Demographic bias in social media language analysis: a case study of African-American English

Talk at Network Science Institute,
Northeastern University, Dec. 14 2016

Brendan O'Connor (http://brenocon.com)
College of Information and Computer Sciences
University of Massachusetts Amherst

Computational Social Science

1900 2000

Computational Social Science

Official social data

Data collection

100 BCE

Data analysis

1829

Data computation

1890

1900 2000

Computational Social Science

Official social data

Data collection | Data analysis | Data computation

100 BCE | *1829* | *1890*

Semi-structured social data

Digitized behavior
Billions of users, messages/day

Digitized news
Thousands of articles/day

Digitized archives
Millions of books/century

1900

2000

TextGenerator(SocialAttributes) → Text

Society *(SocialAttributes)*

Data generation process

Writing *(TextGenerator)*

Text Data *(Text)*

Language for social measurement

P(SocAttr | Text, TextGen)

TextGenerator(SocialAttributes) → *Text*

Society *(SocialAttributes)*

Writing *(TextGenerator)*

Text Data *(Text)*

Data generation process

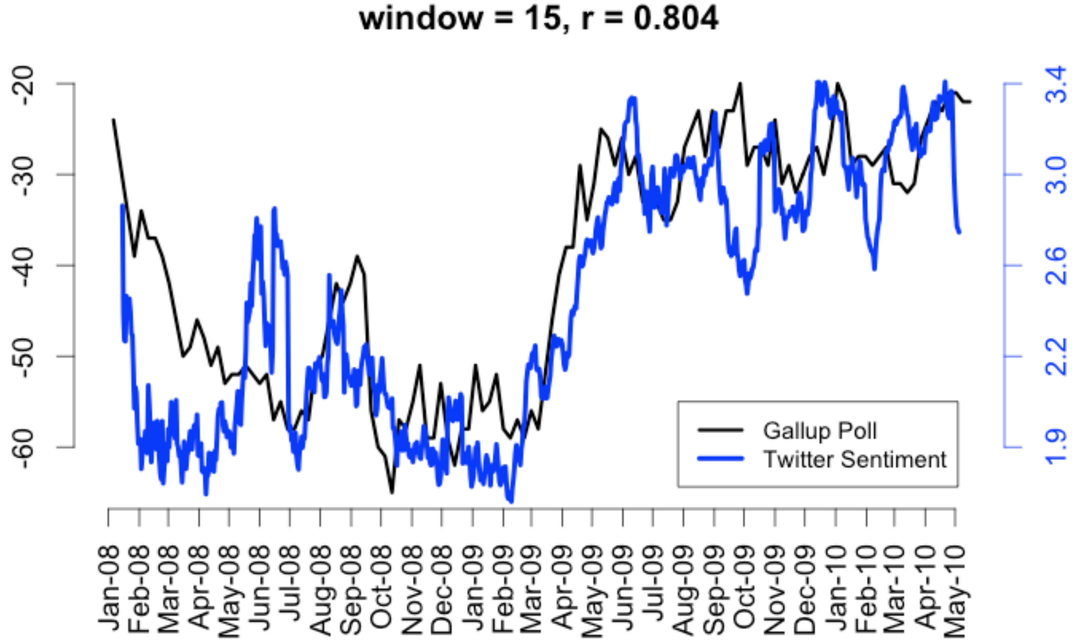
Public opinion

Social media usage

Model assumptions

Language for social measurement

P(SocAttr | Text, TextGen)



[O'Connor et al., ICWSM 2010]

TextGenerator(SocialAttributes) → Text

Society *(SocialAttributes)* | Writing *(TextGenerator)* | Text Data *(Text)*

Data generation process

→

→

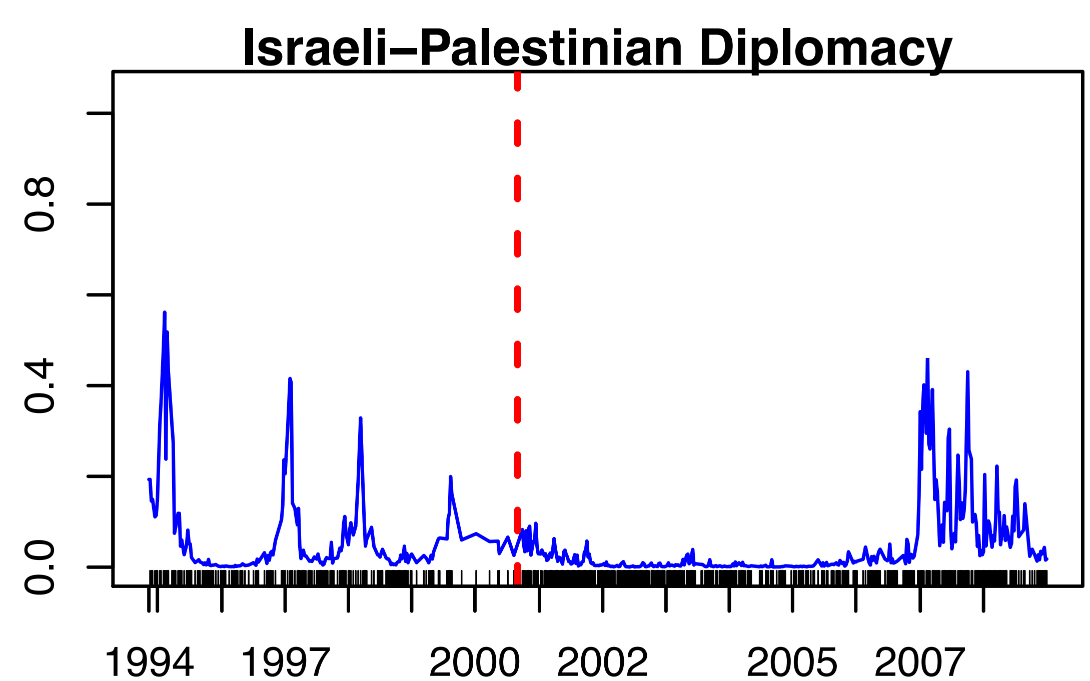
Real-world political events

News media process →

Model assumptions

Language for social measurement
P(SocAttr | Text, TextGen)

[meet with, sign with, praise, say with, arrive in, host, tell, welcome, join, thank, meet, travel to, criticize, leave, take to, begin to, begin with, summon, reach with, hold with...]



