MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

The good and the bad of privacy in Social Media
Part II, the bad

Mainack Mondal

IIT KGP, Social Computing course, August 2017

Roadmap

Previous Lecture

How to **better protect privacy** in Online social media sites (OSMs) – the good of privacy

This lecture

Online abuse: The ill side-effect of privacy and how to defend against the online abuse – the bad of privacy

What do we mean by abusive language?

Anonymity and privacy -- required by many, abused by some

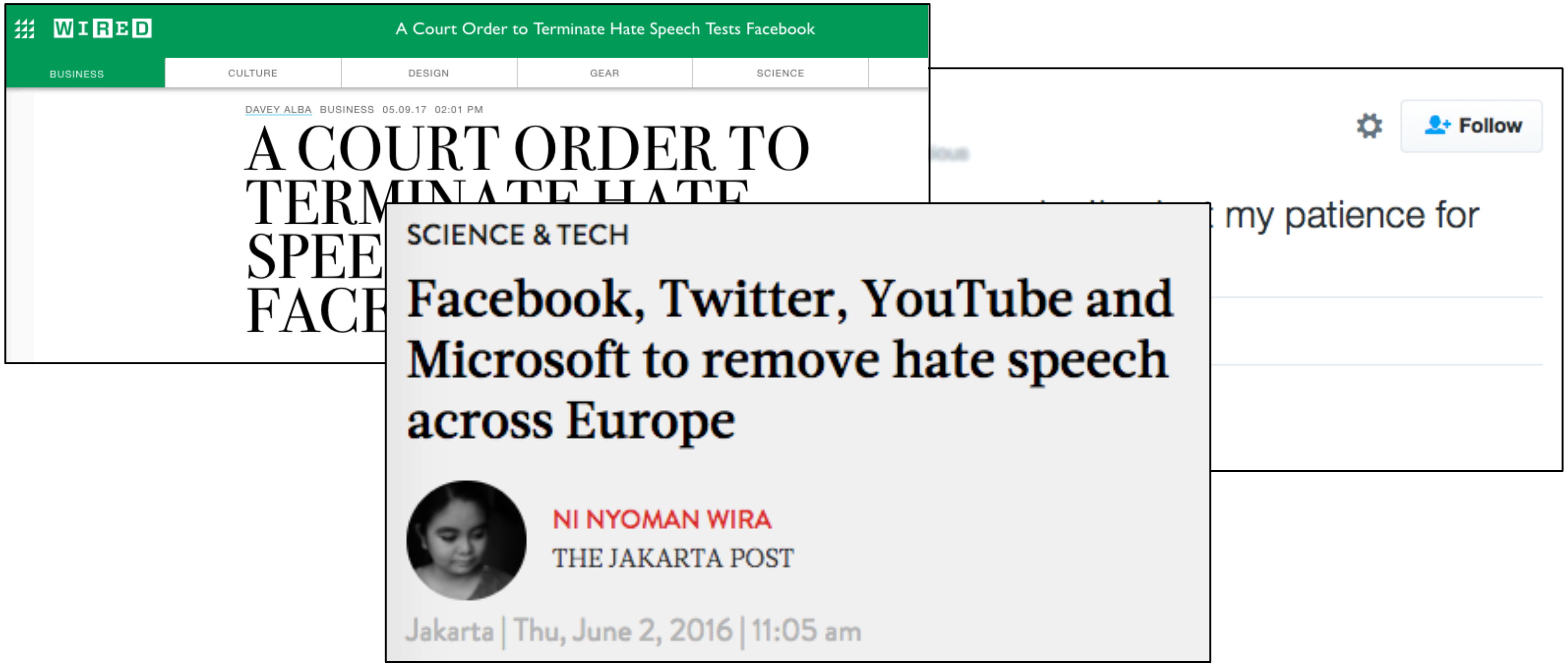
hate speech, cyberbullying, trolling

Types of abusive language

	Explicit	*Implicit*
Directed	"Go kill yourself", "You're a sad little f*ck" (Van Hee et al., 2015a), "@User shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga" (Davidson et al., 2017), "Youre one of the ugliest b*tches Ive ever fucking seen" (Kontostathis et al., 2013).	"Hey Brendan, you look gorgeous today. What beauty salon did you visit?" (Dinakar et al., 2012), "(((@User))) and what is your job? Writing cuck articles and slurping Google balls? #Dumbgoogles" (Hine et al., 2017), "you're intelligence is so breathtaking!!!!!!" (Dinakar et al., 2011)
Generalized	"I am surprised they reported on this crap who cares about another dead n*gger?", "300 missiles are cool! Love to see um launched into Tel Aviv! Kill all the g*ys there!" (Nobata et al., 2016), "So an 11 year old n*gger girl killed herself over my tweets? ^_^ thats another n*gger off the streets!!" (Kwok and Wang, 2013).	"Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home." (Burnap and Williams, 2015), "most of them come north and are good at just mowing lawns" (Dinakar et al., 2011), "Gas the skypes" (Magu et al., 2017)

