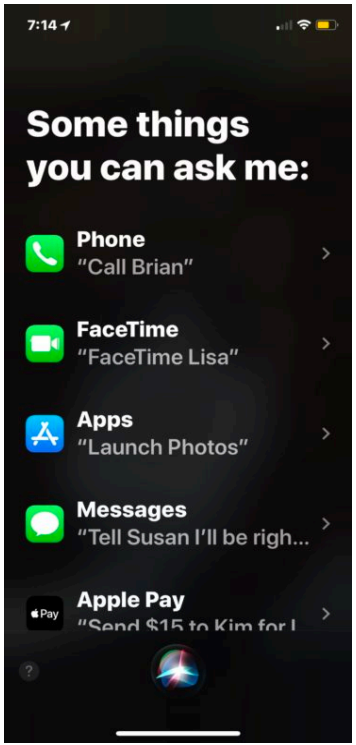
Applying AI for Hate Speech Detection

Xiang Ren
Department of Computer Science
USC

About Me

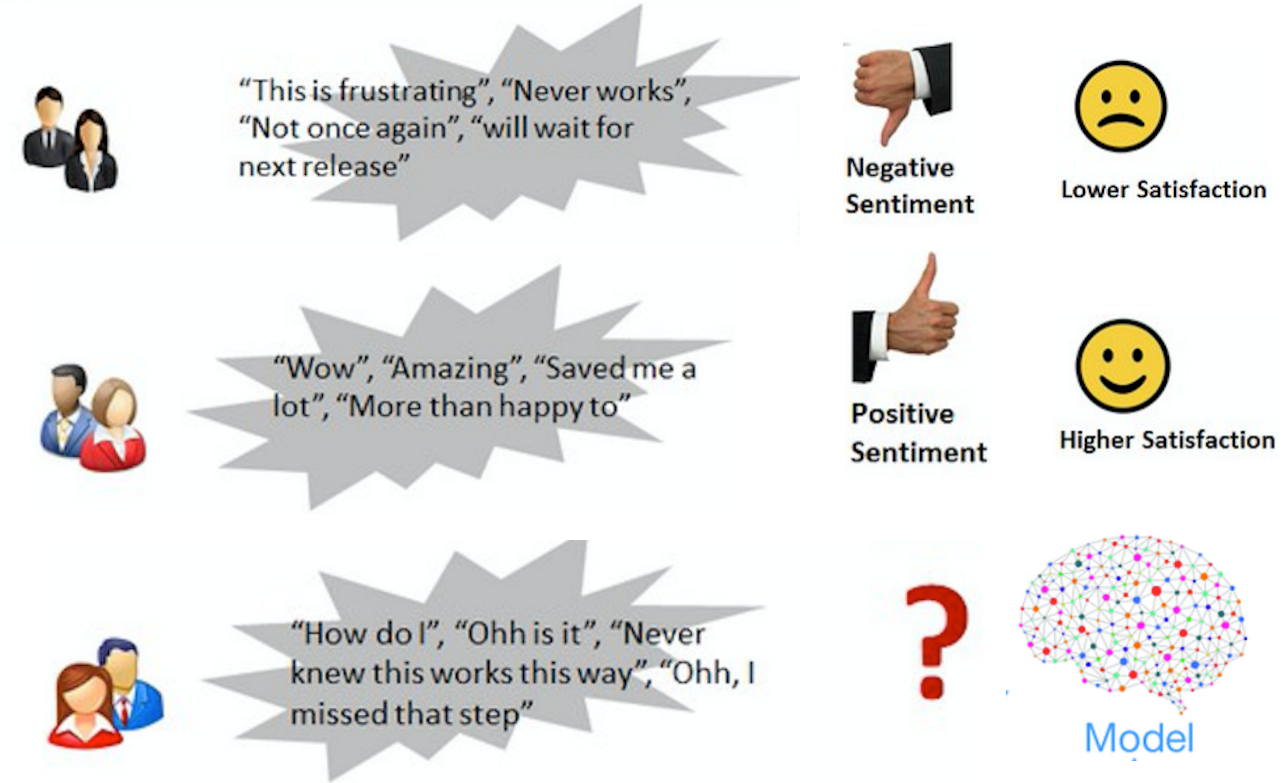
- Assistant professor at USC Computer Science Department
- Did my doctoral study at University of Illinois at Urbana Champaign (Computer Science); spent time at Stanford University; consult for Snapchat
- I teach & do research on ***natural language processing*** **(NLP)**

Example of NLP in "AI" Application: Siri



- Speech recognition: voice → text
- Language analysis
- Question answering
- Dialog processing
- Text to speech