[O'Connor, Stewart, Smith ACL 2013]

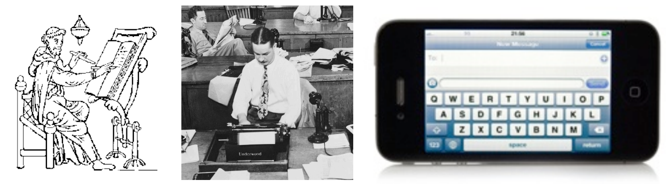
TextGenerator(SocialAttributes) → *Text*

Data generation process

Society *(SocialAttributes)*

→

Writing *(TextGenerator)*



→

Text Data *(Text)*

HATHI TRUST Digital Library

<u>What to analyze:</u>

Social phenomena in social media datasets

- Political speech under Chinese censorship
- Events in international relations
- Social factors in language use

<u>How to analyze:</u>

NLP capabilities we need to do these better

- Part of speech tagging
- Entity extraction
- Syntactic, semantic parsing

What social bias exists in NLP models?

Linguistic/speech act diversity on Twitter

Official announcements

Business advertising

Links to blog and web content

Celebrity self-promotion

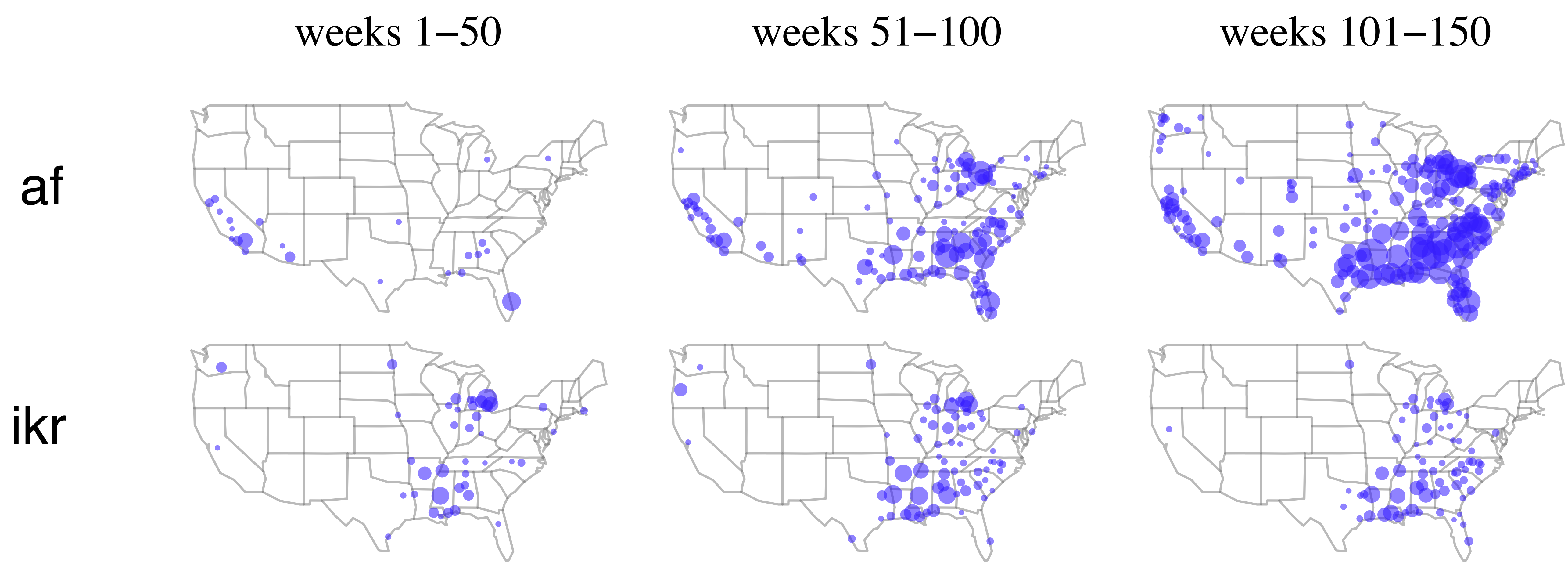
Status messages

Group conversation

Personal conversation

[Slide credit: Jacob Eisenstein]

Kids these days



Do diffusion patterns follow geographic and demographic similarity?
Geolocated Twitter + U.S. Census data
7 TB data, 200 regions, 2600 words, 165 timesteps = 85M parameters

$$n_{w,r,t} \sim \text{Binom}(N_{r,t},\ \sigma(\nu_w + \tau_{r,t} + \eta_{w,*,t} + \eta_{w,r,t}))$$

$$\boldsymbol{\eta}_{w,t} \sim \text{Normal}(\mathbf{A}\boldsymbol{\eta}_{w,t-1},\ \boldsymbol{\Gamma})$$

$\mathbf{A}$ autoregressive coefficients (size $R \times R$)

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

- OK, so socially embedded language exists
- Any implications for natural language processing?

TweetNLP:
Part-of-speech tagging and word clusters
for English-language Twitter
(available at http://www.cs.cmu.edu/~ark/TweetNLP/)

TweetMotif: Exploratory Search and Topic Summarization for Twitter.
Brendan O'Connor, Michel Krieger, and David Ahn.
ICWSM 2010.

Part-of-speech tagging for Twitter: Annotation, Features, and Experiments.
Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan and Noah A. Smith.
ACL 2011.

Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters.
Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider and Noah A. Smith.
NAACL 2013.

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
  - Annotate ~2300 tweets
  - Train word clusters on 56 million tweets, use as features

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"??

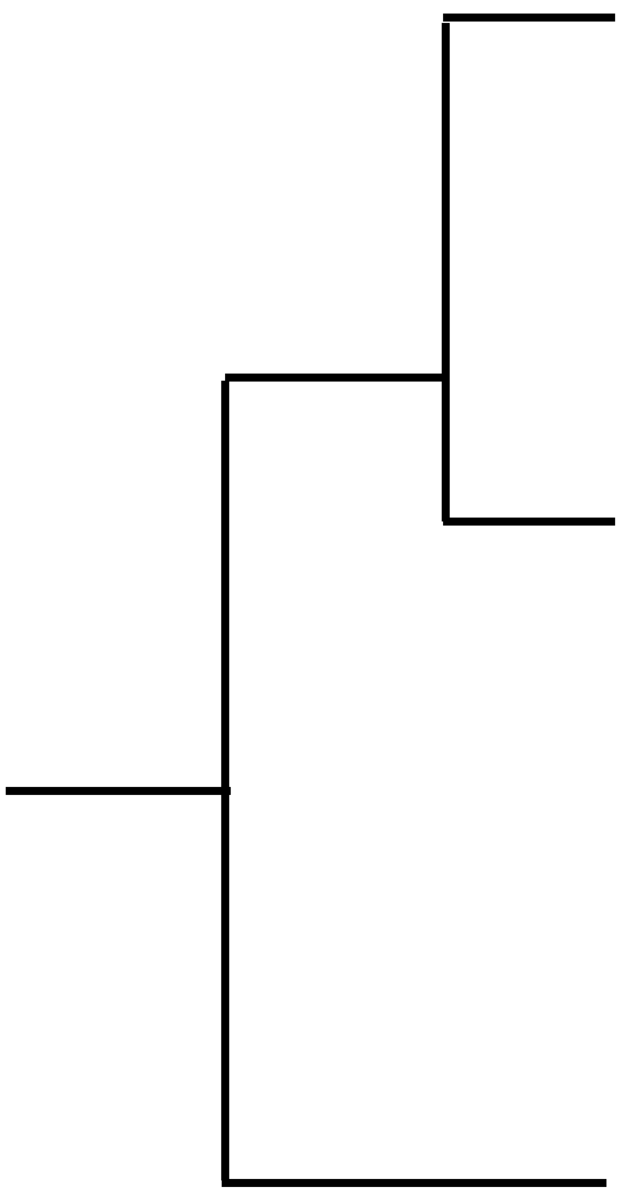
What does it learn?

