



# Hate speech: Detection, Mitigation and beyond

Tutorial at AAAI 2022

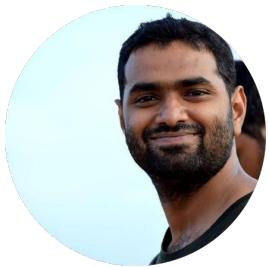
“



This presentation contains material that many will find **offensive** or **hateful**; however this cannot be avoided owing to the nature of the talk.



Animesh Mukherjee  
Twitter: [@Animesh43061078](https://twitter.com/Animesh43061078)



Binny Mathew  
Twitter: [@BinnyM](https://twitter.com/BinnyM)



Punyajoy Saha  
Twitter: [@punyajoysaha](https://twitter.com/punyajoysaha)

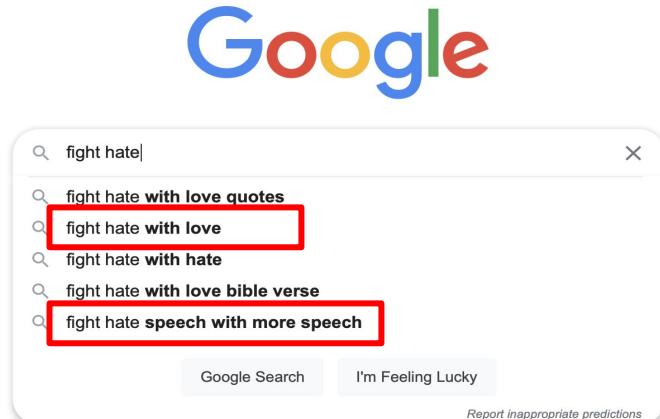
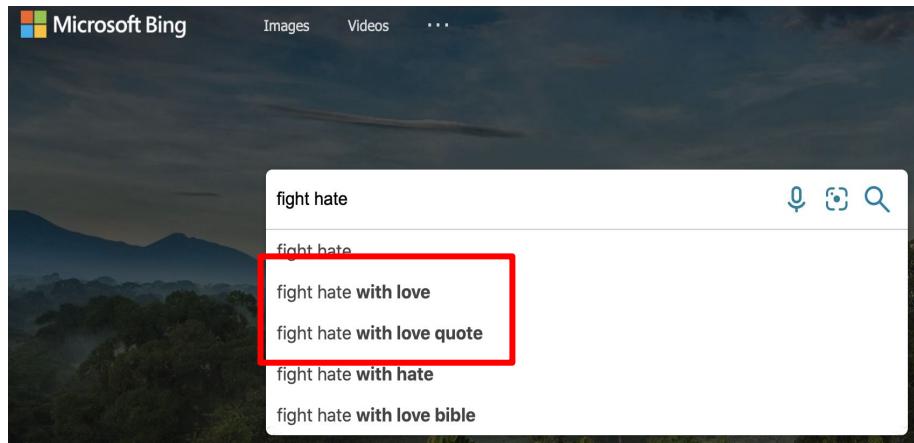
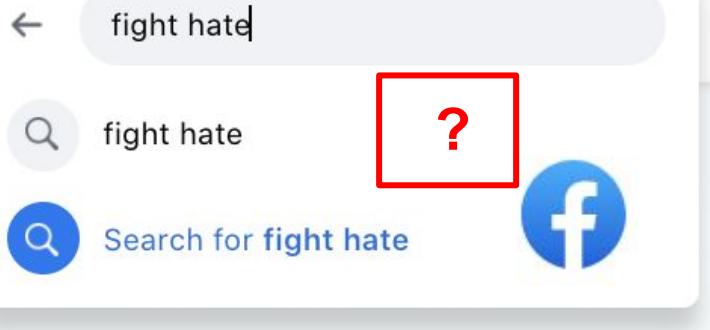
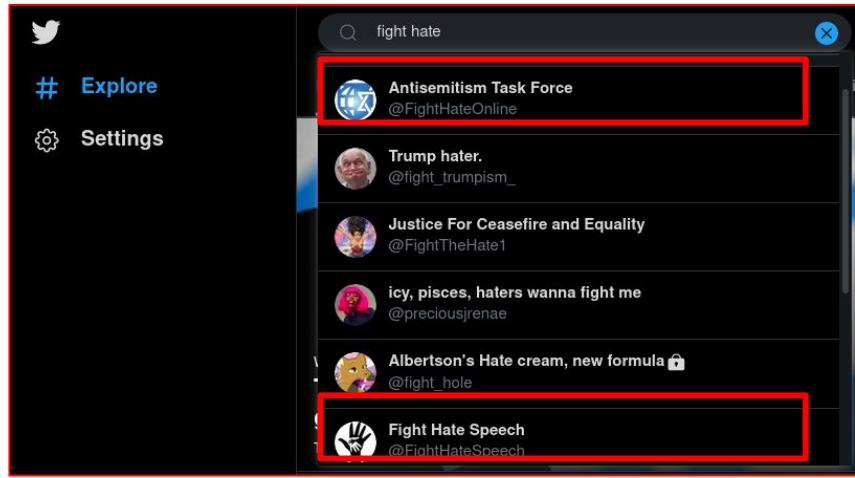


Mithun Das  
Twitter: [@dasmithun92](https://twitter.com/dasmithun92)

## Organisers

Find more about us here!  
<https://hate-alert.github.io/> 3

# Hate speech: A growing concern?



# What to expect from this tutorial?

- What is the problem? Is it really important? How deep are the repercussions?



UNITED NATIONS STRATEGY

## Key commitments

### Foreword

Around the world, we are seeing intolerance – including rising anti-Semitism, anti-Muslimism, anti-black racism, anti-LGBTQ+ discrimination, anti-refugee sentiment, and anti-immigrant attitudes. Social media and other forms of communication have been weaponized for political gain with incendiary language directed at minorities, migrants, refugees, women and other marginalized groups. Tackling hate speech is also crucial to addressing other serious violations of human rights, such as by helping to prevent armed conflict, gender-based violence, and women and other serious violations of human rights.

### Monitoring and analyzing hate speech

### Addressing root causes, drivers and actors of hate speech

### Using technology

### Using education as a tool for addressing and countering hate speech

battling this demon, and so I have done. This Strategy and Plan of Action shows how the United Nations can play its part in upholding freedom of opinion and expression, in partnership with the private sector and other partners.

United Nations Secretary-General  
António Guterres

May 2019

# What to expect from this tutorial?

- Tutorial Part I:
  - UN Key Commitment: Monitoring and analysing hate speech
- How does hate speech spread in the online world?
- Can one comment on the speed and the depth using computational approaches?
- What are the long lasting effects?

# What to expect from this tutorial?

- Tutorial Part II:
  - **UN Key Commitment:** Addressing the root causes/drivers/technology
- What could be the first step to handle this issue? Can we **detect** hate speech using computer algorithms?
- Can the detection results obtained from the model be **explained**?
- Are there **biases** in evaluation? Of what sort?

# What to expect from this tutorial?

- Tutorial Part III:
  - **UN Key Commitment:** Countering hate speech
- How does one contain online hate?
- Conflicts with freedom of speech?
- Can one use more speech to counter hate speech (aka **counterspeech**)?
- Is counterspeech generic or specific to target communities?
- Can one use technology to **automatically generate** counterspeech?

# What to expect from this tutorial?

- Bonus:
  - SWOT analysis
  - Resources: A topically organised notion page consisting of publications, links to codes and dataset.
  - Some hands-on.

# Negative consequences



Bulandshahr Violence



Pittsburg Shooting



Christchurch Shooting



Rohingya Genocide



Sri Lanka Riots



Delhi Riots

# Related tutorials

- The battle against online harmful information: The cases of fake news and hate speech CIKM '20
- Characterization, Detection, and Mitigation of Cyberbullying, ICWSM '18

# Table of contents

- Definitions and related concepts
- Analysis of hate speech
  - Prevalence
  - Effect
- Detection of hate speech
  - Datasets
  - Traditional methods
  - Sequential models
  - Transformer based models
  - Pitfalls of evaluation, explainability, bias
- Mitigation of hate speech
  - Campaigns
  - Counterspeech detection
  - Counterspeech generation
  - Effect of counter speech
- SWOT analysis

# Working definition of hate speech

Direct and serious attacks on any protected category of people based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or disease

**Directed hate:** hate language towards a specific individual or entity.  
Example “@usr4 your a f\*cking queer f\*gg\*t b\*tch”.

**Generalized hate:** hate language towards a general group of individuals who share a common protected characteristic, e.g., ethnicity or sexual orientation.  
Example: “— was born a racist and — will die a racist! — will not rest until every worthless n\*gger is rounded up and hung, n\*ggers are the scum of the earth!! wPww WHITE America”.

# Harmful content online -- a taxonomy

What we will be covering in this tutorial.

Concept	Definition of the concept	Distinction from hate speech
Hate	Expression of hostility without any stated explanation for it [68].	Hate speech is hate focused on stereotypes, and not so general.
Cyberbullying	Aggressive and intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time, against a victim who can not easily defend him or herself [10].	Hate speech is more general and not necessarily focused on a specific person.
Discrimination	Process through which a difference is identified and then used as the basis of unfair treatment [69].	Hate speech is a form of discrimination, through verbal means.
Flaming	Flaming are hostile, profane and intimidating comments that can disrupt participation in a community [35]	Hate speech can occur in any context, whereas flaming is aimed toward a participant in the specific context of a discussion.
Abusive language	The term abusive language was used to refer to hurtful language and includes hate speech, derogatory language and also profanity [58].	Hate speech is a type of abusive language.
Profanity	Offensive or obscene word or phrase [23].	Hate speech can use profanity, but not necessarily.
Toxic language or comment	Toxic comments are rude, disrespectful or unreasonable messages that are likely to make a person to leave a discussion [43].	Not all toxic comments contain hate speech. Also some hate speech can make people discuss more.
Extremism	Ideology associated with extremists or hate groups, promoting violence, often aiming to segment populations and reclaiming status, where outgroups are presented both as perpetrators or inferior populations. [55].	Extremist discourses use frequently hate speech. However, these discourses focus other topics as well [55], such as new members recruitment, government and social media demonization of the in-group and persuasion [62].
Radicalization	Online radicalization is similar to the extremism concept and has been studied on multiple topics and domains, such as terrorism, anti-black communities, or nationalism [2].	Radical discourses, like extremism, can use hate speech. However in radical discourses topics like war, religion and negative emotions [2] are common while hate speech can be more subtle and grounded in stereotypes.

# Hate speech in different contexts

- Targets of hate speech depends on **platform**, **demography** and **language & culture** (Mondal, 2017 and Ousidhoum, 2020)
- Focused research on characterising such diverse types.
  - **Racism** against blacks in Twitter (Kwok, 2013)
  - **Misogyny** across manosphere in Reddit (Farell, 2019)
  - **Sinophobic** behaviour w.r.t COVID-19 (Schild, 2021)
- Often becomes part of different communities
  - **Genetic Testing** Conversations (Mittos, 2020)
  - **QAnon** Conversations (Papasavva, 2021)

# Analysis and Spread

- Definitions and related concepts
- Analysis of hate speech
  - Prevalence
  - Effect
- Detection of hate speech
  - Datasets
  - Traditional methods
  - Sequential models
  - Transformer based models
  - Challenges
- Mitigation of hate speech
  - Campaigns
  - Counterspeech detection
  - Counterspeech generation
- SWOT analysis

# Prevalence of hate speech

- Moderation free platforms like **Gab**, **4chan** and **Bitchute** preferred.



Inside the UK-based site that has become the far right's YouTube

BitChute describes itself as a 'free speech' website but report accuses it of platforming 'hate and terror'.  
Lizzie Dearden reports

Internet Culture  
Gab, the social network that has welcomed Qanon and extremist figures. explained

Gab, a social-networking site popular among the far right, seems to be capitalizing on Twitter bans and Parler being forced offline. It says it's gaining 10,000 new users an hour.

# Prevalence of hate speech

- Gab
- In Gab, early signals show Alt-right, BanIslam as popular hashtags ([Zannettou,2018](#))

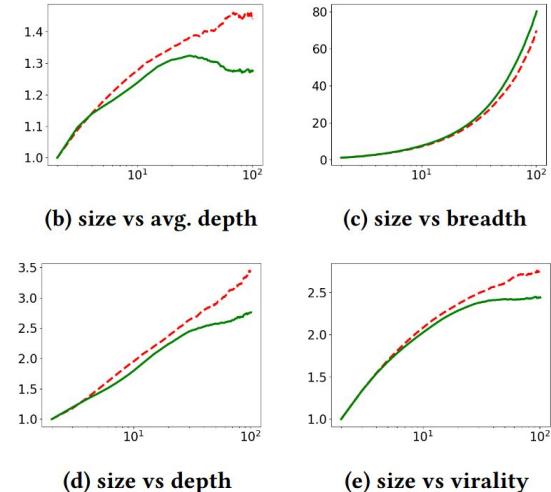
Dataset: collected 22M posts from 336k users, between August 2016 and January 2018

Method: Frequency count

Hashtag	(%)	Mention	(%)
MAGA	6.06%	a	0.69%
GabFam	4.22%	TexasYankee4	0.31%
Trump	3.01%	Stargirlx	0.26%
SpeakFreely	2.28%	YouTube	0.24%
News	2.00%	support	0.23%
Gab	0.88%	Amy	0.22%
DrainTheSwamp	0.71%	RaviCrux	0.20%
AltRight	0.61%	u	0.19%
Pizzagate	0.57%	BlueGood	0.18%
Politics	0.53%	HorrorQueen	0.17%
PresidentTrump	0.47%	Sockalexis	0.17%
FakeNews	0.41%	Don	0.17%
BritFam	0.37%	BrittPettibone	0.16%
2A	0.35%	TukkRivers	0.15%
maga	0.32%	CurryPanda	0.15%
NewGabber	0.28%	Gee	0.15%
CanFam	0.27%	e	0.14%
BanIslam	0.25%	careyetta	0.14%
MSM	0.22%	PrisonPlanet	0.14%
1A	0.21%	JoshC	0.12%

# Prevalence of hate speech

- **Gab**
- In Gab, early signals show Alt-right, BanIslam as popular hashtags. ([Zannettou, 2018](#))
- The posts of hateful users diffuse significantly **farther, wider, deeper** and **faster** than the non hateful users. ([Mathew, 2019](#))