Source: Waseem et al. (https://arxiv.org/pdf/1705.09899.pdf)

Hate speech: A serious problem for OSMs



OSMs and **Governments** are trying hard to **combat hate speech**!

How to detect hate speech?

Survey: http://www.aclweb.org/anthology/W/W17/W17-1101.pdf

Standard workflow

Extract **features**

Learn using unsupervised/supervised method

What features do people use?

Content based

- **Simple surface features**: Bag of words, unigrams
- **Word generalizations**: Use synonyms of relevant words
- **Sentiment** of the content
- **Lexical** resources: lists of swear words
- **Linguistic** features: POS, politeness, type dependency relationships
- **Knowledge-based**: Conceptnet
- **Meta information**: user message history
- **Checking accompanying** image/video data

User based

- **Role of the author** while posting

Network based

- If **people if your network** is posting hate speech

Characterizing Hate speech in OSMs

UNESCO reviewed OSM policies to combat hate speech

- OSMs should **better leverage** the **data**
- Need a **better understanding** of online **hate speech characteristics**

[ICDCIT'12]
[First Monday'15]

Prior work on detecting hate speech **in specific context**

- Used text based, user based, network structure based features
- E.g., hate speech against African-Americans in US

No investigation so far about

- Understanding the **characteristics** of **general OSM hate speech**

Goal

Better characterizing general hate speech in OSMs

Rest of the talk

- ✔ Who are the targets of Hate in OSMs?
- ✔ Does anonymity play any role on hate speech?
- ✔ Does hate speech vary across geography?
- ✔ What is the context of hate speech?

Rest of the talk

- ✔ **Who are the targets of hate in OSMs?**
- ✔ Does anonymity play any role on hate speech?
- ✔ Does hate speech vary across geography?
- ✔ What is the context of hate speech?

Collecting generic hate speech data

Our definition of **hate speech**

*an **offensive post**, motivated, in whole or in a part, by the **writer's bias** against an **aspect of a group of people***

Desirable characteristics of a dataset

Uploader should **express hate** in the post **against a group of people**

The precision should be high for our dataset

Our idea: Leverage the **sentence structure**

Detecting hate speech with sentence structure

I	really	hate	black people
	< intensity >	*< user intent>*	*< hate target>*

Manually collected list of adverbs/ intensifiers or blank

E.g., Really, do

Synonyms of the word "Hate" from a dictionary

E.g., Detest, loathe, abhor

E.g., black people, fat people, n-word

Template 1: " *** people** " , e.g., ghetto people

Template 2: **words from "Hatabase"**, a crowd-sourced hate target database. e.g., n-word

Our hate speech dataset

We used this technique on English **Whisper and Twitter posts**

- Data collected over June '14 to June'15
- Total **20,305 Twitter** and **7,604 Whisper** hate speech posts

Very **high precision** and **not about any specific context**

Classifying the hate targets in OSM

We manually classified the hate targets in nine categories

Categories	Example hate targets
Race	n**ga, black people, white people
Behavior	insecure people, autistic people
Physical aspect	Ugly people, fat people
Sexual orientation	gay people
Class	ghetto people
Gender	pregnant people, c*nt
Ethnicity	Chinese people, paki
Disability	retards
Religion	Jewish people

Who are the top targets of hate?

Twitter		Whisper	
Hate target	% posts	Hate target	% posts
Race	48.73	Behavior	35.81
Behavior	37.05	Race	19.27
Physical aspect	3.38	Physical aspect	14.06
Sexual orientation	1.86	Sexual Orientation	9.32
Class	1.08	Class	3.63

"**Soft targets**" contribute to **a large part** of OSM hate speech

Rest of the talk

- ✔ Who are the targets of hate in OSMs?

  Hate based on Race, physical aspect or behavior is most prominent

- ✔ **Does anonymity play any role on hate speech?**

- ✔ Does hate speech vary across geography?