- Orthographic normalizations

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

soo sooo soooo sooooo soooooo sooooooo soooooooo sooooooooo soooooooooo
sooooooooooo soooooooooooo sooooooooooooo soso soooooooooooooo
sooooooooooooooo soooooooooooooooo sososo superrr sooooooooooooooooo ssooo
so0o superrrr so0 soooooooooooooooooo sosososo sooooooooooooooooooo ssoo
sssooo soooooooooooooooooooo #too s0o ssoooo s00 sooooooooooooooooooooo
so0o0o sosososoSo soooooooooooooooooooooo sssoooo ssooooo superrrrr very2
s000 sooooooooooooooooooooooo soooooooooooooooooooooooo
sooooooooooooooooooooooooo _so_ soooooooooooooooooooooooooo /so/ sssooooo
sososososoSo

- Emoticons etc.
(Clusters/tagger useful for sentiment analysis: NRC-Canada SemEval 2013, 2014)



:d ^^ =d *-* :-d \o/ :dd \m/ 8d *--* *_* u.u :ddd ;;) *.* o/ ;3 =))) *---* \(´▽`)/ n_n b-) (^_^) ^o^ :dddd ;dd *__* :)))))) *----* d/ \o \: =dd n.n -q *___* :33 :ddddd :od -n *-----* xddddd <URL-crunchyroll.com> ^^v (x \= =:) *------* \0/ (˘_˘")

;) :p :-) xd ;-) ;d (; :3 ;p =p :-p =)) ;] xdd 😏 #gno xddd >:) ;-p >:d □ 8-) □ 😜 □ ;-d □ 😝 [; □ :^) =)))) ;-)) <URL-seesmic.com> :pp :~) x'd :op >:p ;^) >:] =))))) :>) <URL-hstl.co> ;)))) ;~) toort >:3 #eden ;pp

:) (: =) :)) :] ☺ :') =] ^_^ :))) ^.^ [: ;)) 😊 ((: ^__^ (= ^-^ :)))) 😁 👍 □ :-)) 😉 ✌ ^___^ (': :} :))))) □ 😃 👌 □ □ 😌 :") :]] □ =]] 😄 □ □ ü ;))) [= (-: ^____^ ;') :-))) (((:

:o o_o « o.o xp ;o ._. t.t t_t #wtf #lol o: x_x =o 0_o dx o_0 :-o ¬¬ --" 0_0 o__o »» u_u #help --' =3 (-_-) -) #confused □ #omg ¬_¬ t^t otl #igetthatalot 😱 xdddd o___o @@ cx t__t d8 □ :{ t___t ---------- #whodoesthat e_e :oo

:(:/ -_- -.- :-(:'(d: :| :s -__- =(=/ >.< -___- :-/ </3 :\ -____- ;(/: ☹ :((>_< =[:[#fml 😒 -_____- =\ >:(😔 -,- >.> >:o ;/ 😳 d; .-. -______- >_> :(((-_-" =s □ ;_; #ugh :-\ =.= □ -_______-

x xx xxx xxxx xxxxx qt xox xxxxxx xxxxxxx xxxxxxxx #pawpawty xxxxxxxxx xxxxxxxxxx #1dfamily #frys #1dqa xxxxxxxxxxx #askliam #dcth xxxxxxxxxxxx #askniall *rt #jbinpoland xxxxxxxxxxxxx #askharry x-x #wiimoms xxxxxxxxxxxxxx oxox #wlf #nipclub +) 1dhq xxxxxxxxxxxxxxx #20peopleilove <URL-paidmodels.com> yart #jedreply #elevensestime <URL-shrtn.us> #askzayn xxxxxxxxxxxxxxxx #wineparty +9 #amwritingparty #tweepletuesday #soumanodomano <URL-today.com> #twfanfriday 22h22

<3 ♥ xoxo <33 xo <333 ❤ ♡ #love s2 <URL-twitition.com> #neversaynever <3333 #swag x3 #believe #100factsaboutme ♥♥ 😘 <3<3 <33333 #blessed xoxoxo 😍 #muchlove #salute xoxox ♥♥♥ #excited ☀ □ #happy #leggo #cantwait <3<3<3 #loveit <333333 #please #dailytweet #thanks 🙏 (˘⌣˘) 💜 #yay #thankyou #loveyou {} ε˘˘) #nsn #iloveyou

Subject-AuxVerb constructs

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

[Contraction splitting?]

