Hate speech: Detection, Mitigation and beyond

Tutorial at WSDM 2023

wsdm 2023
SINGAPORE FEBRUARY 27 - MARCH 3

“

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

Animesh Mukherjee
[@Animesh43061078]

Binny Mathew
[@_BinnyM]

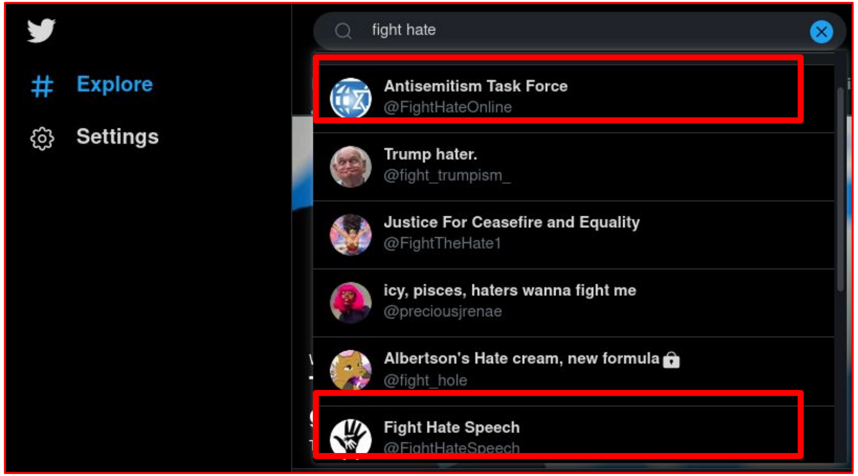
Punyajoy Saha
[@punyajoysaha]

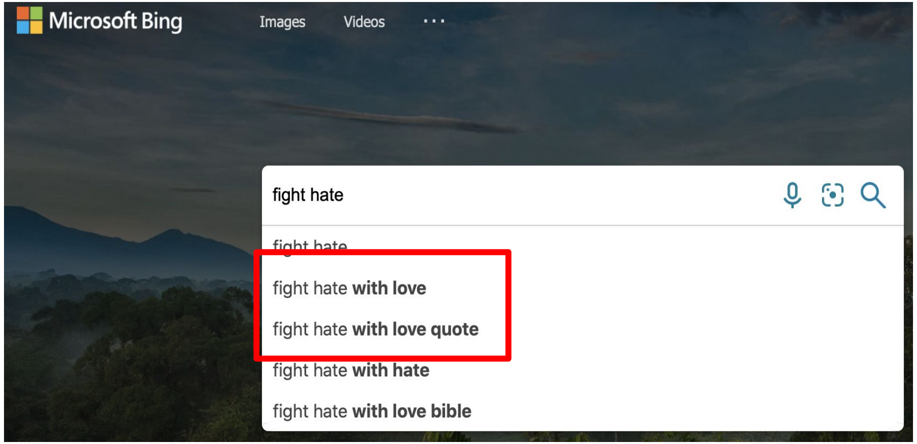
Mithun Das
[@dasmithun92]

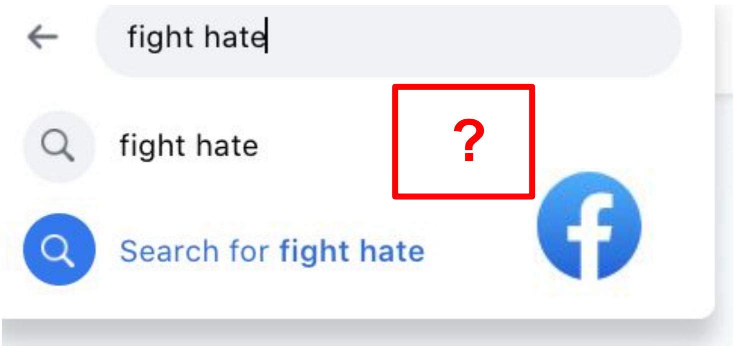
Organisers

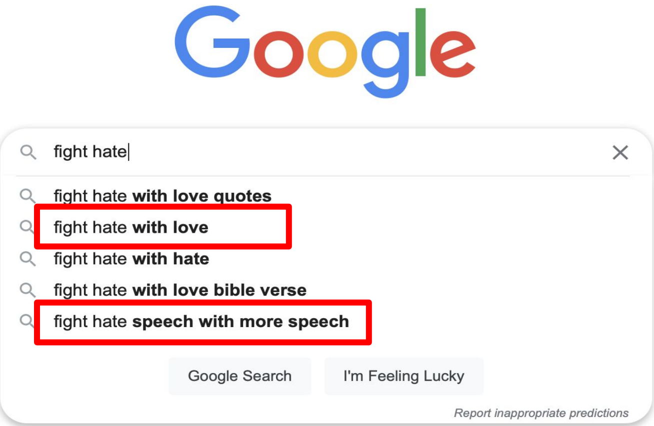
Find more about us here!
[https://hate-alert.github.io/]

Hate speech: A growing concern?



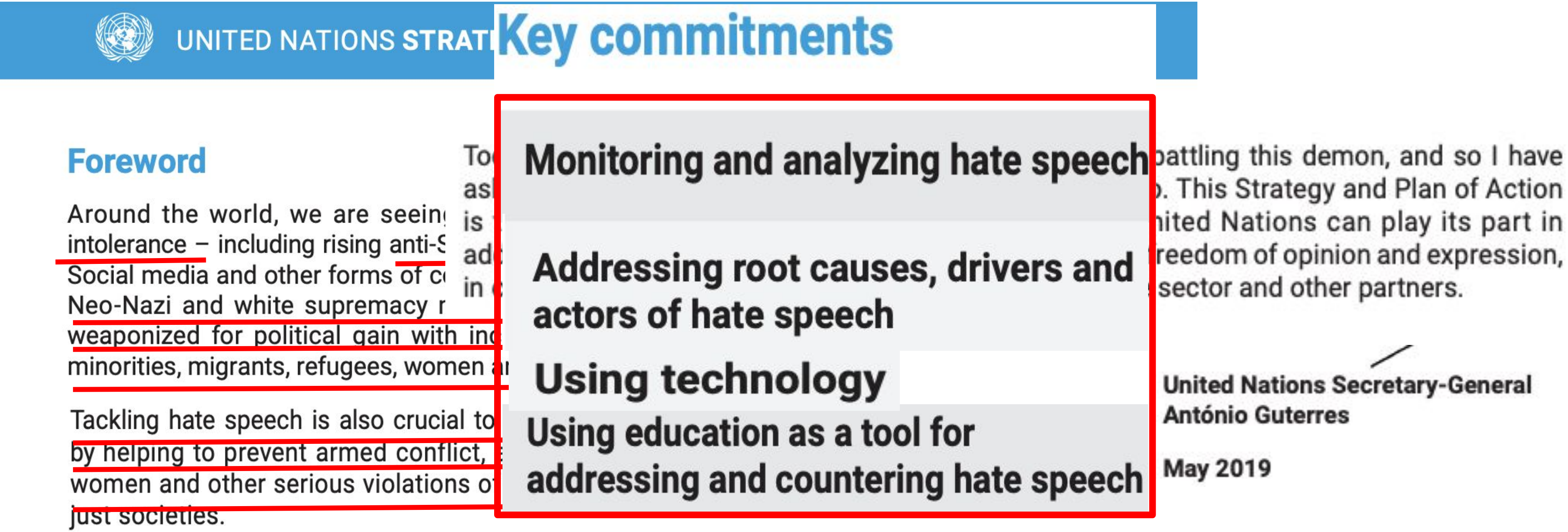






What to expect from this tutorial?

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



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?

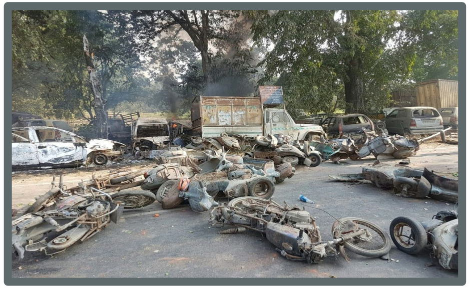
What to expect from this tutorial?

- 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

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

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.

What we will be covering in this tutorial.

Fortuna et al. 2018

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)

right groups have also taken an interest in genetic testing, using them to attack minorities and prove their genetic "purity." In

Over the past few years, the "QAnon" conspiracy has emerged on the anonymous Politically Incorrect (/pol/) board of 4chan. In October 2017, a user going by the nickname "Q" posted numerous threads claiming to be a US government official with a top-secret Q clearance [5]. They explained that Pizzagate was real and that many celebrities, aristocrats, and elected politicians are involved in this vast, satanic pedophile ring. Q further claimed that President Donald Trump is actively working against a satanic pedophile cabal within the US government. QAnon incorporates many theories together into a broadly defined *super-conspiracy theory.*

Analysis and Spread

Prevalence of hate speech

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

gab =

4chan =

BITCHUTE =

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 (Platforms with no moderation)

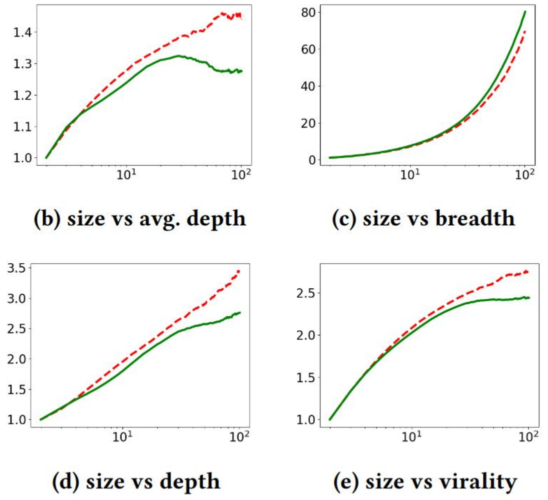
- **Gab**
- In Gab, early signals show **Alt-right**, **BanIslam** as popular hashtags (Zannettou,2018)

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%

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

Prevalence of hate speech (Platforms with no moderation)

- **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)



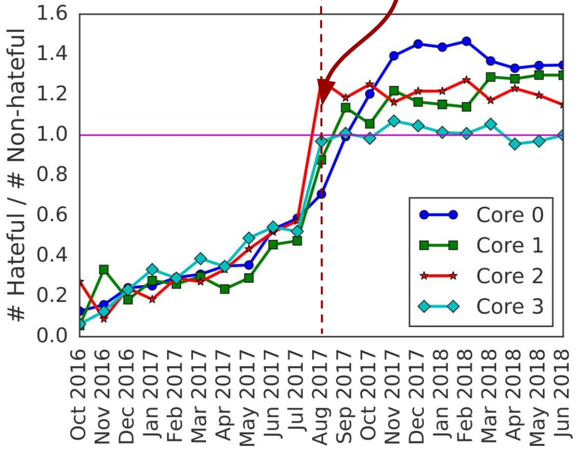
X-axis vs Y-axis

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 (Platforms with no moderation)

- **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)]

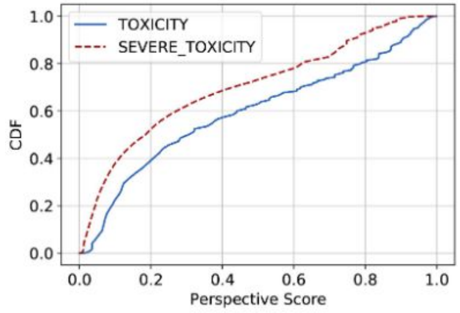
[Unite the right rally]
White supremacist rally at Charlottesville, Virginia

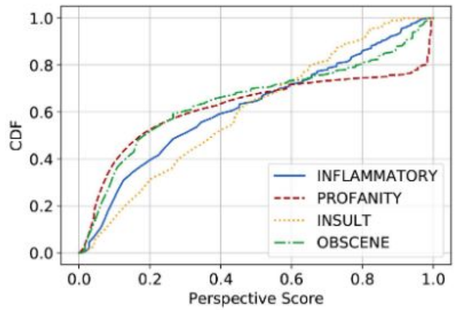


Dataset: collect 21M posts from 340k users, between August 2016 and January 2018
Method: Hate user extraction + Temporal k-core analysis

Prevalence of hate speech (Platforms with no moderation)

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



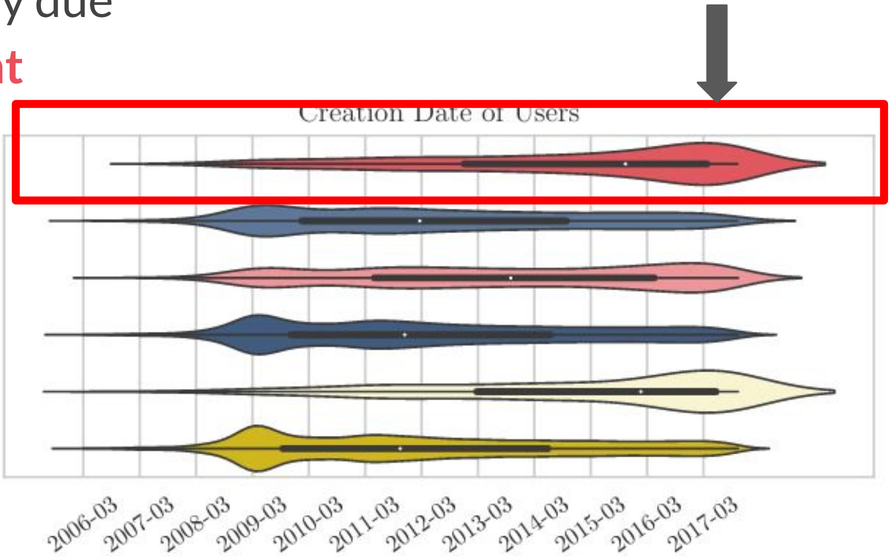


Dataset: Crawling from 4chan's /pol/ thread, June 29, 2016 to November 1, 2019.
Method: Perspective api then CDF

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**

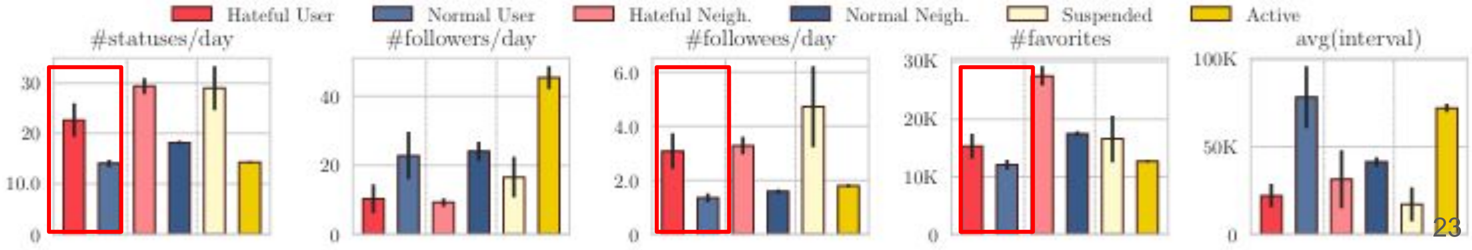


Dataset: Data collected from Twitter, keyword based extraction
Method: Degroot method. Frequency based analysis

Prevalence of hate speech (Platforms with moderation)

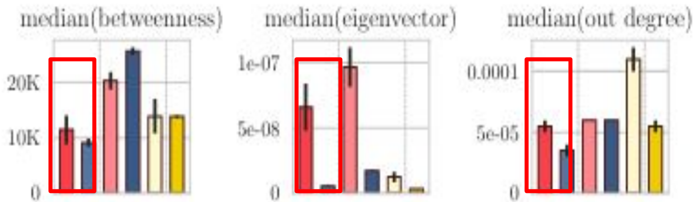
Study on characterising hateful users in Twitter (Riberio,2018)

- 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



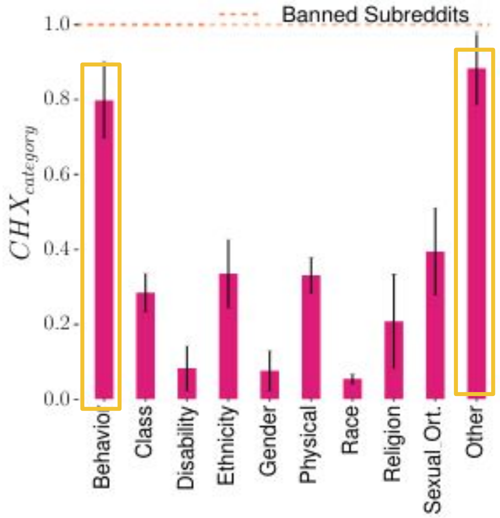
Effect of hate speech

- It is important to understand the **psychological effect** of hate speech
- **Pre-social media -** Interview based study revealed **short-term** → **emotional** & **long term** → **attitudinal** (Leets, 2002)
- **Ignorance** and **repressed hostility** were most common speculated motives(Leets, 2002).
- Most participants prefer **passive response** (Leets, 2002).

Dataset: Interviews with the participants, hate speech (anti-semtism and anti-gay) statements shown as stimulus
Method: Frequency of different codes followed by significance analysis.

Effect of hate speech

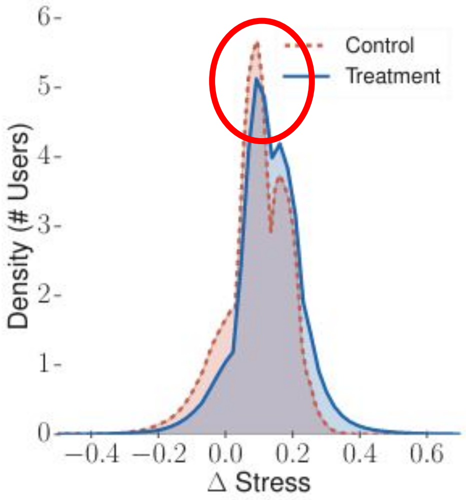
- In a large scale study, the authors found prevalence of hate speech in **college subreddits**. (Saha, 2019)



(a)

Effect of hate speech

- In a large scale study, the authors found prevalence of hate speech in **college subreddits.** (Saha, 2019)
- **Significant difference** exist between the hate exposed (treatment) and not hate exposed group's (control) stress level. (Saha, 2019)



Dataset: Subreddits of different college groups
Method: Hate identifying using keywords, Stress detector used to measure stress between hate exposed vs not group

Dataset: Interviews with the participants based on anti-LGBT hate speech
Method: Coding strategy with significance analysis

Effect of offline events

- An interview based further looked into the pathways of effect and response in a longitudinal study of impact of hate crimes (Patterson, 2018)
- Direct victims were **less empathetic** towards other victims.