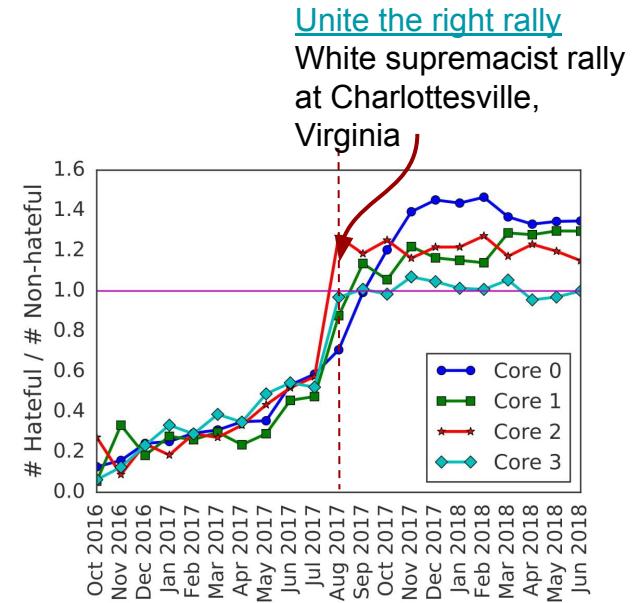
X-axis vs Y-axis

**Dataset:** collect 21M posts from 340k users, between August 2016 and January 2018  
**Method:** Hate user extraction + diffusion method on repost network

# Prevalence of hate speech

- **Gab**
- In Gab, early signals show Alt-right, BanIslam as popular hashtags. ([Zannettou, 2018](#))
- The posts of hateful users diffuse significantly farther, wider, deeper and faster than the non hateful users. ([Mathew, 2019](#))
- Further, **fraction of hateful users** in inner core increased through time in Gab ([Mathew, 2020](#))

**Dataset:** collect 21M posts from 340k users, between August 2016 and January 2018  
**Method:** Hate user extraction + Temporal k-core analysis



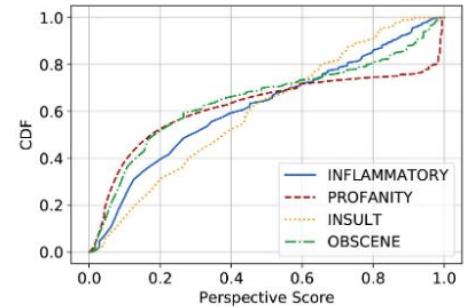
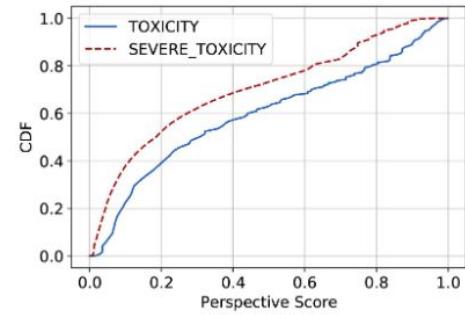
# Prevalence of hate speech

- Gab
- In Gab, early signals show Alt-right, BanIslam as popular hashtags. ([Zannettou, 2018](#))
- The posts of hateful users diffuse significantly farther, wider, deeper and faster than the non hateful users. ([Mathew, 2019](#))
- Further, fraction of hateful users in inner core increased through time in Gab ([Mathew, 2020](#))

# Prevalence of hate speech

- **4chan**
- In 4chan's /pol/ thread ([Papasavva,2020](#))
  - 37% → TOXICITY
  - **27% → SEVERE\_TOXIC**
  - 36% → INFLAMMATORY
  - 33% → PROFANITY
  - 35% → INSULT
  - 30% → OBSCENE

**Dataset:** Crawling from 4chan's /pol/ thread, June 29, 2016 to November 1, 2019.  
**Method:** Perspective api then CDF

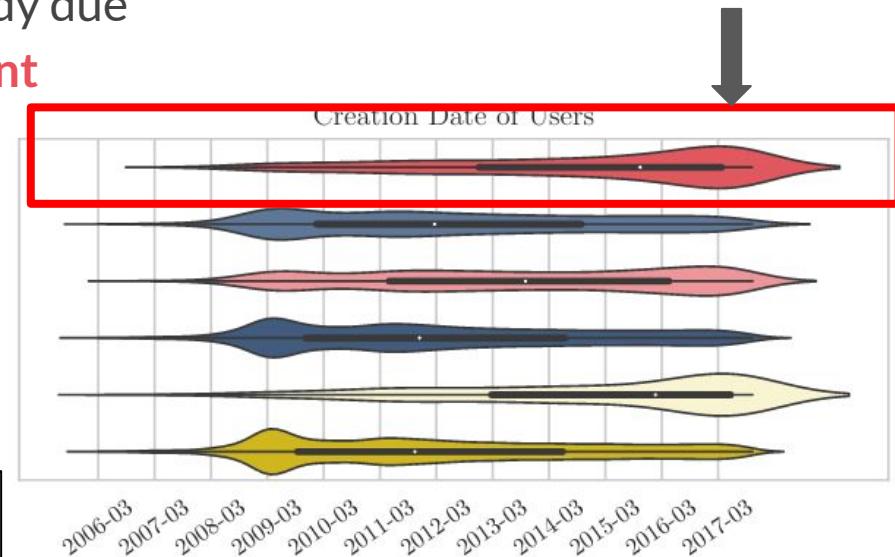


# Prevalence of hate speech (Platforms with moderation)

Study on characterising hateful users in Twitter

([Riberio,2018](#))

- Spread of hatespeech difficult to study due to moderation of **hateful user/content**



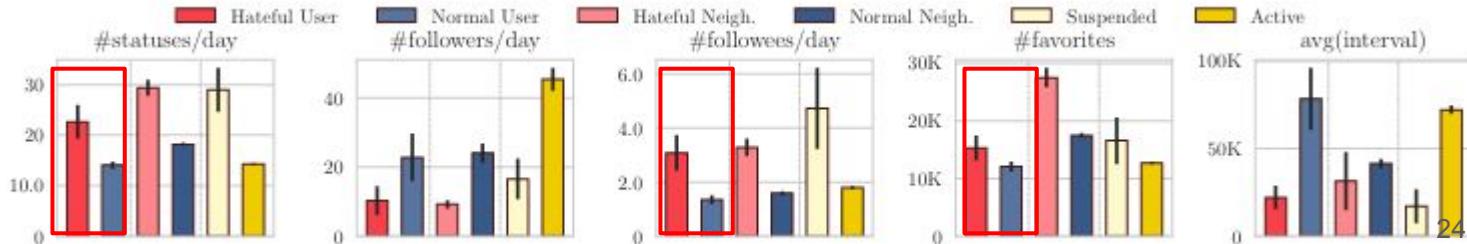
**Dataset:** Data collected from Twitter, keyword based extraction

**Method:** Degroot method. Frequency based analysis

# Prevalence of hate speech (Platforms with moderation)

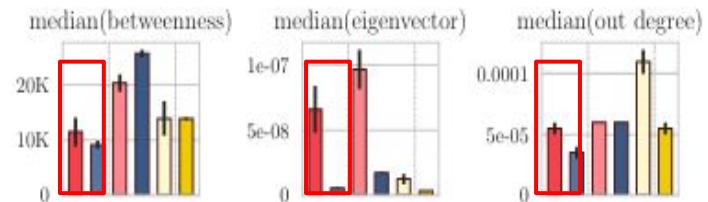
Study on characterising hateful users in Twitter  
([Riberio,2018](#))

- Spread of hatespeech difficult to study due to moderation of hateful user/content
- Hateful users are **power users** (post more, favourite more).



# Prevalence of hate speech (Platforms with moderation)

- Study on characterising hateful users in Twitter ([Riberio,2018](#))
- Spread of hatespeech difficult to study due to moderation of hateful user/content
- Hateful users are power users (post more, favourite more).
- Median hate user is **more central** to the network



# Prevalence of hate speech (Platforms with moderation)

- Study on misogyny in reddit  
[\(Farrell,2019\)](#)
- *r/Braincel*s was the main subreddit after *r/ince*l was banned in 2015

**Dataset:** Pushshift reddit, lexicons, incel subreddits  
**Method:** Degroot method. Frequency based analysis

# Detecting Hate Speech

- Definitions and related concepts
- Analysis of hate speech
  - Prevalence
  - Effect
- Detection of hate speech
  - Datasets
  - Traditional methods
  - Sequential models
  - Transformer based models
  - Challenges
- Mitigation of hate speech
  - Campaigns
  - Counterspeech detection
  - Counterspeech generation
  - Effect of counter speech
- SWOT Analysis

# Datasets

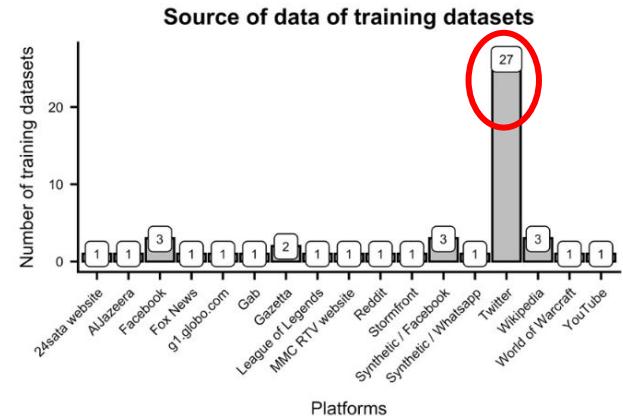
- Different datasets have different **taxonomies**.
  - Binary classification (hate/not, targeting group or not) ([Zampieri,2019](#))
  - Specific binary (Misogyny/not, Racism/not) ([Pamungkas, 2020](#))
  - Multiclass/labels datasets. ([Davidson,2017](#) , [Basile,2019](#))

# Datasets

- Different datasets have different taxonomies.
- Different datasets have different **sources**.

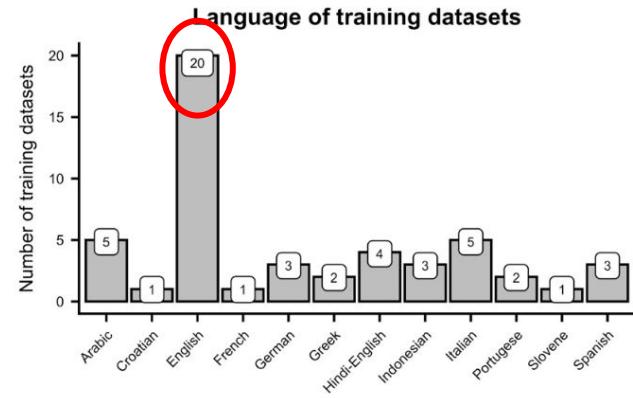
Twitter is one of the major sources.

- The works by Davidson ([Davidson, 2017](#)) and Founta ([Founta, 2018](#)) are two highly used dataset from Twitter
- Twitter is easily accessible.
- Alt-right platforms are often taken down, hence studies are limited ([Voat](#), [Parler](#))



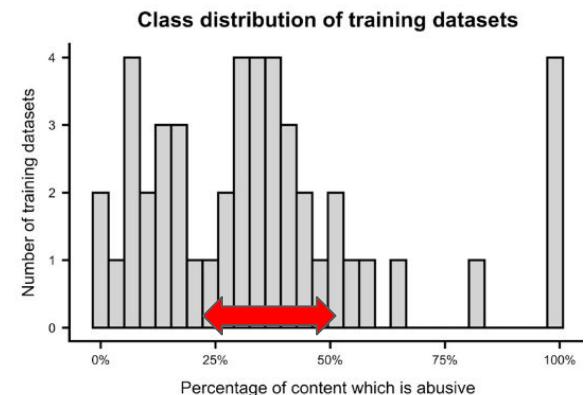
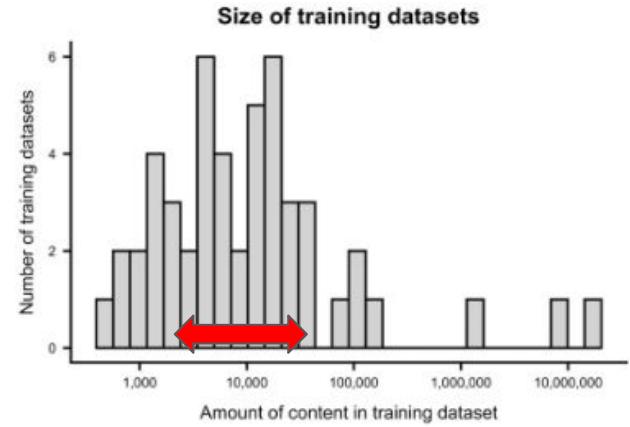
# Datasets

- Different datasets have different taxonomies.
- Different datasets have different sources.  
Twitter is one of the major sources.
- Different datasets have different **languages**,  
English being the prominent one.
  - Arabic ([Mulki,2019](#)), Italian ([Sanguinetti,2018](#)), Spanish ([Basile,2019](#)) and Indonesian ([Ibrohim,2019](#)) has more than 3 datasets
  - Quality is often questionable for these datasets.
  - Can we benefit from english language datasets ?