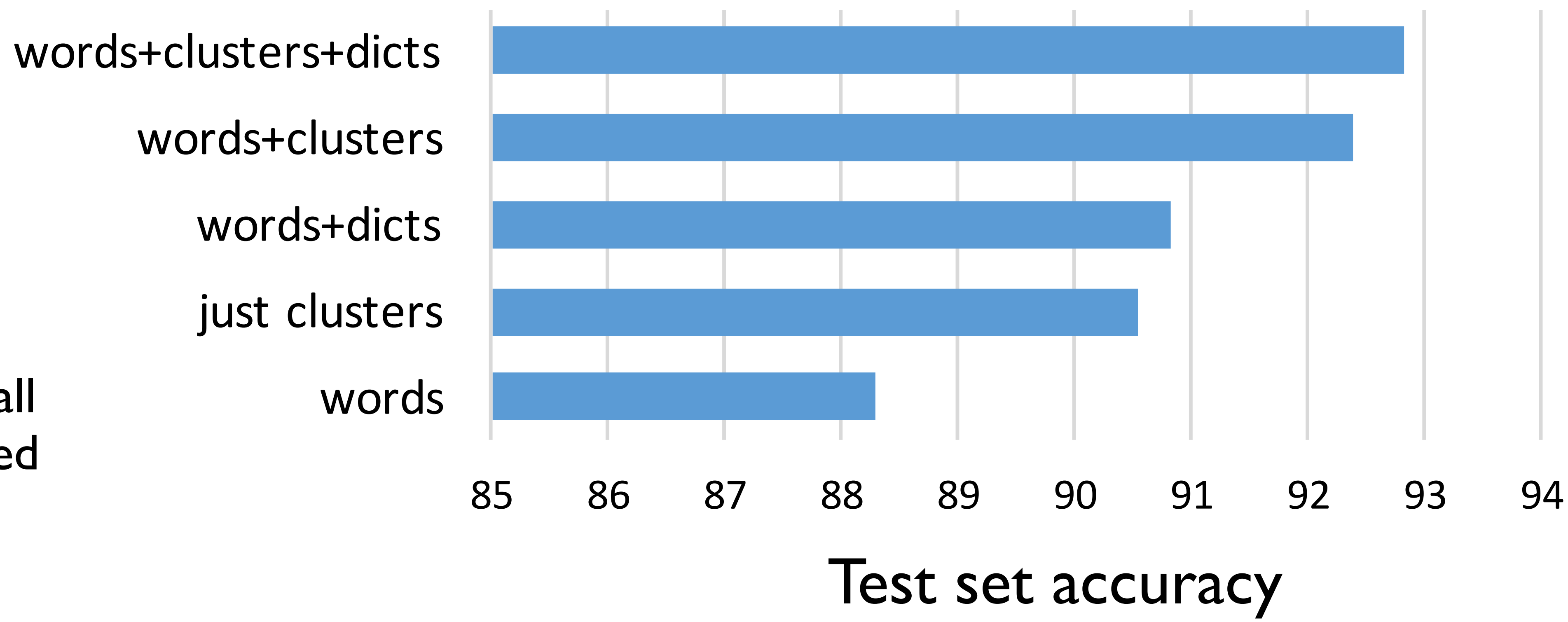
ill ima imma i'ma i'mma ican iwanna umma imaa #imthetypeto iwill amma #menshouldnever igotta #whywouldyou #iwishicould #sometimesyouhaveto #thoushallnot #ihatewhenpeople illl #thingspeopleshouldnotdo #howdareyou #thingsgirlswantboystodo im'a #womenshouldnever #thingsblackgirlsdo immma iima #ireallyhatewhenpeople 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

Clusters help POS tagging



"words": all handcrafted features

- A little annotation + lots of data
- Unsupervised word representation learning (clusters, embeddings) is a crucial technique in NLP

- 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

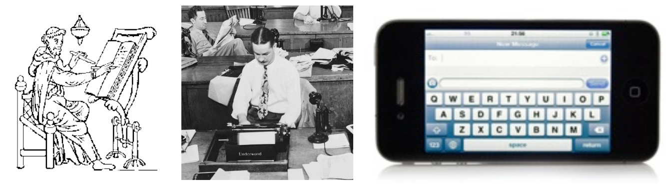
@Charismatic_Cee ay shoot me those pictures you calm you got of us n **imma** post em on the gram !

TextGenerator(SocialAttributes) → *Text*

Society *(SocialAttributes)*

Data generation process

Writing *(TextGenerator)*



Text Data *(Text)*

HATHI TRUST Digital Library

<u>What to analyze:</u>

Social phenomena in social media datasets

- Political speech under Chinese censorship
- Events in international relations
- Social factors in language use

<u>How to analyze:</u>

NLP capabilities we need to do these better

- Part of speech tagging
- Entity extraction
- Syntactic, semantic parsing

What social bias exists in NLP models?

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

Su Lin Blodgett

Lisa Green

Brendan O'Connor

EMNLP 2016

What social bias exists in NLP models?

Dialect

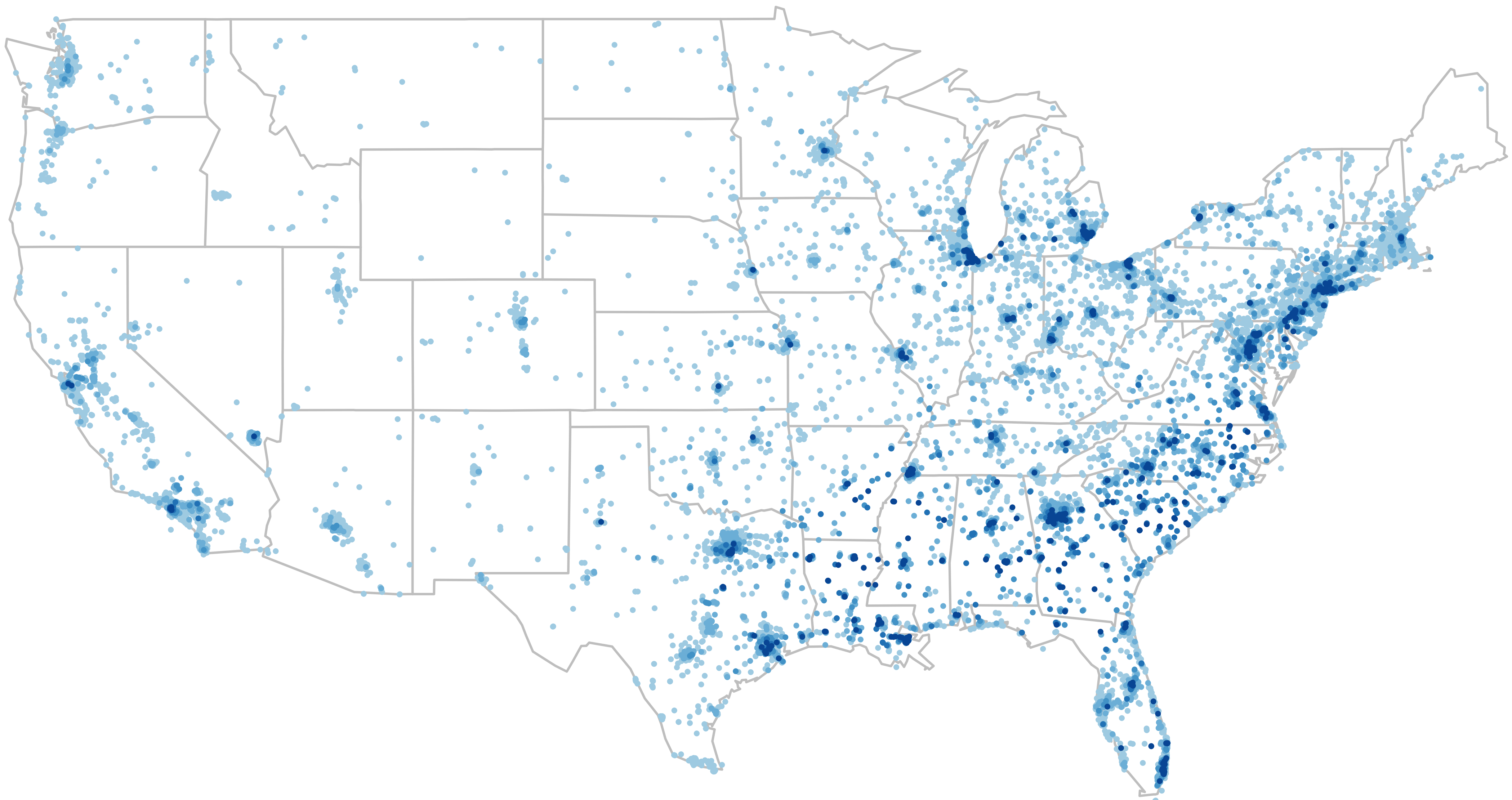

he woke af smart af educated af daddy af coconut oil using af GOALS AF & shares food af
RETWEETS
3
LIKES
42
1:08 AM - 8 Jul 2016
3
42

Dialect

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

Why is social media different?