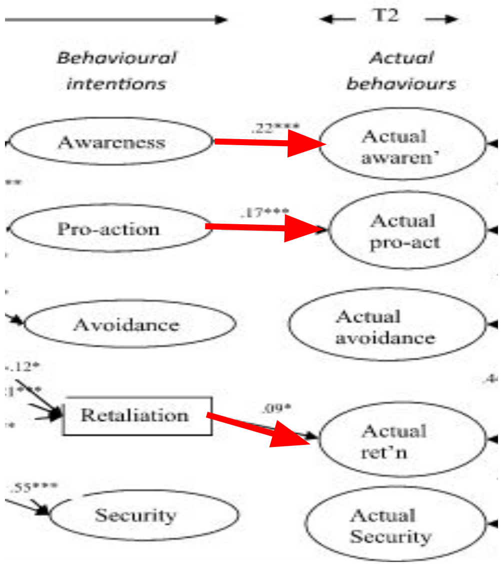
Effect of offline events

Dataset: Interviews with the participants based on anti-LGBT hate speech
Method: Coding strategy with significance analysis

- An interview based looked into the pathways of effect and response in a longitudinal study of impact of **hate crimes** [(Patterson, 2018)]
- Direct victims were **less empathetic** towards other victims.
- Longitudinal study show not all **behavioural intentions** transformed to **actual actions**

Gap of 3 months

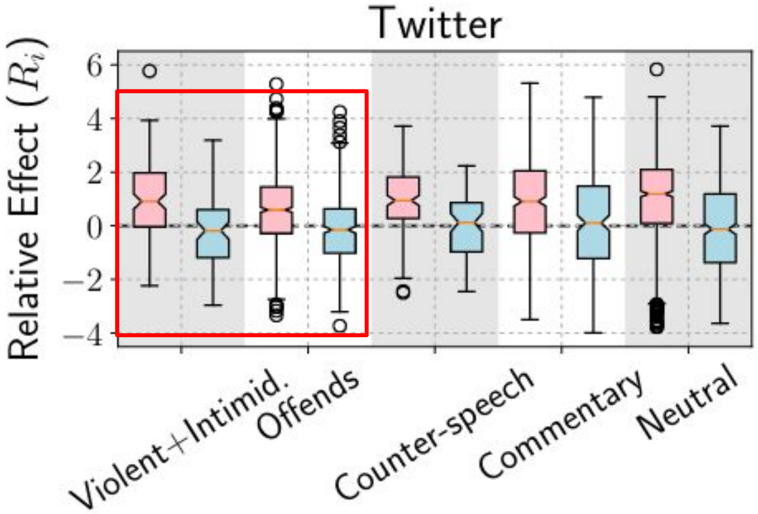
Time 1 **Time 2**



Effect of offline events

Dataset: Collected from twitter using islamic keywords
Method: Framing annotations with impact analysis

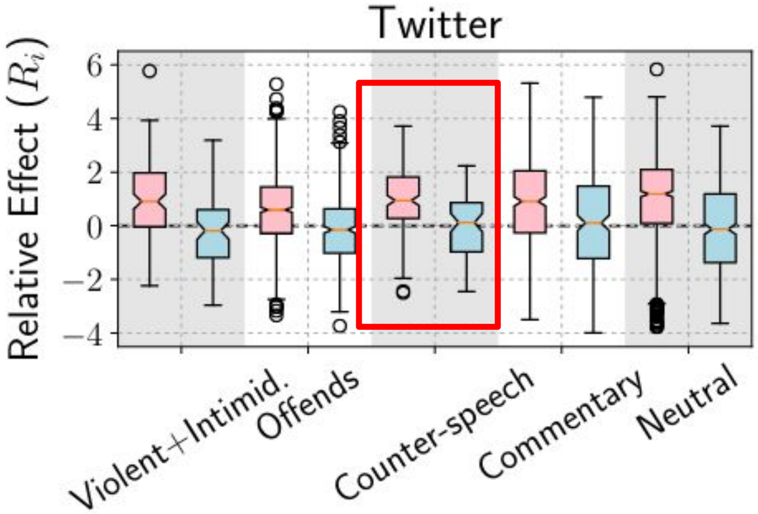
- A study on different social media platforms measured the effect of **hate crime** and **terrorism** on **hate** and **counter speech** (Olteanu, 2018).
- Terms with violence and offense increased after **terrorism** but **not after hate crime**



Effect of offline events

Dataset: Collected from twitter using islamic keywords
Method: Framing annotations with impact analysis

- A study on different social media platforms measured the effect of **hate crime** and **terrorism** on **hate** and **counter speech** (Olteanu, 2018).
- Terms with violence and offense increased after terrorism but **not after** hate crime
- Terms with counterspeech increased after **terrorism** but **not after hate crime**

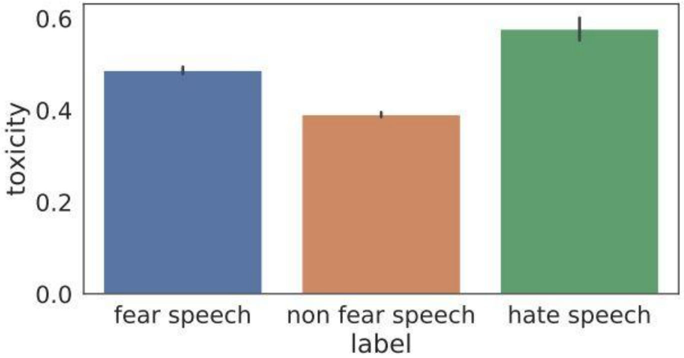


Not Hateful?? Not Normal?? What's Then ?

Text (translated from Hindi)	Label
Leave chatting and read this post or else all your life will be left in chatting. In 1378, a part was separated from India, became an Islamic nation - named Iran …and now Uttar Pradesh, Assam and Kerala are on the verge of becoming an Islamic state …People who do *love jihad* – is a Muslim. Those who think of ruining the country – Every single one of them is a Muslim !!!! Everyone who does not share this message forward should be a Muslim. If you want to give muslims a good answer, please share!! We will finally know how many Hindus are united today !!	FS
That's why I hate Islam! See how these mullahs are celebrating. Seditious traitors!!	HS
A child's message to the countrymen is that Modi ji has fooled the country in 2014, distracted the country from the issues of inflationary job development to Hindu-Muslim and patriotic issues.	NFS

Not Hateful?? Not Normal?? ~~What's Then ?~~ Fear speech

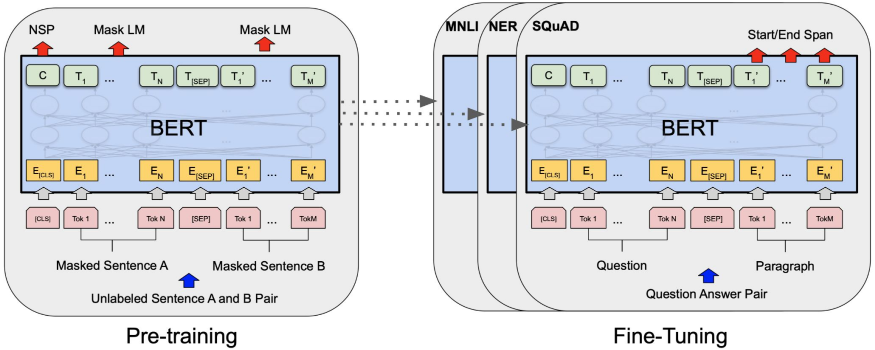
- Fear speech used elements from **history**, and contains **misinformation** to vilify Muslims. At the end, they ask the readers, to take action by **sharing the post**(Saha,2021).



Text (translated from Hindi)	Label
Leave chatting and read this post or else all your life will be left in chatting. In 1378, a part was separated from India, became an Islamic nation - named Iran …and now Uttar Pradesh, Assam and Kerala are on the verge of becoming an Islamic state …People who do *love jihad* – is a Muslim. Those who think of ruining the country – Every single one of them is a Muslim !!!! Everyone who does not share this message forward should be a Muslim. If you want to give muslims a good answer, please share!! We will finally know how many Hindus are united today !!	FS
That's why I hate Islam! See how these mullahs are celebrating. Seditious traitors!!	HS
A child's message to the countrymen is that Modi ji has fooled the country in 2014, distracted the country from the issues of inflationary job development to Hindu-Muslim and patriotic issues.	NFS

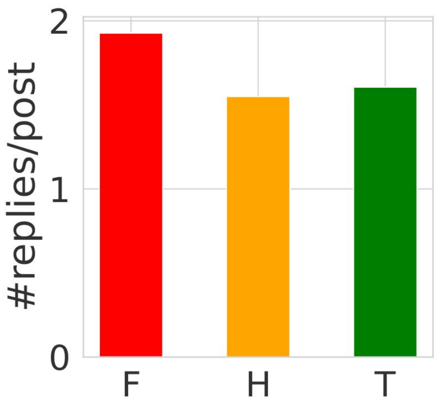
Fear speech datasets and models

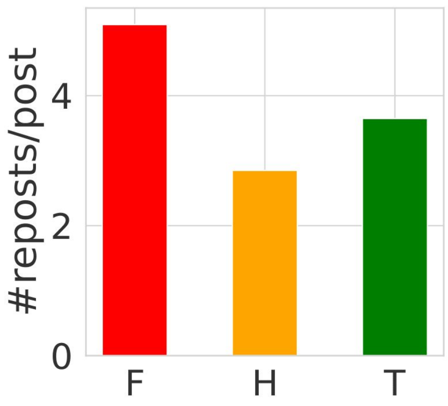
Source	Annotators	Fear speech	Hate speech	Normal
Whatsapp	in-house	1,142	--	3640
Gab	Mturk annotators	1,800	4000	4200

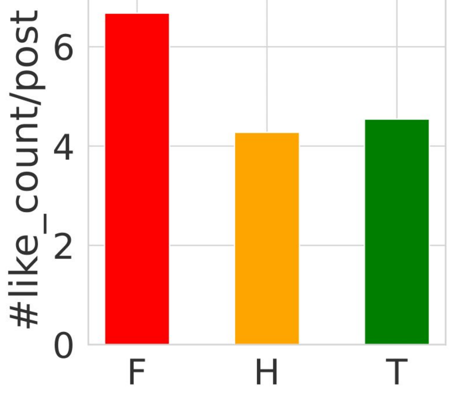


Fear speech vs Hate speech: Reactions on posts

We observe that the average level of engagement of users with fear speech posts is much higher than hate speech posts.

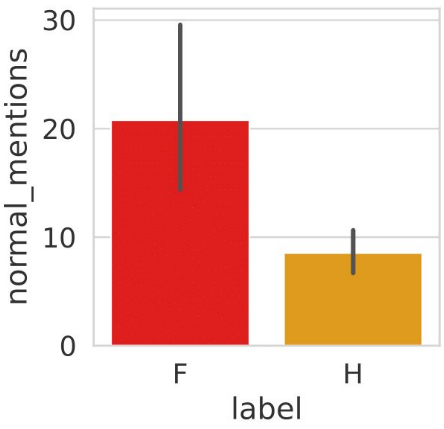


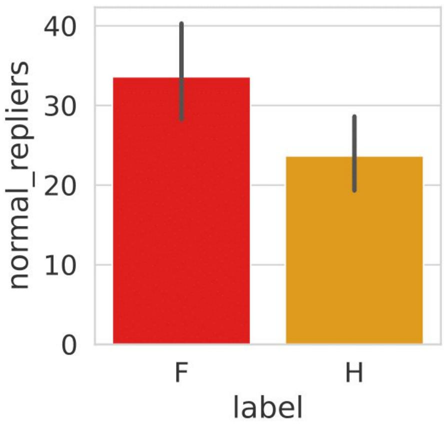


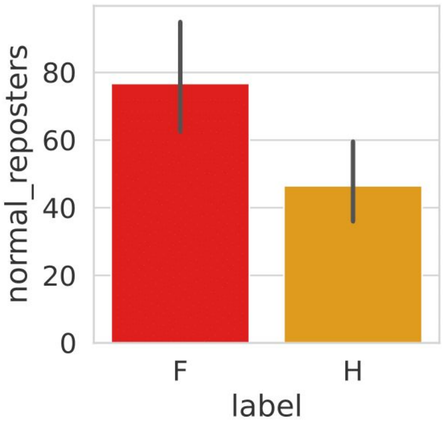


Fear speech vs Hate speech: Effect on normal users?

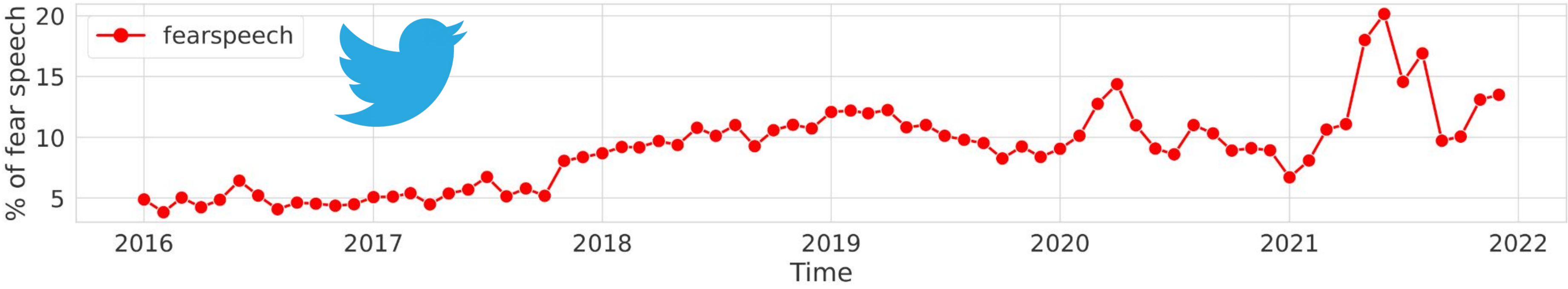
Normal users get mentioned more, reply more and repost more to fear speech than hate speech

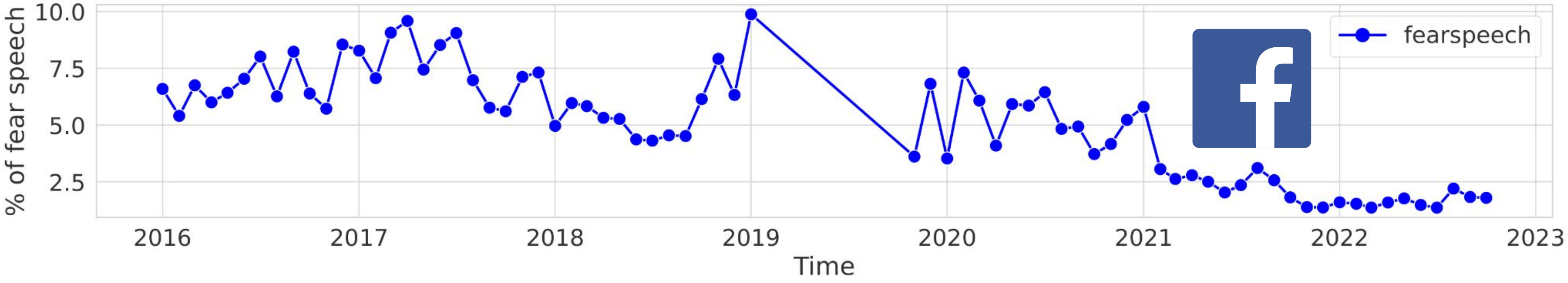






What about other platforms?





Harmful speech

Harmful speech consists of a range of phenomenon that often overlap and intersect, and includes a variety of types of speech that cause different harms.

Hate speech

Cyberbullying

Fear speech

Call for violence

Trolling

Offensive speech

[1] Faris, R., Ashar, A., Gasser, U., & Joo, D. (2016). Understanding harmful speech online. Berkman Klein Center Research Publication, (2016-21).

Detecting Hate Speech

Datasets

- Different dimensions datasets.
  - Unimodal (textual) (Zampieri,2019, Davidson,2017 , Basile,2019)
  - Multimodal (audio, memes, videos) (Gomez,2019, Goswami.2021, Pramanick.2021)

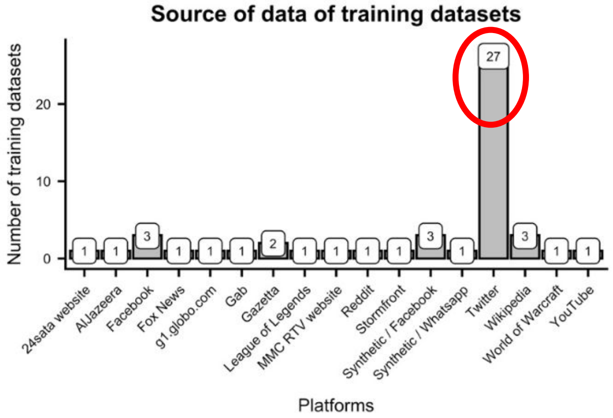
Datasets

- Different dimensions 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)

Vidgen B, Derczynski L (2020) *Directions in abusive language training data, a systematic review: Garbage in, garbage out.* PLoS ONE 15(12): e0243300. https://doi.org/10.1371/journal.pone.0243300.