# Datasets

- Different datasets have different taxonomies.
- Different datasets have different sources.  
Twitter is one of the major sources.
- Different datasets have different languages,  
English being the prominent one.
- **Training size and amount of hate/abuse** also  
varies across datasets



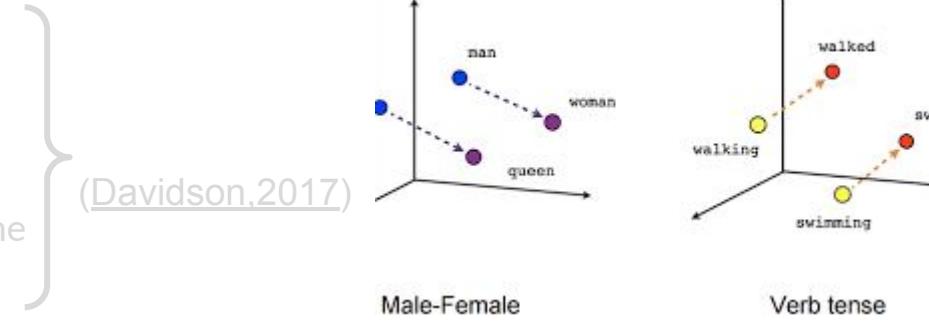
# Earlier Detection Methods

- Features used :-
    - TF-IDF vectors
    - Parts-of-speech tags
    - Linguistic features
      - Sentiment lexicons
      - Frequency counts of URL, username
      - Readability scores
- 
- ([Davidson,2017](#))

# Earlier Detection Methods

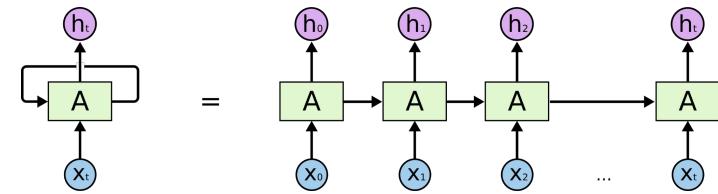
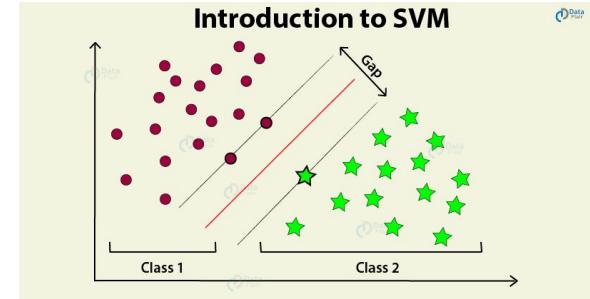
- Features used :-
  - TF-IDF vectors
  - Parts-of-speech tags
  - Linguistic features
    - Sentiment lexicons
    - Frequency counts of URL, username
    - Readability scores
  - **Word embeddings**
    - Twitter word embeddings ([Zimmerman, 2018](#)). [Click here](#)
  - **Sentence embeddings**
    - Google's universal embeddings ([Saha, 2018](#)). [Click here](#)

(Davidson, 2017)



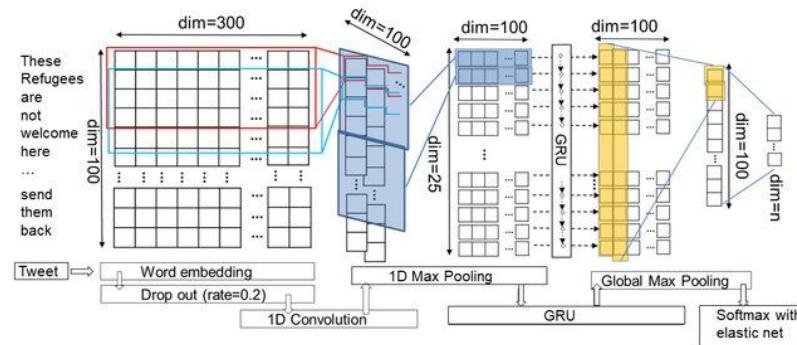
# Earlier Detection Methods

- Features used
- Detection method
  - Logistic regression
  - **SVM** ([Canós, 2018](#))
  - XGboost ([Saha, 2018](#))
  - **LSTM/GRU** ([Gao, 2017](#))
  - CNN-GRU ([Zhang, 2018](#))



# Earlier Detection Methods

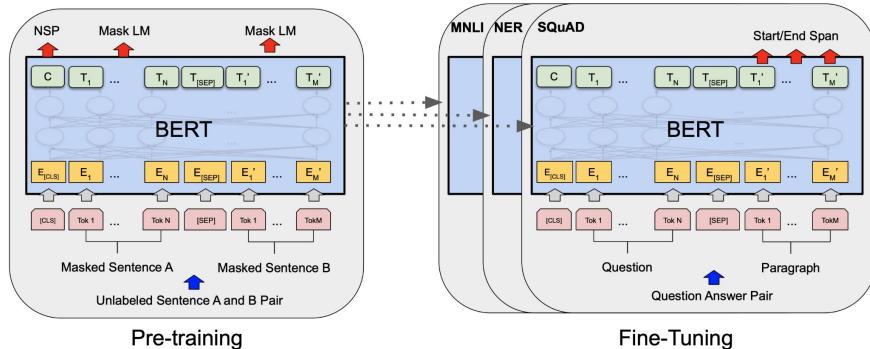
- Features used
- Detection method
  - Logistic regression
  - SVM ([Canós, 2018](#))
  - XGboost ([Saha, 2018](#))
  - LSTM/GRU ([Gao, 2017](#))
  - **CNN-GRU ([Zhang, 2018](#))**



Dataset	SVM	SVM+ CNN	CNN+ GRU <sub>E</sub>	CNN+ GRU <sub>I</sub>	State of the art
WZ-L	0.74	0.74	0.80	0.81	<b>0.82</b> Waseem <a href="#">[26]</a> , best F1
WZ-S.amt	0.86	0.87	0.91	<b>0.92</b>	0.84 Waseem <a href="#">[25]</a> , Best features
WZ-S.exp	0.89	0.90	0.90	0.91	<b>0.92</b> Waseem <a href="#">[25]</a> , Best features
WZ-S.gb	0.86	0.87	0.91	0.92	<b>0.93</b> Gamback <a href="#">[10]</a> , best F1
WZ-LS	0.72	0.73	0.81	0.81	0.82 Park <a href="#">[20]</a> , WordCNN 0.81 Park <a href="#">[20]</a> , CharacterCNN <b>0.83</b> Park <a href="#">[20]</a> , HybridCNN
DT	0.87	0.89	<b>0.94</b>	<b>0.94</b>	0.87 SVM, Davidson <a href="#">[7]</a>
RM	0.86	0.89	0.90	0.91	<b>0.92</b> 0.86 SVM, Davidson <a href="#">[7]</a>

# Current Models

- Earlier models cannot completely capture context
- **BERT** and other transformers model helped in getting improved performance across different datasets ([Mozafari,2019](#))



Method	Datasets	Precision(%)	Recall(%)	F1-score(%)
Waseem and Hovy [22]	Waseem	72.87	77.75	73.89
Davidson et al. [3]	Davidson	91	90	90
Waseem et al. [23]	Waseem	-	-	80
	Davidson	-	-	89
BERT <sub>base</sub>	Waseem	81	81	81
	Davidson	91	91	91
BERT <sub>base</sub> + Nonlinear Layers	Waseem	73	85	76
	Davidson	76	78	77
BERT <sub>base</sub> + LSTM	Waseem	87	86	86
	Davidson	91	92	92
BERT <sub>base</sub> + CNN	Waseem	<b>89</b>	<b>87</b>	<b>88</b>
	Davidson	<b>92</b>	<b>92</b>	<b>92</b>

# Current Models

- Earlier models cannot completely capture context
- BERT and other transformers model helped in getting improved performance across different datasets ([Mozafari,2019](#))
- Incorporating lexicon into the BERT architecture → HurtBERT ([Koufakou,2020](#)).

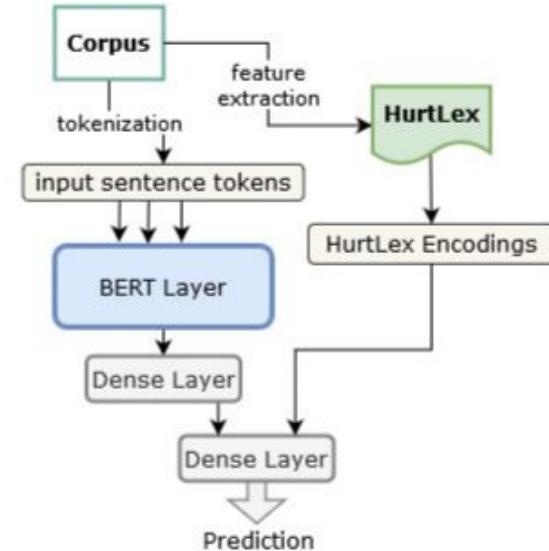
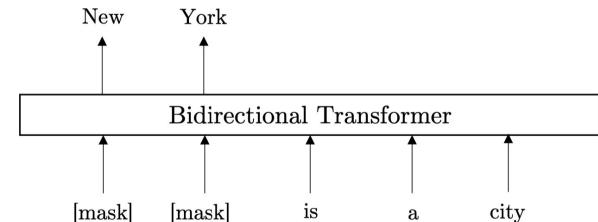


Figure 1: HurtBERT-Enc, our model using HurtLex Encodings

# Current Models

- Earlier models cannot completely capture context
- BERT and other transformers model helped in getting improved performance across different datasets ([Mozafari,2019](#))
- Incorporating lexicon into the BERT architecture → HurtBERT ([Koufakou,2020](#)).
- Re-training BERT with banned subreddit data → HateBERT ([Caselli,2021](#)).

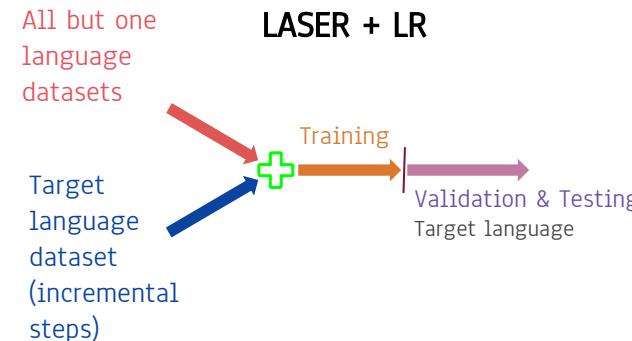
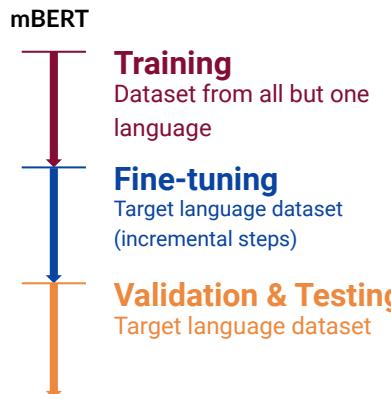


Dataset	Model	Macro F1 Pos. class - F1	
OffensEval 2019	BERT	.803±.006	.715±.009
	HateBERT	<b>.809±.008</b>	<b>.723±.012</b>
	Best	.829	.599
AbusEval	BERT	.727±.008	.552±.012
	HateBERT	<b>.765±.006</b>	<b>.623±.010</b>
	Caselli et al. (2020)	.716±.034	.531
HatEval	BERT	.480±.008	.633±.002
	HateBERT	<b>.516±.007</b>	<b>.645±.001</b>
	Best	.651	-

# Multilingual Hate speech

- Analysis of multilingual models across 9 different languages and 16 datasets ([Aluru,2020](#)).

Language	Low resource	High resource
Arabic	Monolingual, LASER + LR	Multilingual, mBERT
English	Multilingual, LASER + LR	Multilingual, mBERT
German	Monolingual, LASER + LR	Translation + BERT
Indonesian	Multilingual, LASER + LR	Monolingual, mBERT
Italian	Multilingual, LASER + LR	Monolingual, mBERT
Polish	Multilingual, LASER + LR	Translation + BERT
Portuguese	Multilingual, LASER + LR	Monolingual, LASER+LR
Spanish	Monolingual, LASER + LR	Multilingual, mBERT
French	Monolingual, LASER + LR	Translation + BERT

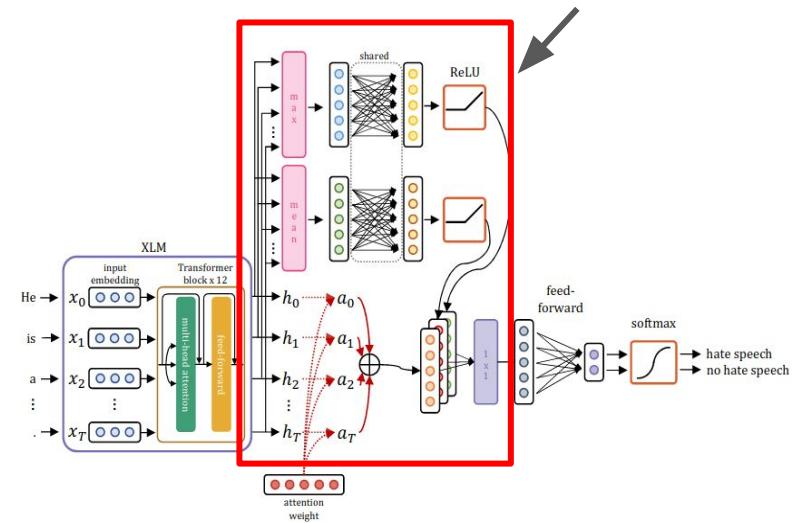


Click logo for demo

# Multilingual Hate speech

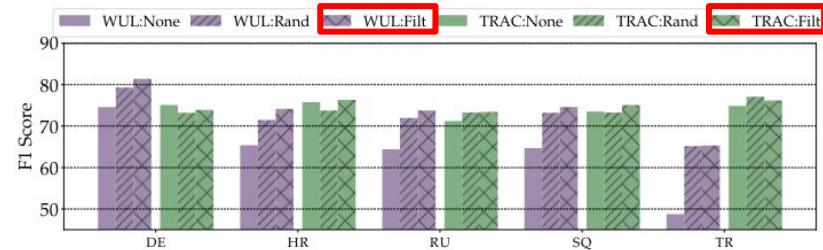
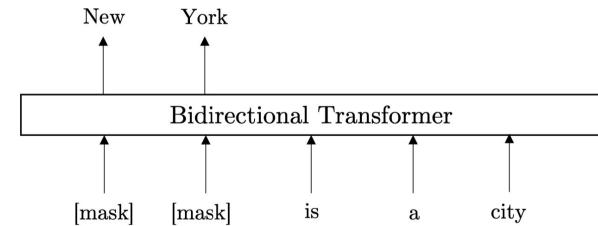
- Benchmarking multilingual models across 9 different languages and 16 datasets ([Aluru,2020](#)).
- A novel classification block -AXEL to improve cross lingual transfer ([Stappen,2020](#)) on Hateval data.

	Dense	Att	AXEL
EN⇒ES	41.31	34.37	<b>53.42</b>
ES⇒EN	<b>60.83</b>	48.47	52.48
ES⇒EN-S	49.38	39.10	<b>53.24</b>
EN⇒(ES→EN)	60.59	62.40	<b>64.39</b>
ES⇒(EN→ES)	56.89	49.17	<b>58.31</b>
ES⇒(EN-S→ES)	56.57	49.17	<b>65.04</b>



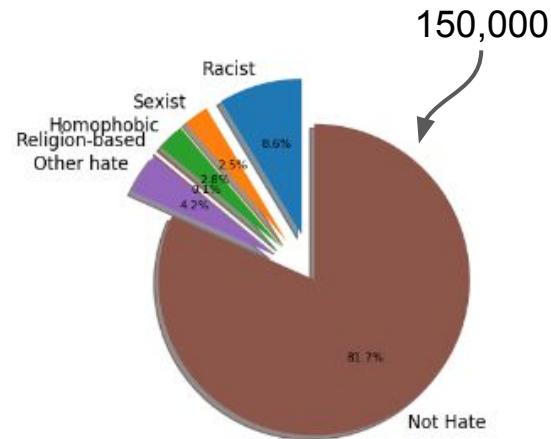
# Multilingual Hate speech

- Benchmarking multilingual models across 9 different languages and 16 datasets ([Aluru,2020](#)).
- A novel classification block -AXEL to improve cross lingual transfer ([Stappen,2020](#)) on Hateval data.
- **Pre-training** on keyword based filtered data also can help in cross lingual transfer ([Glavaš,2020](#))



# More Modalities

- **MMHS150K** is one of the largest dataset. image-text pair in hate speech research ([Gomez,2019](#)).
- Text based models are at par with multimodal models.