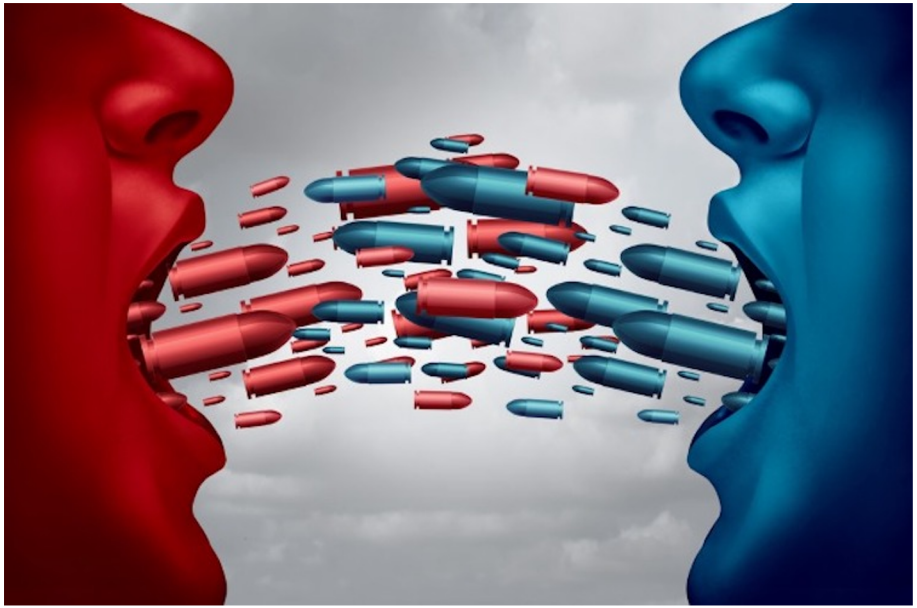
Applying NLP to determine sentiment of text



Can we apply NLP to help moderate hate speech on social media?

First, what is Hate Speech?

Hate speech expresses *prejudice* against someone's race, ethnicity, gender identity, religion, sexual orientation, nationality, or mental and physical disability.

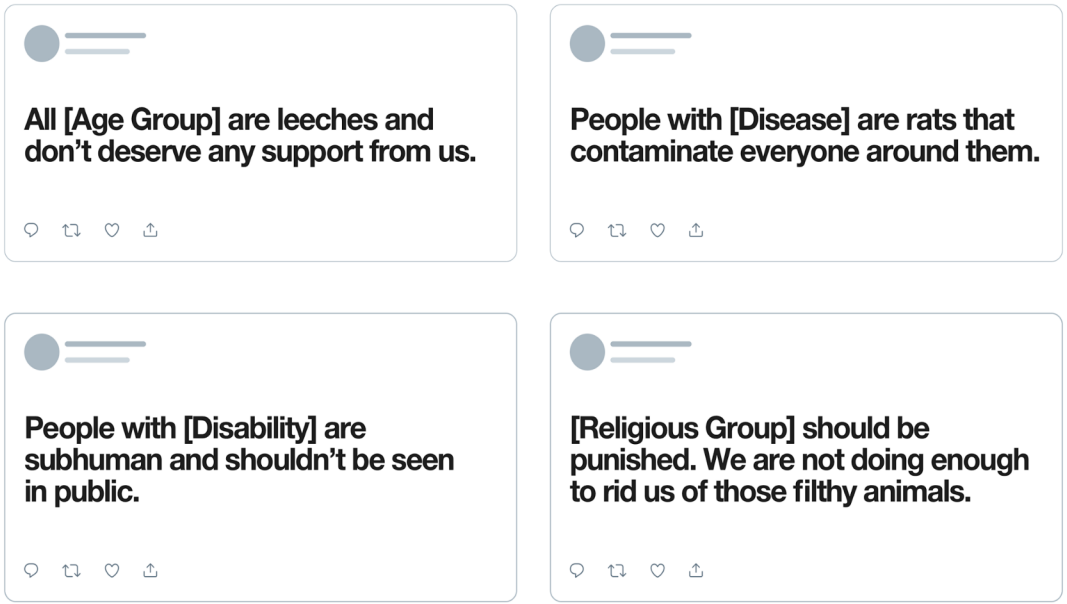


Warning: This presentation contains offensive language

Examples: hate speech on twitter

Source: twitter hate conduct policy
https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Examples: hate speech on twitter



Source: twitter hate conduct policy
https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Hate speech online is hard to combat...