Datasets

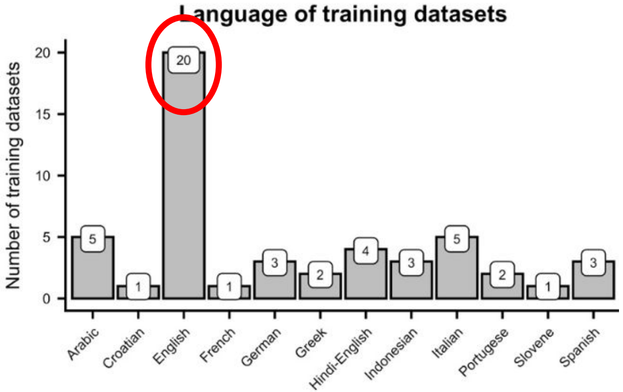
- Different datasets have different **taxonomies**.
- Different datasets have different **sources**.
  Twitter is one of the major sources.
  - The works by Davidson ([Davidson,2017]()) and Founta ([Founta, 2018]()) are two highly used dataset from Twitter
  - Twitter is easily accessible.
  - Alt-right platforms are often taken down, hence studies are limited ([Voat](), [Parler]())



Vidgen B, Derczynski L (2020) *Directions in abusive language training data, a systematic review: Garbage in, garbage out*. PLoS ONE 15(12): e0243300. https://doi.org/10.1371/journal.pone.0243300.

Datasets

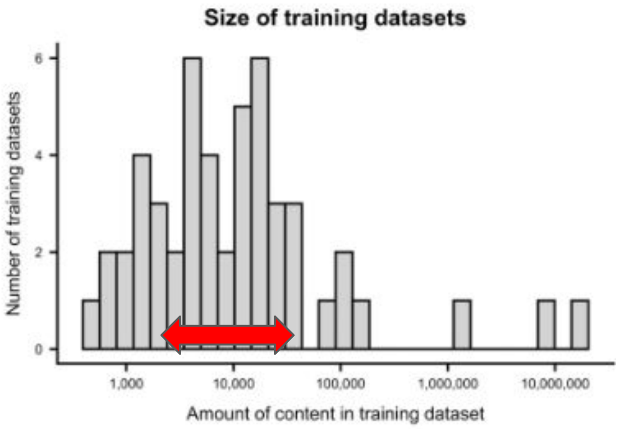
- Different datasets have different **taxonomies.**
- Different datasets have different **sources.**
  Twitter is one of the major sources.
- Different datasets have different **languages**, English being the prominent one.
  - Arabic (Mulki,2019), Italian (Sanguinetti,2018), Spanish (Basile,2019) and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

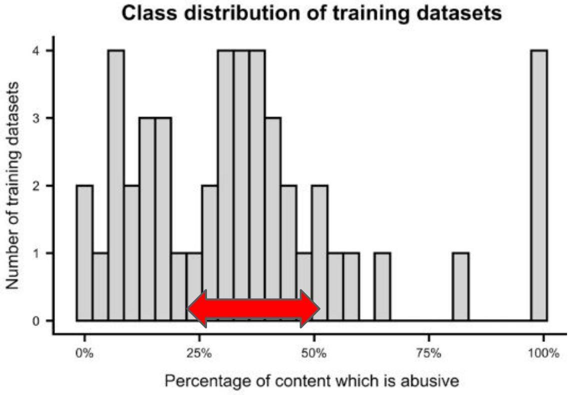


Vidgen B, Derczynski L (2020) *Directions in abusive language training data, a systematic review: Garbage in, garbage out.* PLoS ONE 15(12): e0243300. https://doi.org/10.1371/journal.pone.0243300.

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





Vidgen B, Derczynski L (2020) *Directions in abusive language training data, a systematic review: Garbage in, garbage out.* PLoS ONE 15(12): e0243300. https://doi.org/10.1371/journal.pone.0243300.

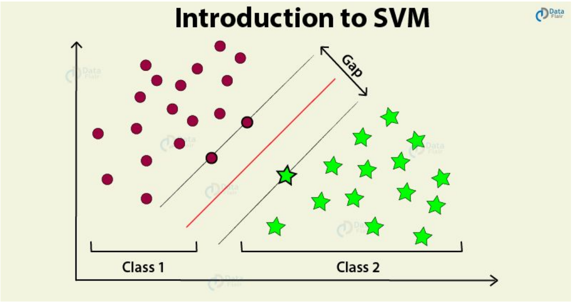
Earlier Detection Methods

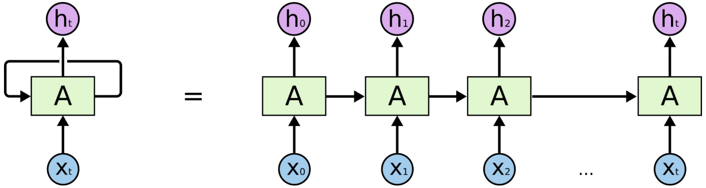
- Features used :-
  - TF-IDF vectors
  - Parts-of-speech tags
  - Linguistic features
    - Sentiment lexicons
    - Frequency counts of URL, username
    - Readability scores
  - **Word embeddings**
    - Twitter word embeddings (Zimmerman, 2018). Click here
  - **Sentence embeddings**
    - Google's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

(Davidson,2017)

Earlier Detection Methods

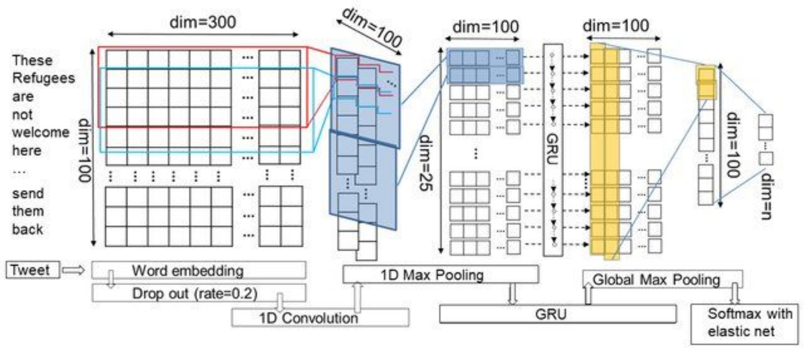
- Features used
- Detection method
  - Logistic regression
  - **SVM** (Canós,2018)
  - XGboost (Saha, 2018)
  - **LSTM/GRU** (Gao,2017)
  - CNN-GRU (Zhang, 2018)





Earlier Detection Methods

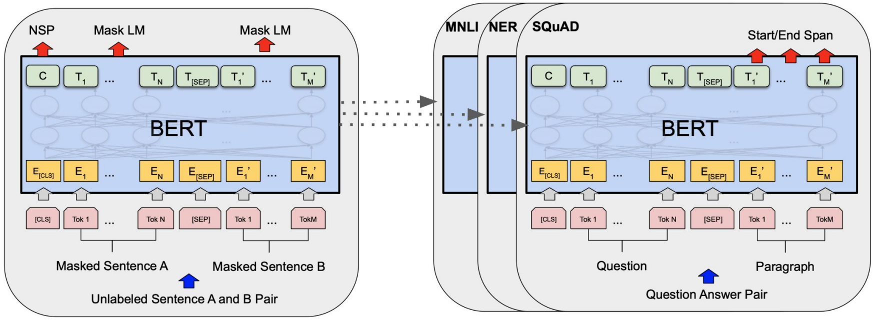
- Features used
- Detection method
  - Logistic regression
  - SVM (Canós,2018)
  - XGboost (Saha, 2018)
  - LSTM/GRU (Gao,2017)
  - CNN-GRU (Zhang, 2018)



Dataset	SVM	SVM+	CNN	CNN+ GRU$_B$	CNN+ GRU	State of the art
WZ-L	0.74	0.74	0.80	0.81	**0.82**	.74 Waseem [26], best F1
WZ-S.amt	0.86	0.87	0.91	**0.92**	**0.92**	.84 Waseem [25], Best features
WZ-S.exp	0.89	0.90	0.90	0.91	**0.92**	.91 Waseem [25], Best features
WZ-S.gb	0.86	0.87	0.91	0.92	**0.93**	.90 Gamback [10], best F1
WZ-LS	0.72	0.73	0.81	0.81	0.82	.82 Park [20], WordCNN .81 Park [20], CharacterCNN **.83** Park [20], HybridCNN
DT	0.87	0.89	**0.94**	**0.94**	**0.94**	.87 SVM, Davidson [7]
RM	0.86	0.89	0.90	0.91	**0.92**	.86 SVM, Davidson [7]

Current Models

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



Pre-training
Fine-Tuning

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
$\mathrm{BERT}_{\mathrm{base}}$	Waseem	81	81	81
	Davidson	91	91	91
$\mathrm{BERT}_{\mathrm{base}}$ + Nonlinear Layers	Waseem	73	85	76
	Davidson	76	78	77
$\mathrm{BERT}_{\mathrm{base}}$ + LSTM	Waseem	87	86	86
	Davidson	91	92	92
$\mathrm{BERT}_{\mathrm{base}}$ + CNN	Waseem	**89**	**87**	**88**
	Davidson	**92**	**92**	**92**

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).

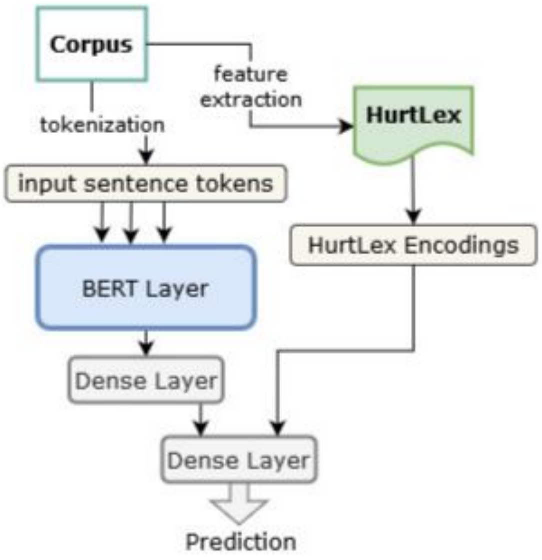
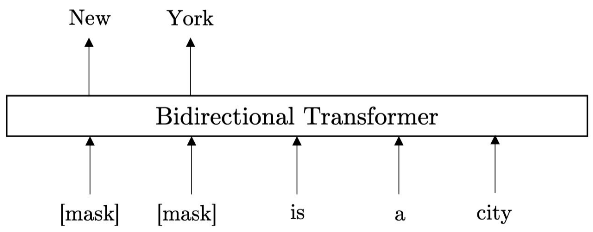


Figure 1: HurtBERT-Enc, our model using HurtLex Encodings

Current Models



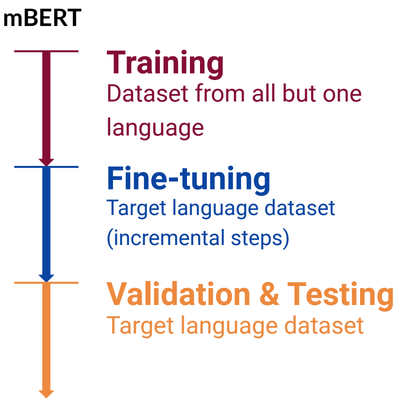
- Earlier models cannot completely capture context
- **BERT** and other transformers model helped in getting improved performance across different datasets (Mozafari,2019)
- Incorporating lexicon into the BERT architecture → HurtBERT (Koufakou,2020).
- Re-training BERT with banned subreddit data → HateBERT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

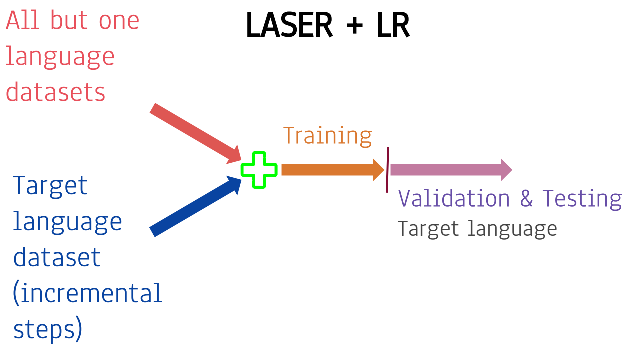
Dataset	Model	Macro F1	Pos. class - F1
OffensEval 2019	BERT	.803±.006	.715±.009
	HateBERT	**.809±.008**	**.723±.012**
	Best	.829	.599
AbusEval	BERT	.727±.008	.552±.012
	HateBERT	**.765±.006**	**.623±.010**
	Caselli et al. (2020)	.716±.034	.531
HatEval	BERT	.480±.008	.633±.002
	HateBERT	**.516±.007**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



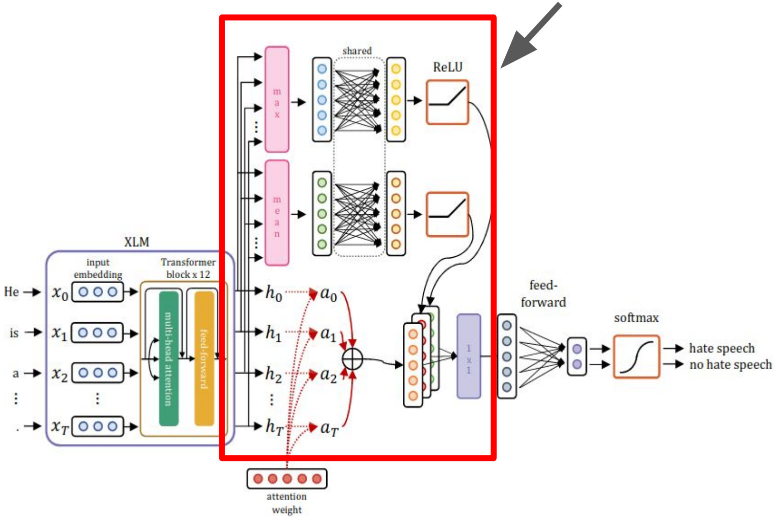


Click logo for demo

Multilingual Hate speech

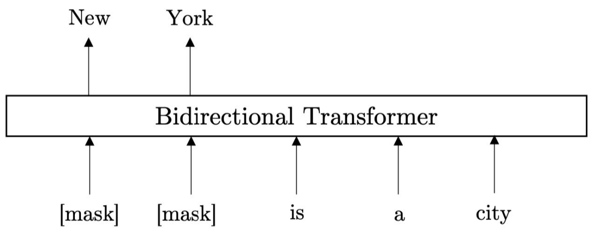
- Benchmarking multilingual models across 9 different languages and 16 datasets (Aluru,2020).
- A novel classification block -AXEL to improve cross lingual transfer (Stappen,2020) on Hateval data.

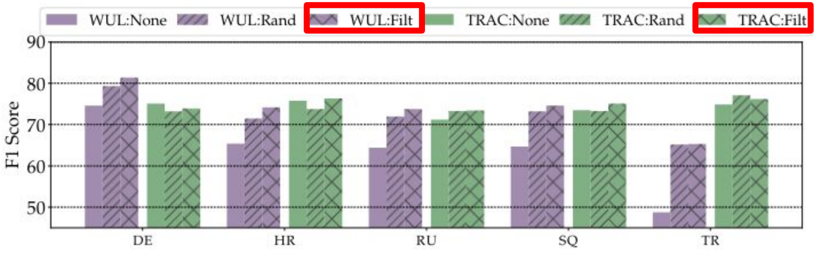
	Dense	Att	AXEL
EN⇒ES	41.31	34.37	**53.42**
ES⇒EN	**60.83**	48.47	52.48
ES⇒EN-S	49.38	39.10	**53.24**
EN⇒(ES→EN)	60.59	62.40	**64.39**
ES⇒(EN→ES)	56.89	49.17	**58.31**
ES⇒(EN-S→ES)	56.57	49.17	**65.04**



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)

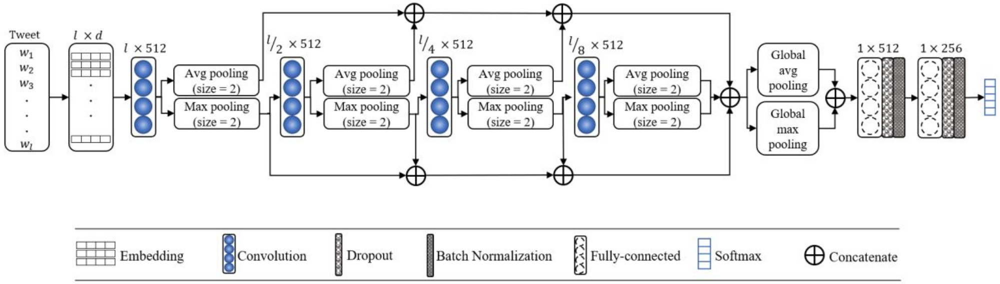




Code-Mixed/Roman Hate speech

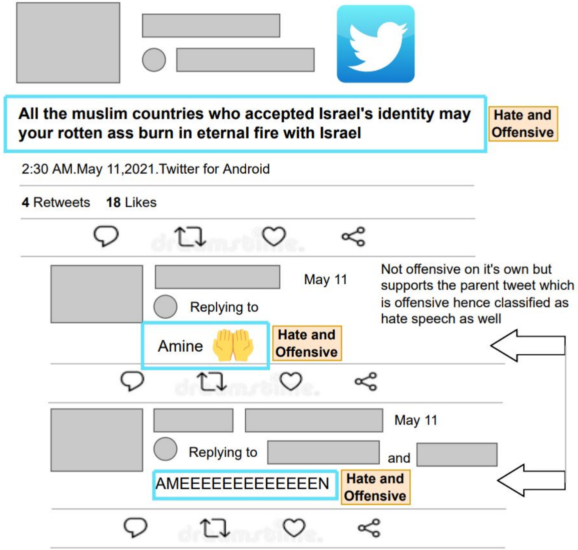
Tweet	Translation	Target Label
randi ke bache tu apne hashar ki fikar kar	you son of a prostitute, you should worry for what will happen to you.	Abusive/Offensive
Hindu bhenchod hi ki gaand ma hi keerra hota hay Tum hindu ho hi harami tumhara kabhi 1 baap nhi hota	There are always insects in asses of Hindu sisterfu**kers. These hindus have multple fathers instead of 1	Religious Hate
No wonder you can't make it to First Lady. At least you managed to grab the title of FIRST RANDDI	No wonder you can't make it to First Lady. At least you managed to grab the title of FIRST PROSTITUTE	Sexism
bahria central park karachi forms sold out in two days. Abhi tax maango bhenchodo ka rona shru hojayega	bahria central park karachi forms sold out in two days. Now ask them for tax these motherf**kers start crying.	Profane
pakistan me ptv news or ptv parliment ne hi mulk k liye acha kam kia	in pakistan, only ptv news and ptv parliment has done good work for the country	Neutral

	Accuracy	Precision	Recall	F1-score
LSTM+GBDT	0.54	0.58	0.51	0.38
BERT+LASER+GBDT	0.89	0.89	0.89	0.89
FastText+CNN	0.87	0.87	0.87	0.87
SVM+RF+AB	**0.90**	**0.90**	**0.90**	**0.90**
BERT+LAMB	**0.90**	**0.90**	0.89	0.89
Domain Embeddings+CNN	0.88	0.89	0.88	0.88
BiLSTM with Attention	0.86	0.86	0.85	0.85
BERT+CNN-gram	**0.90**	**0.90**	**0.90**	**0.90**
XLM-RoBERTa+CNN-gram	0.88	0.88	0.88	0.88
FastText+CNN-gram	0.81	0.81	0.80	0.80
RomUrEm+CNN-gram	0.89	0.89	0.89	0.89

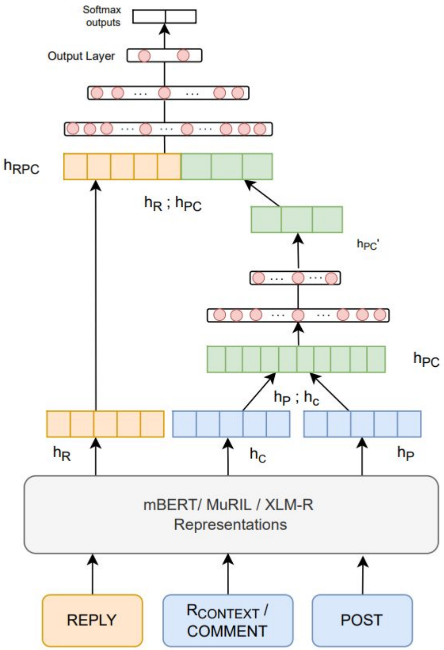


Rizwan, H., Shakeel, M. H., & Karim, A. (2020, November). Hate-speech and offensive language detection in roman Urdu. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 2512-2522).