# Shared tasks timeline

AMI'18   SemEval'19   HASOC'19   VLSP'19



## EVALITA AMI 2018

Task- Misogyny  
Best- Feature  
based XGBoost

## SemEval-2019

Task-Multilingual  
Best- SVM with  
RBF

## HASOC 2019

Task-  
Hate/Offensive  
Best- Ensemble

## VLSP HSD 2019

Task- Hate  
Speech  
Best- LR + ngram

# Shared tasks timeline

AMI'18 SemEval'19 HASOC'19 VLSP'19 EVALITA'20 SemEval'20 HASOC'20



## EVALITA HSD 2020

**Task-**  
HateSpeech  
**Best-** BERT

## SemEval-2020

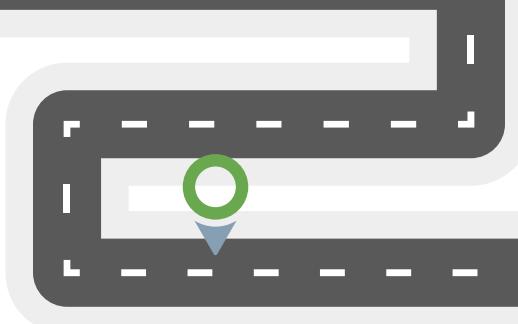
**Task-**Multilingual  
**Best-** BERT,  
m-BERT

## HASOC 2020

**Task-**  
Multilingual  
**Best-** CNN, BERT

## OSACT4 HSD 2020

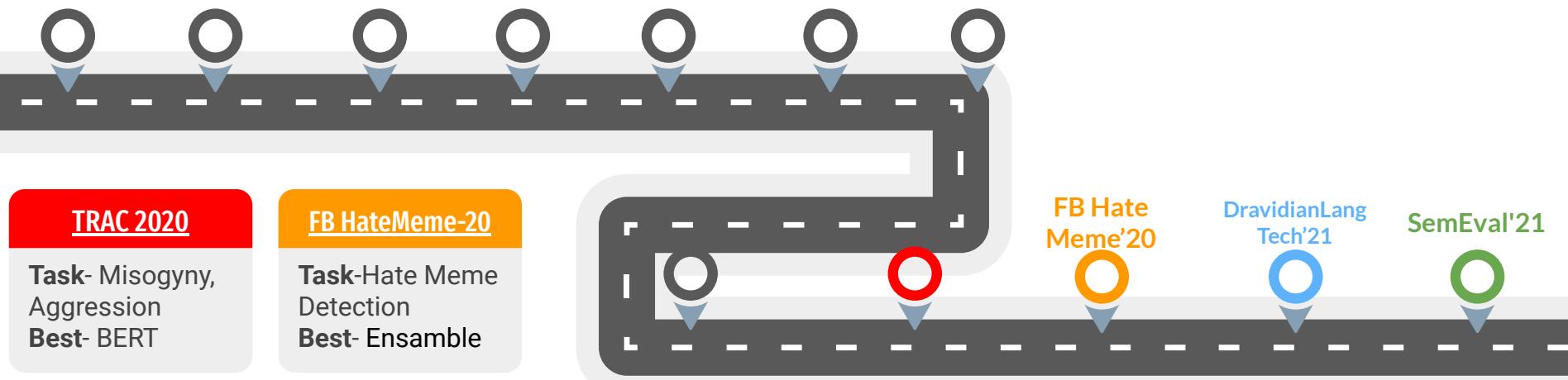
**Task-** Arabic  
Hate Speech  
**Best-** CNN



OSACT4'20

# Shared tasks timeline

AMI'18 SemEval'19 HASOC'19 VLSP'19 EVALITA'20 SemEval'20 HASOC'20



# Pitfalls of Model Evaluation

- Two of the previous studies had spurious evaluations ([Badjatiya,2017](#) and [Agrawal,2018](#))
- Types of **wrong evaluations**
  - Oversampling before train-test split ([Agrawal,2018](#))
  - Feature extraction using the whole train and test split ([Badjatiya,2017](#))

**Dataset:** Waseem and Hovy dataset  
**Method:** LSTM+GBDT , BiLSTM with attention

Method	Class	Prec.	Rec.	F1
Badjatiya et al. [2] Emb. over all dataset	Neither	95.5	96.8	96.1
	Racist	94.5	93.5	94.0
	Sexist	91.2	87.5	89.3
	Micro avg.	94.6	94.6	94.6
	Macro avg.	93.7	92.6	93.1
Agrawal and Awekar [1] Oversamp. all dataset	Neither	95.1	91.7	93.4
	Racist	94.9	96.0	95.4
	Sexist	92.5	97.0	94.6
	Micro avg.	94.4	94.4	94.4
	Macro avg.	94.2	94.9	94.5

After correcting  
the errors

Drop of 20% in Macro F1!

Method	Class	Prec.	Rec.	F1
Badjatiya et al. [2] Emb. over train set	Neither	82.3	94.7	88.1
	Racist	78.0	64.0	70.2
	Sexist	84.5	47.8	60.9
	Micro avg.	82.3	82.1	80.7
	Macro avg.	81.6	68.9	73.1
Agrawal and Awekar [1] Oversamp. train set	Neither	90.3	86.5	88.3
	Racist	69.6	81.3	75.0
	Sexist	74.0	77.4	75.5
	Micro avg.	84.7	84.1	84.3
	Macro avg.	78.0	81.7	79.6

# Pitfalls of Model Evaluation

- Two of the previous studies had spurious evaluations (Badjatiya,2017 and Agrawal,2018)
- Wrong evaluations
  - Oversampling before train-test split (Agrawal,2018)
  - Feature extraction using the whole train and test split (Badjatiya,2017)
- **Removing user overlap** between train and test set.

**Dataset:** Waseem and Hovy dataset  
**Method:** LSTM+GBDT , BiLSTM with attention

Method	Class	Prec.	Rec.	F1
Bajjatiya et al. [2]	None	49.6	93.4	64.3
	Hateful	68.8	15.4	23.5
	Micro avg.	63.8	54.1	46.1
	Macro avg.	59.2	54.4	43.9
Agrawal and Awekar [1]	None	47.5	98.0	63.0
	Hateful	75.3	03.5	06.7
	Micro avg.	62.3	48.4	35.1
	Macro avg.	61.4	50.8	34.9

# Pitfalls of Model Evaluation

- Datasets lack testing in the **wild**, train-test comes from the same distribution.
- Different test suites generated to test the classifiers. [\(Röttger,2020\)](#)
- **Error in neutral and positive statement about group**

## Models

DistilBERT-Davidson - **DB-D**

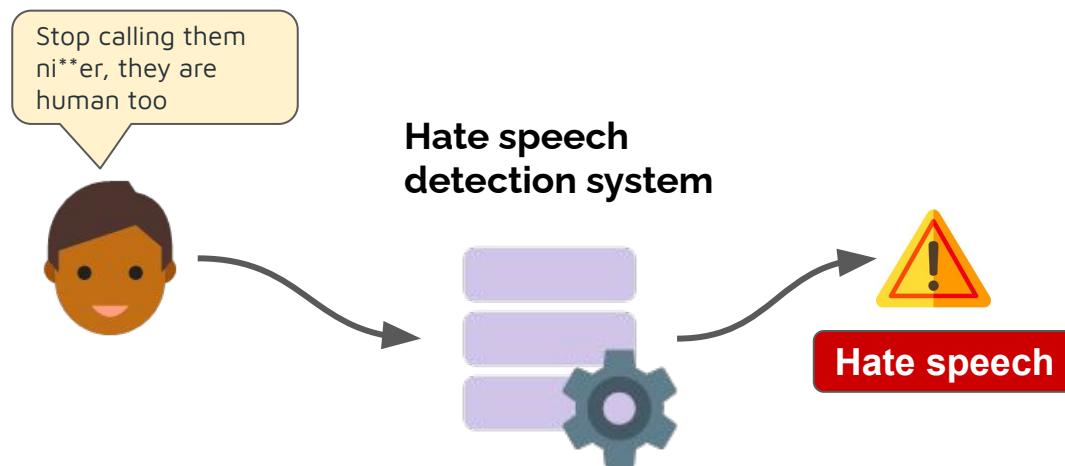
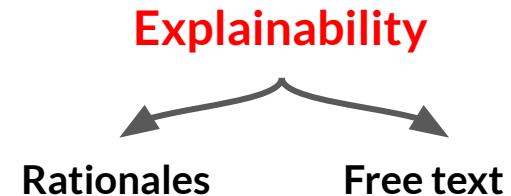
DistilBERT-Founta - **DB-F**

Perspective api - **PERS**

Functionality	Example	Gold Label	n	Accuracy (%)		
				DB-D	DB-F	PERS
Group identifier	<b>F18:</b> Neutral statement using protected group identifier	"We are a group of [IDENTITY]." "I live with two [IDENTITY]."	non-hateful	126	61.1	76.2
	<b>F19:</b> Positive statement using protected group identifier	"I love [IDENTITY]." "[IDENTITY] are great."	non-hateful	189	86.2	79.9

# Explainability of Models

- **Explainability** is a crucial aspect more so in these social dimensions

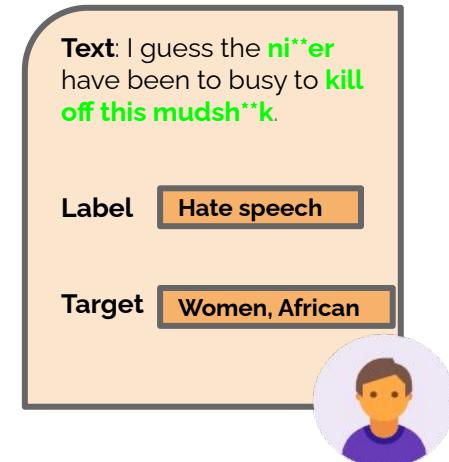


# Explainability of Models

- **Explainability** is a crucial aspect more so in these social dimensions
- **HateXplain** - first dataset to include rationales along with labels. ([Mathew,2020](#))

Models	Accuracy	F1 Score	AUROC
CNN-GRU	0.627	0.606	0.793
BERT	0.690	0.674	0.843
<b>BERT-HateXplain</b>	<b>0.698</b>	<b>0.687</b>	<b>0.851</b>

Models performance is better !

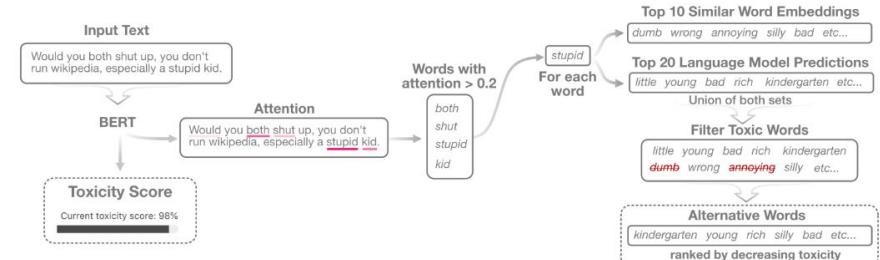


Click logo for demo

colab

# Explainability of Models

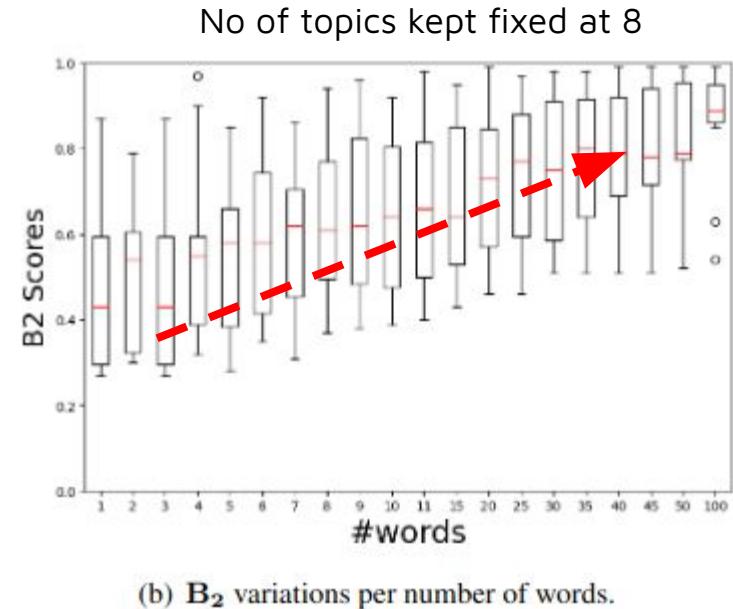
- **Explainability** is a crucial aspect more so in these social dimensions
- **Hatexplain** - first dataset to include rationales as well as target along with labels. (Mathew,2020)
- **RECAST** - tool to suggest alt wordings based on attention scores. (Wright,2021)



**Advantage** - reduce toxicity, way of debugging model  
**Disadvantage** - malicious users might game the system.

# Bias in Data/Models

- Bias from different directions
  - How is **data selected** ?
  - Who is the annotator?
  - Who is the speaker/target ?
- Often hate speech dataset can carry bias related to some identity words  
[\(Ousidhoum,2020\)](#)
- Increase in semantic relatedness between corpus and keywords as number of keywords are increased



**B2 measures how frequently keyword appear in topics**

# Bias in Data/Models

- Bias from different directions
  - How is data selected ?
  - Who is the **annotator**?
  - Who is the speaker/target ?
- Data using expert annotators (activists) performs better than amateurs (crowdsouce)  
[\(Waseem,2016\)](#)