- People can act anonymously and without retribution online
- Governments have been slow to develop laws against online hate speech

The harms of online hate speech

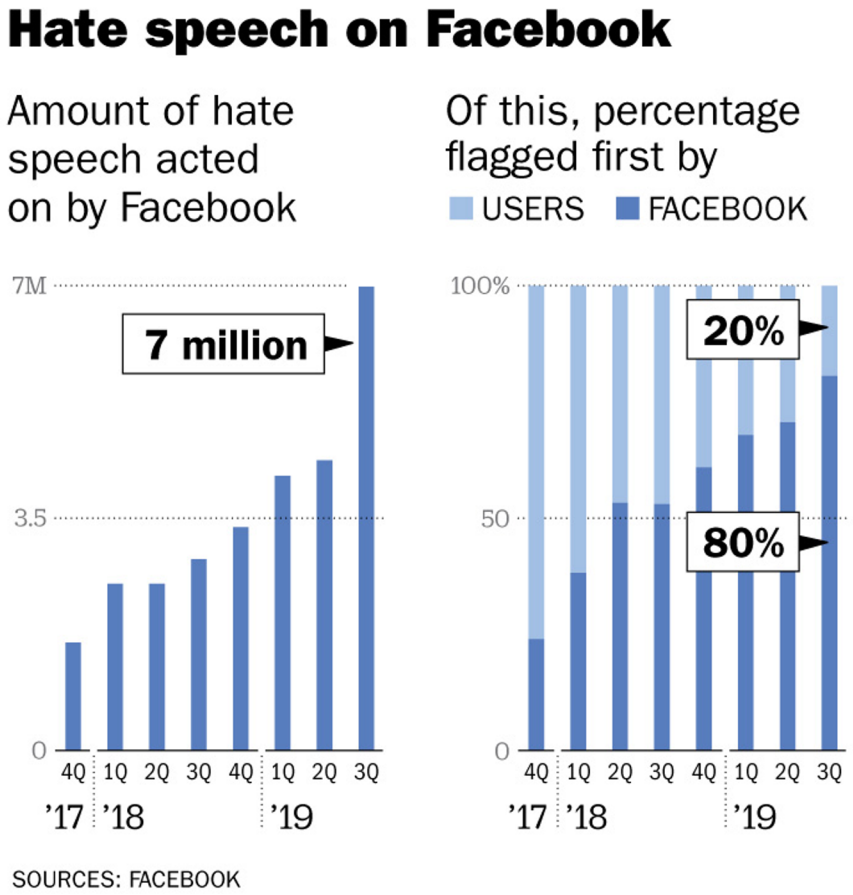
In 2018, a white supremacist posted his rants against Jews on his “Gab” social media account shortly before murdering 11 people in a synagogue.

Shannon Martinez of the Free Radicals Project, says:

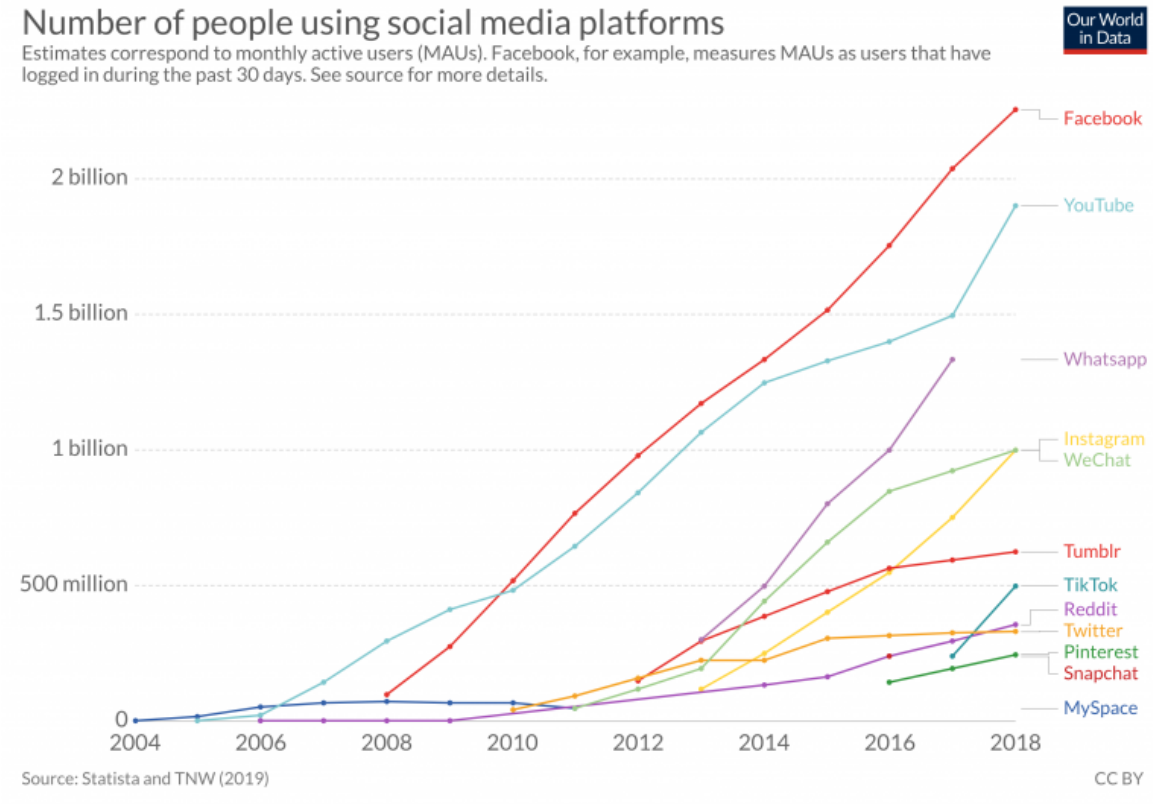
> The digital world gives white supremacists a safe space to explore extreme ideologies and intensify their hate without consequence.