Conversational Hate-Speech

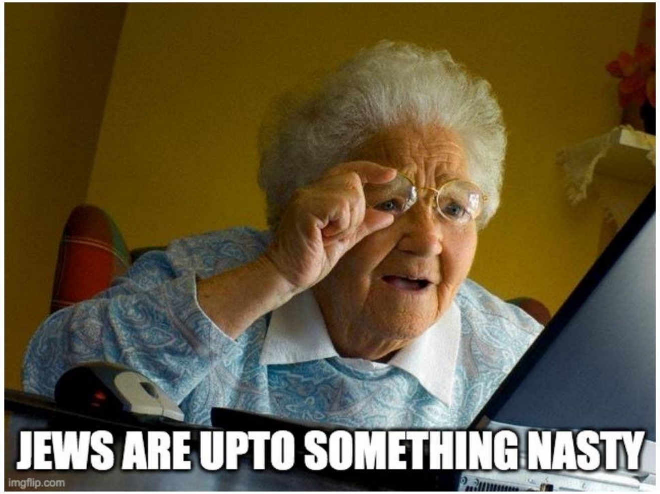


Model	Method	Accuracy					F1 Score				
		RF	LR	XGB	VC	Direct FT	RF	LR	XGB	VC	Direct FT
CM-XLMR	XLM-R + Norm	-	-	-	-	0.61	-	-	-	-	0.46
SCB	mBERT	0.55	0.61	0.49	0.57	0.56	0.55	0.60	0.57	0.49	0.50
	MuRIL	0.50	0.40	0.45	0.46	0.57	0.50	0.29	0.45	0.45	0.51
	XLM-R	0.55	0.58	0.52	0.58	0.40	0.54	0.49	0.50	0.53	0.27
WBC	mBERT	0.62	0.59	0.61	0.62	0.66	0.61	0.57	0.60	0.61	0.64
	MuRIL	0.59	0.41	0.54	0.53	0.40	0.55	0.29	0.52	0.53	0.29
	XLM-R	0.64	0.64	0.59	0.64	0.66	0.60	0.62	0.57	0.61	0.65
SLCB	mBERT	0.64	0.55	0.60	0.62	0.66	0.58	0.57	0.54	0.58	0.61
	MuRIL	0.64	0.60	0.55	0.62	0.62	0.57	0.56	0.54	0.57	0.55
	XLM-R	0.64	0.62	0.61	0.65	0.40	0.62	0.60	0.59	0.63	0.27
CAB	mBERT	0.57	0.58	0.55	0.58	0.58	0.57	0.58	0.55	0.58	0.53
	MuRIL	0.60	0.59	0.61	0.65	0.58	0.60	0.58	0.61	0.64	0.54
	XLM-R	0.62	0.64	0.59	0.64	0.63	0.61	0.64	0.59	0.64	0.60
Hierarchial	mBERT	0.54	0.58	0.60	0.62	0.60	0.52	0.54	0.56	0.62	0.65
	MuRIL	0.59	0.63	0.62	0.64	0.63	0.55	0.61	0.60	0.64	0.67
	XLM-R	0.63	0.61	0.64	0.66	**0.68**	0.62	0.60	0.62	0.63	**0.72**



Bagora, A., Shrestha, K., Maurya, K., & Desarkar, M. S. (2022, June). Hostility Detection in Online Hindi-English Code-Mixed Conversations. In 14th ACM Web Science Conference 2022 (pp. 390-400).

More Modalities



Multimodal Datasets

- **MMHS150K** is one of the largest dataset. image-text pair in hate speech research (Gomez,2019).
- **Hateful Memes** is another dataset of 10K+ posts created by Facebook AI. (Gomez,2019, Goswami.2021, Pramanick.2021)
- **TamilMemes** is another dataset of 2,969 memes created by Suryawanshi et al. for Tamil Troll meme classification (Suryawanshi,2020)
- Automated multimodal detection of online **antisemitism**.(Chandra.2021)
- **HarMeme** is another dataset consisting of 3,544 memes related to COVID-19.(Pramanick.2021)

Models

- **Text Based**
  - Glove, Fastext Embedding with Dense ANN layer
  - BERT, RoBERTa
- **Image Based model**
  - ResNet-152, VGG19, ResNeXt-101 etc.
- **Multimodal model**
  - ViLBERT CC, V-BERT COCO
  - VisualBERT, MMBT, UNITER, CLIP

Modality	Model	2-Class Classification					
		Acc ↑	P ↑	R ↑	F1 ↑	MAE ↓	MMAE ↓
	Human†	90.68	84.35	84.19	83.55	0.1760	0.1723
	Majority	64.76	32.38	50.00	39.30	0.3524	0.5000
Text Only	TextBERT	70.17	65.96	66.38	66.25	0.3173	0.2911
Image Only	VGG19	68.12	60.25	61.23	61.86	0.3204	0.3190
	DenseNet-161	68.42	61.08	62.10	62.54	0.3202	0.3125
	ResNet-152	68.74	61.86	62.89	62.97	0.3188	0.3114
	ResNeXt-101	69.79	62.32	63.26	63.68	0.3175	0.3029
Image + Text (Unimodal Pre-training)	Late Fusion	73.24	70.28	70.36	70.25	0.3167	0.2927
	Concat BERT	71.82	71.58	72.23	71.82	0.3033	0.3156
	MMBT	73.48	68.89	68.95	67.12	0.3101	0.3258
Image + Text (Multimodal Pre-training)	ViLBERT CC	78.53	78.62	**81.41**	78.06	0.2279	0.1881
	V-BERT COCO	**81.36**	**79.55**	81.19	**80.13**	**0.1972**	**0.1857**

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

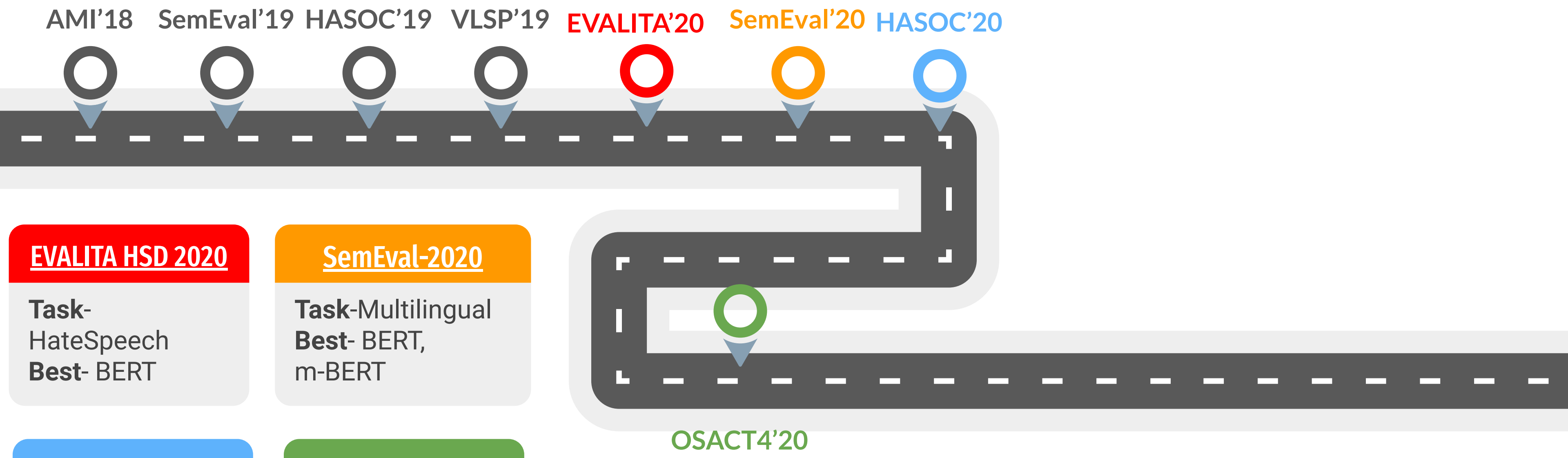
HASOC 2019

Task- Hate/Offensive
Best- Ensemble

VLSP HSD 2019

Task- Hate Speech
Best- LR + ngram

Shared tasks timeline



EVALITA HSD 2020

Task- HateSpeech
Best- BERT

SemEval-2020

Task-Multilingual
Best- BERT, m-BERT

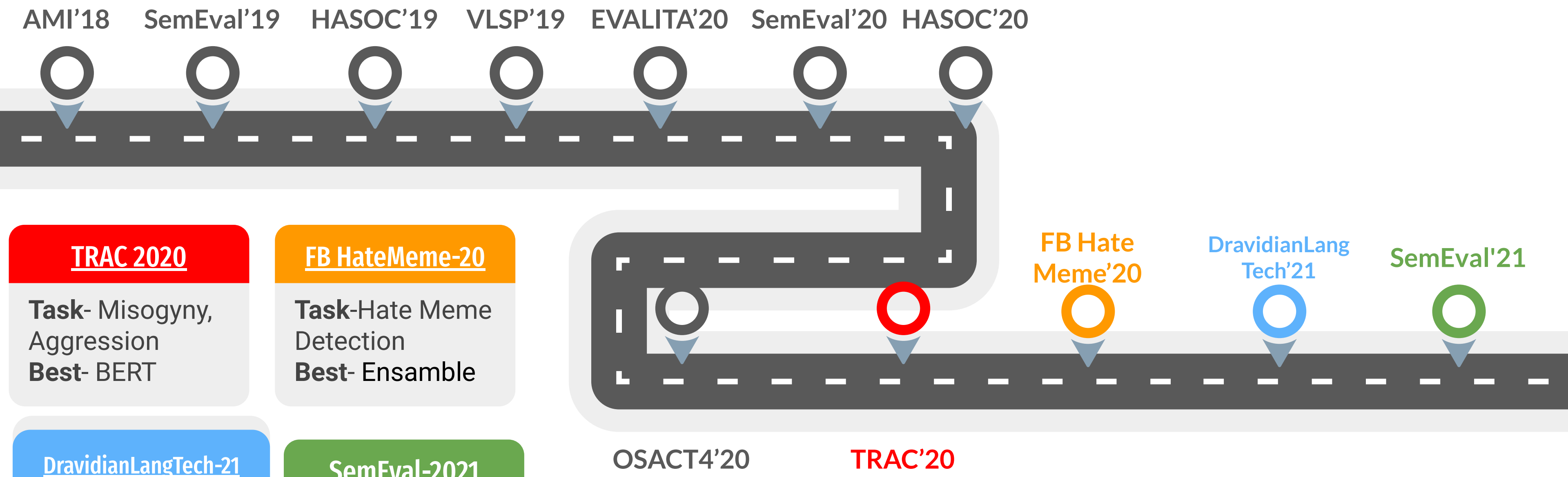
HASOC 2020

Task- Multilingual
Best- CNN, BERT

OSACT4 HSD 2020

Task- Arabic Hate Speech
Best- CNN

Shared tasks timeline



TRAC 2020

Task- Misogyny, Aggression
Best- BERT

FB HateMeme-20

Task-Hate Meme Detection
Best- Ensamble

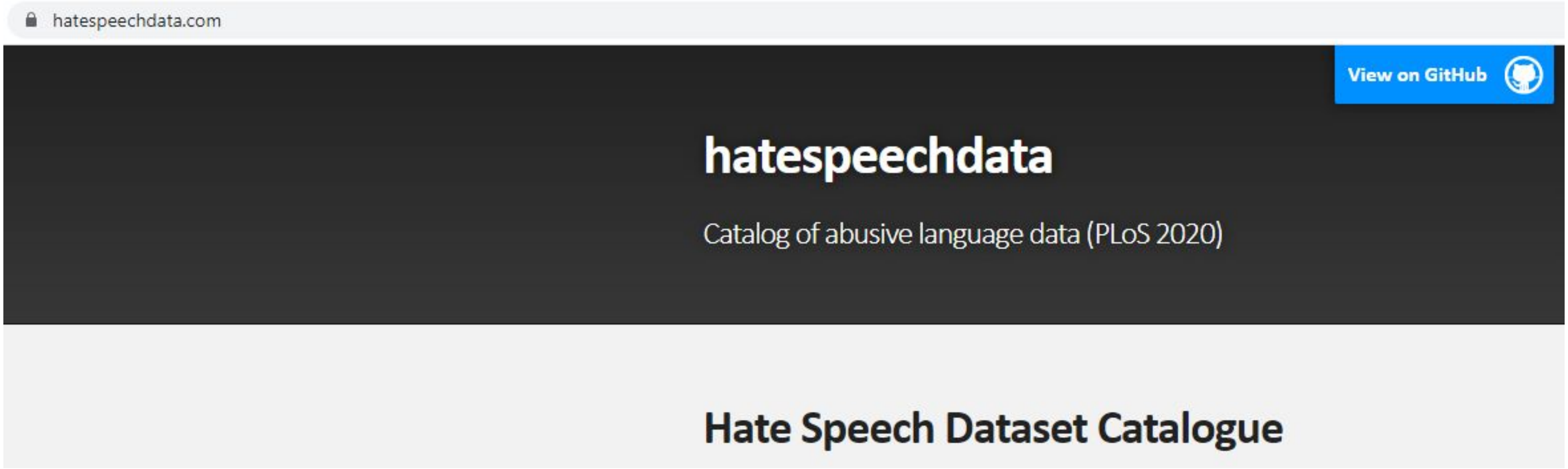
DravidianLangTech-21

Task-Multilingual
Best- BERT-ensemble

SemEval-2021

Task- Toxic Span Detection
Best- BERT

Where to find dataset?



https://hatespeechdata.com/

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)

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	$\boxed{93.1}$
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	$\boxed{73.1}$
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

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

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation. *SIGIR'19*

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.

Method	Class	Prec.	Rec.	F1
Badjatiya 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

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

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. *Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation. SIGIR'19*

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**

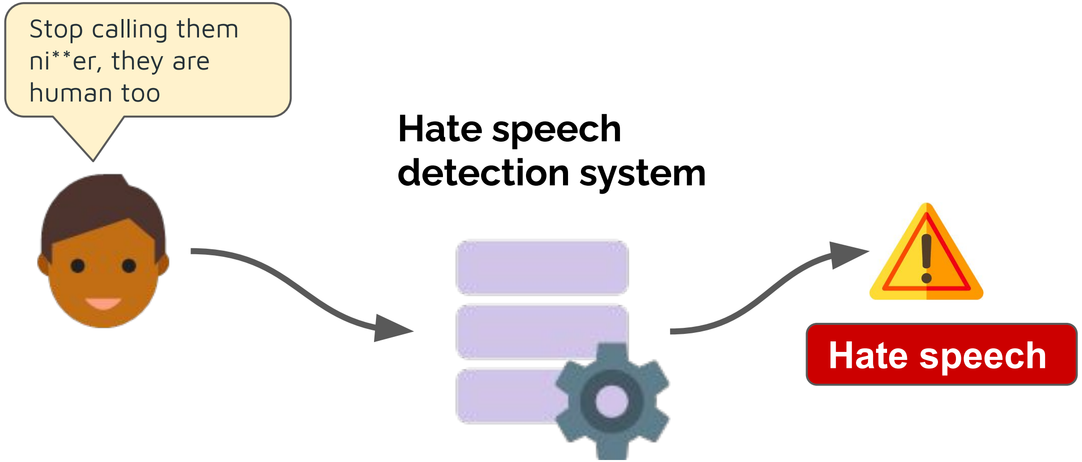
	Functionality	Example	Gold Label	n	Accuracy (%) DB-D	DB-F	PERS
Group identifier	**F18**: Neutral statement using protected group identifier	"We are a group of [IDENTITY]." "I live with two [IDENTITY]."	non-hateful	126	61.1	76.2	**84.1**
	F19: Positive statement using protected group identifier	"I love [IDENTITY]." "[IDENTITY] are great."	non-hateful	189	**86.2**	79.9	54.0

Explainability of Models

Explainability

Rationales | Free text

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

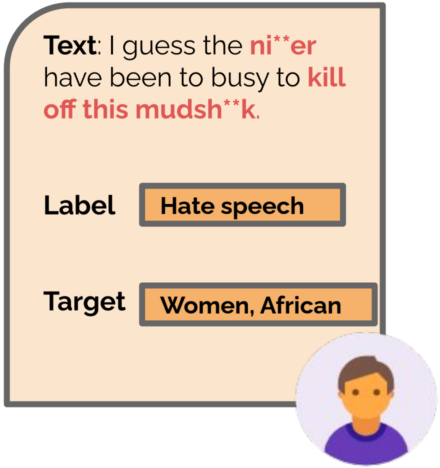


Explainability of Models

- **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
BERT-HateXplain	**0.698**	**0.687**	**0.851**

Models performance is better !