- ✔ What is the context of hate speech?

How to measure the effect of anonymity?

Challenge: Need to **compare** the behavior of **anonymous and non-anonymous** accounts in **same** OSM

We used **posts from Twitter** for our purpose

- **Twitter** have **weak identity**
- Twitter accounts are **not** needed to be associated with a **real identity**

We leverage **Facebook's real name** policy

- Dataset of 100 million unique Facebook names
- Create database of **millions of personal names**

Twitter account without **personal names** are "**anonymous**"

Does anonymity enable hate speech?

Category	%tweets posted anonymously (Without personal names)
Random tweets (Baseline)	40%
Race	55%
Sexual orientation	54%
Physical	49%
Behavior	46%

Users post **more hate speech anonymously**!

Rest of the talk

- ✔ Who are the targets of hate in OSMs?

  Hate based on Race, physical aspect or behavior is most prominent

- ✔ Does anonymity play any role on hate speech?

  Users post more hate speech anonymously

- ✔ **Does hate speech vary across geography?**

- ✔ What is the context of hate speech?

Top hate speech categories across countries

Whisper posts contain city level location

- We focus on hate speech posted from US, UK, Canada
- Contributed total 92% of hate speech in our Whisper dataset

Top 3 hate categories		
US	**Canada**	**UK**
Behavior	Behavior	Behavior
Race	Physical aspect	Physical aspect
Physical aspect	Race	Sexual orientation

Hate speech **categories** from **different countries** are **different**

What are the corresponding top hate targets?

Top hate targets across countries

We focus on top hate targets in US, UK, Canada

Top 3 hate targets		
US	**Canada**	**UK**
Black people	Fat people	Fat people
Fake people	Stupid people	Gay people
Stupid people	Fake people	Stupid people

There are **country specific biases** even for **hate speech**

What about hate speech within a country?

Comparing hate speech across US states

Raw volume of hate speech from a state may be biased

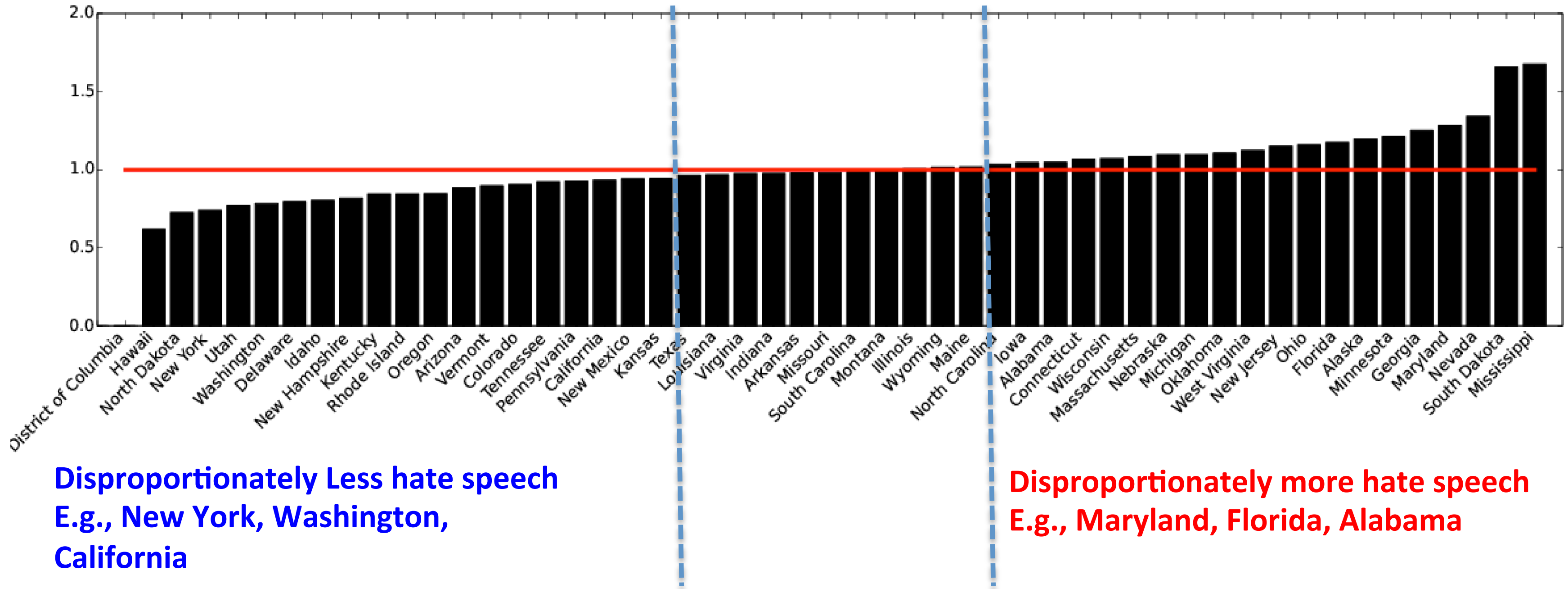
- More hate speech might simply be due to more uploaded posts