Upward Trend

- More hate speech on Facebook is being flagged over time
- Even in 2020, companies have had a difficult time handling the problem of hate speech being used on their platforms (Vox, 2020)



Can we only rely on human moderators?



The human cost of flagging hate speech

When human annotators are charged with poring over hate speech and depictions or threats of violence, the results can be catastrophic.

In 2019, many human content moderators reported severe psychological trauma, PTSD, and one worker even committed suicide [The Verge, 2019]

The logistic cost of flagging hate speech

The sheer volume of content that human moderators must process is impossible under even healthy conditions:

- 7 millions posts per quarter
- 28 million per year
- If one post requires ~10 seconds to moderate, this amounts to **> 75,000 hours** of emotionally draining labor per year.

“Automating” Hate Speech Detection?

Jews must be expelled, by force if need be	?
Jewish holidays occur on the same dates every year in the Hebrew calendar	?
African music consists of complex rhythmic patterns	?
Africans will always be savages	?

***Note**: Hate examples taken from the “Gab” Hate Corpus (GHC; Kennedy et al., 2020)*

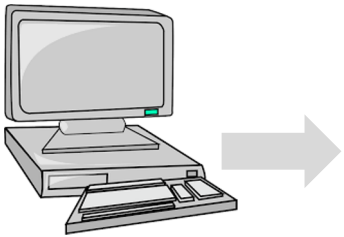
“Automating” Hate Speech Detection?

[Jews] must be expelled, by force if need be	Hate
Jewish holidays occur on the same dates every year in the Hebrew calendar	Not-hate
African music consists of complex rhythmic patterns	Not-hate
[Africans] will always be savages	Hate

***Note**: Hate examples taken from the “Gab” Hate Corpus (GHC; Kennedy et al., 2020)*

"Automating" Hate Speech Detection?

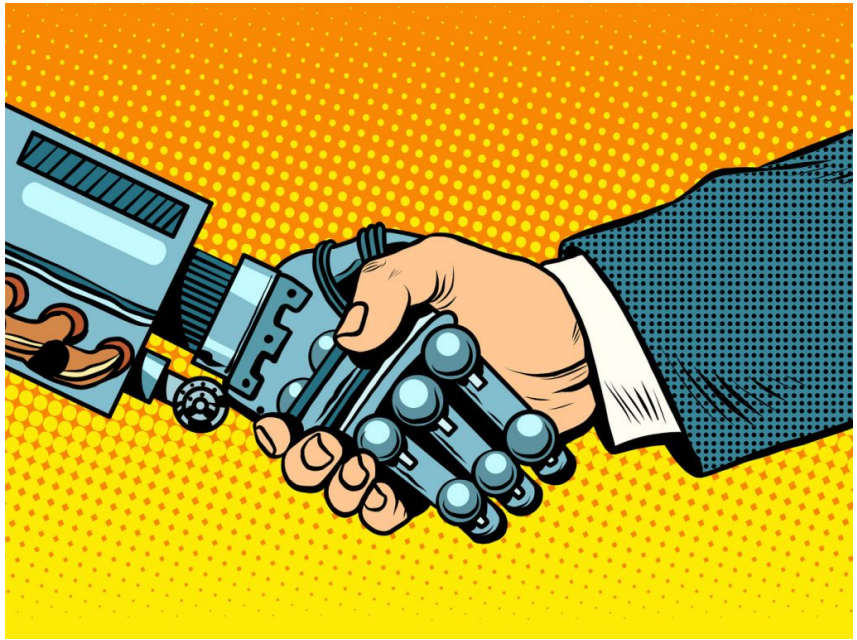
Instead of having human laborers label millions of messages, we build AI algorithms to detect hate speech



[Jews] must be expelled, by force if need be	Hate
Jewish holidays occur on the same dates every year in the Hebrew calendar	Not-hate
African music consists of complex rhythmic patterns	Not-hate
[Africans] will always be savages	Hate

***Note**: Hate examples taken from the "Gab" Hate Corpus (GHC; Kennedy et al., 2020)*

How can AI (NLP) help on this problem?