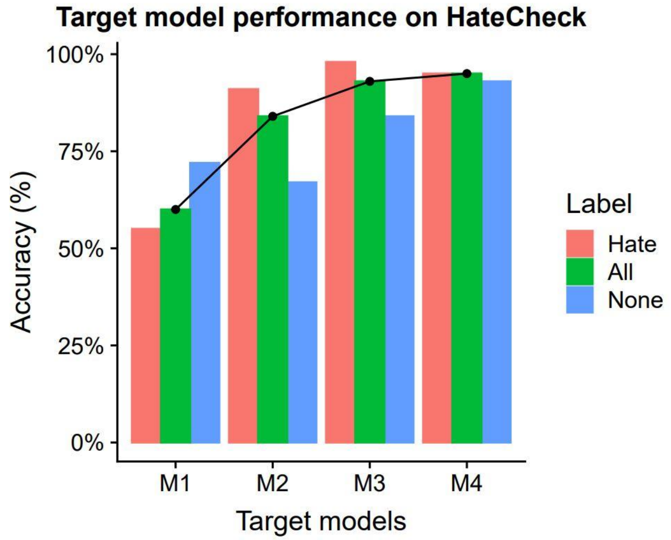


Click logo for demo

Building more Robust Models

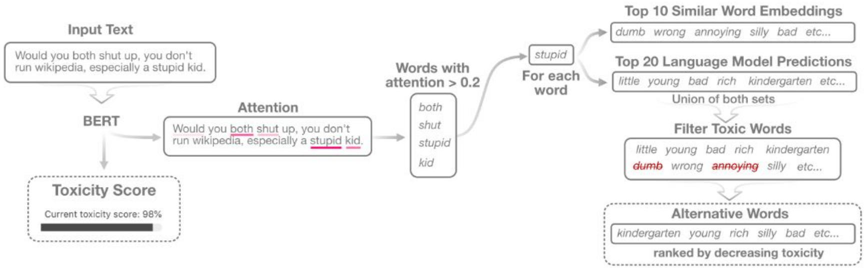
- A human-and model-in-the-loop process for training online hate detection models. (Vidgen.2021)

Round	Total	Not	Hate
R1	54.7%	64.6%	49.2%
R2	34.3%	38.9%	29.7%
R3	27.8%	20.5%	35.1%
R4	27.7%	23.7%	31.7%



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)]

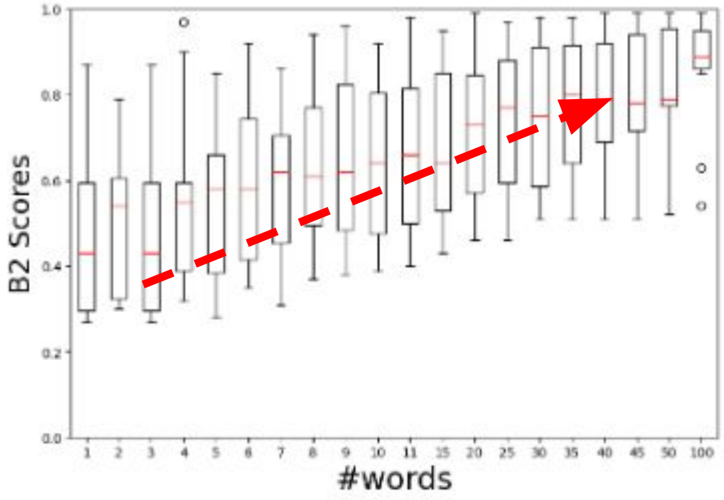


Advantage - reduce toxicity, way of debugging model
Disadvantage - malicious users might game the system.

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

No of topics kept fixed at 8



(b) $\mathbf{B_2}$ variations per number of words.

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 (crowdsource) [(Waseem,2016)]

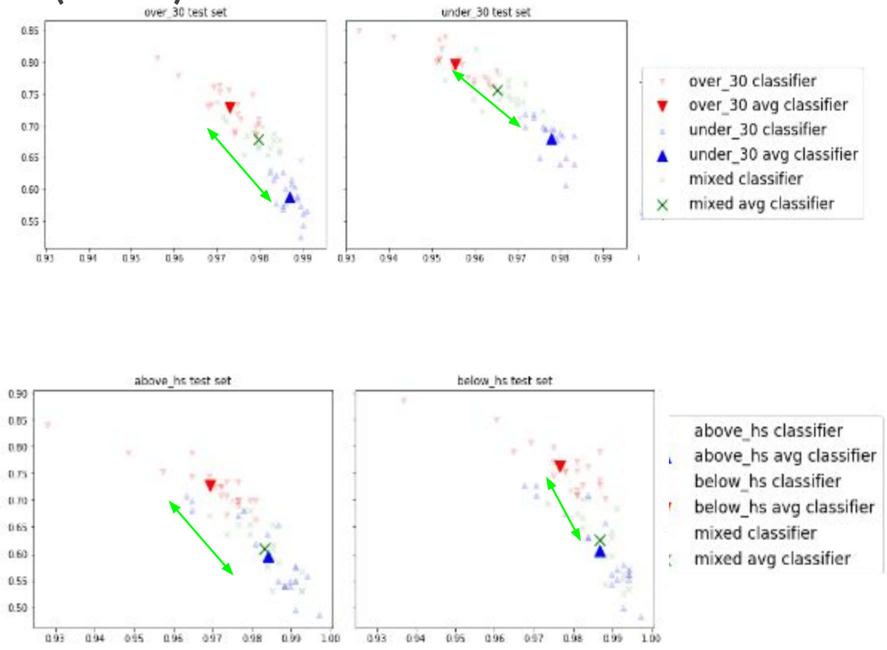
Feature Set	Amateur F1	Amateur Recall	Amateur Precision	Expert F1	Expert Recall	Expert Precision
Close	86.39	88.60%	87.59%	91.24	92.49%	92.67%
Middling	84.07	86.76%	85.43%	87.81	90.10%	88.53%
Distant	71.71	80.17%	82.05%	77.77	84.76%	71.85%
All	86.39	88.60%	87.59%	90.77	92.20%	92.23%
Best	83.88	86.68%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

Community not annotated

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

Table 2: **Experiment 1**

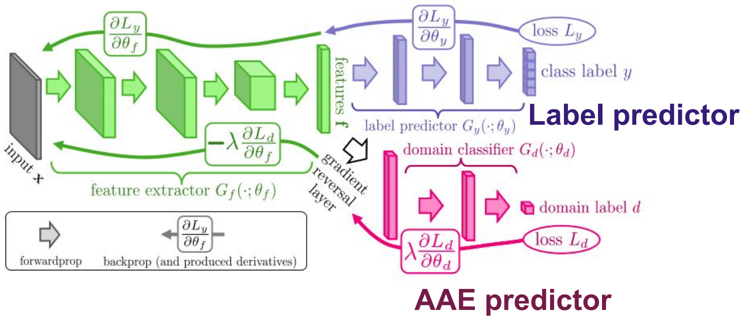
Values greater than 1 indicate that black-aligned tweets are classified as belonging to class at a higher rate than white

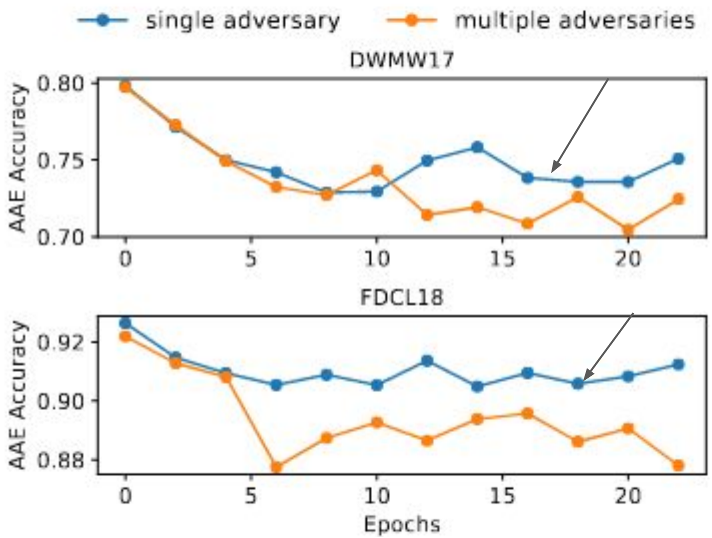
Dataset and model used for dialect identification (Blodgett,2016)

Community not annotated

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)](#).





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

Community annotated

Bias in Data/Models

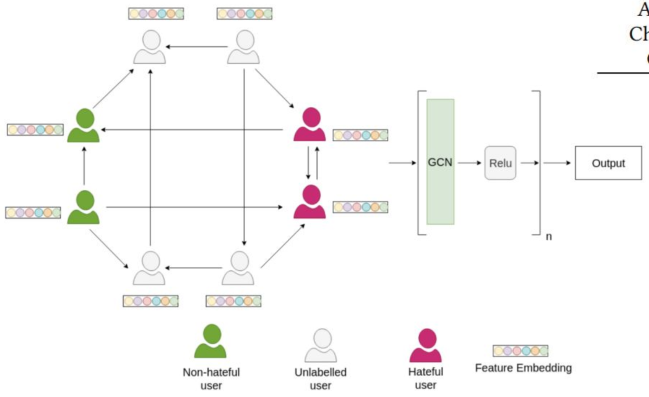
- Bias from different directions
  - How is data selected ?
  - Who is the annotator?
  - Who is the **speaker/target** ?
- Often hate speech model can detect false positives for tweets written by different community (Davidson,2019).
- Training with adversarial loss can help reduce the bias (Xia,2020).
- 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
BERT-HateXplain	**0.807**	**0.745**	**0.763**

Models less biased !

Hateful Users Detection

Method	Inputs	Gab						Twitter					
		5%	10%	15%	20%	50%	80%	5%	10%	15%	20%	50%	80%
fastText	Y, X_L	0.492	0.537	0.571	0.603	0.690	0.709	0.624	0.634	0.648	0.651	0.670	0.676
Glove	Y, X_L	0.695	0.720	0.745	0.750	0.778	0.784	0.650	0.666	0.674	0.681	0.691	0.695
LSTM	Y, X_L	0.579	0.600	0.605	0.608	0.622	0.645	0.514	0.487	0.567	0.564	0.592	0.608
Doc2vec	Y, X_L	0.733	0.767	0.783	0.779	0.779	0.781	0.715	0.715	0.719	0.729	0.749	0.758
BERT	Y, X_L	0.631	0.660	0.682	0.701	0.740	0.764	0.603	0.665	0.690	0.709	0.729	0.740
TSVM	Y, X	0.686	0.704	0.712	0.712	0.739	0.753	0.480	0.520	0.533	0.533	0.585	0.611
DeepWalk	Y, G	0.652	0.676	0.700	0.713	0.723	0.734	0.757	0.764	0.767	0.767	0.773	0.779
Node2vec	Y, G	0.647	0.672	0.695	0.704	0.725	0.744	0.692	0.720	0.732	0.734	0.749	0.748
GraphSAGE	Y, X, G	0.778	**0.808**	0.806	0.811	0.827	0.828	0.762	0.773	0.774	0.780	0.782	0.777
GCN	Y, X, G	0.721	0.735	0.730	0.738	0.751	0.758	0.756	0.759	0.767	0.773	0.776	0.770
AGNN	Y, X, G	**0.791**	0.796	**0.818**	**0.824**	**0.830**	**0.833**	**0.780**	**0.785**	**0.785**	**0.790**	0.786	**0.787**
ARMA	Y, X, G	0.765	0.778	0.783	0.797	0.809	0.805	0.757	0.760	0.761	0.762	0.770	0.769
ChebNet	Y, X, G	0.778	0.802	0.796	0.798	0.805	0.812	0.746	0.750	0.754	0.762	0.761	0.766
GAT	Y, X, G	0.683	0.718	0.725	0.726	0.745	0.758	0.757	0.774	0.781	0.777	**0.787**	0.782

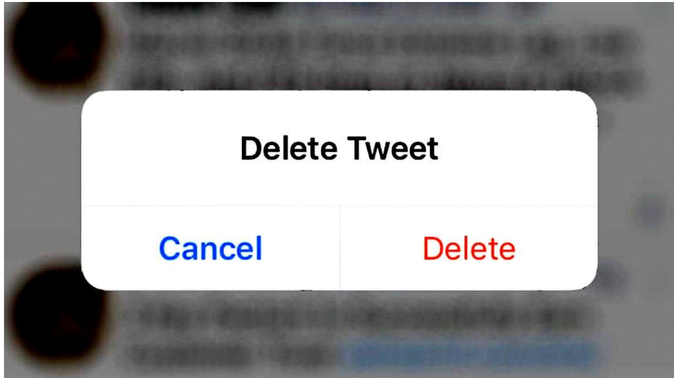


Textual and network features together can improve the performance of hateful users detection.

Mitigating Hate Speech

What is done after detecting hate speech?

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

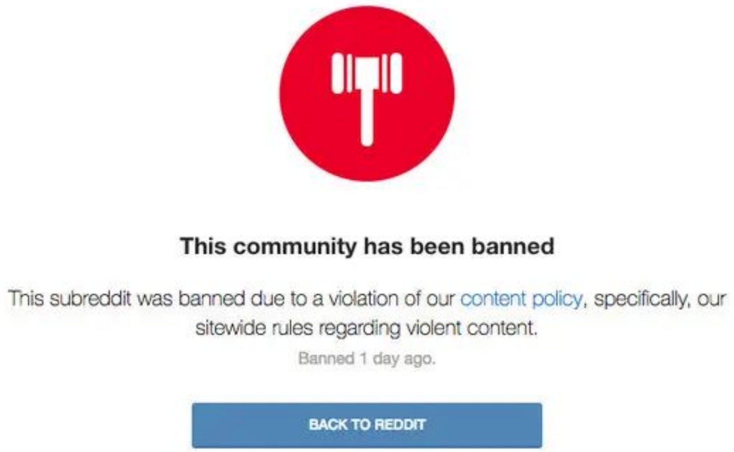


Is banning effective?

Is banning effective?

Case study of Reddit[2015]

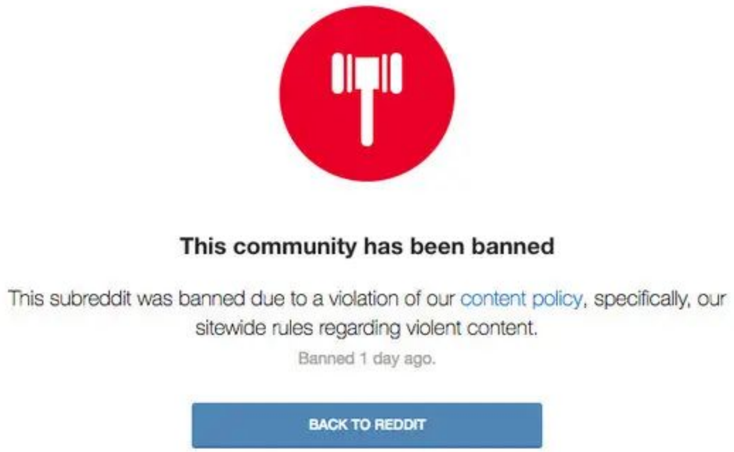
- In 2015, Reddit closed several subreddits due to violations of Reddit's anti-harassment policy.
- Foremost among them were r/fatpeoplehate and r/CoonTown
- How effective was the ban?



Is banning effective ?

Case study of Reddit[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?



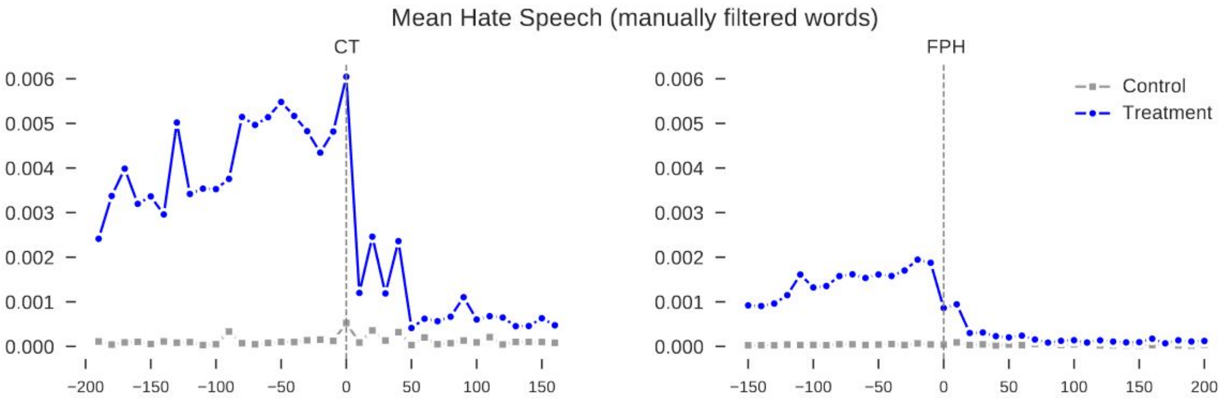
***You Can't Stay Here*: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech [Chandrasekharan 2017]**

The Efficacy of Reddit's 2015 Ban

- User-level - Following Reddit's 2015 ban, a large, significant percentage of users from banned communities left Reddit. Others migrated to other sub-reddits where hate was prominent

The Efficacy of Reddit's 2015 Ban

- **User-level -** Following Reddit's 2015 ban, a large, significant percentage of users from banned communities **left Reddit.** Others migrated to other sub-reddits where hate was prominent
- **Community-level -** 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".**

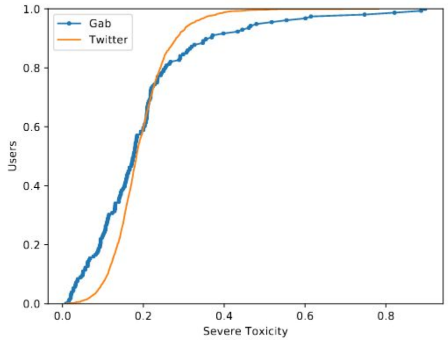


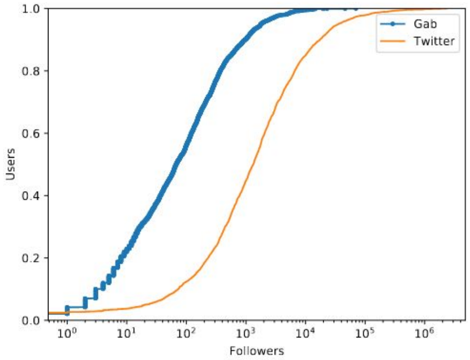
What about the users who left?