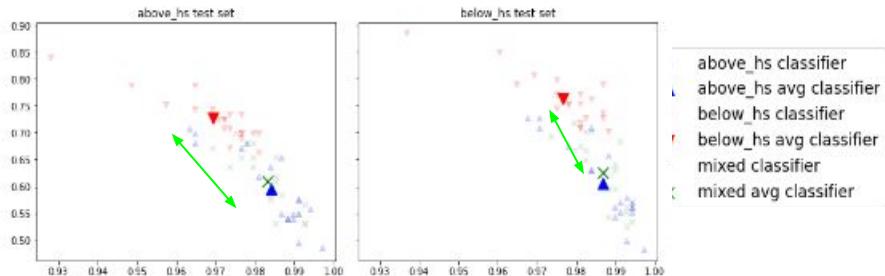
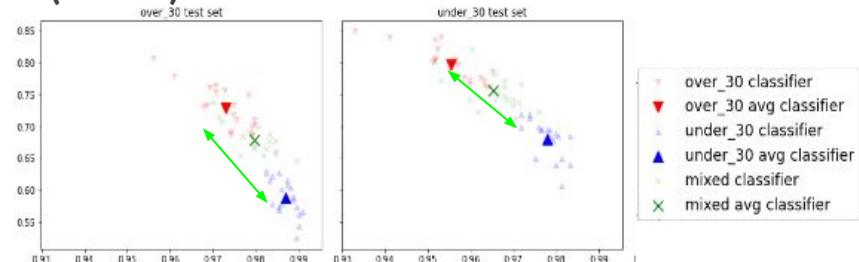
Feature Set	F1	Amateur			Expert		
		Recall	Precision	F1	Recall	Precision	
Close	86.39	88.60%	87.59%	91.24	92.49%	92.67%	
Middling	84.07	86.76%	85.43%	87.81	90.10%	88.53%	
Distant	71.71	80.17%	82.05%	77.77	84.76%	71.85%	
All	86.39	88.60%	87.59%	90.77	92.20%	92.23%	
Best	83.88	86.68%	85.54%	91.19	92.49%	92.50%	
Baseline	70.84	79.80%	63.69%	77.77	84.76%	71.85%	

**Table 5:** Scores obtained for each of the feature sets.

# Bias in Data/Models

- Bias from different directions
  - How is data selected ?
  - Who is the **annotator**?
  - Who is the speaker/target ?
- Data using expert annotators (activists) performs better than amateurs (crowdsource)  
(Waseem,2016)
- A study found significant bias for age and education of the annotators. (Kuwatly,2020)

Specificity (X-axis) vs sensitivity (Y-axis)



**Method** - Trained different classifiers on data annotated by different group and evaluated them

# Bias in Data/Models

- Bias from different directions
  - How is data selected ?
  - Who is the annotator?
  - Who is the **speaker/target** ?
- Often hate speech model can detect false positives for tweets written by different community ([Davidson,2019](#))

Dataset	Class	$\widehat{p_{i black}}$	$\widehat{p_{i white}}$	t	p	$\frac{\widehat{p_{i black}}}{\widehat{p_{i white}}}$
<i>Waseem and Hovy</i>	Racism	0.001	0.003	-20.818	***	0.505
	Sexism	0.083	0.048	101.636	***	1.724
	Racism	0.001	0.001	0.035		1.001
	Sexism	0.023	0.012	64.418	***	1.993
<i>Waseem</i>	Racism and sexism	0.002	0.001	4.047	***	1.120
	Hate	0.049	0.019	120.986	***	2.573
	Offensive	0.173	0.065	243.285	***	2.653
<i>Davidson et al.</i>	Harassment	0.032	0.023	39.483	***	1.396
	Hate	0.111	0.061	122.707	***	1.812
<i>Golbeck et al.</i>	Abusive	0.178	0.080	211.319	***	2.239
	Spam	0.028	0.015	63.131	***	1.854

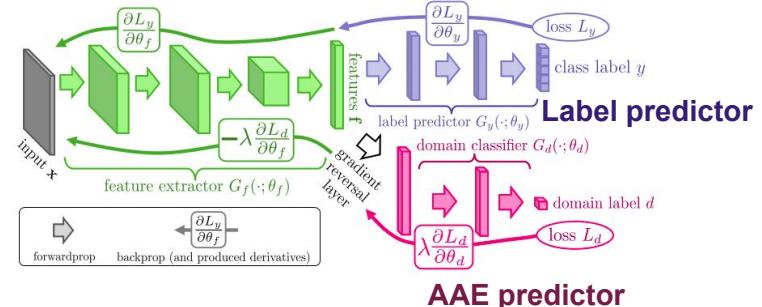
Table 2: **Experiment 1**

Values greater than 1 indicate that black-aligned tweets are classified as belonging to class at a higher rate than white

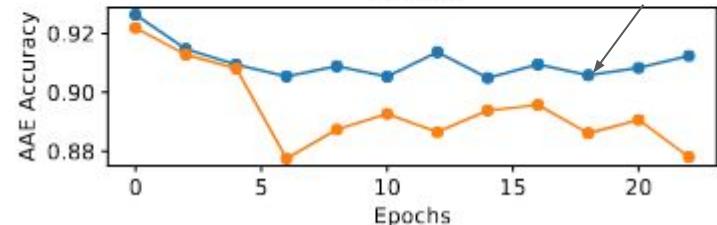
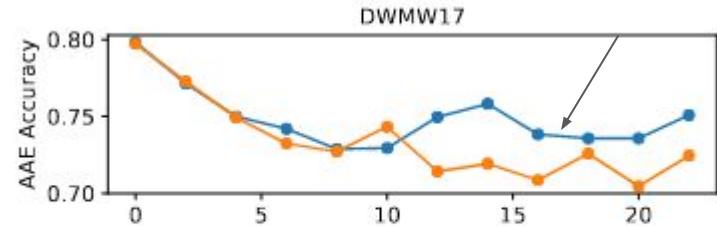
Dataset and model used for dialect identification ([Blodgett,2016](#))

# Bias in Data/Models

- Bias from different directions
  - How is data selected ?
  - Who is the annotator?
  - Who is the **speaker/target** ?
- Often hate speech model can detect false positives for tweets written by different community ([Davidson,2019](#)).
- Training with adversarial loss can help reduce the bias ([Xia,2020](#)).



● single adversary      ● multiple adversaries



Dataset and model used for dialect identification ([Blodgett,2016](#))

# Bias in Data/Models

- Bias from different directions
  - How is data selected ?
  - Who is the annotator?
  - Who is the **speaker/target** ?
- Often hate speech model can detect false positives for tweets written by different community ([Davidson,2019](#)).
- Training with adversarial loss can help reduce the bias ([Xia,2020](#)).
- Using rationales can make the models less biased towards different targets ([Mathew,2020](#))

Models	GMB-Sub	GMB-BPSN	GMB-BNSP
CNN-GRU	0.654	0.623	0.659
BERT	0.762	0.709	0.757
<b>BERT-HateXplain</b>	<b>0.807</b>	<b>0.745</b>	<b>0.763</b>

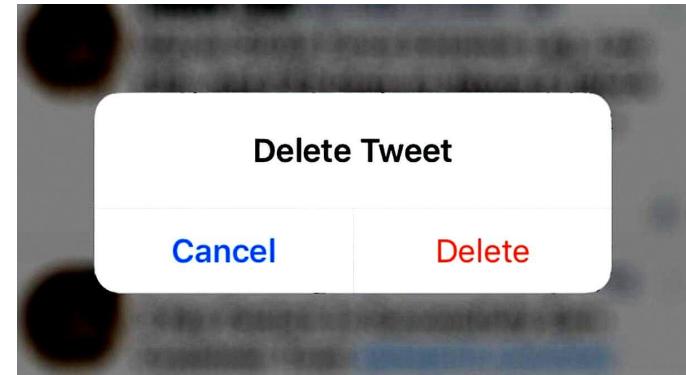
Models less biased !

# Mitigating Hate Speech

- Definitions and related concepts
- Analysis of hate speech
  - Prevalence
  - Effect
- Detection of hate speech
  - Datasets
  - Traditional methods
  - Sequential models
  - Transformer based models
  - Challenges
- Mitigation of hate speech
  - Campaigns
  - Counterspeech detection
  - Counterspeech generation
  - Effect of counter speech
- SWOT analysis

# What is done after detecting hate speech?

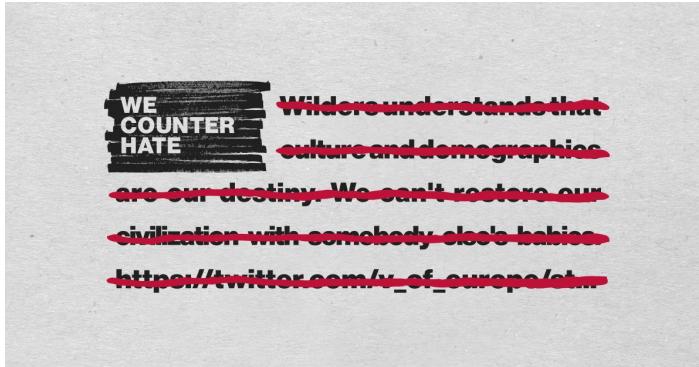
- **Deletion** of posts
- **Suspension** of user accounts
- **Shadow banning**



# Campaign to deter hate

FACEBOOK

Counterspeech.fb



WeCounterHate



ADL



NO HATE  
SPEECH  
MOVEMENT

NoHateSpeechMovement

# Hate speech laws

- Several countries have **laws** that prohibit hate speech
- The **definition** of hate speech varies according to the country
- Models which detect hate speech will need to take these **nuances** into account



# Reddit Ban [2015]

- In 2015, Reddit closed several subreddits due to **violations** of Reddit's anti-harassment policy.
- Foremost among them were **r/fatpeoplehate** and **r/CoonTown**
- How **effective** was the ban?



**This community has been banned**

This subreddit was banned due to a violation of our [content policy](#), specifically, our sitewide rules regarding violent content.

Banned 1 day ago.

[BACK TO REDDIT](#)

# Reddit Ban [2015]

- In 2015, Reddit closed several subreddits due to **violations** of Reddit's anti-harassment policy.
- Foremost among them were r/**fatpeoplehate** and r/**CoonTown**
- How **effective** was the ban?

**You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech [Chandrasekharan 2017]**



**This community has been banned**

This subreddit was banned due to a violation of our [content policy](#), specifically, our sitewide rules regarding violent content.

Banned 1 day ago.

[BACK TO REDDIT](#)

# The Efficacy of Reddit's 2015 Ban

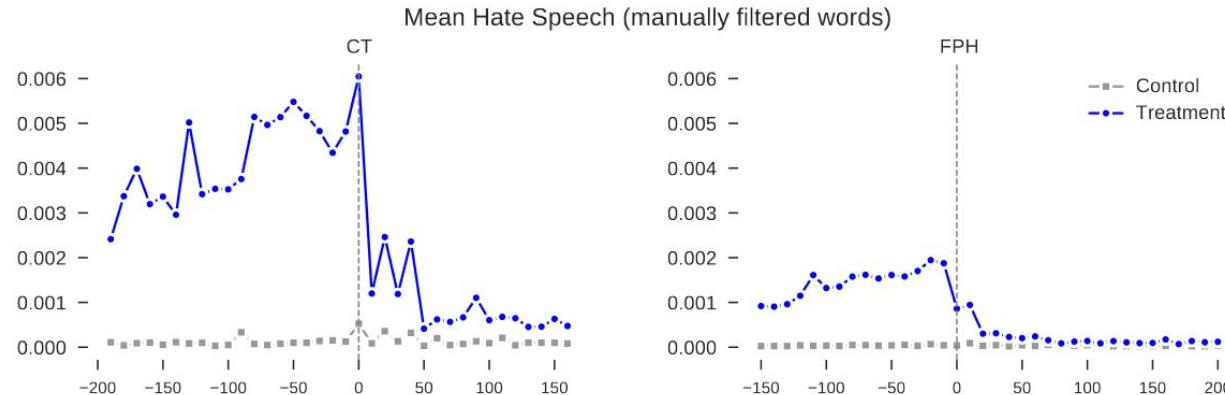
## User-level Effects of the Ban

- Following Reddit's 2015 ban, a large, significant percentage of users from banned communities left Reddit
- Following the ban, Reddit saw a decrease of over 80% in the usage of hate words by r/fatpeoplehate and r/CoonTown users

# The Efficacy of Reddit's 2015 Ban

## User-level Effects of the Ban

- For the banned community users that remained active, the ban drastically reduced the amount of hate speech they used across Reddit by a large and significant amount.



# The Efficacy of Reddit's 2015 Ban

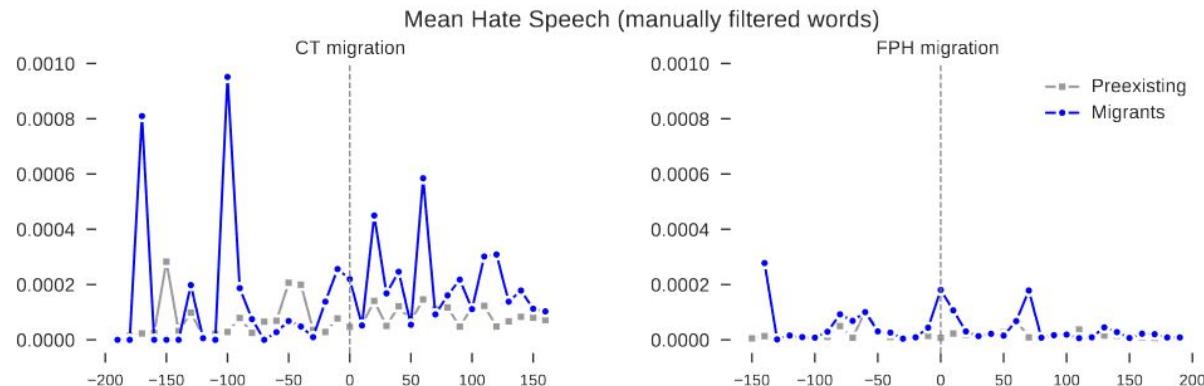
## Community-level Effects of the Ban

- Following the banning of r/fatpeoplehate and r/CoonTown, the affected users migrated to other parts of Reddit.
- The majority of r/CoonTown users migrated to other subreddits (like r/The\_Donald, r/homeland, r/BlackCrimeMatters) where racist behavior has either been noted or is prevalent.

# The Efficacy of Reddit's 2015 Ban

## Community-level Effects of the Ban

- The migrant users did not bring hate speech with them to their new communities, nor did the longtime residents pick it up from them. **Reddit did not “spread the infection”.**



# **Doctrine of Counterspeech/Counter-Narrative**

- The counterspeech doctrine posits that the proper response to negative speech is to **counter it with positive expression**.
- Combating hate speech in this way has some advantages: it is **faster**, more **flexible** and **responsive**, capable of dealing with extremism from anywhere and in any language and it does not form a barrier against the principle of free and open public space for debate.

# Counterspeech Examples

Hate Speech



[REDACTED] Cowardly attack on innocent people as it has happened in Gujurat carnage and various lynching in different regions of India. Cowards everywhere attack on unarmed civilians. Violence must be condemned at every level.

Like · Reply · 12 July at 00:53

[REDACTED] Muslims are not terrorists brother it's just because of few Muslims the name of the entire community is getting spoilt please learn to respect the religion.

