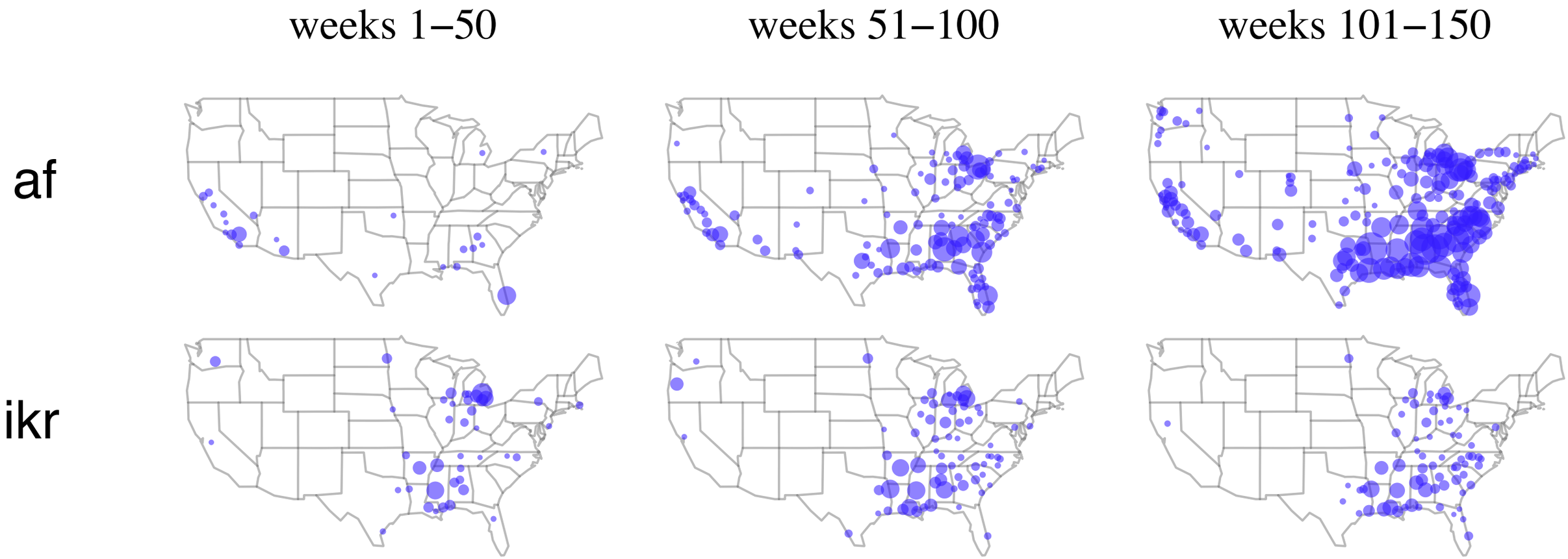
Example: gender bias in YouTube autocaptions *[Tatman 2017, EACL Ethics in NLP workshop]*



- What information can a user access?
- Whose voices are heard?

- Implications of socially embedded language for natural language processing?
- Twitter POS tagging: Race matters!
- African-American English: Language ID and parsing
- Goal: have NLP tools work well across dialects and genres/mediums
  - Language technology to *serve* dialect speakers

Kids these days



Do diffusion patterns follow geographic and demographic similarity?
Geolocated Twitter

[Eisenstein, O'Connor, Smith, Xing, PLOS ONE 2014]

NLP on social media's own terms

ikr	smh	he	asked	fir	yo	last
name	so	he	can	add	u	on
fb	lololol					

- Is this "noisy text"?
- Any NLP system, starting with POS tagging, needs different models/resources than traditional written English
- Word clusters on unlabeled tweets (56 million)

[Owoputi et al. NAACL 2013, Gimpel et al. ACL 2011, www.cs.cmu.edu/~ark/TweetNLP]

NLP on social media's own terms

ikr	smh	he	asked	fir	yo	last
!	G	O	V	P	D	A
name	so	he	can	add	u	on
N	P	O	V	V	O	P
fb	lololol					
^	!					

w fo fa fr fro ov fer **fir** whit abou aft serie fore fah fuh w/her w/that fron isn agains

"non-standard prepositions"

yeah yea nah naw yeahh nooo yeh noo noooo yeaa **ikr** nvm yeahhh nahh nooooo

"interjections"

facebook **fb** itunes myspace skype ebay tumblr bbm flickr aim msn netflix pandora

"online service names"

smh jk #fail #random #fact smfh #smh #winning #realtalk smdh #dead #justsaying

"hashtag-y interjections"??

[Owoputi et al. NAACL 2013, Gimpel et al. ACL 2011, www.cs.cmu.edu/~ark/TweetNLP]

What does it learn?

- Orthographic normalizations

so s0 -so so- $o /so //so

soo sooo soooo sooooo soooooo sooooooo soooooooo sooooooooo soooooooooo
soooooooooo sooooooooooo soooooooooooo soso sooooooooooooo
sooooooooooooo sooooooooooooooo sososo superrr sooooooooooooooo ssooo
so0o superrrr so0 sooooooooooooooo sosososo sooooooooooooooooo ssoo
sssooo sooooooooooooooooooo #too s0o ssoooo s00 sooooooooooooooooooo
so0o0o sososososo sooooooooooooooooooooo sssoooo ssooooo superrrrr very2
s000 sooooooooooooooooooooo sooooooooooooooooooooooo
sooooooooooooooooooooo _so_ sooooooooooooooooooooooooo /so/ ssssooooo
sosososososo