What about the users who left ?

Users who get banned on Twitter/Reddit exhibit an **increased level** of activity and toxicity on Gab, although the **audience** they potentially reach **decreases**

***Understanding the Effect of Deplatforming on Social Networks* [[Ali 2021]]**





Are there any alternatives?

One of the alternatives

Counterspeech: Directly intervening with textual response that counter the hate-content.

Taxonomy of counterspeech Benesch 2016

1. Presenting facts to correct misstatements or mis-perceptions
2. Pointing out hypocrisy or contradictions
3. Affiliation
4. Visual Communication
5. Humor and sarcasm
6. Denouncing hateful or dangerous speech
7. Tone

Taxonomy of counterspeech Benesch 2016

1. Presenting facts to correct misstatements or mis-perceptions
2. Pointing out hypocrisy or contradictions
3. **Affiliation**
4. Visual Communication
5. Humor and sarcasm
6. Denouncing hateful or dangerous speech
7. Tone

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

1. Presenting facts to correct misstatements or mis-perceptions
2. Pointing out hypocrisy or contradictions
3. Affiliation
4. **Visual Communication**
5. Humor and sarcasm
6. Denouncing hateful or dangerous speech
7. Tone

Taxonomy of counterspeech Benesch 2016

1. Presenting facts to correct misstatements or mis-perceptions
2. Pointing out hypocrisy or contradictions
3. Affiliation
4. Visual Communication
5. Humor and sarcasm
6. Denouncing hateful or dangerous speech
7. Tone

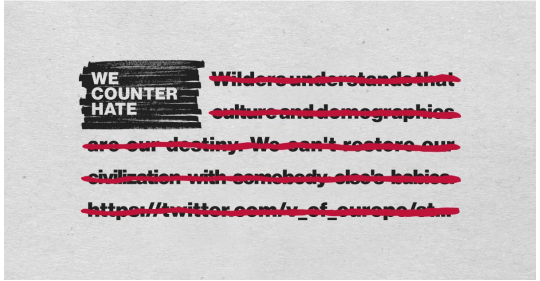
Taxonomy of counterspeech Benesch 2016

1. Presenting facts to correct misstatements or mis-perceptions
2. Pointing out hypocrisy or contradictions
3. Affiliation
4. Visual Communication
5. Humor and sarcasm
6. Denouncing hateful or dangerous speech
7. **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"

Mitigation of harmful speech

Counterspeech: Directly intervening with textual response that counter the hate-content.



WeCounterHate

NoHateSpeechMovement

Mitigation of harmful speech

Counterspeech: Directly intervening with textual response that counter the hate-content.

Why counter speech?

- Suspension/removal of posts is a threat to doctrine of free speech.
- Can act as a first line of response before other intervention techniques

Mitigation of harmful speech

Counterspeech: Directly intervening with textual response that counter the hate-content.

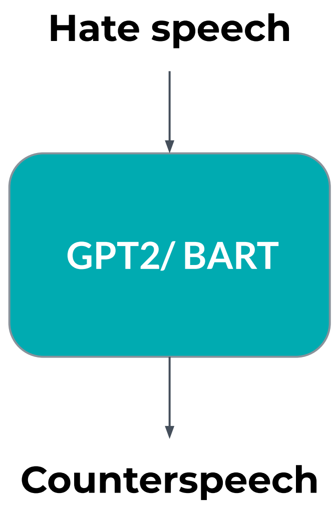
Why counter speech?

- Suspension/removal of posts is a threat to doctrine of free speech.
- Can act as a first line of response before other intervention techniques

Adds to the challenges of content moderation. Can we use NLGs to help the moderators?

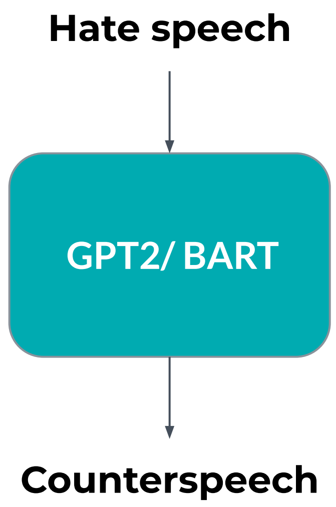
Generation of counterspeech

- A **response generation** problem
- **Research challenges**
  - Quality and diverse dataset
  - Building the generation framework



Generation of counterspeech

- A **response generation** problem
- **Research challenges**
  - **Quality and diverse dataset**
  - Building the generation framework



Counterspeech datasets

Dataset	Annotators	Unique hate speech	Source of hate	Target Community
Qian et al., '19	Crowdworkers	3,847	**REDDIT**	**Mixed**
		11,169	**GAB**	**Mixed**
Chung et al. '19	Expert Annotators	408	**SYNTHETIC**	**Muslims**

[8] Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech. arXiv preprint arXiv:1909.04251.
[9] Chung, Y. L., Kuzmenko, E., Tekiroglu, S. S., & Guerini, M. (2019). CONAN--COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. arXiv preprint arXiv:1910.03270.

Counterspeech datasets

Dataset	Annotators	Unique hate speech	Source of hate	Target Community
Qian et al., '19	Crowdworkers	3,847	**REDDIT**	**Mixed**
		11,169	**GAB**	**Mixed**
Chung et al. '19	Expert Annotators	408	**SYNTHETIC**	**Muslims**

Crowdworkers generally write simple content like - Don't say that slur word

Counterspeech datasets

Dataset	Annotators	Unique hate speech	Source of hate	Target Community
Qian et al., '19	Crowdworkers	3,847	REDDIT	Mixed
		11,169	GAB	Mixed
Chung et al. '19	Expert Annotators	408	SYNTHETIC	Muslims

Synthetic hate speech may not represent the real world hate speech

Counterspeech datasets

Dataset	Annotators	Unique hate speech	Source of hate	Target Community
Qian et al., '19	Crowdworkers	3,847	**REDDIT**	**Mixed**
		11,169	**GAB**	**Mixed**
Chung et al. '19	Expert Annotators	408	**SYNTHETIC**	**Muslims**

Cannot scale the dataset with **expert annotators.**

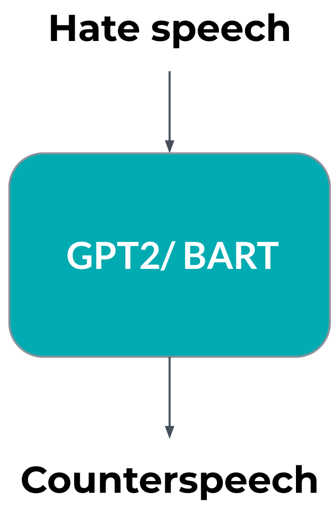
Counterspeech datasets

Dataset	Annotators	Unique hate speech	Source of hate	Target Community
Qian et al., '19	Crowdworkers	3,847	**REDDIT**	**Mixed**
		11,169	**GAB**	**Mixed**
Chung et al. '19	Expert Annotators	408	**SYNTHETIC**	**Muslims**

Challenge: How to build a **quality** and **diverse** counterspeech dataset at **scale**?

Generation of counterspeech

- A response generation problem
- Research challenges
  - Quality and diverse dataset
  - **Building the generation framework**



Generation models

VAE - RNN

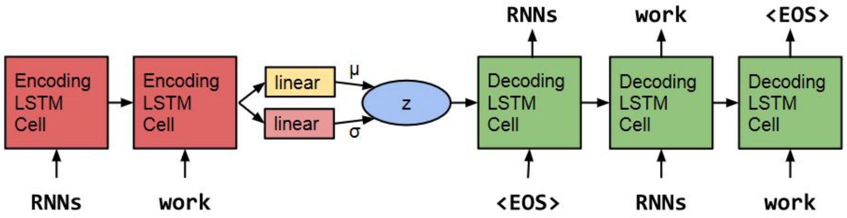
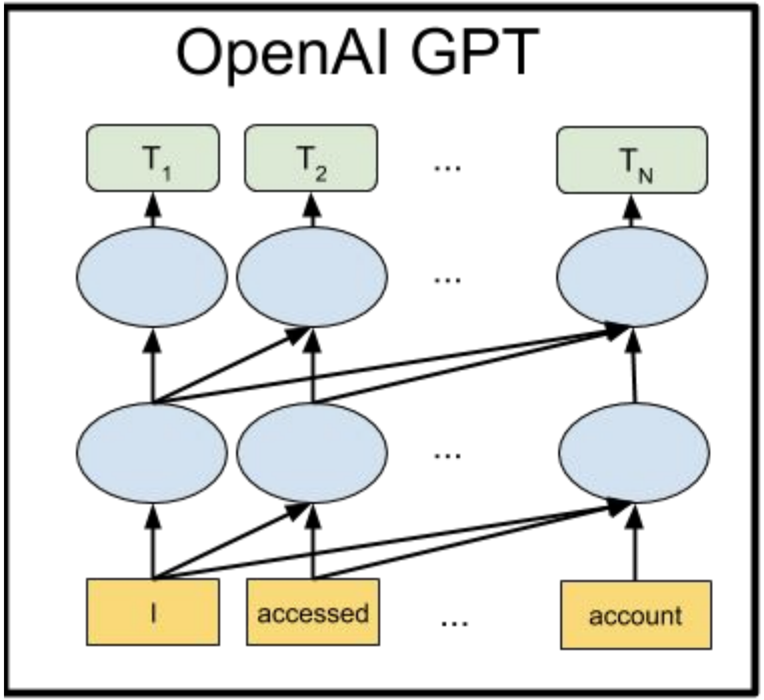


Figure 1: The core structure of our variational autoencoder language model. Words are represented using a learned dictionary of embedding vectors.

Generation models



Different generation models and decoding methods

Models
- BERT
- GPT-2
- DialoGPT
- BART
- T5

Decoding methodologies
- Beamsearch
- Top-p
- Top-k
- Combining Top-p and Top-k

[10] Tekiroğlu, Serra Sinem, et al. "Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: a Comparative Study." Findings of the Association for Computational Linguistics: ACL 2022. 2022

Automatic evaluation

	Overlap				Diversity		Toxicity
	ROU	**B-1**	**B-3**	**B-4**	**RR**	**NOV**	**-**
BART	0.268	0.277	0.085	**0.051**	20.722	0.560	0.420
BERT	0.237	0.277	0.073	0.037	24.747	0.605	0.406
T5	**0.274**	**0.302**	0.083	0.042	8.548	**0.655**	0.359
DialoGPT	**0.273**	**0.304**	**0.093**	**0.052**	**8.248**	0.643	**0.343**
GPT-2	0.264	0.297	**0.088**	0.050	**7.736**	**0.653**	**0.342**

Models

DialogGPT is better

Decoding methods

	Overlap				Diversity		Toxicity
	ROU	**B-1**	**B-3**	**B-4**	**RR**	**NOV**	**-**
BS	**0.287**	0.299	0.096	0.059	21.579	0.561	0.398
Top_{pk}	**0.287**	**0.320**	**0.106**	0.059	11.404	0.639	**0.352**
Top_k	0.282	0.314	**0.106**	**0.060**	**10.076**	**0.652**	**0.374**
Top_p	0.285	**0.319**	0.105	**0.060**	**11.270**	**0.640**	0.381

[10] Tekiroğlu, Serra Sinem, et al. "Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: a Comparative Study." Findings of the Association for Computational Linguistics: ACL 2022. 2022

Automatic evaluation

	Overlap				Diversity		Toxicity
	ROU	B-1	B-3	B-4	RR	NOV	-
BART	0.268	0.277	0.085	**0.051**	20.722	0.560	0.420
BERT	0.237	0.277	0.073	0.037	24.747	0.605	0.406
T5	**0.274**	**0.302**	0.083	0.042	8.548	**0.655**	0.359
DialoGPT	**0.273**	**0.304**	**0.093**	**0.052**	**8.248**	0.643	**0.343**
GPT-2	0.264	0.297	**0.088**	0.050	**7.736**	**0.653**	**0.342**

Models

Decoding methods

BS is repetitive

	Overlap				Diversity		Toxicity
	ROU	B-1	B-3	B-4	RR	NOV	-
BS	**0.287**	0.299	0.096	0.059	21.579	0.561	0.398
Top_{pk}	**0.287**	**0.320**	**0.106**	0.059	11.404	0.639	**0.352**
Top_k	0.282	0.314	**0.106**	**0.060**	**10.076**	**0.652**	**0.374**
Top_p	0.285	**0.319**	0.105	**0.060**	**11.270**	**0.640**	0.381

[10] Tekiroğlu, Serra Sinem, et al. "Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: a Comparative Study." Findings of the Association for Computational Linguistics: ACL 2022. 2022

Automatic evaluation

	Overlap				Diversity		Toxicity
	ROU	B-1	B-3	B-4	RR	NOV	-
BART	0.268	0.277	0.085	**0.051**	20.722	0.560	0.420
BERT	0.237	0.277	0.073	0.037	24.747	0.605	0.406
T5	**0.274**	**0.302**	0.083	0.042	8.548	**0.655**	0.359
DialoGPT	**0.273**	**0.304**	**0.093**	**0.052**	**8.248**	0.643	**0.343**
GPT-2	0.264	0.297	**0.088**	0.050	**7.736**	**0.653**	**0.342**

Models

	Overlap				Diversity		Toxicity
	ROU	B-1	B-3	B-4	RR	NOV	-
BS	**0.287**	0.299	0.096	0.059	21.579	0.561	0.398
Top_{pk}	**0.287**	**0.320**	**0.106**	0.059	11.404	0.639	**0.352**
Top_k	0.282	0.314	**0.106**	**0.060**	**10.076**	**0.652**	**0.374**
Top_p	0.285	**0.319**	0.105	**0.060**	**11.270**	**0.640**	0.381

Decoding methods

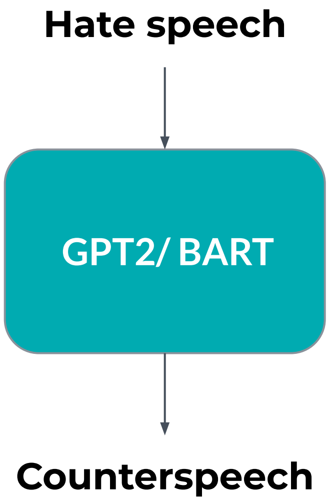
BS is repetitive

Stochastic methods are novel

[10] Tekiroğlu, Serra Sinem, et al. "Using Pre-Trained Language Models for Producing Counter Narratives Against Hate Speech: a Comparative Study." Findings of the Association for Computational Linguistics: ACL 2022. 2022

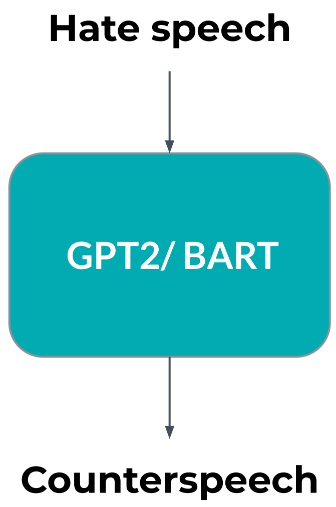
Generation of counterspeech

- **A response generation problem**
- **Research challenges**
    - Quality and diverse dataset
    - Building the generation framework
    - **How can we improve ?**
        - Improve the generation
        - Hallucination
        - Personalisation

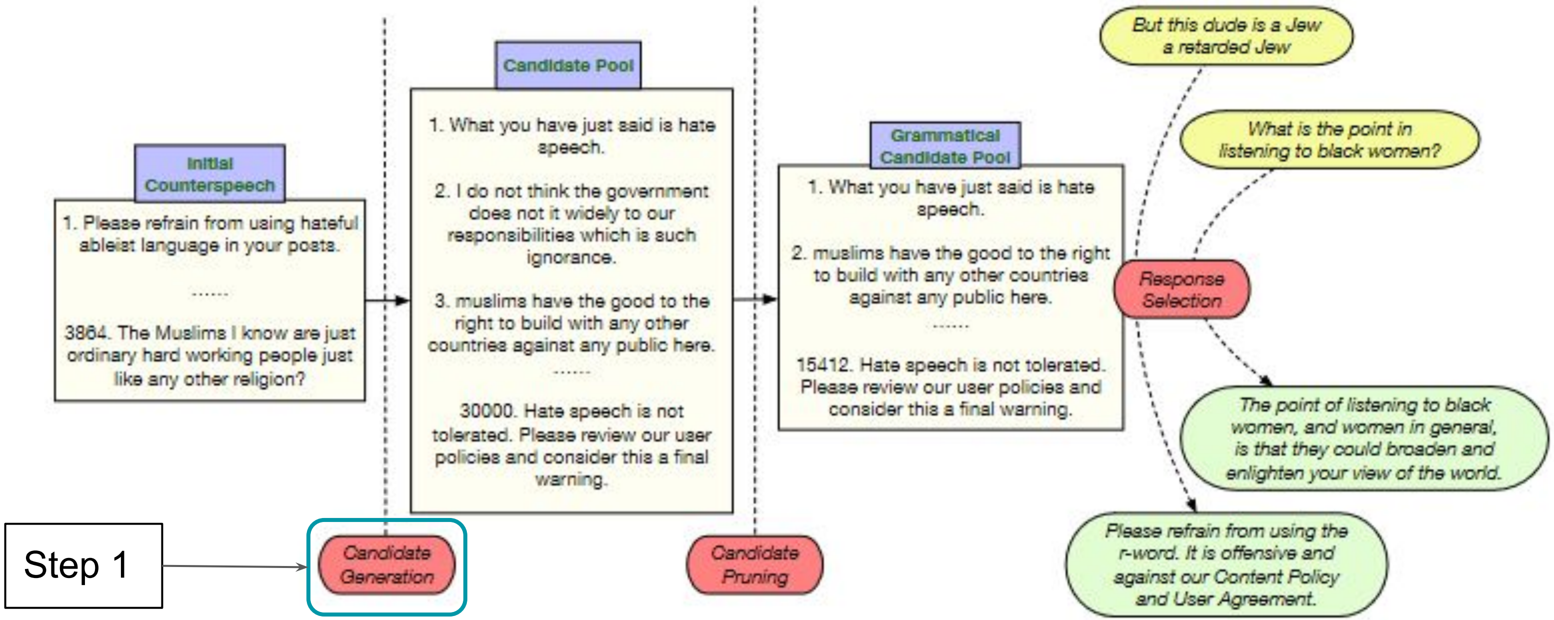


Generation of counterspeech

- A **response generation** problem
- **Research challenges**
    - Quality and diverse dataset
    - Building the generation framework
    - **How can we improve ?**
        - **Improve the generation** - select better counter speech.
        - Hallucination
        - Personalisation