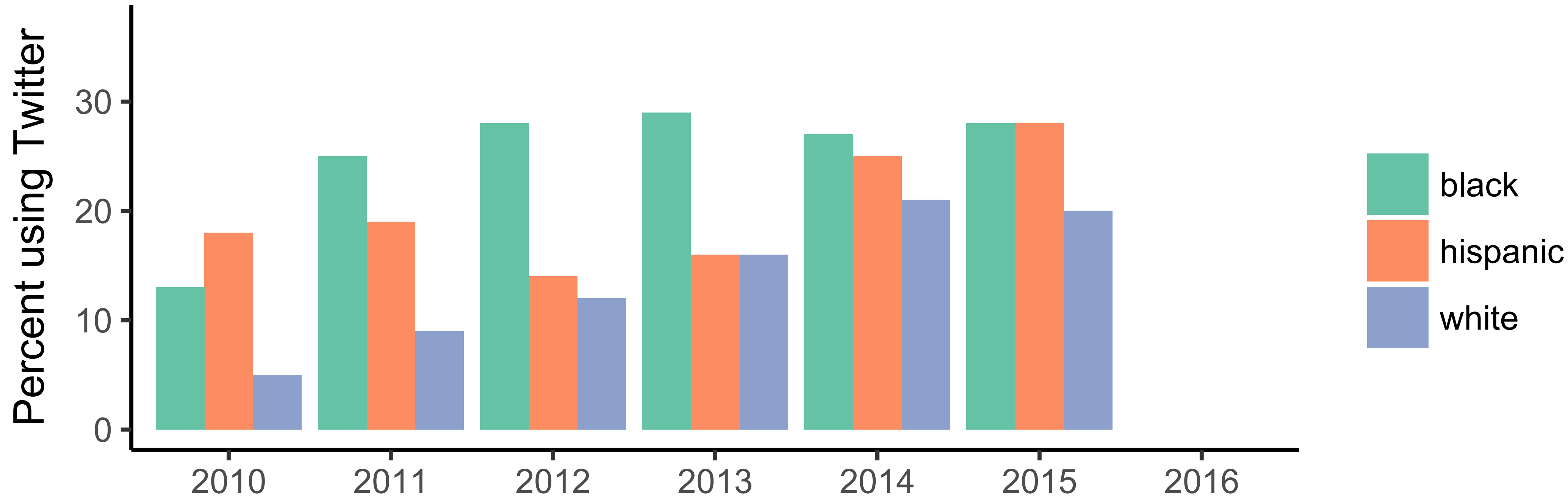
- Internet speech?
- Pre-existing dialectal English?
  - Geographic patterns of word usage often reveal relationships to race, ethnicity etc.
  - African-American English in Twitter
    [Eisenstein 2013, Jorgensen et al. 2015, Jones 2015]



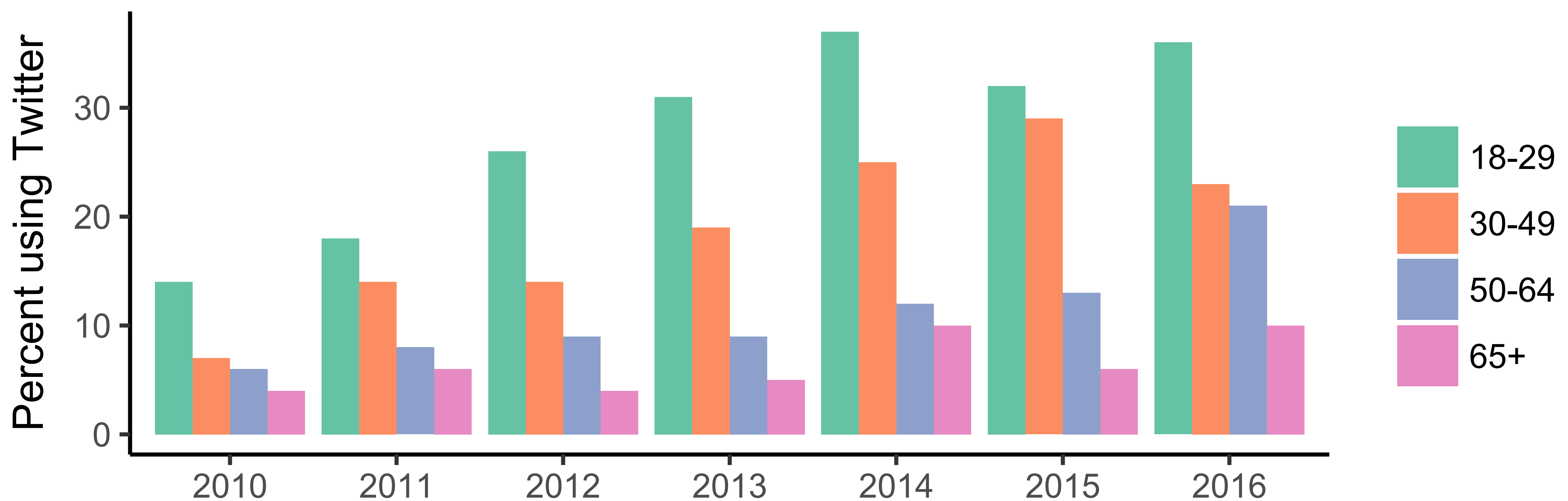
Youth, minorities on Twitter

[Pew Research]

P(use twitter | race)

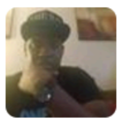


P(use twitter | age)



- From U.S. Census data and geo-located tweets: identify demographic-specific terms and messages via probabilistic model
- Validate African-American-associated corpus against linguistics literature on African-American English
- Investigate racial disparities in natural language processing tools

Associating geolocated tweets with demographics



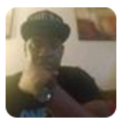
King Me
@tmac_datboi

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

block group 010730039001

Bored af den my phone finna die!!!