We compute **relative amount of hate speech** for each US state

- Divide **% of hate speech from state X** with **% of total posts from X**
- Value **more than one** implies comparatively **more hate speech**
- Value **less than one** implies comparatively **less hate speech**

Does some **US state** post **relatively more/less** hate speech?

Comparison of related amount of hate speech posted by US states



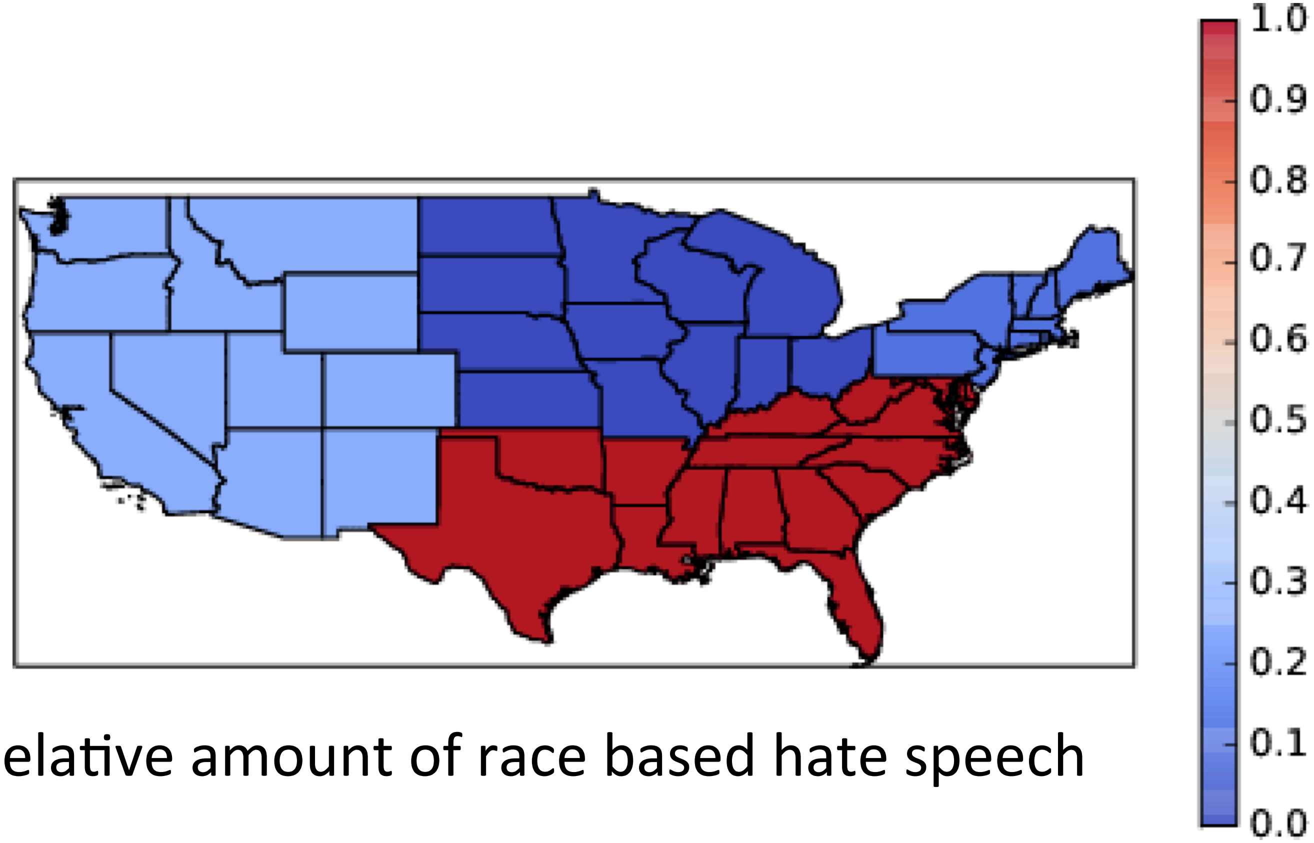
Disproportionately Less hate speech
E.g., New York, Washington, California

Disproportionately more hate speech
E.g., Maryland, Florida, Alabama

US **states** upload **disproportionately more/less** hate speech

How **hate speech from different categories** are posted?