Like · Reply · 8 · July 28, 2016 at 12:40am

Counterspeech



patriargate ✅  
@patriargate

Follow

So #Muslims do not seem to care so much about having a nice place to live. Or maybe they just believe that white (christian) slave should do the job.



MorgothLives @LivesMorgoth

Hackney in London is just 30% white yet a photo of volunteer litter pickers looks like this?  
But if they ask Diane Abbott to represent them as much as the black community she'll block them

1:41 PM - 16 Nov 2018

1 1 1 1 1 1

Tweet your reply

More replies



We Counter Hate @we\_counter\_hate · 14m

Replies to @patriargate

This hate tweet is now being countered. Think twice before retweeting. For every retweet, a donation will be committed to a non-profit fighting for inclusion, equality and diversity. [tinyurl.com/ybv4expb](https://tinyurl.com/ybv4expb)

WE COUNTER HATE  
We understand that culture and demographies are our destiny. We can't restore our civilization with somebody else's babies.  
[https://twitter.com/u\\_of\\_europe/stw](https://twitter.com/u_of_europe/stw)

# Taxonomy of counterspeech Benesch 2016

## 1. Presenting facts to correct misstatements or mis-perceptions

*“Actually homosexuality is natural. Nearly all known species of animal have their gay commu-nities. Whether it be a lion or a whale, they have or had(if they are endangered) a gay community. Also marriageis an unnatural act. Although there are some species that do have longer relationships with a partner most known do not”.*

This comment was in response to a interview video in which the interviewee says that homosexuality is unnatural, detrimental and destructive to the society

# Taxonomy of counterspeech Benesch 2016

1. Presenting facts to correct misstatements or mis-perceptions
2. Pointing out hypocrisy or contradictions



...

Whn Muslims tweet #KillAllChristians its called a  
terrorist threat,whn a Christian say #KillAllMuslims its  
called freedm of speech  
Hypocrisy

1:39 PM · Nov 14, 2015 · Twitter for BlackBerry

# Taxonomy of counterspeech Benesch 2016

1. Presenting facts to correct misstatements or mis-perceptions
2. Pointing out hypocrisy or contradictions
3. Warning of offline or online consequences

*"I'm not gay but nevertheless, whether You are beating up some-one gay or straight, it is still an assault and by all means, this **preacher should be arrested for sexual harassment and instigating!!!**"*

# Taxonomy of counterspeech Benesch 2016

1. Presenting facts to correct misstatements or mis-perceptions
2. Pointing out hypocrisy or contradictions
3. Warning of offline or online consequences
4. **Affiliation**

*“Hey I'm Christian and I'm gay and this guy is so wrong. Stop the justification and start the accepting. I know who my heart and soul belong to and that's with God: creator of heaven and earth. We all live in his plane of consciousness so it's time we started accepting one another. That's all”*

# Taxonomy of counterspeech Benesch 2016

## 5. Visual Communication



# Taxonomy of counterspeech Benesch 2016

5. Visual Communication
6. Denouncing hateful or dangerous speech

“Maybe you are not a racist. But *that's a racist thing to say*”

“#KillAllMuslims is literally the most disgraceful thing I've seen on Twitter”

# Taxonomy of counterspeech Benesch 2016

5. Visual Communication
6. Denouncing hateful or dangerous speech
7. Humor and sarcasm



Replies [redacted] ...

Replies to [redacted]

[redacted] ISIS leaders: We urgently call upon every Muslim to join the fight, especially those in the land of the two shrines (Saudi Arabia), rise.

9:14 PM · Dec 26, 2015 · TweetDeck

Replies [redacted] ...

Replies to [redacted]

[redacted] Too busy being part of a civilised and functioning society. Also, Sherlock S04 in 4 days. I can't miss the first episode.

7:45 PM · Dec 28, 2015 · Twitter for Android

---

53 Retweets 341 Likes

# Taxonomy of counterspeech Benesch 2016

5. Visual Communication
6. Denouncing hateful or dangerous speech
7. Humor and sarcasm
8. Tone

*"I am a Christian, and I believe we're to love everyone!! No matter age, race, religion, sex, size, disorder... whatever!! I LOVE PEOPLE!! treat EVERYONE with respect"*

## **Considerations for Successful Counterspeech. Benesch 2016**

- When do you call a counterspeech as **successful**?

## Considerations for Successful Counterspeech. Benesch 2016

- When do you call a counterspeech as successful?
- First is when the speech has a favorable impact on the original (hateful) user, shifting his or her discourse if not also his or her beliefs. This is usually indicated by an apology or recanting, or the deletion of the original tweet or account.



✓

▼

Today I was reminded of some past insensitive tweets,  
and I am deeply sorry to anyone I offended. I have since  
deleted those tweets as they do not reflect my views or  
who I am today.

3:08 PM · Nov 20, 2019 · Twitter for iPhone

## **Considerations for Successful Counterspeech.** Benesch 2016

- When do you call a counterspeech as successful?
- First is when the speech has a favorable impact on the original (hateful) user, shifting his or her discourse if not also his or her beliefs. This is usually indicated by an apology or recanting, or the deletion of the original tweet or account.
- Second type of success is to **positively affect the discourse norms of the ‘audience’** of a counterspeech conversation: all of the other users or ‘cyberbystanders’ who read one or more of the relevant exchange of tweets.

# **Considerations for Successful Counterspeech.** Benesch 2016

## **Recommended Strategies**

- Warning of Consequences
- Shaming/Labeling
- Empathy and Affiliation
- Humor
- Images

## **Discouraged Strategies**

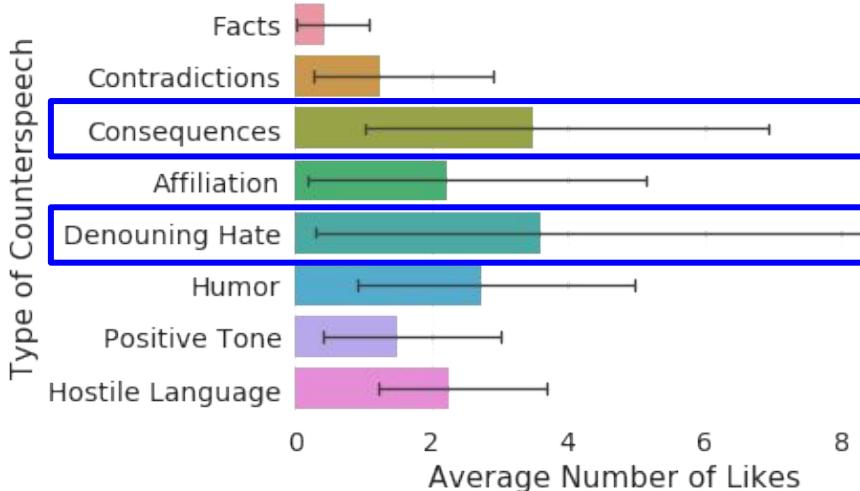
- Hostile or Aggressive Tone, Insults
- Fact-Checking
- Harassment and Silencing

# Thou Shalt Not Hate: Countering Online Hate Speech

## [Mathew 2019]

Click logo for demo

colab



In case of the African-American community, the counterspeakers call out for racism and talk about consequences of their actions

Example:

"i hope these cops got fired! this is bullshit"

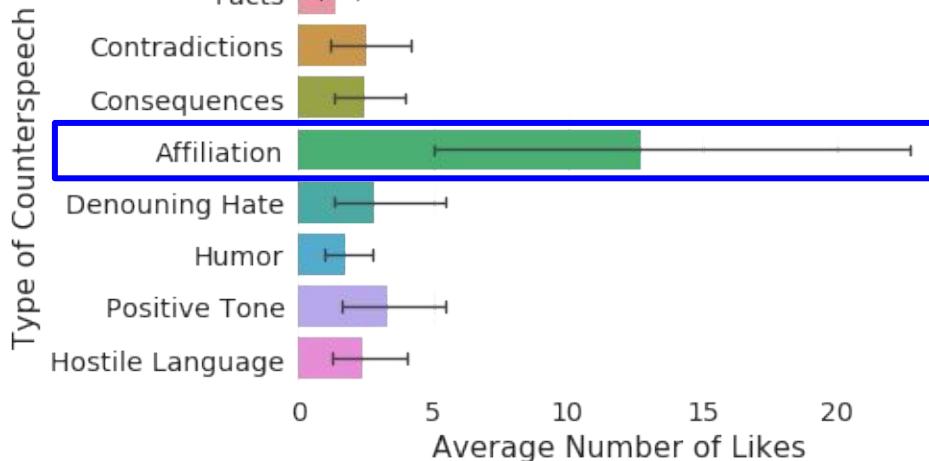
"Sad to see the mom teaching her children to be racist and hateful. The way the guy handled it was great."

# Thou Shalt Not Hate: Countering Online Hate Speech

## [Mathew 2019]

Click logo for demo

colab



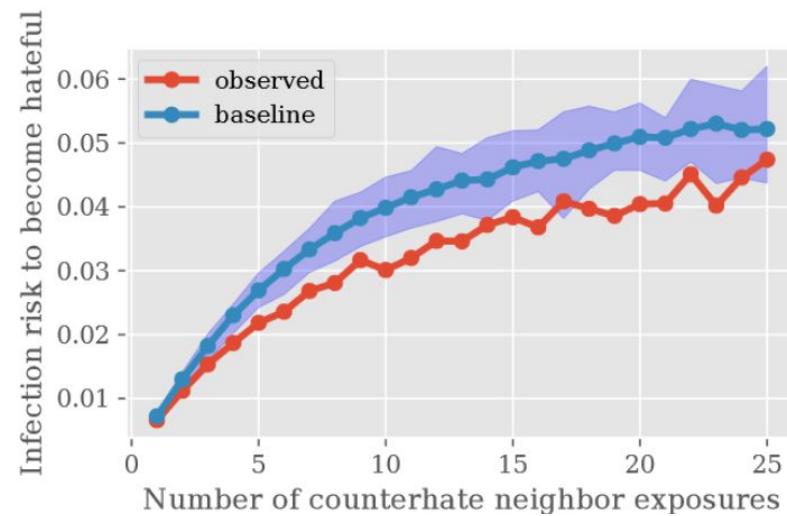
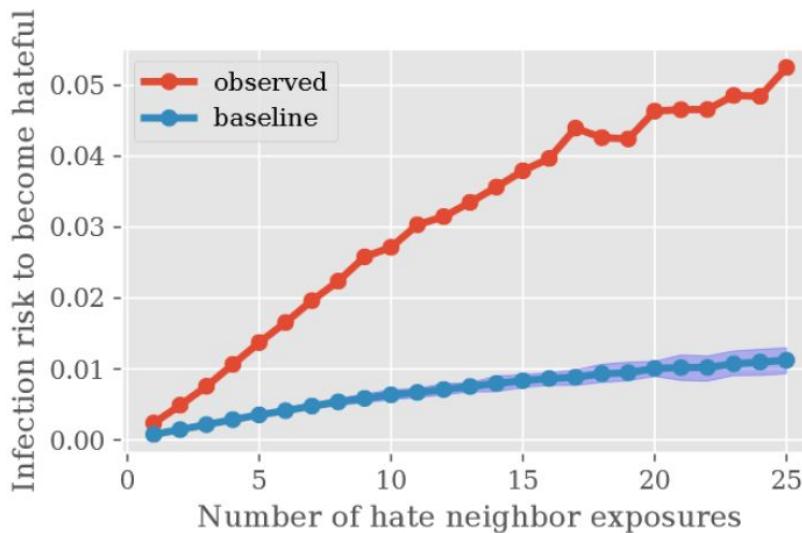
In case of the Jews community, we observe that the people affiliate with both the target and the source community ('Muslims', 'Christians') to counter the hate message.

Example:

"I'm Jewish And I'm really glad there some people that stand up for us And I have no problems with Muslims. We're all brothers and sisters"

# Racism is a Virus: Anti-Asian Hate and Counterhate in Social Media during the COVID-19 Crisis [[Ziems 2020](#)]

Analysis reveals that counterhate messages can discourage users from turning hateful in the first place.



# Datasets

- Counterspeech YouTube [[Mathew 2019](#)]
- Counterspeech Twitter Dataset [[Ziems 2020](#), [Mathew 2020](#), [Garland 2020](#)]
- Hope Speech and Help Speech [[Palakodety 2019](#)] (YouTube Comments)
- CONAN Dataset [[Chung 2019](#)] (NGO Trainers)
- Intervene Dataset [[Qian 2019](#)] (Gab & Reddit)



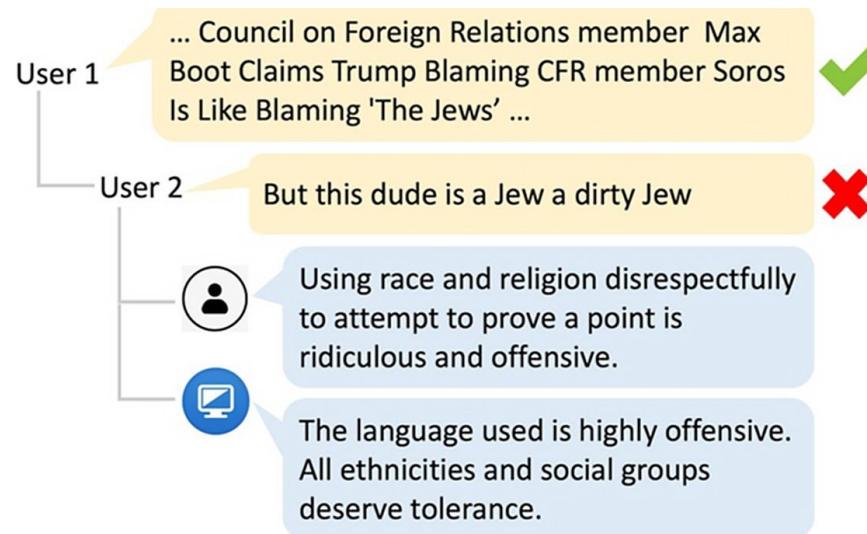
# Counterspeech Generation

The core idea is to **directly intervene** in the discussion with textual responses that are **meant to counter the hate content** and prevent it from further spreading

Manual intervention against hate speech is **not scalable**

# Counterspeech Generation

The core idea is to directly intervene in the discussion with textual responses that are meant to counter the hate content and prevent it from further spreading



# Counterspeech Generation

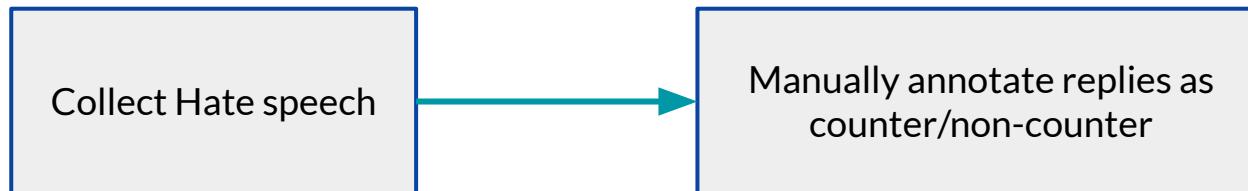
The core idea is to directly intervene in the discussion with textual responses that are meant to counter the hate content and prevent it from further spreading

**Issues:** lack of sufficient amount of **quality data** and tend to produce **generic/repetitive responses.**

# Counterspeech collection Strategy Tekiroglu 2020

Crawling (CRAWL) :Mathew 2019 focuses on the intuition that Counterspeech can be found on social media as responses to hateful expressions. The proposed approach is a mix of automatic hate speech collection via linguistic patterns, and a manual annotation of replies to check if they are responses that counter the original hate content.