(Immediate?) future auxiliaries

gonna gunna gona gna guna gnna ganna qonna gonnna gana qunna gonne goona gonnaa g0nna goina gonnah goingto gunnah gonaa gonan gunnna going2 gonnnna gunnaa gonny gunaa quna goonna qona gonns goinna gonnae qnna gonnaaa gnaa

tryna gon finna bouta trynna boutta gne fina gonn tryina fenna qone trynaa qon boutaa funna finnah bouda boutah abouta fena bouttah boudda trinna qne finnaa fitna aboutta goin2 bout2 finnna trynah finaa ginna bouttaa fna try'na g0n trynn tyrna trna bouto finsta fnna tranna finta tryinna finnuh tryingto boutto

- finna ~ “fixing to”
- tryna ~ “trying to”
- bouta ~ “about to”

Subject-AuxVerb constructs

[Contraction splitting?]

i'd you'd we'd he'd they'd she'd who'd i’d u'd youd you’d iwould theyd icould we’d i`d #whydopeople he’d i´d #iusedto they’d i'ld she’d #iwantsomeonewhowill i'de imust a:i'd you`d yu'd icud l'd

ill ima imma i'ma i'mma ican iwanna umma imaa #imthetypeto iwill amma #menshouldnever igotta #whywouldyou #iwishicould #sometimesyouhaveto #thoushallnot #ihatewhenpeople illl #thingspeopleshouldnotdo #howdareyou #thingsgirlswantboystodo im'a #womenshouldnever #thingsblackgirlsdo immma iima #ireallyhatewhenpeople ishould #thingspeopleshouldntdo #irefuseto itl #howtospoilahoodrat iwont imight #thingsweusedtodoaskids ineeda #thingswhitepeopledo we'l #whycantyoujust #whydogirls #everymanshouldknowhowto #ushouldnt #howtopissyourgirloff #amanshouldnot #uwannaimpressme #realfriendsdont immaa #ilovewhenyou

[Mixed]

you'll we'll it'll he'll they'll she'll it'd that'll u'll that'd youll ull you’ll itll there'll we’ll itd there'd theyll this'll thatd thatll they’ll didja he’ll it’ll yu'll she’ll youl you`ll you'l you´ll yull u'l it'l we´ll we`ll didya that’ll it’d he'l shit'll they'l theyl she'l everything'll he`ll things'll u’ll this'd

i'll i’ll i'l i`ll i´ll i'lll l'll i\'ll i''ll -i'll /must @pretweeting she`ll

- Where do nonstandard terms come from?

https://twitter.com/search?q=imma&src=typd&vertical=default&f=tweets

imma

https://twitter.com/search?q=imma&src=typd&vertical=default&f=tweets

r.i.p spikeee @xEnvyme_alex · 2m

Feel like **imma** have it up way more my sophomore year than I did freshman year

Angelica @BrowncoatAnge · 2m

Imma start unfollowing all these celebs crying about a Lion and not even remotely interested in the struggle of the Black community.

Billi3 J3an @Misskooki3 · 2m

Niggas thought they wanted to pimp me before they noticed bitch **imma** boss.. I pimp the nigga who brought u here

kye @hippys0ul · 2m

K imma just scroll my tweets to the day abel followed. 😁❤️

LifeΣtyleTr3 @LifeStyleTr3 · 2m

@Charismatic_Cee ay shoot me those pictures you calm you got of us n **imma** post em on the gram !

Demographic Dialectal Variation in Social Media: A Case Study of African-American English

Su Lin Blodgett

Lisa Green

Brendan O'Connor

EMNLP 2016

Dialect

he woke af smart af educated af daddy af
coconut oil using af GOALS AF & shares food af

RETWEETS 3 LIKES 42

1:08 AM - 8 Jul 2016

3 42

Dialect

SAE:
*he **is** woke af*

he woke af smart af educated af daddy af
coconut oil using af GOALS AF & shares food af