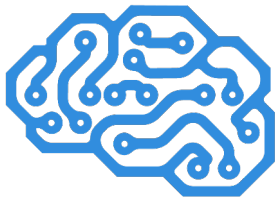


NLP for hate speech detection

Data collection

Text	Label
We respect the elderly	Non-hate
We respect the deaf	Non-hate
...	...
We hate the elderly	Hate

Model “training"



Prediction

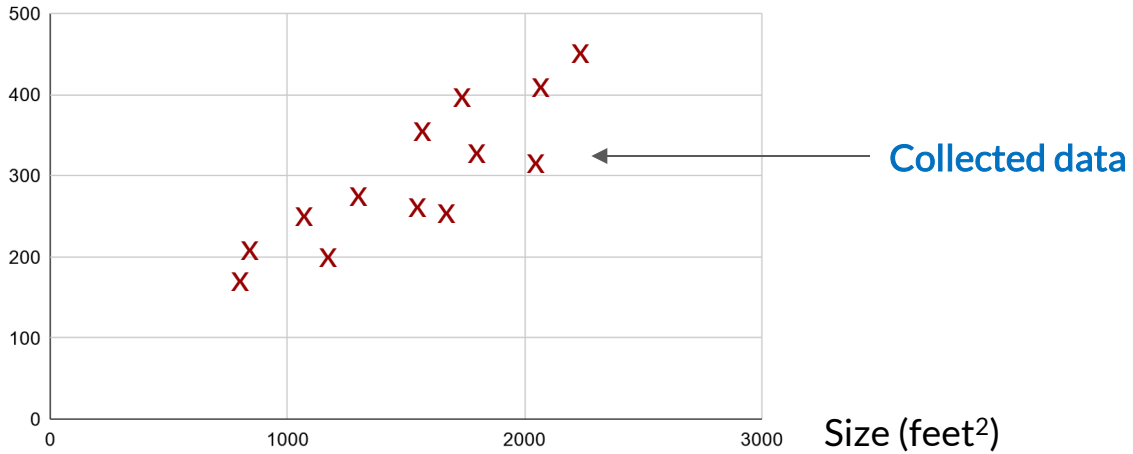
We hate the deaf

Hate speech

The science and math in NLP

- Machine “learns” from experience/examples

Housing prices
in 1,000 dollars



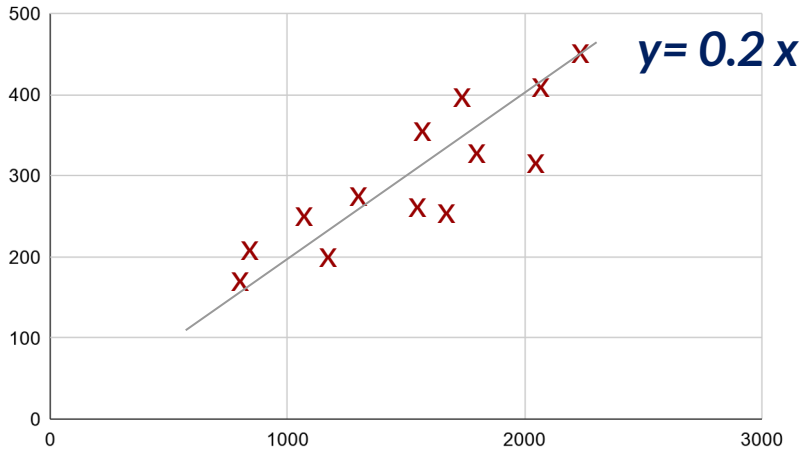
Can we estimate the price of a house of size 1,000 feet²?