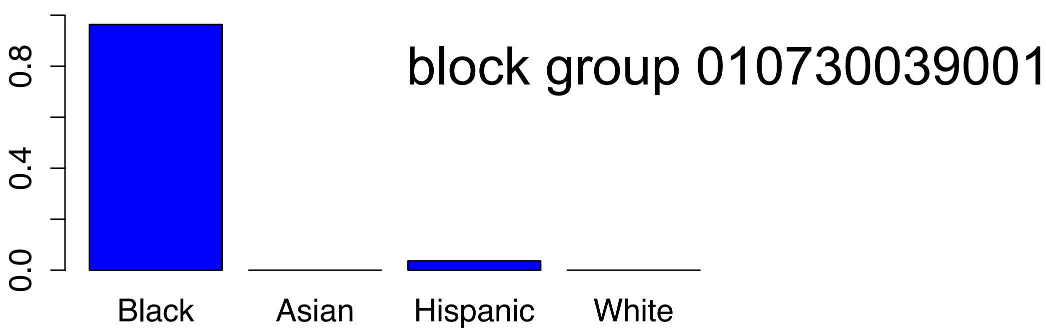
block group 010730058003

Associating geolocated tweets with demographics

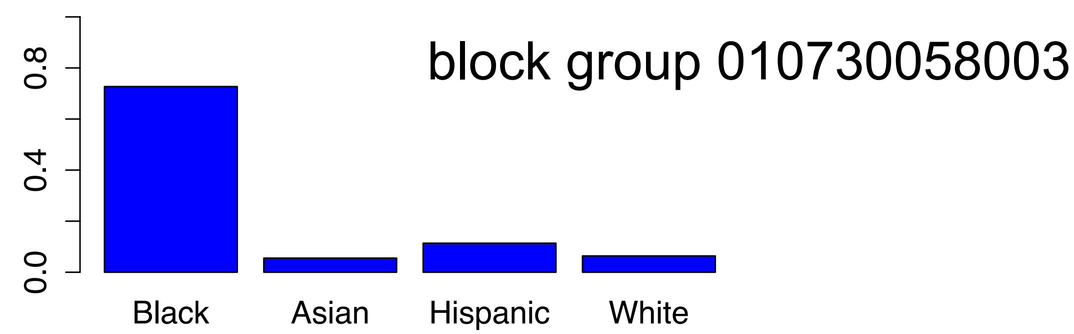
King Me
@tmac_datboi

Follow

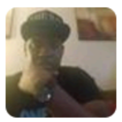
he woke af smart af educated af daddy af
coconut oil using af GOALS AF & shares food af



Bored af den my phone finna die!!!



Associating geolocated tweets with demographics

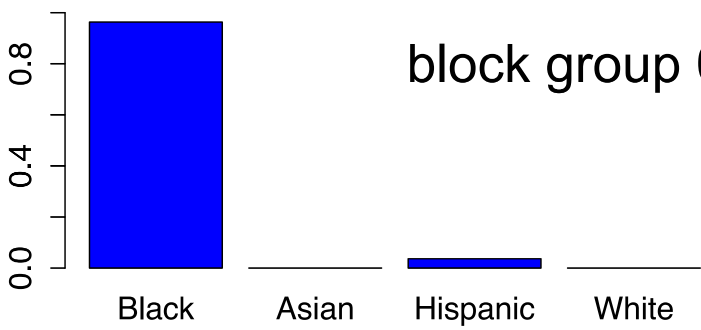


King Me
@tmac_datboi

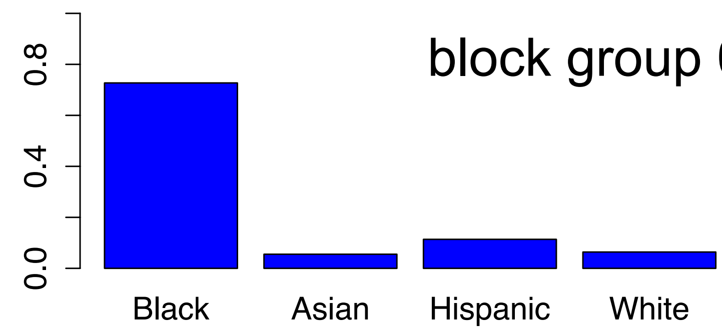
Follow

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!!!