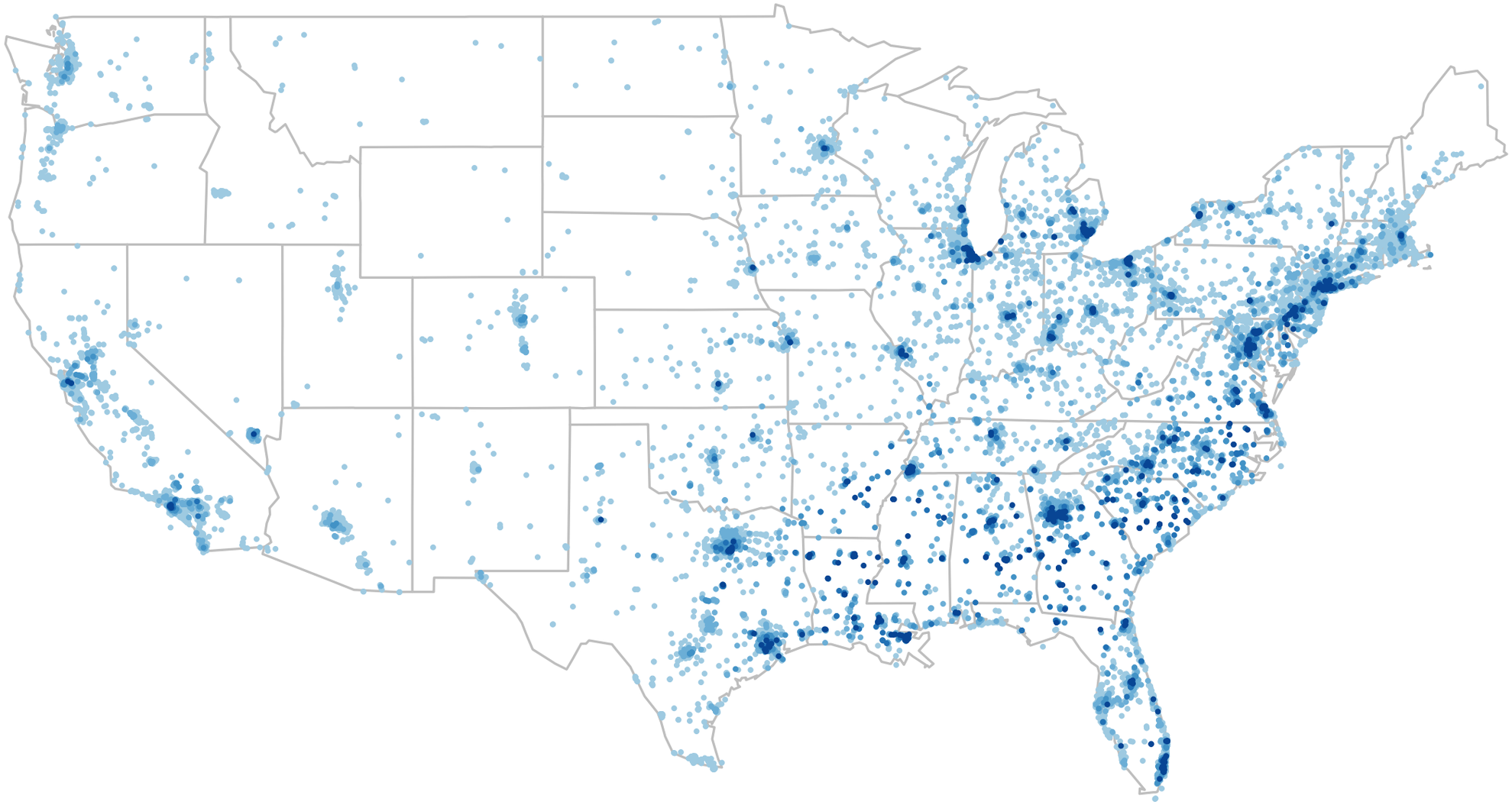
RETWEETS 3 LIKES 42

1:08 AM - 8 Jul 2016

3 42

Why is social media different?

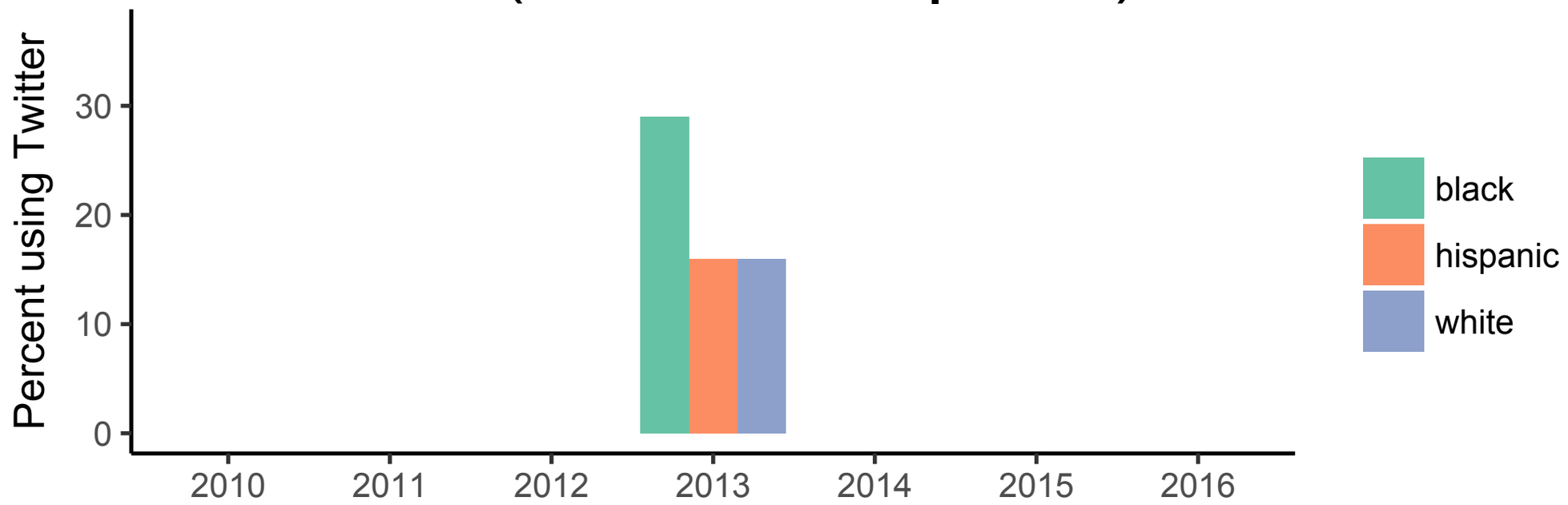
- Internet speech?
- Pre-existing dialectal English?
  - Geographic patterns of word usage often reveal relationships to race, ethnicity etc.
  - African-American English in Twitter
    [Eisenstein 2013, Jorgensen et al. 2015, Jones 2015]