Example from machine learning course by Andrew Ng on coursera.org

The science and math in NLP

- We can build a “model” using the collected data

Housing prices
in 1,000 dollars

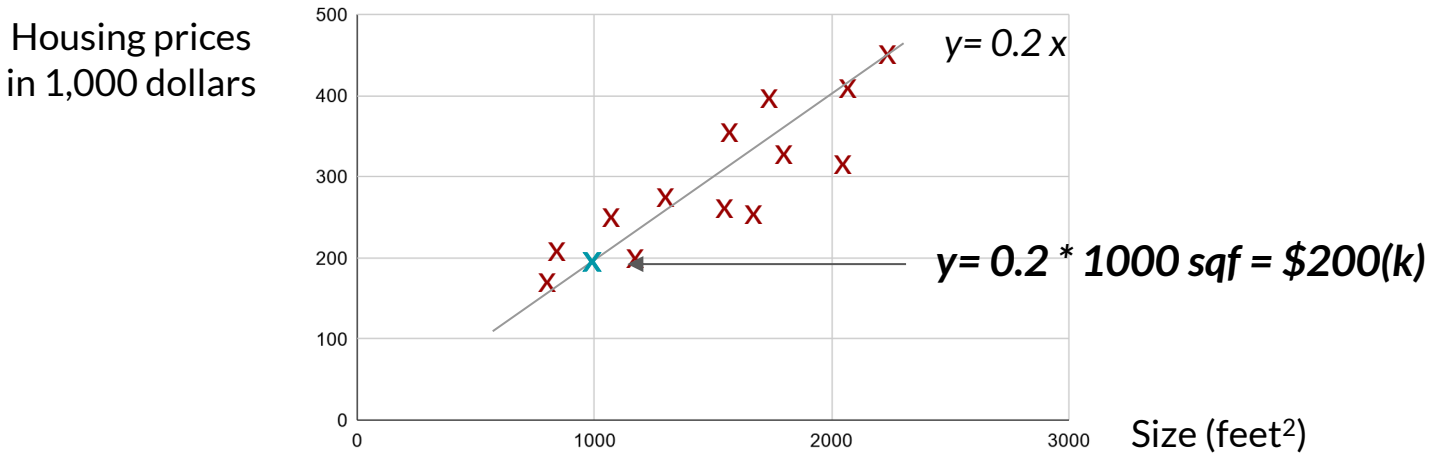


Size (feet^2)

Example from machine learning course by Andrew Ng on coursera.org

The science and math in NLP

- With the model, we can infer / predict the housing price on the requested size



Similarly, we can predict whether the housing price is above $200,000 or below *(classification)*

Example from machine learning course by Andrew Ng on coursera.org

Similarly, for hate speech detection

- A model based on *word presence*
  - The prediction is sum of binary "weights" of the words in a sentence

$$y = w_{we}x_{we} + w_{the}x_{the} + w_{deaf}x_{deaf} + w_{elderly}x_{elderly} + w_{respect}x_{respect} + w_{hate}x_{hate}$$

Similarly, for hate speech detection

"*We hate the **elderly***" →

Word	We	The	Deaf	Elderly	Respect	Hate
Presence (X)	1	1	1	0	0	1

$$y = w_{we} * 1 + w_{the} * 1 + w_{deaf} * 0 + w_{elderly} * 1 + w_{respect} * 0 + w_{hate} * 1$$

Similarly, for hate speech detection

A simple model:

Classify a sentence as "hate" if the word "*hate*" is mentioned in the sentence

Word	We	The	Deaf	Elderly	Respect	Hate
Weight (W)	0	0	0	0	0	**1**

Similarly, for hate speech detection

A simple model:

Classify a sentence as "hate" if the word "*hate*" is mentioned in the sentence

Word	We	The	Deaf	Elderly	Respect	Hate
Weight (W)	0	0	0	0	0	**1**

"We hate the **elderly**" →

$$y = w_{we} * 1 + w_{the} * 1 + w_{deaf} * 0 + w_{elderly} * 1 + w_{respect} * 0 + w_{hate} * 1$$

$$1\ (\text{hate}) = 0 * 1 + 0 * 1 + 0 * 0 + 0 * 1 + 0 * 0 + 1 * 1$$

So, is hate speech detection “solved”?

Issues of hate speech detection models

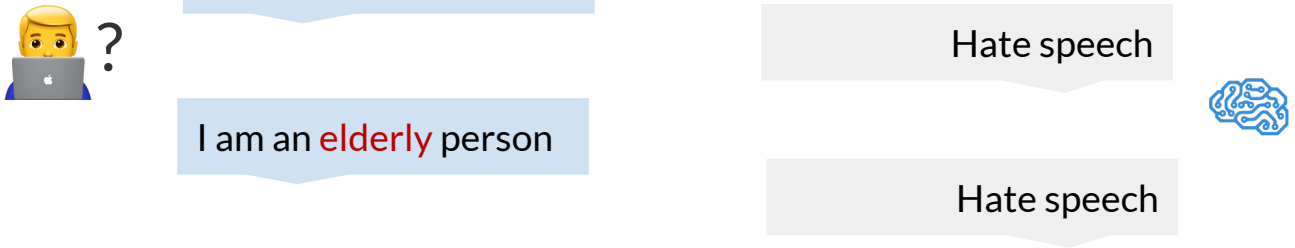
Issues of hate speech detection models

- Some social group terms frequently co-occur with the “hate” sentences

Text	Label
We hate the elderly	Hate
Being elderly is awful	Hate
...	...

Bias in hate speech detection

the model **performs differently** on text related to different social groups, and the difference may **cause harms** to the corresponding group



Negative consequence of bias

- Disportionate removal of posts mentioning / written by certain communities, impeding their participation in online platforms

- Yield negative perceptions of text mentioning / written by certain communities

- ...

Bias in hate speech detection - real world

- Bias in hate speech detection is an open problem
- Even very powerful hate speech detectors
  - Classify **28%** of New York Times articles with social group names as hate speech
- Hate speech detectors can also be biased in many other ways
  - Overly sensitive to some dialects, mentions of gender identity, etc.