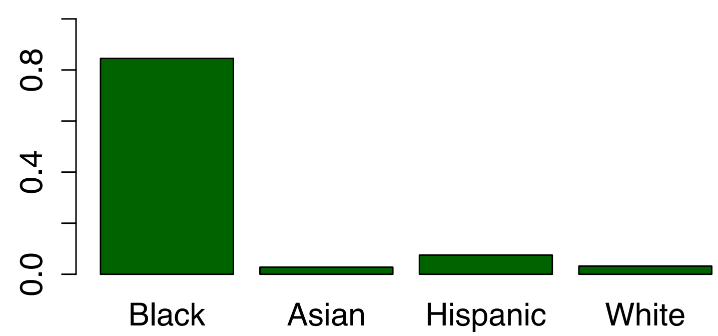


block group 010730039001

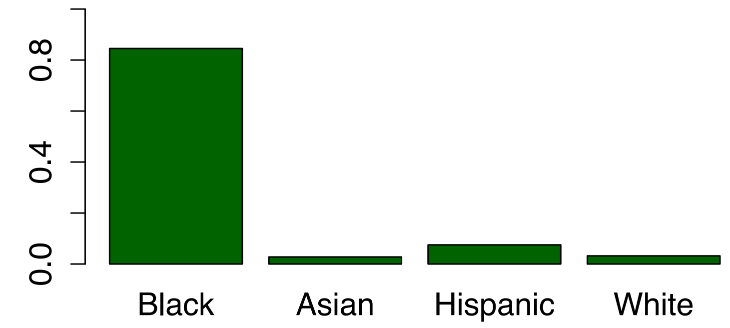


block group 010730058003

$\pi_{user} =$



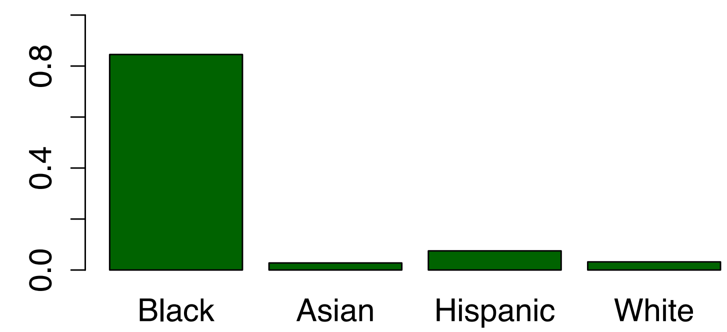
Mixed membership model

$\pi_{user} =$ 

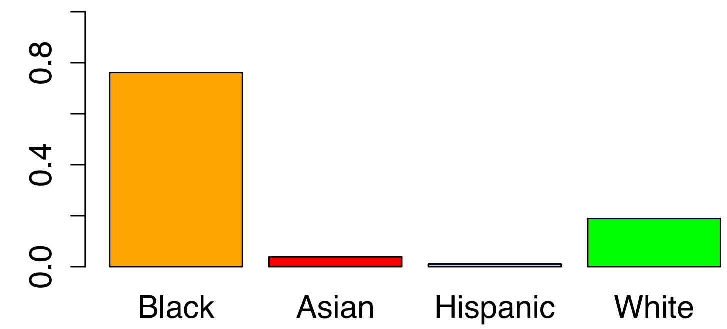
Mixed membership model

$\theta_{msg} \sim \mathrm{Dir}(\alpha\pi)$

$\pi_{user} =$



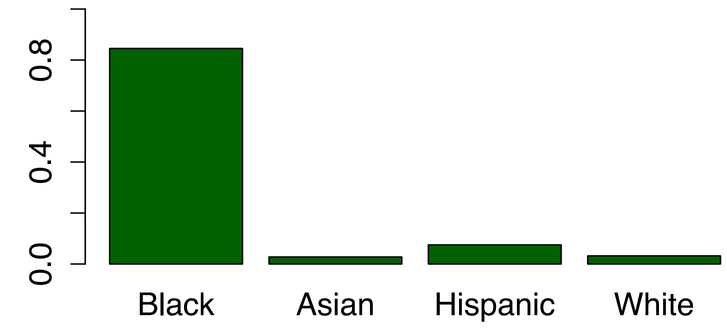
$\theta_{msg} =$



Mixed membership model

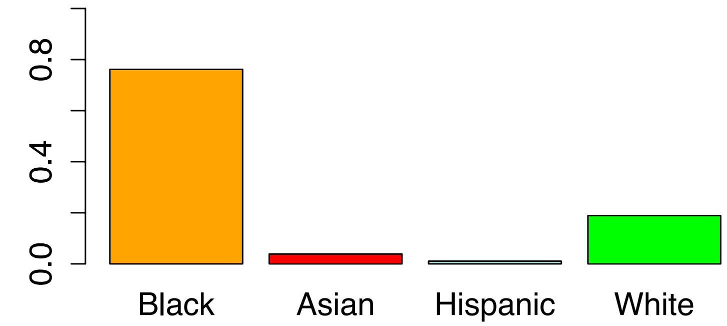
$\theta_{msg} \sim \mathrm{Dir}(\alpha\pi)$, $z \sim \theta_{msg}$, $w \sim \phi_z$

$\pi_{user} =$



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

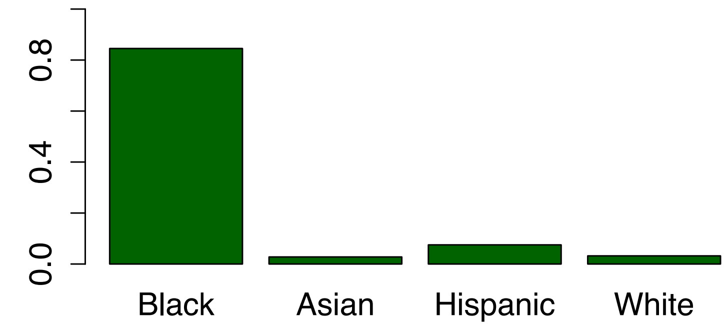
$\theta_{msg} =$



Mixed membership model

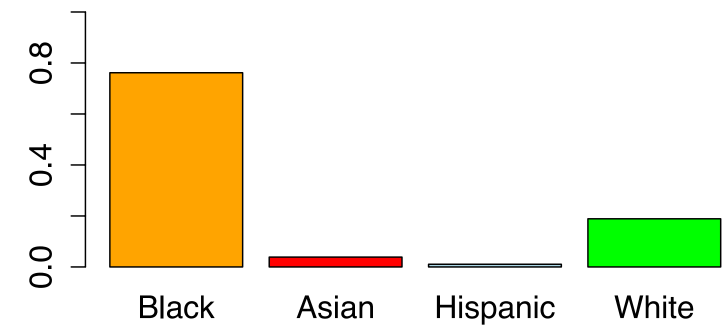
$\theta_{msg} \sim Dir(\alpha\pi),\ \ z \sim \theta_{msg},\ w \sim \phi_z$

$\pi_{user} =$



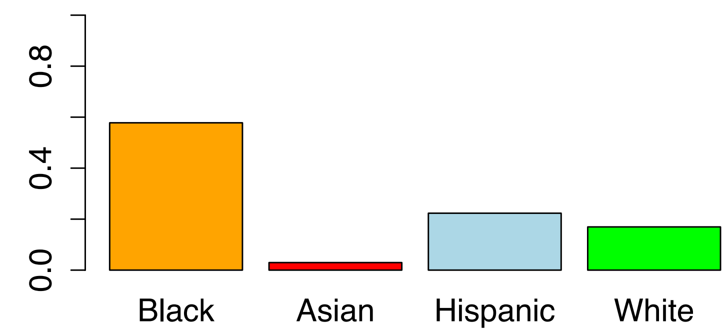
he woke af smart af educated af daddy af coconut oil using af GOALS AF & shares food af

$\theta_{msg} =$



Bored af den my phone finna die!!!

$\theta_{msg} =$



Mixed membership model

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

m_1

Bored af den my phone finna die!!!

m_2

Mixed membership model

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

m_1

Bored af den my phone finna die!!!

m_2

Word	AA	Asian	Hisp.	White
woke	1	0	0	0
af	6	0	0	0
educated	0	0	0	1
…				

Mixed membership model

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

m_1

Bored af den my phone finna die!!!

m_2

Word	AA	Asian	Hisp.	White
woke	1	0	0	0
af	6	0	0	0
educated	0	0	0	1
…				

Message	AA	Asian	Hisp.	White
m_1	7	0	0	2
m_2	2	0	1	1

Mixed membership model

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

m_1

Bored af den my phone finna die!!!