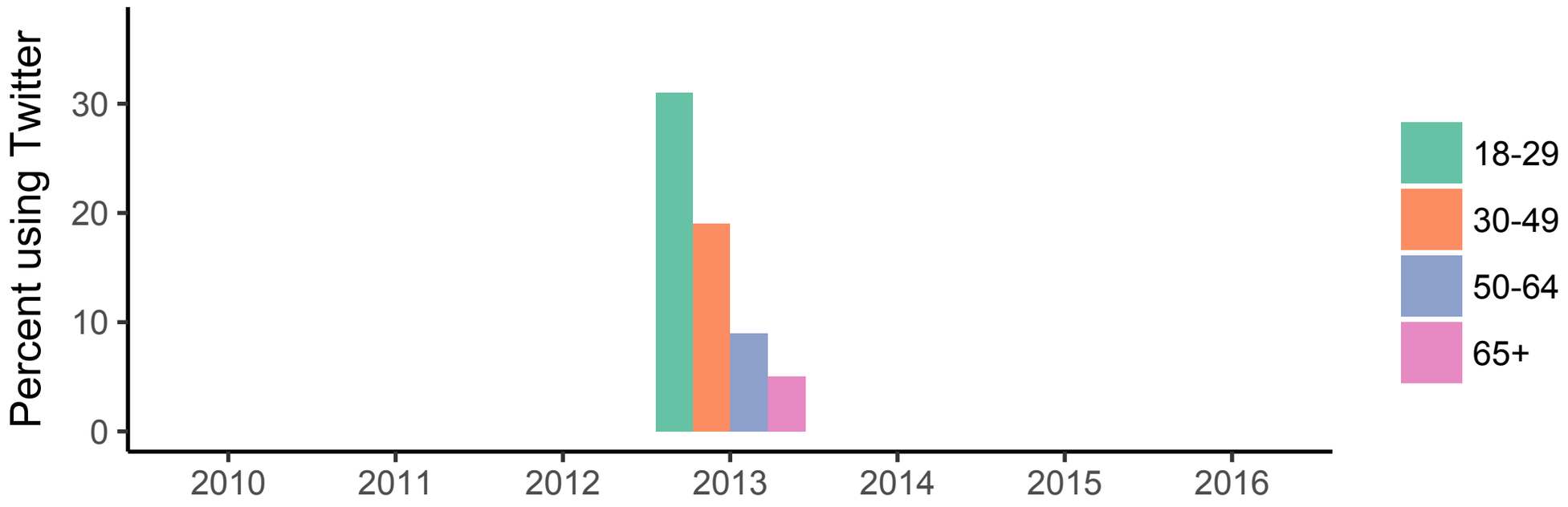
Youth, minorities on Twitter

[Pew Research]

P(use twitter | race)



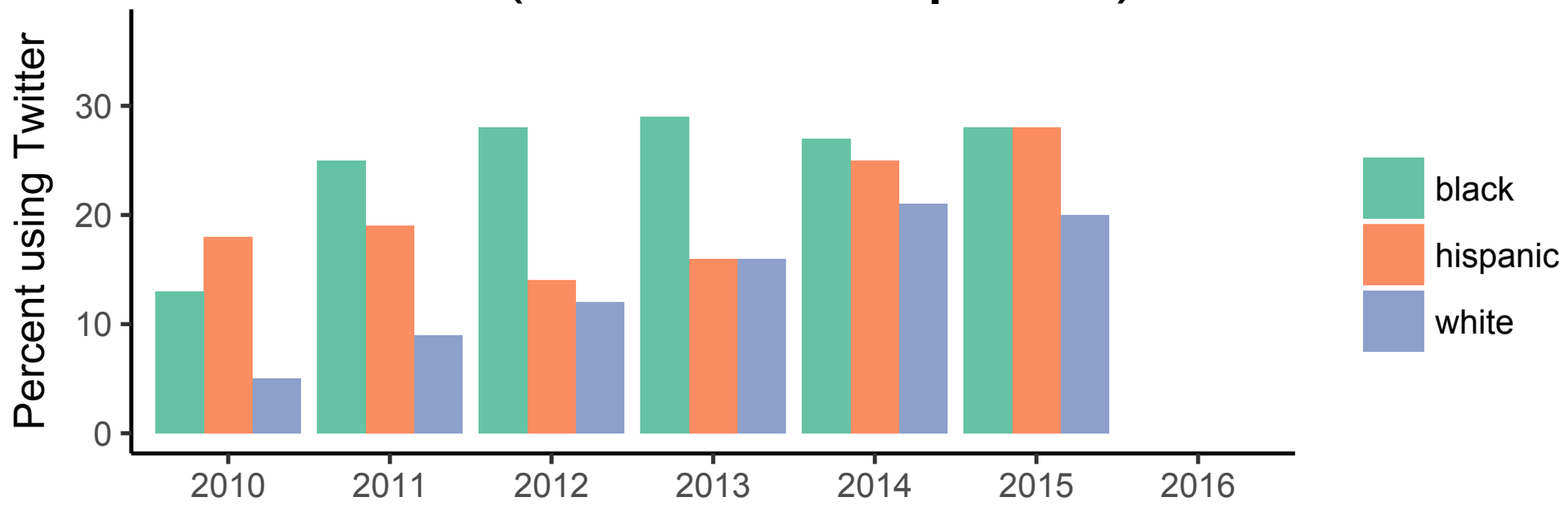
P(use twitter | age)