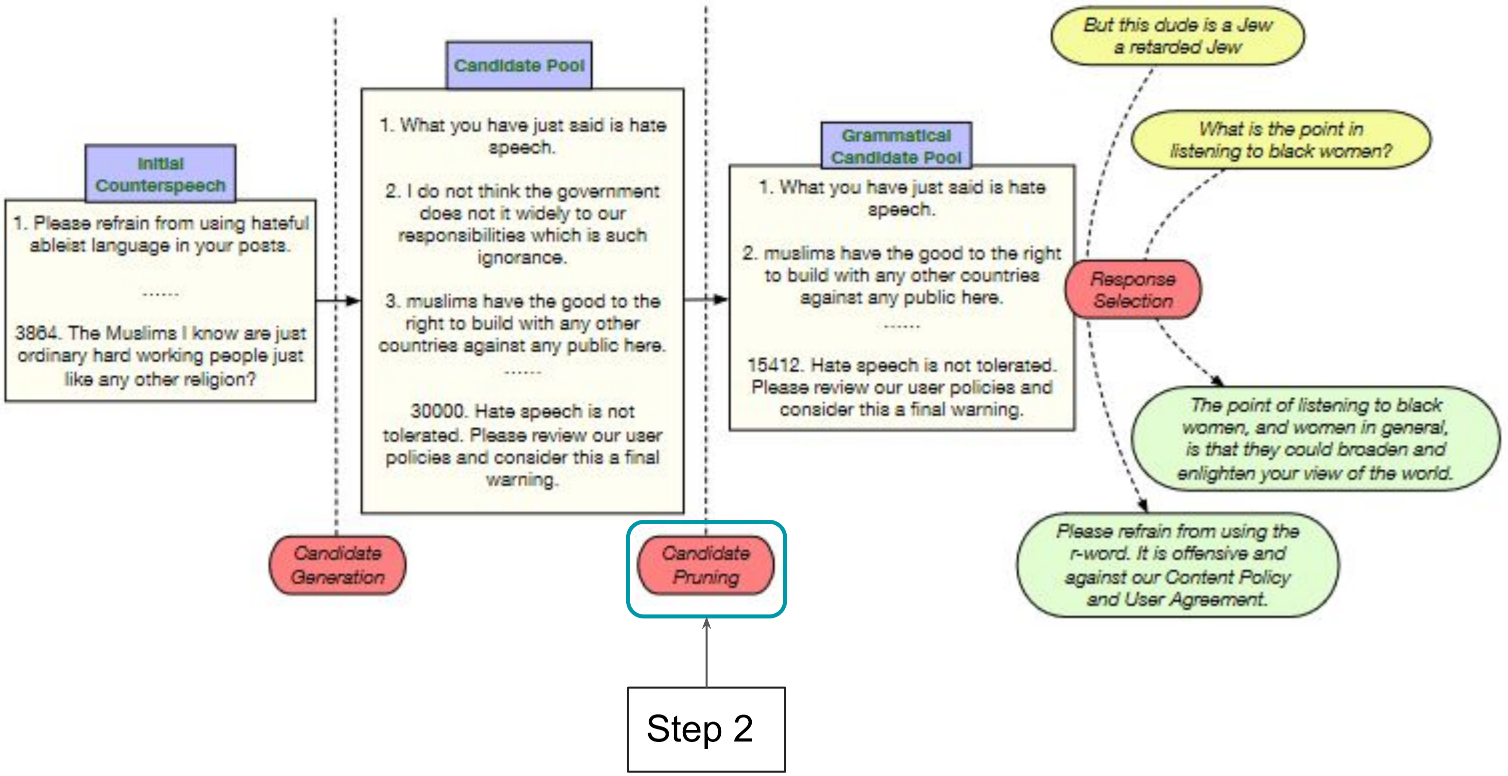


Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech [Zhu 2021]

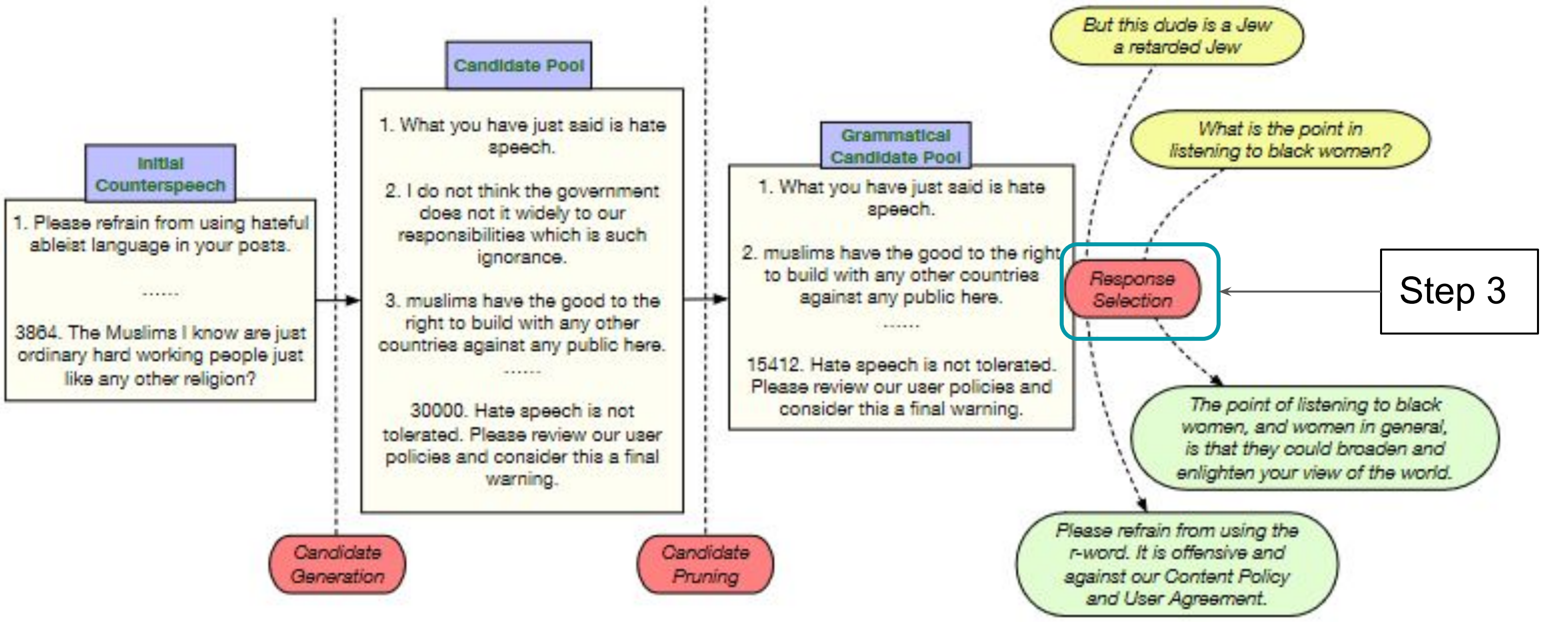


VAE - RNN

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]

		Diversity						Relevance					LQ.
		Dist-1	Dist-2	Ent-1	Ent-2	SB1*	SB2*	B2	R2	MS	BS	BM25	GR
CONAN	Seq2Seq	0.06	0.23	5.12	6.63	0.54	0.30	3.4	3.0	4.4	0.83	2.66	0.38
	MMI	0.06	0.23	4.88	6.41	0.57	0.35	2.9	2.3	3.9	0.82	1.63	0.33
	SpaceFusion	0.00	0.00	1.06	1.86	0.98	0.98	0.0	0.0	-14.2	0.76	0.12	0.38
	BART	0.04	0.23	5.98	7.80	0.52	0.26	3.9	3.6	7.1	0.84	1.86	0.71
	GPS	0.06	0.27	5.77	7.41	0.43	0.19	7.1	6.5	10.9	0.85	5.43	0.71
Reddit	Seq2Seq	0.04	0.24	5.07	6.61	0.58	0.31	6.5	4.0	6.8	0.85	0.14	0.64
	MMI	0.05	0.32	5.11	6.76	0.56	0.29	6.4	4.0	6.9	0.85	0.14	0.56
	SpaceFusion	0.00	0.02	2.73	4.16	0.87	0.76	0.9	0.0	-2.5	0.79	0.16	0.26
	BART	0.03	0.19	5.08	6.63	0.69	0.55	7.8	6.9	7.8	0.86	0.83	0.72
	GPS	0.09	0.53	5.74	7.61	0.41	0.15	8.1	7.1	7.8	0.87	2.58	0.75
Gab	Seq2Seq	0.02	0.17	5.14	6.71	0.56	0.30	7.5	5.0	6.7	0.86	0.14	0.67
	MMI	0.02	0.17	5.28	6.82	0.55	0.30	5.8	3.6	6.2	0.85	0.18	0.65
	SpaceFusion	0.00	0.01	3.72	4.84	0.81	0.73	1.8	0.1	0.0	0.82	0.17	0.21
	BART	0.03	0.17	5.42	7.25	0.60	0.38	6.9	6.4	6.8	0.86	0.81	0.72
	GPS	0.06	0.40	5.82	7.83	0.39	0.15	7.6	6.4	6.8	0.87	1.94	0.76

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

		Diversity						Relevance					LQ.
		Dist-1	Dist-2	Ent-1	Ent-2	SB1*	SB2*	B2	R2	MS	BS	BM25	GR
CONAN	Seq2Seq	**0.06**	0.23	5.12	6.63	0.54	0.30	3.4	3.0	4.4	0.83	2.66	0.38
	MMI	**0.06**	0.23	4.88	6.41	0.57	0.35	2.9	2.3	3.9	0.82	1.63	0.33
	SpaceFusion	0.00	0.00	1.06	1.86	0.98	0.98	0.0	0.0	-14.2	0.76	0.12	0.38
	BART	0.04	0.23	**5.98**	**7.80**	0.52	0.26	3.9	3.6	7.1	0.84	1.86	**0.71**
	GPS	**0.06**	**0.27**	5.77	7.41	**0.43**	**0.19**	**7.1**	**6.5**	**10.9**	**0.85**	**5.43**	**0.71**
Reddit	Seq2Seq	0.04	0.24	5.07	6.61	0.58	0.31	6.5	4.0	6.8	0.85	0.14	0.64
	MMI	0.05	0.32	5.11	6.76	0.56	0.29	6.4	4.0	6.9	0.85	0.14	0.56
	SpaceFusion	0.00	0.02	2.73	4.16	0.87	0.76	0.9	0.0	-2.5	0.79	0.16	0.26
	BART	0.03	0.19	5.08	6.63	0.69	0.55	7.8	6.9	**7.8**	0.86	0.83	0.72
	GPS	**0.09**	**0.53**	**5.74**	**7.61**	**0.41**	**0.15**	**8.1**	**7.1**	**7.8**	**0.87**	**2.58**	**0.75**
Gab	Seq2Seq	0.02	0.17	5.14	6.71	0.56	0.30	7.5	5.0	6.7	0.86	0.14	0.67
	MMI	0.02	0.17	5.28	6.82	0.55	0.30	5.8	3.6	6.2	0.85	0.18	0.65
	SpaceFusion	0.00	0.01	3.72	4.84	0.81	0.73	1.8	0.1	0.0	0.82	0.17	0.21
	BART	0.03	0.17	5.42	7.25	0.60	0.38	6.9	**6.4**	**6.8**	0.86	0.81	0.72
	GPS	**0.06**	**0.40**	**5.82**	**7.83**	**0.39**	**0.15**	**7.6**	**6.4**	**6.8**	**0.87**	**1.94**	**0.76**

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

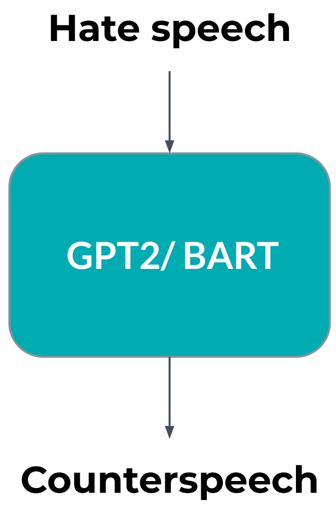
		Diversity						Relevance					LQ.
		Dist-1	Dist-2	Ent-1	Ent-2	SB1*	SB2*	B2	R2	MS	BS	BM25	GR
CONAN	Seq2Seq	**0.06**	0.23	5.12	6.63	0.54	0.30	3.4	3.0	4.4	0.83	2.66	0.38
	MMI	**0.06**	0.23	4.88	6.41	0.57	0.35	2.9	2.3	3.9	0.82	1.63	0.33
	SpaceFusion	0.00	0.00	1.06	1.86	0.98	0.98	0.0	0.0	-14.2	0.76	0.12	0.38
	BART	0.04	0.23	**5.98**	**7.80**	0.52	0.26	3.9	3.6	7.1	0.84	1.86	**0.71**
	GPS	**0.06**	**0.27**	5.77	7.41	**0.43**	**0.19**	**7.1**	**6.5**	**10.9**	**0.85**	**5.43**	**0.71**
Reddit	Seq2Seq	0.04	0.24	5.07	6.61	0.58	0.31	6.5	4.0	6.8	0.85	0.14	0.64
	MMI	0.05	0.32	5.11	6.76	0.56	0.29	6.4	4.0	6.9	0.85	0.14	0.56
	SpaceFusion	0.00	0.02	2.73	4.16	0.87	0.76	0.9	0.0	-2.5	0.79	0.16	0.26
	BART	0.03	0.19	5.08	6.63	0.69	0.55	7.8	6.9	**7.8**	0.86	0.83	0.72
	GPS	**0.09**	**0.53**	**5.74**	**7.61**	**0.41**	**0.15**	**8.1**	**7.1**	**7.8**	**0.87**	**2.58**	**0.75**
Gab	Seq2Seq	0.02	0.17	5.14	6.71	0.56	0.30	7.5	5.0	6.7	0.86	0.14	0.67
	MMI	0.02	0.17	5.28	6.82	0.55	0.30	5.8	3.6	6.2	0.85	0.18	0.65
	SpaceFusion	0.00	0.01	3.72	4.84	0.81	0.73	1.8	0.1	0.0	0.82	0.17	0.21
	BART	0.03	0.17	5.42	7.25	0.60	0.38	6.9	6.4	6.8	0.86	0.81	0.72
	GPS	**0.06**	**0.40**	**5.82**	**7.83**	**0.39**	**0.15**	**7.6**	**6.4**	**6.8**	**0.87**	**1.94**	**0.76**

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.