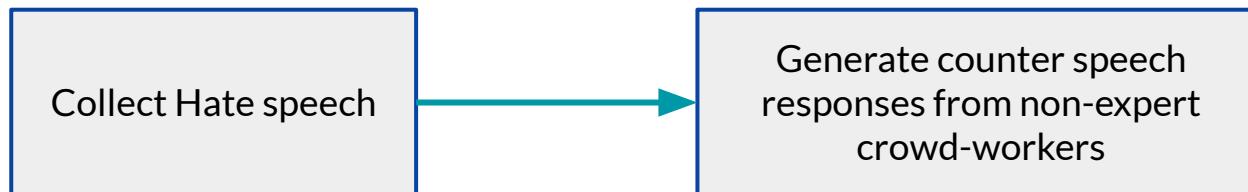
All the material collected is made of natural/real occurrences of hate-counter pairs.



# Counterspeech collection Strategy [Tekiroglu 2020](#)

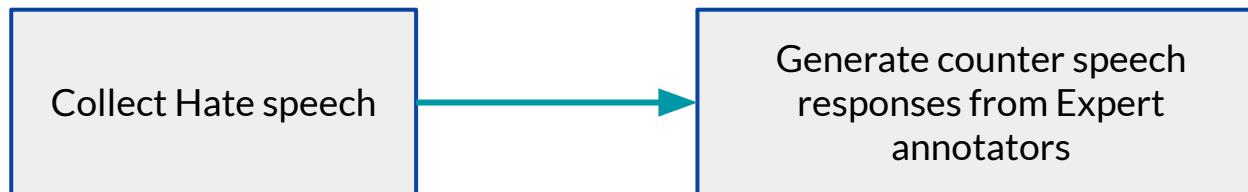
Crowdsourcing (CROWD) : [Qian 2019](#) propose that once a list of hate speech is collected and manually annotated, we can briefly instruct crowd-workers (non-expert) to write possible responses to such hate content.

In this case the content is obtained in controlled settings as opposed to crawling approaches.



# Counterspeech collection Strategy [Tekiroglu 2020](#)

Nichesourcing (NICHE): [Chung 2019](#) still relies on the idea of outsourcing and collecting counterspeech in controlled settings. However, in this case the **counterspeech is written by NGO operators**, i.e. persons specifically trained to fight online hatred via textual responses that can be considered as experts in counterspeech production.

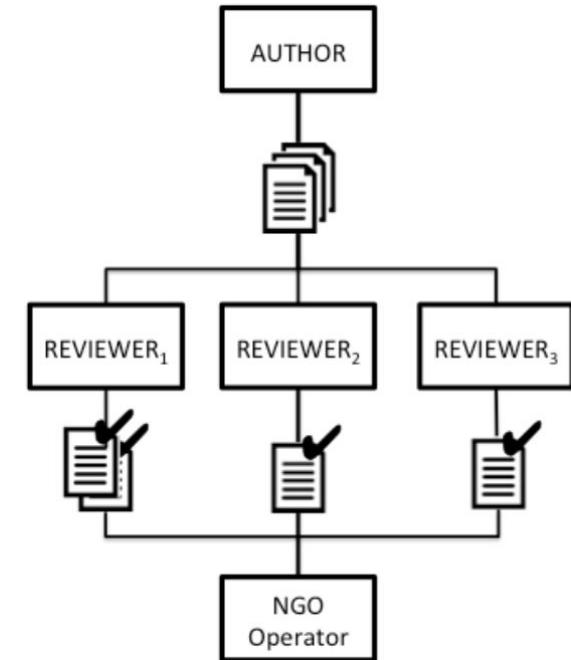


# Counterspeech collection Strategy [Tekiroglu 2020](#)

Author-Reviewer framework [[Tekiroglu 2020](#)]: An author is tasked with text generation and a reviewer can be a human or a classifier model that filters the produced output.

A validation/post-editing phase is conducted with NGO operators over the filtered data.

This framework is scalable allowing to obtain datasets that are suitable in terms of diversity, novelty, and quantity.



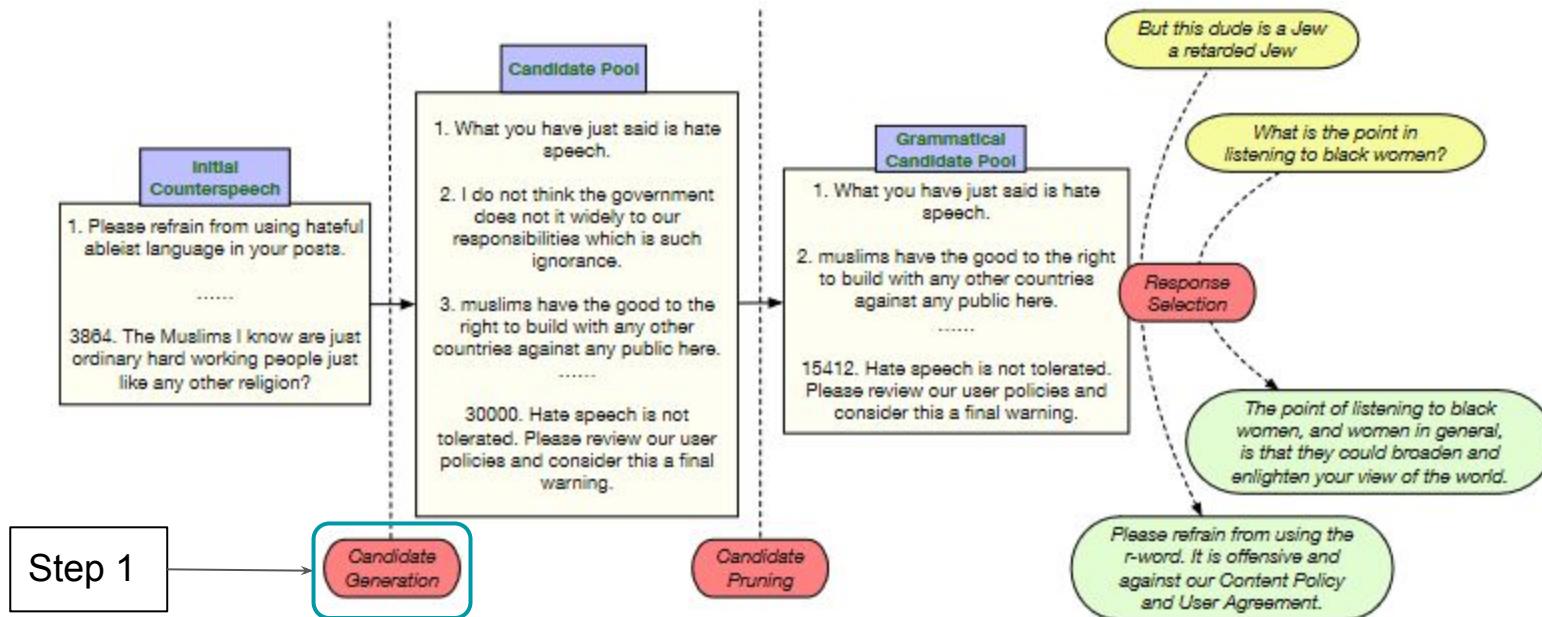
# Machine Generated Counterspeech

Conversation	Seq2Seq	VAE	RL
User 1: Ok. Keep White Knighting for a sand ni**er who f***ks tons of White women. You're entitled to your opinion.	This is is not allowed and un time to treat people by their skin color	Use of the r-word is unacceptable in our discourse as it demeans and insults people with mental disabilities	I do not use derogatory racial slurs
User 1: Feminists are f***king ret***ed, why are they so pissed about manspreading shit makes no difference in space. I'm gonna start throwing things in womens cleavage as they walk by.	Using talking about someones character to describe a woman is insulting to hateful	I don't think using words that are sexist in nature contribute to a productive conversation	Your post uses speech that denigrates sex and gender, which is in violation of our Community Guidelines and TOS Agreement
User 1: Only a nig****ick could reach that c***t User 2: I bet a midget football team could hide under her overhang during a rainstorm and not get wet. User 3: Ni**er probably uses her as a waterbed	Use of the c-word is unacceptable in our discourse as it demeans and insults women	Please do not use derogatory language for intellectual disabilities	If you don't agree with you, there 's no need to resort to name calling

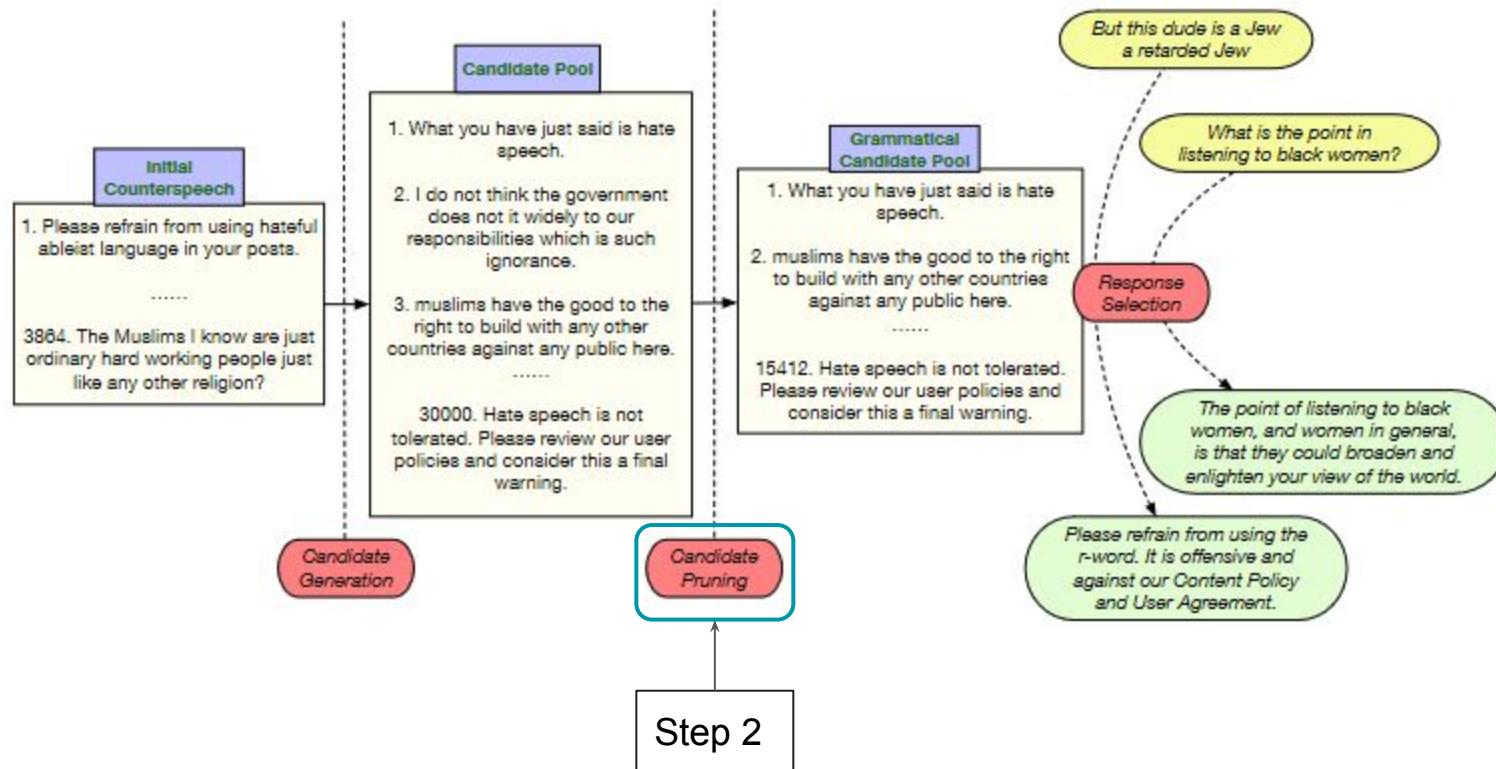
# Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech [Zhu 2021]

<b>Hate Speech:</b>	I am done with Islam and isis. All Muslims should be sent to their homeland. Britain will be better without their violence and ideology.
<b>Expert:</b>	I agree that ISIS is an evil aberration, but to extend this to include up to 3 million people just in the UK is just plain silly.
<b>Common-place:</b>	Hate speech is not tolerated. Please review our user policies. Thank you for your cooperation.
<b>Not relevant:</b>	Use of the r-word is unacceptable as it de-means and insults people with disabilities.