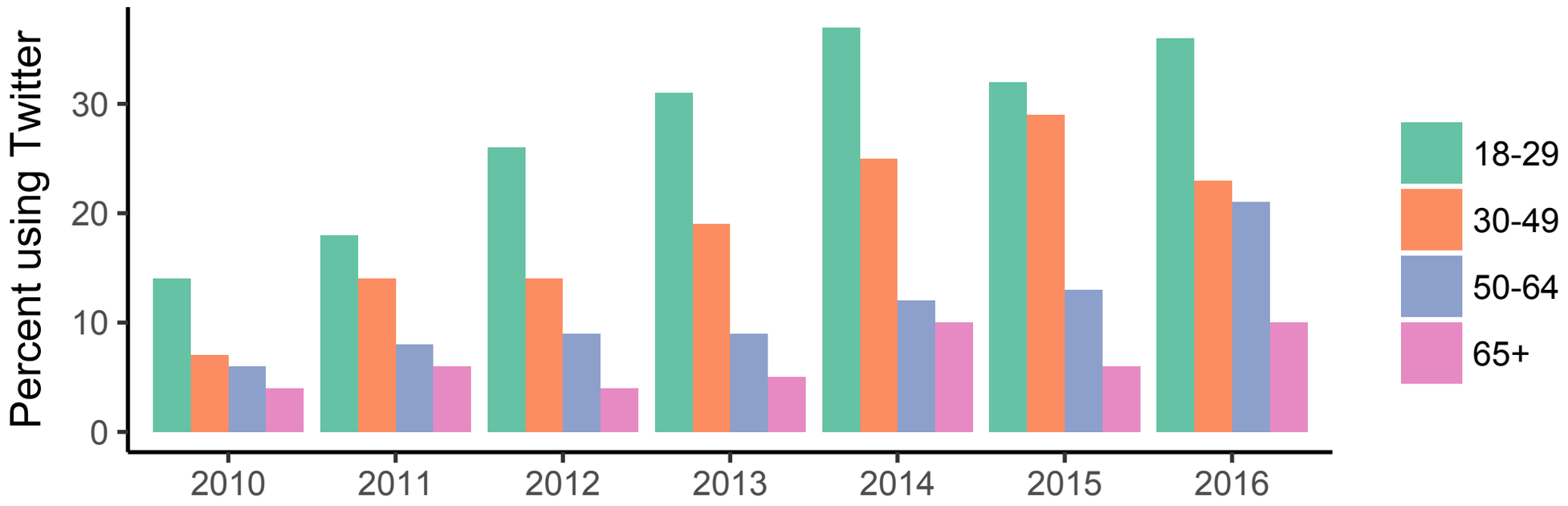
Youth, minorities on Twitter

[Pew Research]

P(use twitter | race)



P(use twitter | age)



- **Learn** language models correlated with U.S. Census racial demographics
- **Validate** against sociolinguistic knowledge
- **Investigate** racial disparities in natural language processing tools
- **Adapt/create** fair NLP tools

Associating geolocated tweets with demographics

King Me
@tmac_datboi

Follow

he woke af smart af educated af daddy af
coconut oil using af GOALS AF & shares food af

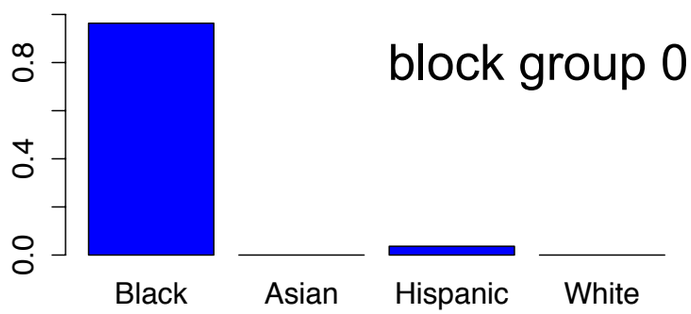
Bored af den my phone finna die!!!

Associating geolocated tweets with demographics

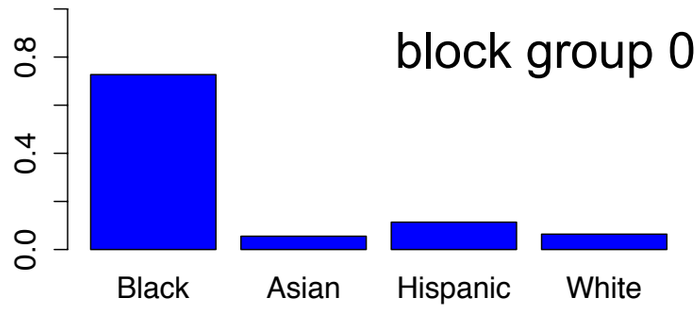
King Me
@tmac_datboi

Follow

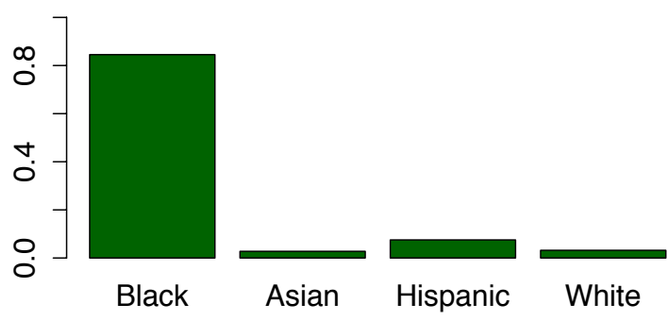
he woke af smart af educated af daddy af
coconut oil using af GOALS AF & shares food af



Bored af den my phone finna die!!!