Comparison for hate categories across US states



Relative amount of race based hate speech

More race based **hate speech** is uploaded **from southern states**

Rest of the talk

✔ Who are the targets of hate in OSMs?

Hate based on Race, physical aspect or behavior is most prominent

✔ Does anonymity play any role on hate speech?

Users post more hate speech anonymously.

✔ Does hate speech vary across geography?

Hate speech varies inter as well as intra country

✔ **What is the context of hate speech?**

Identifying the context of hate speech

I hate black people, their point of view is subhuman

Our detected
hate speech pattern

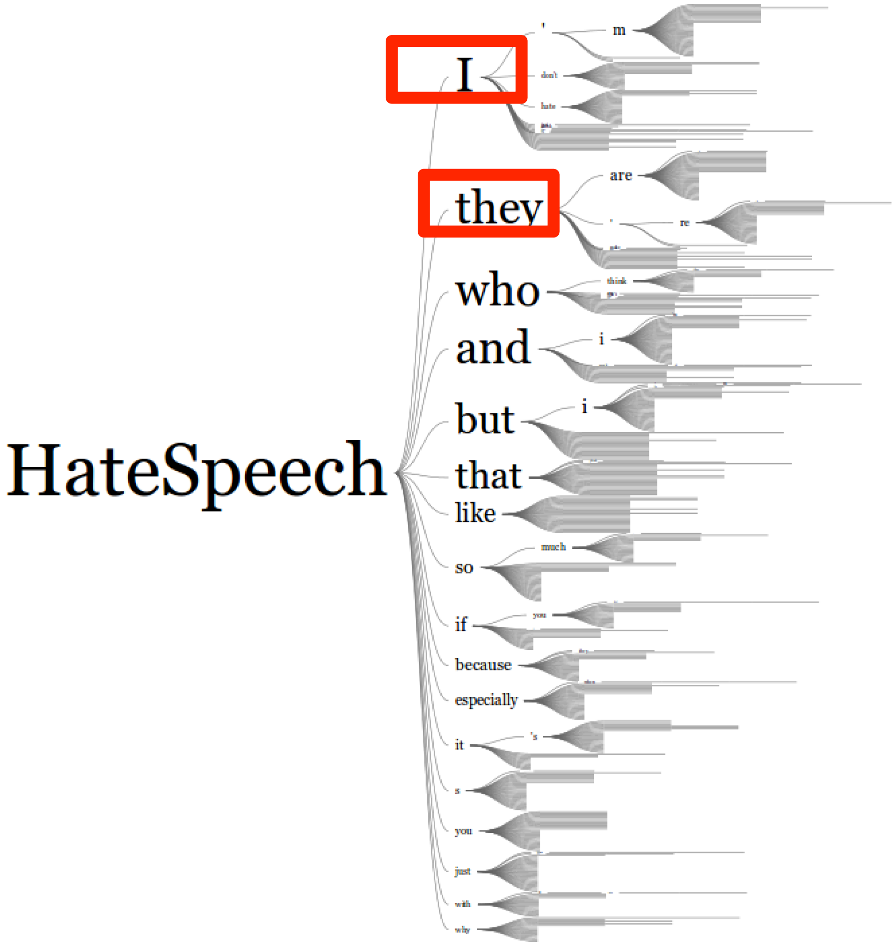
context of hate speech

For each of the hate speech post in our database

We **removed** the **detected hate speech pattern**

The **resulting** text gives us the **context**

Understanding the context of hate speech



Uploaders justify hate speech by including **personal opinions**

Summary

Created **a high precision hate speech dataset** from OSM posts

Hate based on **Race, physical aspect or behavior** is most prominent

Checked if anonymity have a correlation with posting hate speech

More **hate speech** is posted **anonymously**

Investigated Geographic variation in hate speech

There are both **inter and intra country variations** in hate speech

Uploaders **justify hate speech** by stating their **personal beliefs**

https://people.mpi-sws.org/~mainack/

Thanks

https://people.mpi-sws.org/~mainack/