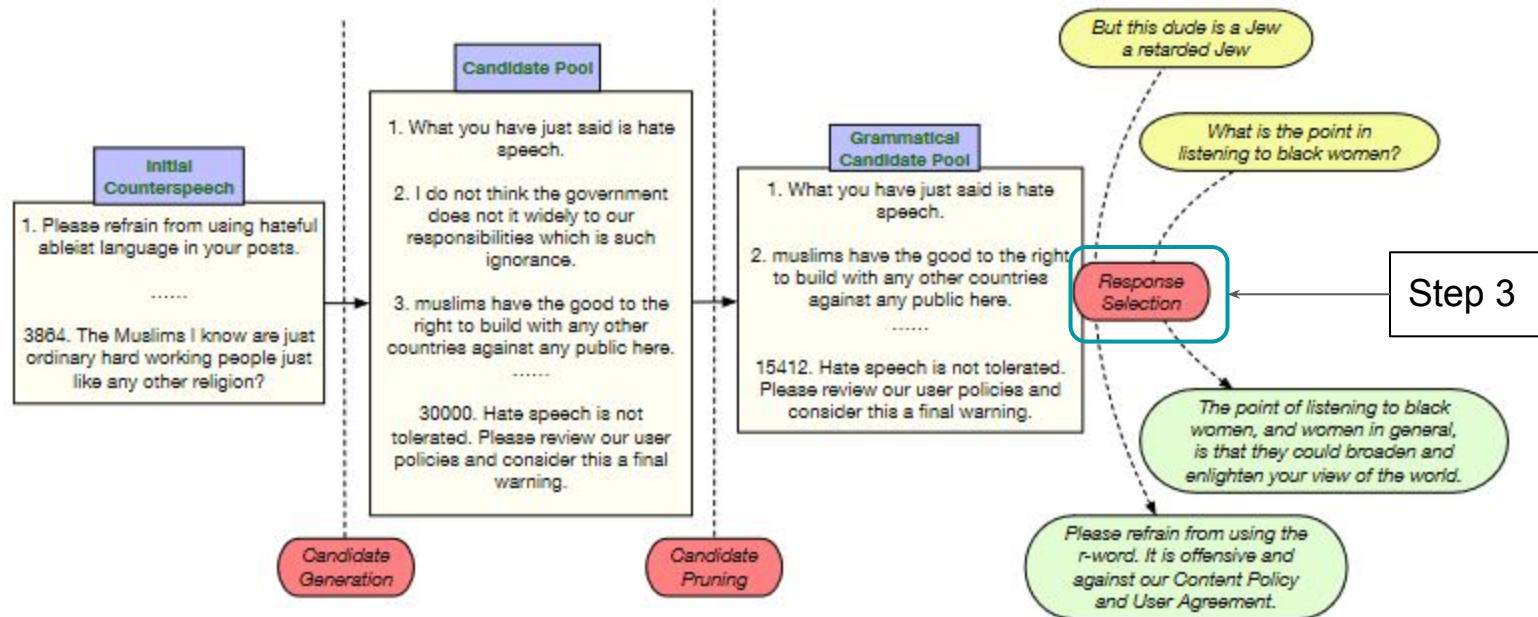
# Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech [Zhu 2021]



# Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech [Zhu 2021]



# Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech [Zhu 2021]



# Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech [Zhu 2021]

Hate Speech	Generated Counterspeech by GPS
(a): All the Muslims should have been punished for yesterday, not only a few. Citizen will take action if the government does not do anything.	You cannot blame all people for actions of a few. Banning something altogether will not solve anything.
(b): The world would be a better place without Muslims. They are only killing and raping our children.	The world would actually be a very dark place without Muslims who contribute a lot to our society. What about our Muslim doctors, scientists, philanthropists, actors, job-creators?
(c): What is the point in listening to black women? They are all retarded.	The point of listening to black women, and women in general, is that they could broaden and enlighten your view of the world.

# Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment [Munger 2016](#)

- Collected a sample of Twitter users who have harassed other users and **use control accounts** (“bots”) to sanction the harassers.

# Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment Munger 2016

- By varying the identity of the bots between in-group (white man) and out-group (black man) and by varying the number of Twitter followers each bot has, the author found that subjects who were countered by a **high-follower white male** significantly **reduced their use of a racist slur**.



@[REDACTED] Hey man, just remember that there are real people who are hurt when you harass them with that kind of language

Greg [REDACTED] (@Greg [REDACTED]) · New York, NY

TWEETS 70 FOLLOWING 39 FOLLOWERS 2

Tweets Tweets & replies

Greg [REDACTED] Retweeted SportsCenter (@SportsCenter) · 11m Michael Oher's "Blind Side" family joined him on the field to celebrate his team advancing to the Super Bowl. [es.pn/1QwVGnw](https://es.pn/1QwVGnw)

510 895 View summary

Greg [REDACTED] (@Greg [REDACTED]) · 15 Sep 2015 Hey man, just remember that there are real people who are hurt when you harass them with that kind of language

View conversation Trends Change

# SWOT

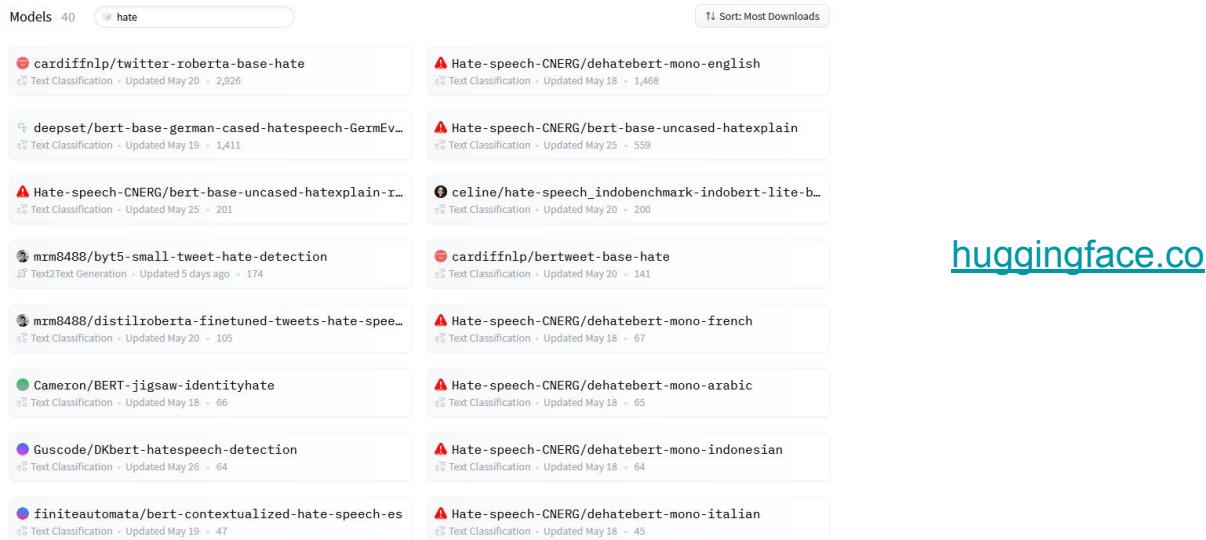
- Definitions and related concepts
- Analysis of hate speech
  - Prevalence
  - Effect
- Detection of hate speech
  - Datasets
  - Traditional methods
  - Sequential models
  - Transformer based models
  - Challenges
- Mitigation of hate speech
  - Campaigns
  - Counterspeech detection
  - Counterspeech generation
  - Effect of counter speech
- SWOT analysis

# **Strengths**

- Growing interest in the scientific community across different disciplines

# Strengths

- Growing interest in the scientific community across different disciplines
- Deep neural architectures specially engineered for hate speech detection, e.g., HateBERT, HateXplain etc.



# Strengths

- Growing interest in the scientific community across different disciplines
- Deep neural architectures specially engineered for hate speech detection, e.g., HateBERT, HateXplain etc.
- **Extensions to multiple languages, e.g., DE-LIMIT**

# Strengths

- Growing interest in the scientific community across different disciplines
- Deep neural architectures specially engineered for hate speech detection, e.g., HateBERT, HateXplain etc.
- Extensions to multiple languages, e.g., DE-LIMIT
- **Datasets becoming available multiple modes, e.g., image, video, text, etc.**

# Strengths

- Growing interest in the scientific community across different disciplines
- Deep neural architectures specially engineered for hate speech detection, e.g., HateBERT, HateXplain etc.
- Extensions to multiple languages, e.g., DE-LIMIT
- Datasets becoming available multiple modes, e.g., image, video, text, etc.
- **Counterspeech initiatives by various NGOs and tech giants**

# Strengths

- Growing interest in the scientific community across different disciplines
- Deep neural architectures specially engineered for hate speech detection, e.g., HateBERT, HateXplain etc.
- Extensions to multiple languages, e.g., DE-LIMIT
- Datasets becoming available multiple modes, e.g., image, video, text, etc.
- Counterspeech initiatives by various NGOs and tech giants
- **Theme research grants, competitions, shared tasks and dedicated workshops**

# **Weakness**

- Inconsistent annotations

# Weakness

- Inconsistent annotations
- Lack of standardization across datasets
  - Different dataset have different class labels --  
abusive/non-abusive, hate/non-hate, toxic/non-toxic

# **Weakness**

- Inconsistent annotations
- Lack of standardization across datasets
  - Different dataset have different class labels --  
abusive/non-abusive, hate/non-hate, toxic/non-toxic
- **Lack of generalisability of the models**

# Weakness

- Inconsistent annotations
- Lack of standardization across datasets
  - Different dataset have different class labels -- abusive/non-abusive, hate/non-hate, toxic/non-toxic
- Lack of generalisability of the models
- **Scarce multilingual and multi-modal data**

# Weakness

- Inconsistent annotations
- Lack of standardization across datasets
  - Different dataset have different class labels -- abusive/non-abusive, hate/non-hate, toxic/non-toxic
- Lack of generalisability of the models
- Scarce multilingual and multi-modal data
- **Bias in data as well as in models**

# Weakness

- Inconsistent annotations
- Lack of standardization across datasets
  - Different dataset have different class labels -- abusive/non-abusive, hate/non-hate, toxic/non-toxic
- Lack of generalisability of the models
- Scarce multilingual and multi-modal data
- Bias in data as well as in models
- **Lack of explainability in models**

# Opportunities

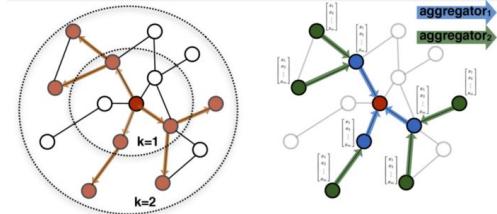
- Hateful user detection

*Timesup, yall getting w should have happened long ago*

Which was in reply to another tweet that mentioned the holocaust. Although the tweet, whose author's profile contained white-supremacy imagery, incited violence, it is hard to conceive how this could be detected as hateful with only textual features. Furthermore, the lack of hate-related words makes it difficult for this kind of tweet to be sampled.

# Opportunities

- **User** as another aspect
  - Helps in contextualising some tweets
  - User moderation more feasible from a practical perspective
  - Issue - Annotation guidelines
- On twitter dataset , GraphSage is the best model  
[\(Riberio,2018\)](#).

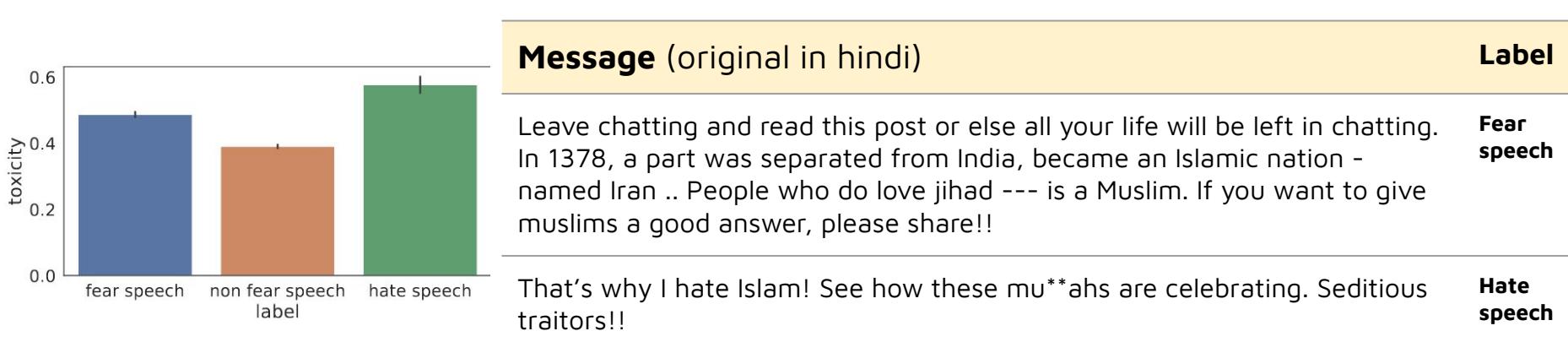


Graph sage algorithm

		Hateful/Normal			Suspended/Active		
Model	Features	Accuracy	F1-Score	AUC	Accuracy	F1-Score	AUC
GradBoost	user+glove	$84.6 \pm 1.0$	$52.0 \pm 2.2$	$88.4 \pm 1.3$	$81.5 \pm 0.6$	$48.4 \pm 1.1$	$88.6 \pm 0.1$
	glove	$84.4 \pm 0.5$	$52.0 \pm 1.3$	$88.4 \pm 1.3$	$78.9 \pm 0.7$	$44.8 \pm 0.7$	$87.0 \pm 0.5$
AdaBoost	user+glove	$69.1 \pm 2.4$	$37.6 \pm 2.4$	$85.5 \pm 1.4$	$70.1 \pm 0.1$	$38.3 \pm 0.9$	$84.3 \pm 0.5$
	glove	$69.1 \pm 2.5$	$37.6 \pm 2.4$	$85.5 \pm 1.4$	$69.7 \pm 1.0$	$37.5 \pm 0.8$	$82.7 \pm 0.1$
GraphSage	user+glove	$90.9 \pm 1.1$	$67.0 \pm 4.1$	$95.4 \pm 0.2$	$84.8 \pm 0.3$	$55.8 \pm 4.0$	$93.3 \pm 1.4$
	glove	$90.3 \pm 1.9$	$65.9 \pm 6.2$	$94.9 \pm 2.6$	$84.5 \pm 1.0$	$54.8 \pm 1.6$	$93.3 \pm 1.5$

# Opportunities

- Lot of new problems coming up
  - Interaction of **fake news with hate speech** ([Ameur,2021](#))
  - Emergence of **fear speech** ([Saha, 2021](#)), **dangerous speech** ([Alsheri,2020](#))



# Threats

- Newer methods of promoting hate -- e.g., hate codes which are very difficult to identify automatically

# Threats

- Newer methods of promoting hate -- e.g., hate codes which are very difficult to identify automatically
- Many new platforms cropping up as alternatives -- [Parler](#) (used to be a small scale initiative, but from the last week of June 2020, 1.5M daily users)

# Threats

- Newer methods of promoting hate -- e.g., hate codes which are very difficult to identify automatically
- Many new platforms cropping up as alternatives -- [Parler](#) (used to be a small scale initiative with few million users, but from the last week of June 2020, 1.5M daily users)
- **Govt agencies and political parties weaponizing hate speech**

# Resources

- [Notion page](#) containing hate speech papers.
- [Demo codes](#) for using our open source models
- A dataset resource created and maintained by Leon Derczynski and Bertie Vidgen. Click the link [here](#)
- This resource collates all the resources and links used in this information hub, for both teachers and young people. Click the link [here](#)



# Thank You

Contacts:

<https://hate-alert.github.io>

[https://twitter.com/hate\\_alert](https://twitter.com/hate_alert)