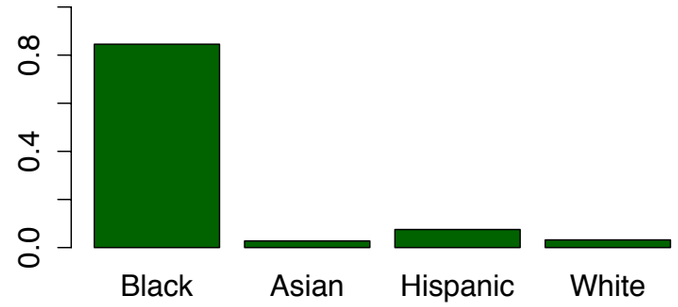
π_{user} =



Mixed membership model

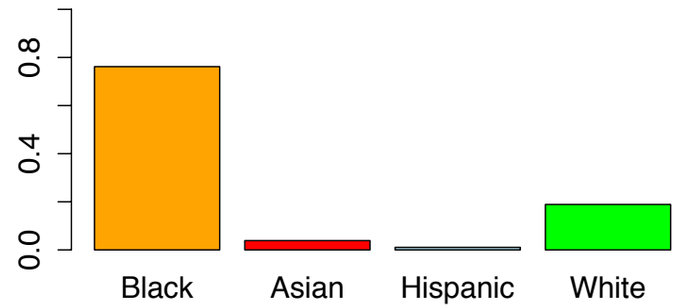
$\theta_{msg} \sim Dir(\alpha\pi)$, $z \sim \theta_{msg,}$ $w \sim \phi_z$

π_{user} =



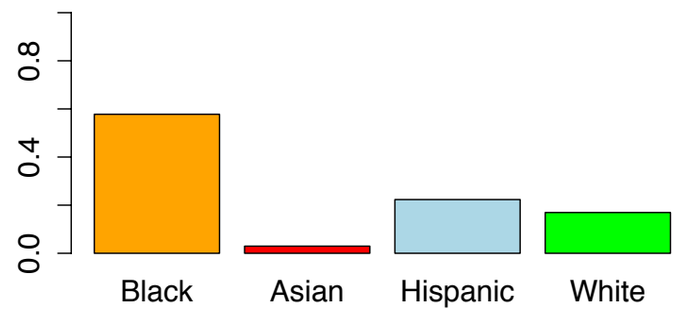
he woke af smart af educated af daddy af coconut oil using af GOALS AF & shares food af

θ_{msg} =



Bored af den my phone finna die!!!

θ_{msg} =



Corpus creation and linguistic validation

- Beyond unigrams: creation of user-level topic-aligned corpora
- How do we linguistically validate them?
  - Lexicon
  - Phonology (Jones, Jorgensen et al.)
  - Syntax (Stewart)

Lexical analysis

- For every word in vocabulary *w* and topic *k*, calculate

$$r_k(w) = \frac{p(w|z = k)}{p(w|z \neq k)}$$

- Examine *w* where $r_{AA}(w) \geq 2$, $r_{white}(w) \geq 2$: AA- and white-aligned words

- 79% of AA-aligned words, 58% of white-aligned words not in a standard English dictionary

Phonological analysis

- Calculate $r_{AA}(w)$ for 31 phonological variants illustrated through nonstandard spellings

- For 30/31 variants: $r \geq 1$

AAE	Ratio	SAE
sholl	1802.49	sure
iont	930.98	I don't
wea	870.45	where
talmbout	809.79	talking about
sumn	520.96	something

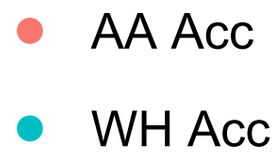
Examining NLP tools - language identification

- Language identification - key step in NLP pipelines

```
>>> s = '''he woke af smart af educated af daddy af coconut oil using af
... GOALS AF & shares food af'''
>>> langid.classify(s)
('da', 0.9999999993212958)

>>> s = 'Bored af den my phone finna die!!!'
>>> langid.classify(s)
('da', 0.9999968001354156)
```