Where does bias arise?

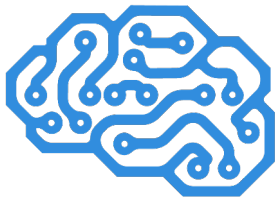
Bias arises

Data collection

Text	Label
We respect the elderly	Non-hate
We respect the deaf	Non-hate
...	...
We hate the elderly	Hate

Bias arises

Model training



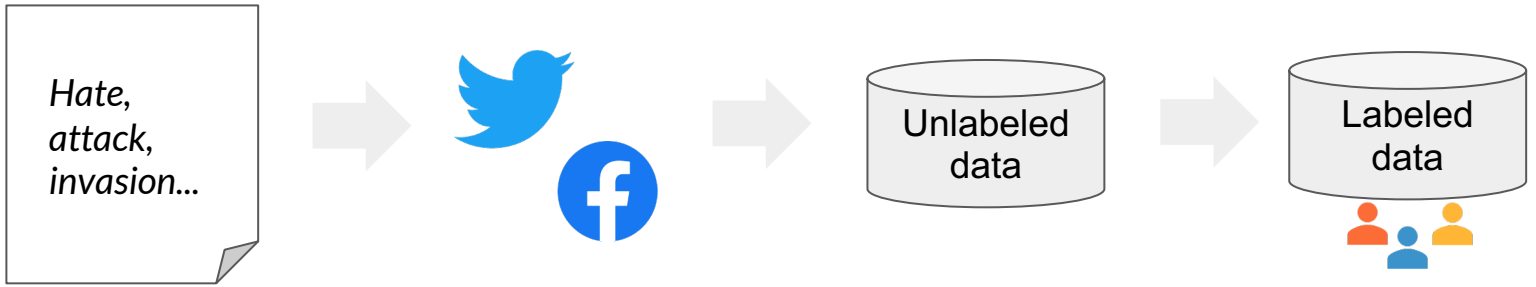
Predictions

We hate the deaf

Hate speech

Bias in data collection

- Hate speech is scarce in practice - to effectively constructing a hate speech detection dataset, data collection relies on a predefined set of keywords



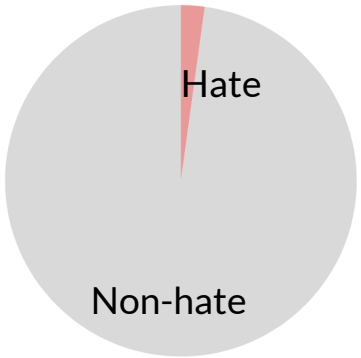
List of manually collected keywords

Find social media post containing these keywords

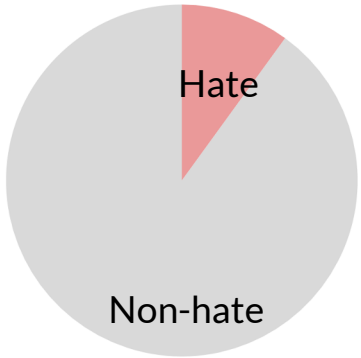
Annotate the dataset (hate vs. non-hate)

Bias in data collection

- The portion of hate speech related to a certain group can be higher / lower than the “real data distribution”