m_2

Word	AA	Asian	Hisp.	White
woke	1	0	0	0
af	6	0	0	0
educated	0	0	0	1
…				

Message	AA	Asian	Hisp.	White
m_1	7	0	0	2
m_2	2	0	1	1

User	AA	Asian	Hisp.	White
u_1	9	0	1	3

Corpus creation and linguistic validation

- Beyond unigrams: creation of user-level topic-aligned corpora

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

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

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

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

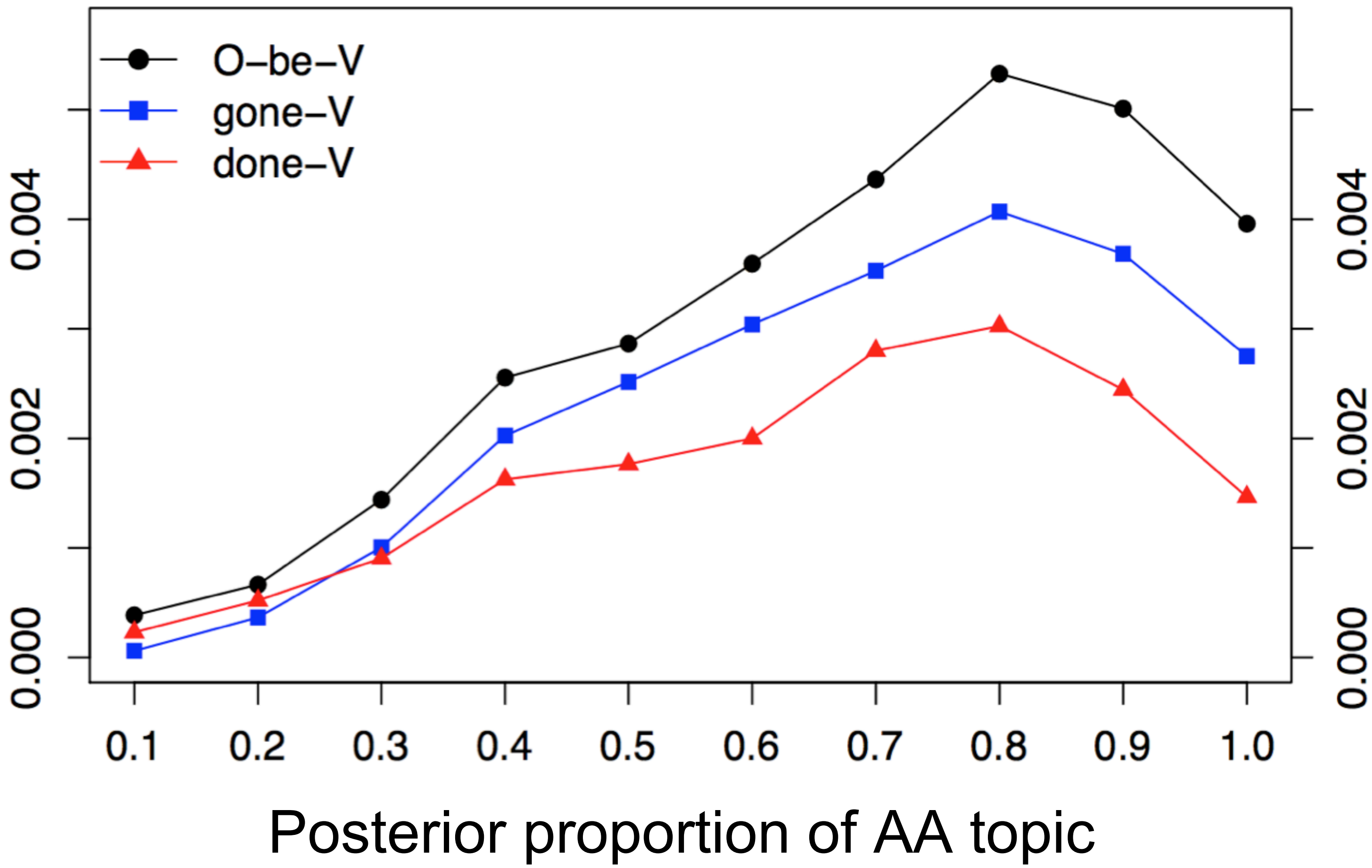
Syntactic analysis

- Select 3 well-known AAE verbal markers
- Search for sequences of unigrams and POS tags

Construction	Example
O-*be/b*-V	*I be tripping bruh*
gone/gne/gon-V	*Then she gon be single Af*
done/dne-V	*I done laughed so hard that I'm weak*

Syntactic analysis

Proportion of tweets with construction



Historical vs. Online?

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

Historical vs. Online?

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

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

Historical vs. Online?

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

Examining NLP tools - dependency parsing

- Compare annotated parses to systems' output parses

Examining NLP tools - dependency parsing

- Compare annotated parses to systems' output parses

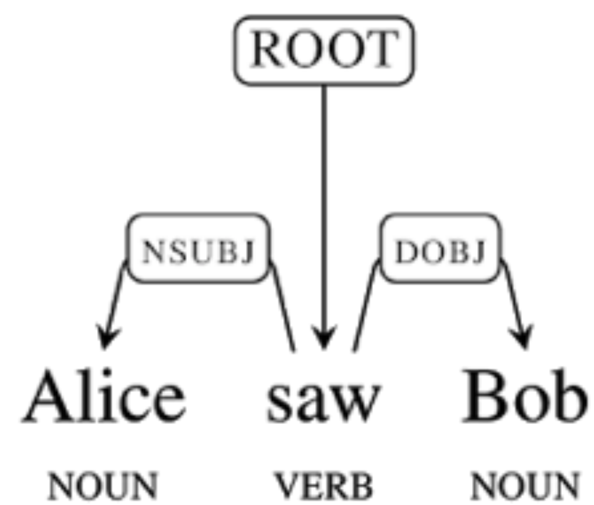
WIRED Google Has Open Sourced SyntaxNet, Its AI for Understanding Language

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)

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

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.

Examining NLP tools - language identification

- Language identification - key step in NLP pipelines

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)
```

Examining NLP tools - language identification

- Language identification - key step in NLP pipelines

	AA-Aligned	**White-Aligned**
langid.py	13.2%	7.6%

Proportion of messages classified as non-English

Examining NLP tools - language identification

- Language identification - key step in NLP pipelines

	AA-Aligned	**White-Aligned**
langid.py	13.2%	7.6%
Twitter	24.4%	17.6%

Proportion of messages classified as non-English

Examining NLP tools - language identification

- Solution: build ensemble classifier to augment langid.py
- Given a message, classifier:
    - Calculates langid.py's prediction
    - If prediction is English, return English
    - If not English, return English if our model's (AA + white + Hispanic) posterior probabilities ≥ 0.9
    - Otherwise, return langid.py's prediction

Examining NLP tools - language identification

- Solution: build ensemble classifier to augment langid.py

Message set	***langid.py***	**Ensemble**
High AA	80.1%	99.5%
High White	96.8%	99.9%
General	*88.0%*	*93.4%*

Imputed recall of English messages for
2014 messages

- Develop a model leveraging demographic correlations to generate dialectal corpora
- Corpus reproduces well-known dialectal phenomena
- Demonstrate disparity in performance by two kinds of NLP tools
- Provide ensemble classifier augmenting existing tools with our model

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