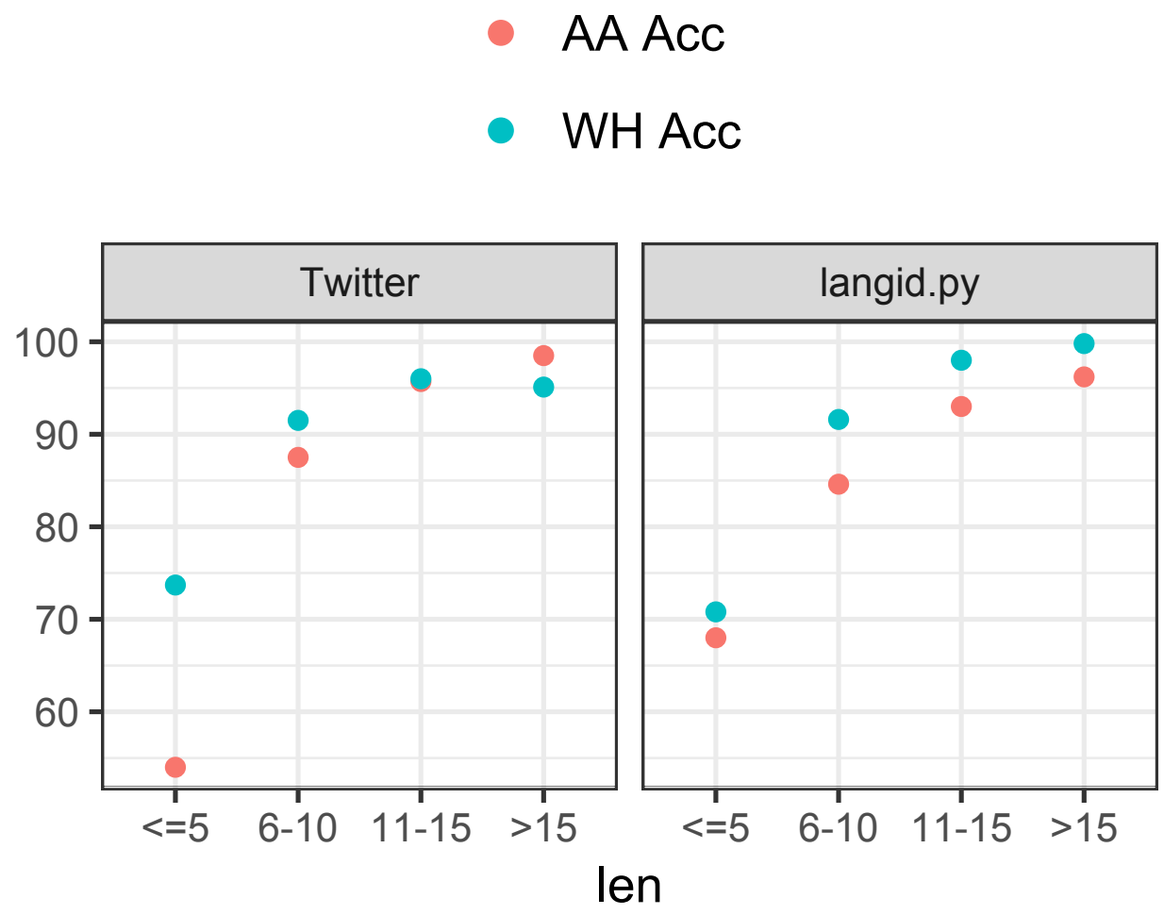
Racial disparity in accuracy



- p(correct | Wh) vs p(correct | AA)
- Assess disparity in
  - langid.py: popular open-source system *[Lui and Baldwin, 2012]*
  - Twitter (in metadata)
  - IBM, Microsoft

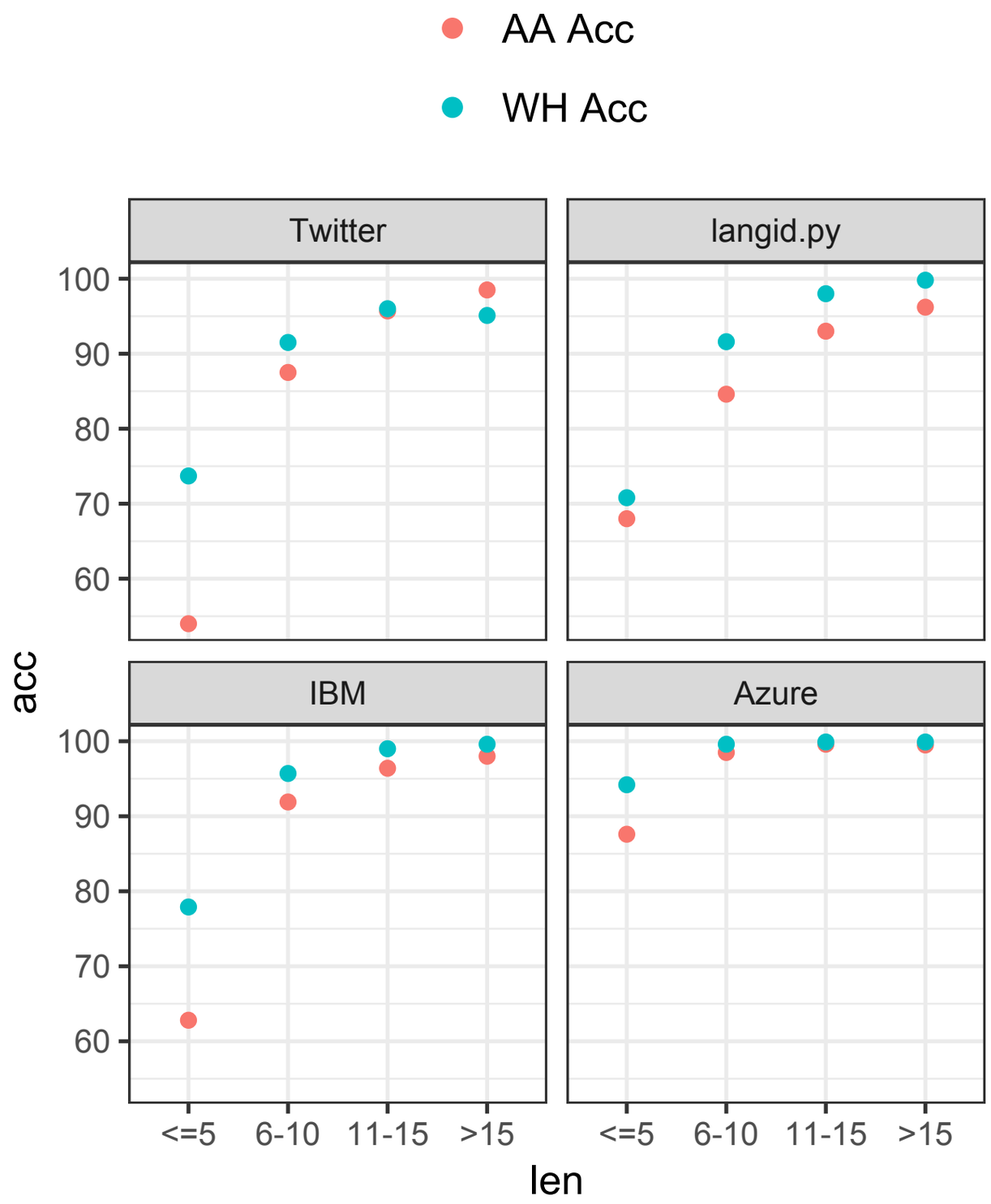
Racial disparity in accuracy

- p(correct | Wh) vs p(correct | AA)
- Assess disparity in
- langid.py: popular open-source system *[Lui and Baldwin, 2012]*
- Twitter (in metadata)
- IBM, Microsoft