Real data mentioning “*elderly*” that are hate speech

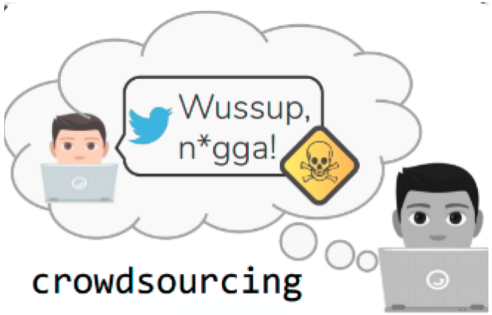


Data mentioning “*elderly*” that are hate speech *in the collected data*

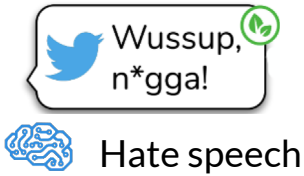
Bias in data collection

- Labels themselves can be incorrect

None of them are hate speech in right situations



Annotators imagine speakers, audience, and situations by themselves

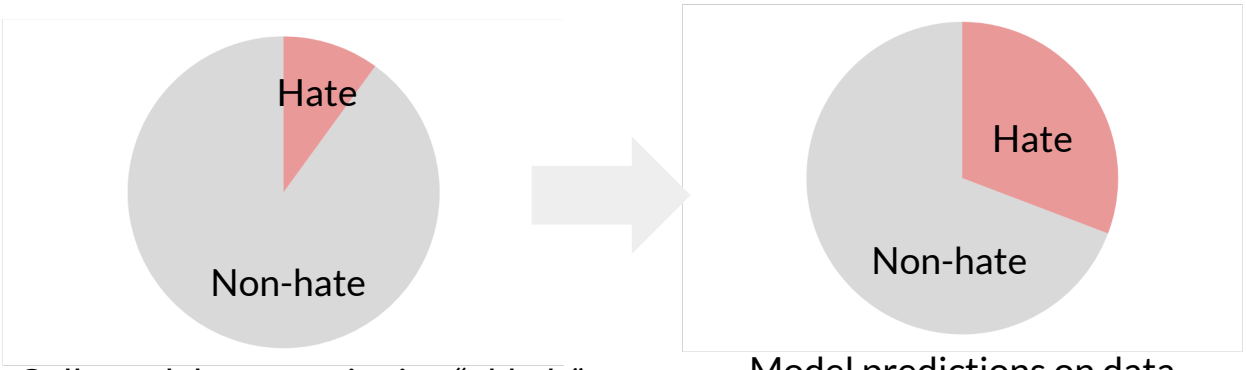


Model inherit mistakes from annotators

Sap et al. The risk of racial bias in hate speech detection, 2019

Bias in model training

- **Amplification of Bias**
  - Model may amplify a small differences in the dataset for different groups



Collected data mentioning "*elderly*"

Model predictions on data mentioning "*elderly*"

Why?

It depends on multiple factors: the choice of model type, data distribution...

Mitigating bias in model

- While how bias arises is complicated, there can be (simple) ways to mitigate bias
- For example: reducing the weights of tokens related to some protected groups

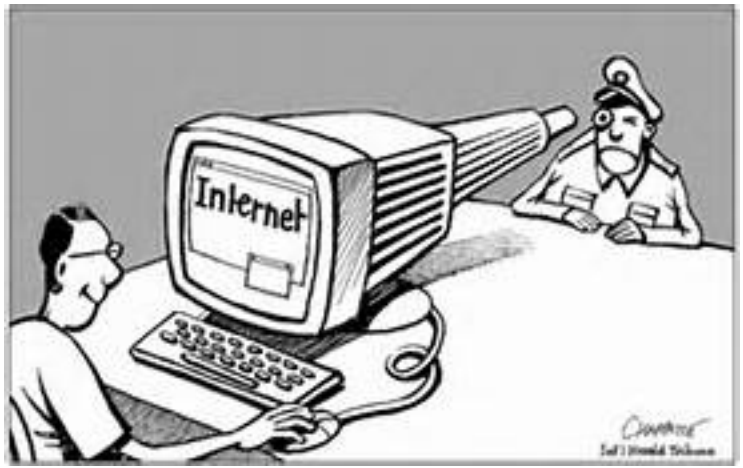
Word	We	The	Deaf	Elderly	Respect	Hate
Weight	0	0	**0**	**0**	0	**1**

Fix weights as 0 for group related terms

Implications & Discussion

Issues of automatic hate speech detection

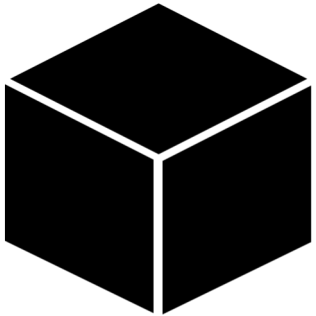
- Bias in hate speech detection model will cause negative societal impact
- Freedom of speech?



NLP models are like “black box”

Today’s NLP models are often called “Black Boxes”:

- They learn “features” and their “weights” from data and can perform very well, but often it is hard to understand *how* they are performing so well



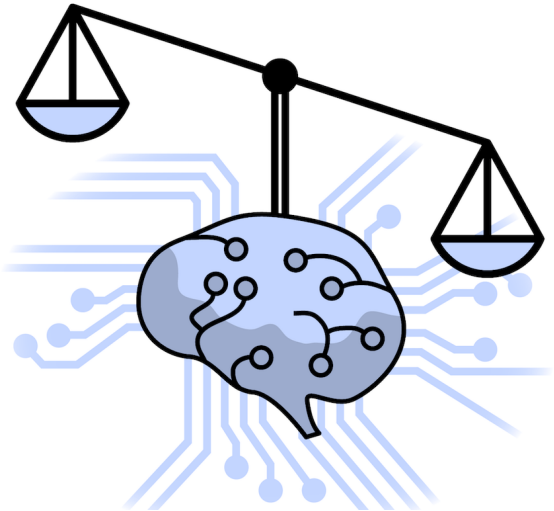
Emerging topics:

- Explaining what the models are doing --> open the black box
- Examine the potential bias in the model
- Mitigate the bias

"Fairness" of NLP models

Take bias and fairness as major metrics for evaluating a model

- instead of pursuing only accuracy



FAQ

xiangren@usc.edu