Racial disparity in accuracy

- p(correct | Wh) vs p(correct | AA)
- Assess disparity in
  - langid.py: popular open-source system *[Lui and Baldwin, 2012]*
  - Twitter (in metadata)
  - IBM, Microsoft



Racial disparity in accuracy

		AA Acc.	WH Acc.	Diff.
langid.py	$t \leq 5$	68.0	70.8	2.8
	$5 < t \leq 10$	84.6	91.6	7.0
	$10 < t \leq 15$	93.0	98.0	5.0
	$t > 15$	96.2	99.8	3.6
IBM Watson	$t \leq 5$	62.8	77.9	15.1
	$5 < t \leq 10$	91.9	95.7	3.8
	$10 < t \leq 15$	96.4	99.0	2.6
	$t > 15$	98.0	99.6	1.6
Microsoft Azure	$t \leq 5$	87.6	94.2	6.6
	$5 < t \leq 10$	98.5	99.6	1.1
	$10 < t \leq 15$	99.6	99.9	0.3
	$t > 15$	99.5	99.9	0.4
Twitter	$t \leq 5$	54.0	73.7	19.7
	$5 < t \leq 10$	87.5	91.5	4.0
	$10 < t \leq 15$	95.7	96.0	0.3
	$t > 15$	98.5	95.1	-3.0

Less disparate language ID

- Our approach: use our model's demographic inferences as additional signal of English-ness: Allows for racial and online dialect
  - Effectively expands identifier's training set
- Improves English recall, even for international tweets (!) *[Blodgett and O'Connor, WNUT 2017]*
- *Jurgens et al., ACL 2017:* Variety of methods for broader training data also improve English recall

Examining NLP tools - dependency parsing

- Compare annotated parses to systems' output parses

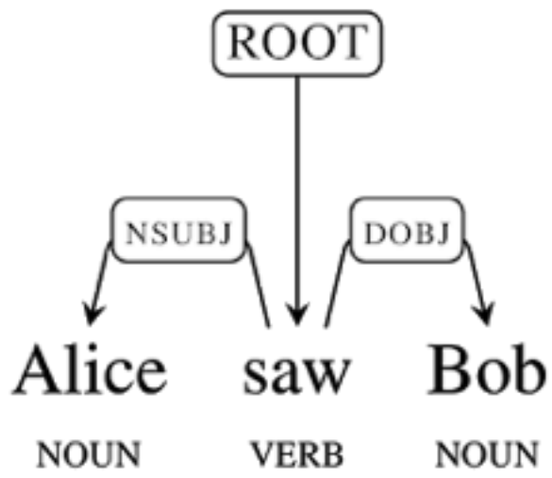
WIRED **Google Has Open Sourced SyntaxNet, Its AI for Understanding Language**

Google Research Blog

Announcing SyntaxNet: The World's Most Accurate Parser Goes Open Source

Thursday, May 12, 2016

Posted by Slav Petrov, Senior Staff Research Scientist



Examining NLP tools - dependency parsing

- Compare annotated parses to systems' output parses
- AAE-like tweets are much harder than SAE-like tweets

Parser	AA	Wh.	Difference
SyntaxNet	64.0 (2.5)	80.4 (2.2)	16.3 (3.4)
CoreNLP	50.0 (2.7)	71.0 (2.5)	21.0 (3.7)

Recall for annotated edges for each message set, bootstrapped standard errors in parentheses.

Ongoing work: create a fair parser

Online language captures historical AAE variation (Uniformly sampled Twitter = lots of AAE!)

1914: reported speech

(Elizabeth Waties Allston Pringle, "A Woman Rice Planter," *First-Person Narratives of the American South Collection)*

dey b'longs to **dat** gent'man ahaid

Online language captures historical AAE variation (Uniformly sampled Twitter = lots of AAE!)

1914: reported speech

(Elizabeth Waties Allston Pringle, "A Woman Rice Planter," *First-Person Narratives of the American South Collection)*

> **dey** b'longs to **dat** gent'man ahaid

2013: Twitter data

$$\frac{P(\text{dat} \mid \text{AA})}{P(\text{dat} \mid \neg\text{AA})} = 5.9 \qquad \frac{P(\text{dey} \mid \text{AA})}{P(\text{dey} \mid \neg\text{AA})} = 6.8$$

Online language captures historical AAE variation (Uniformly sampled Twitter = lots of AAE!)

1914: reported speech

(Elizabeth Waties Allston Pringle, "A Woman Rice Planter," *First-Person Narratives of the American South Collection)*

dey b'longs to **dat** gent'man ahaid

2013: Twitter data

$$\frac{P(\text{dat} \mid \text{AA})}{P(\text{dat} \mid \neg\text{AA})} = 5.9 \qquad \frac{P(\text{dey} \mid \text{AA})}{P(\text{dey} \mid \neg\text{AA})} = 6.8$$

POS taggers: standard vs. designed for Twitter

CoreNLP	**dey**/NN(PRP) b/NN(VBZ) '/Punct longs/NNS(VBZ) to/TO **dat**/VB(DT) gent/JJ '/Punct man/NN ahaid/VBN(RB)
ARK	**dey**/Pro b'longs/Verb to/Prep **dat**/Det gent'man/Noun ahaid/Adv

NLP and social bias

- Natural language processing (NLP) resources are typically designed for standard English or other major languages
  - But non-standard languages correlates with social background
- How do social confounds affect other language technologies?
  - Sentiment measurement? Political science? Digital humanities?
  - Search? Translation?
- How to adapt NLP systems
- Online data from social processes reproduces social phenomena, and algorithms re-learn it