Generation of counterspeech

- A **response generation** problem
- **Research challenges**
  - Quality and diverse dataset
  - Building the generation framework
  - **How can we improve ?**
    - Improve the generation
    - **Hallucination** - use knowledge sentences while fine-tuning
    - Personalisation

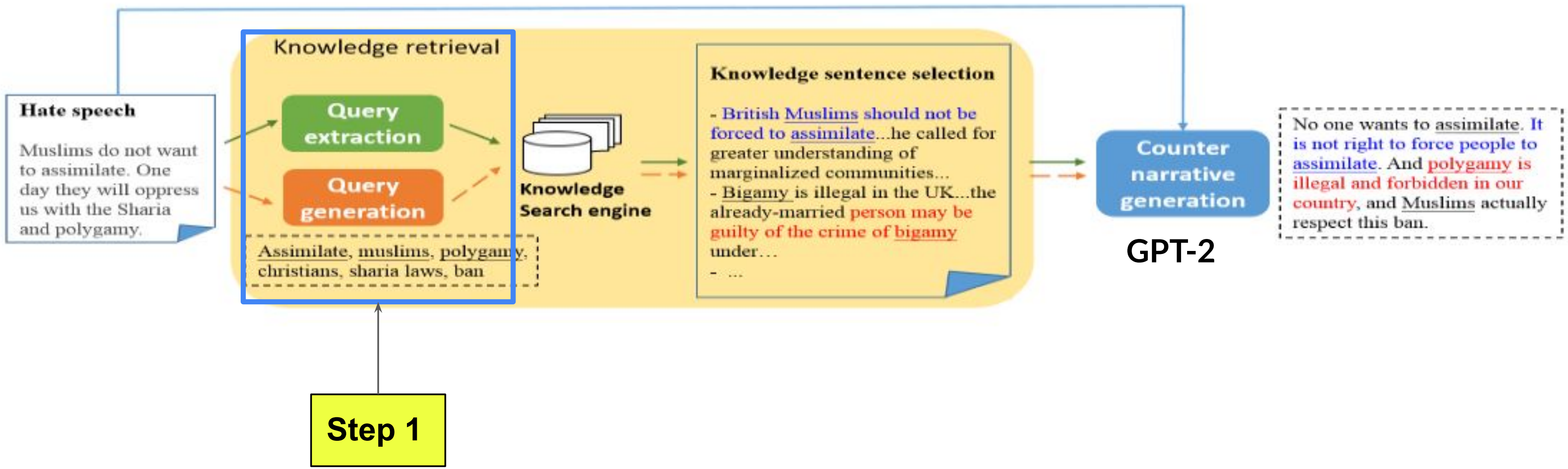


Problems: Hallucination

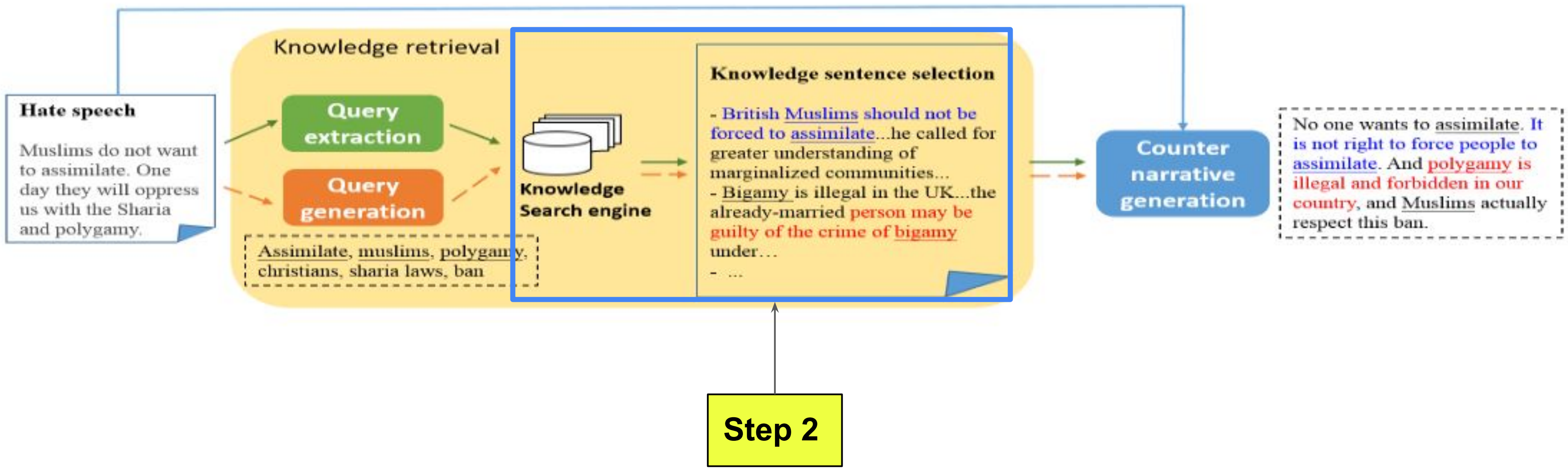
HS:	*Muslims are invading our country*
CN:	*Actually, there were 16,938,000 Muslims in the UK in 2016, so if you exclude London, that is actually increasing by 2%, which doesn't seem very significant.*

Table 2: Hallucinated CN generated by GPT-2 that is fluent and credible (according to Office for National Statistics, the Muslim population is just above 3M).

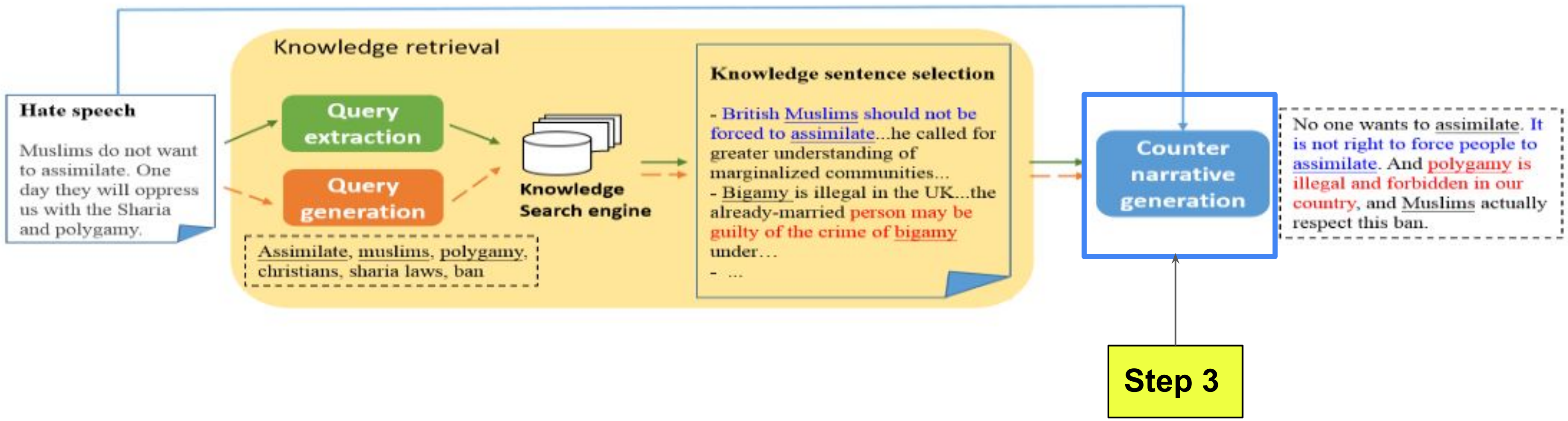
Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech [Chung 2021]



Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech [Chung 2021]



Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech [Chung 2021]



Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech [Chung 2021]

Models	Nov.	RR	B-2	R-L	#Word	#Sent.	KN overlap (ngram)		
							1	2	3
without knowledge									
TRF	0.467	7.72	0.082	0.094	21.47	1.70	-	-	-
GPT-2	0.688	9.04	0.045	0.100	15.95	1.35	-	-	-
Train_{cn}	-	3.91	-	-	21.79	1.87	0.307	0.054	0.016
with knowledge									
Candela (Q_{hs})	0.692	21.87	0.040	0.098	23.85	2.47	0.173	0.008	0.001
GPT-2_{KN}									
w/ Q_{hs}	0.723	8.13	0.082	0.094	15.60	1.32	0.258	0.023	0.008
w/ Q_{gen}	0.728	7.48	0.067	0.091	12.75	1.17	0.260	0.050	0.019
w/ $Q_{hs \cup gen}$	0.735	6.30	0.085	0.103	15.35	1.59	0.358	0.068	0.024
w/ $Q_{hs \cup cn}$	0.727	7.17	**0.166**	0.110	13.10	1.16	0.282	0.058	0.022
$\text{GPT-2}_{KN,MT}$									
w/ Q_{hs}	0.744	11.69	0.050	0.090	13.35	1.17	0.269	0.049	0.017
w/ Q_{gen}	0.731	10.37	0.052	0.092	13.34	1.14	0.253	0.044	0.017
w/ $Q_{hs \cup gen}$	0.747	7.59	0.091	0.090	16.91	1.26	0.269	0.033	0.009
w/ $Q_{hs \cup cn}$	0.731	9.56	0.048	0.107	13.05	1.13	0.276	0.057	0.023
XNLG									
w/ Q_{hs}	**0.824**	14.42	0.073	0.084	55.51	3.71	0.841	0.650	0.558
w/ Q_{gen}	0.819	6.88	0.097	0.084	55.64	3.64	0.849	0.656	0.558
w/ $Q_{hs \cup gen}$	0.812	6.98	0.074	0.089	57.58	3.00	0.828	0.579	0.475
w/ $Q_{hs \cup cn}$	0.819	**5.69**	0.076	**0.116**	55.69	3.42	0.840	0.631	0.529

Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech [Chung 2021]

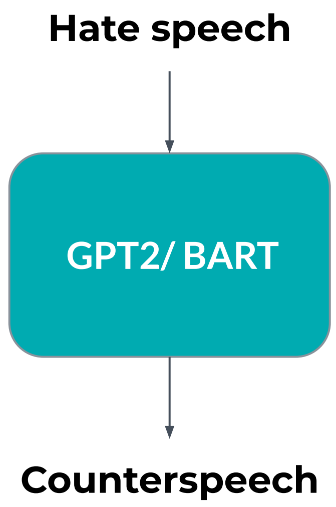
Models	Nov.	RR	B-2	R-L	#Word	#Sent.	KN overlap (ngram) 1	2	3
without knowledge									
TRF	0.467	7.72	0.082	0.094	21.47	1.70	-	-	-
GPT-2	0.688	9.04	0.045	0.100	15.95	1.35	-	-	-
Train_{cn}	-	3.91	-	-	21.79	1.87	0.307	0.054	0.016
with knowledge									
Candela (Q_{hs})	0.692	21.87	0.040	0.098	23.85	2.47	0.173	0.008	0.001
GPT-2$_{KN}$									
w/ Q_{hs}	0.723	8.13	0.082	0.094	15.60	1.32	0.258	0.023	0.008
w/ Q_{gen}	0.728	7.48	0.067	0.091	12.75	1.17	0.260	0.050	0.019
w/ $Q_{hs \cup gen}$	0.735	6.30	0.085	0.103	15.35	1.59	0.358	0.068	0.024
w/ $Q_{hs \cup cn}$	0.727	7.17	**0.166**	0.110	13.10	1.16	0.282	0.058	0.022
GPT-2$_{KN,MT}$									
w/ Q_{hs}	0.744	11.69	0.050	0.090	13.35	1.17	0.269	0.049	0.017
w/ Q_{gen}	0.731	10.37	0.052	0.092	13.34	1.14	0.253	0.044	0.017
w/ $Q_{hs \cup gen}$	0.747	7.59	0.091	0.090	16.91	1.26	0.269	0.033	0.009
w/ $Q_{hs \cup cn}$	0.731	9.56	0.048	0.107	13.05	1.13	0.276	0.057	0.023
XNLG									
w/ Q_{hs}	**0.824**	14.42	0.073	0.084	55.51	3.71	0.841	0.650	0.558
w/ Q_{gen}	0.819	6.88	0.097	0.084	55.64	3.64	0.849	0.656	0.558
w/ $Q_{hs \cup gen}$	0.812	6.98	0.074	0.089	57.58	3.00	0.828	0.579	0.475
w/ $Q_{hs \cup cn}$	0.819	**5.69**	0.076	**0.116**	55.69	3.42	0.840	0.631	0.529

Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech [Chung 2021]

Models	Nov.	RR	B-2	R-L	#Word	#Sent.	KN overlap (ngram) 1	2	3
without knowledge									
TRF	0.467	7.72	0.082	0.094	21.47	1.70	-	-	-
GPT-2	0.688	9.04	0.045	0.100	15.95	1.35	-	-	-
Train_{cn}	-	3.91	-	-	21.79	1.87	0.307	0.054	0.016
with knowledge									
Candela (Q_{hs})	0.692	21.87	0.040	0.098	23.85	2.47	0.173	0.008	0.001
GPT-2$_{KN}$									
w/ Q_{hs}	0.723	8.13	0.082	0.094	15.60	1.32	0.258	0.023	0.008
w/ Q_{gen}	0.728	7.48	0.067	0.091	12.75	1.17	0.260	0.050	0.019
w/ $Q_{hs \cup gen}$	0.735	6.30	0.085	0.103	15.35	1.59	0.358	0.068	0.024
w/ $Q_{hs \cup cn}$	0.727	7.17	**0.166**	0.110	13.10	1.16	0.282	0.058	0.022
GPT-2$_{KN,MT}$									
w/ Q_{hs}	0.744	11.69	0.050	0.090	13.35	1.17	0.269	0.049	0.017
w/ Q_{gen}	0.731	10.37	0.052	0.092	13.34	1.14	0.253	0.044	0.017
w/ $Q_{hs \cup gen}$	0.747	7.59	0.091	0.090	16.91	1.26	0.269	0.033	0.009
w/ $Q_{hs \cup cn}$	0.731	9.56	0.048	0.107	13.05	1.13	0.276	0.057	0.023
XNLG									
w/ Q_{hs}	**0.824**	14.42	0.073	0.084	55.51	3.71	0.841	0.650	0.558
w/ Q_{gen}	0.819	6.88	0.097	0.084	55.64	3.64	0.849	0.656	0.558
w/ $Q_{hs \cup gen}$	0.812	6.98	0.074	0.089	57.58	3.00	0.828	0.579	0.475
w/ $Q_{hs \cup cn}$	0.819	**5.69**	0.076	**0.116**	55.69	3.42	0.840	0.631	0.529

Generation of counterspeech

- A response generation problem
- Research challenges
  - Quality and diverse dataset
  - Building the generation framework
  - **How can we improve ?**
    - Improve the generation
    - Hallucination
    - **Personalisation -** use controller modules to make the framing of counterspeech better



Is controlling the tone important ?

- Sentimental or casual tone received 83% more response [1]
- Mathew et al. [2] found that different communities find different types of counterspeech effective .
- Empathy based counter speech can help reduce the racist comments [3]

[1] Ross Frenett and Moli Dow. One to one online interventions: A pilot cve methodology. Institute for Strategic Dialogue, 2015.
[2] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. Thou shalt not hate: Countering online hate speech. In ICWSM, 2019.
[3] Hangartner, Dominik, et al. "Empathy-based counterspeech can reduce racist hate speech in a social media field experiment." Proceedings of the National Academy of Sciences 118.50 (2021).

Add control to counterspeech datasets ?

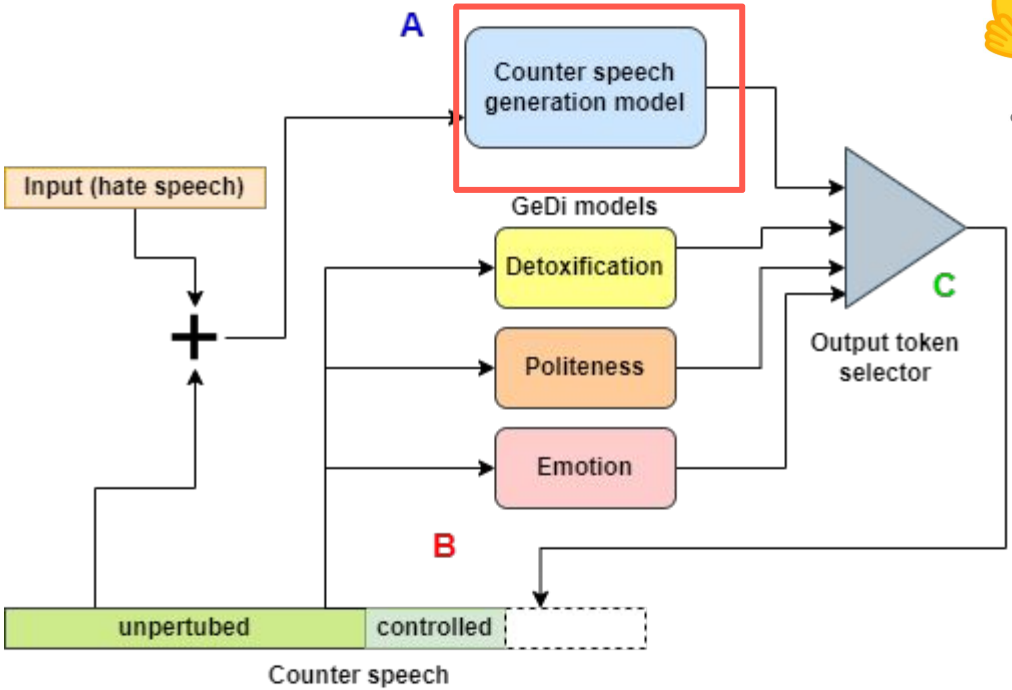
Dataset	Annotators	Unique hate speech	Source of hate	Target Community
Qian et al., '19	Crowdworkers	3,847	**REDDIT**	**Mixed**
		11,169	**GAB**	**Mixed**
Chung et al. '19	Expert Annotators	408	**SYNTHETIC**	**Muslims**

Note:- None of these dataset have additional labels to **control the tone** of the counter speech by supervision. Adding the tone might be **costly annotation task.**

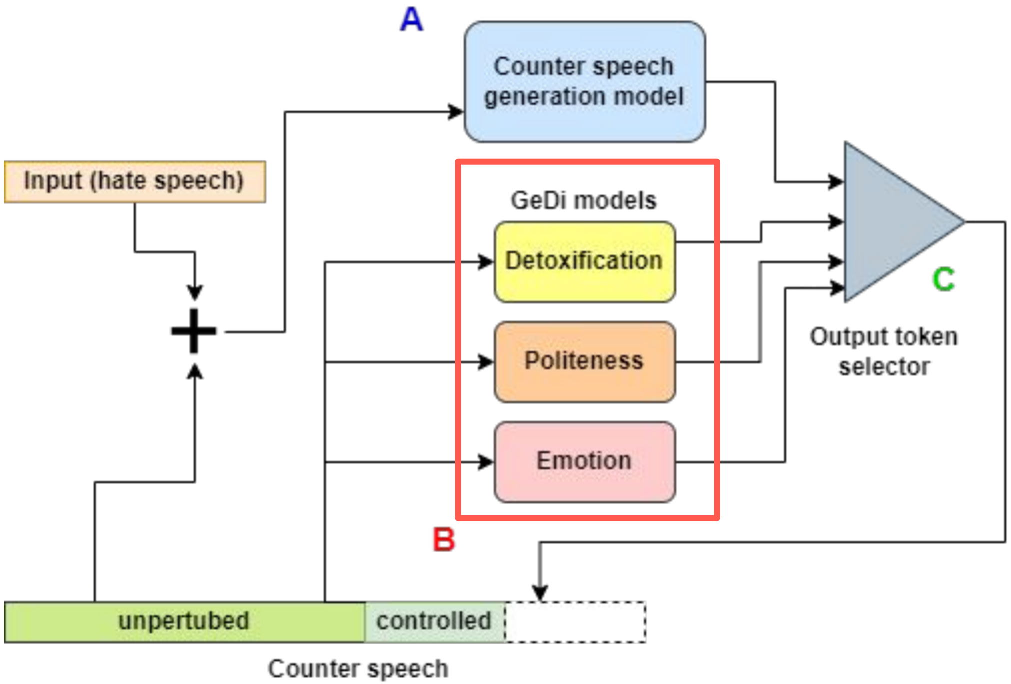
Our proposal - CounterGEDI

DialoGPT

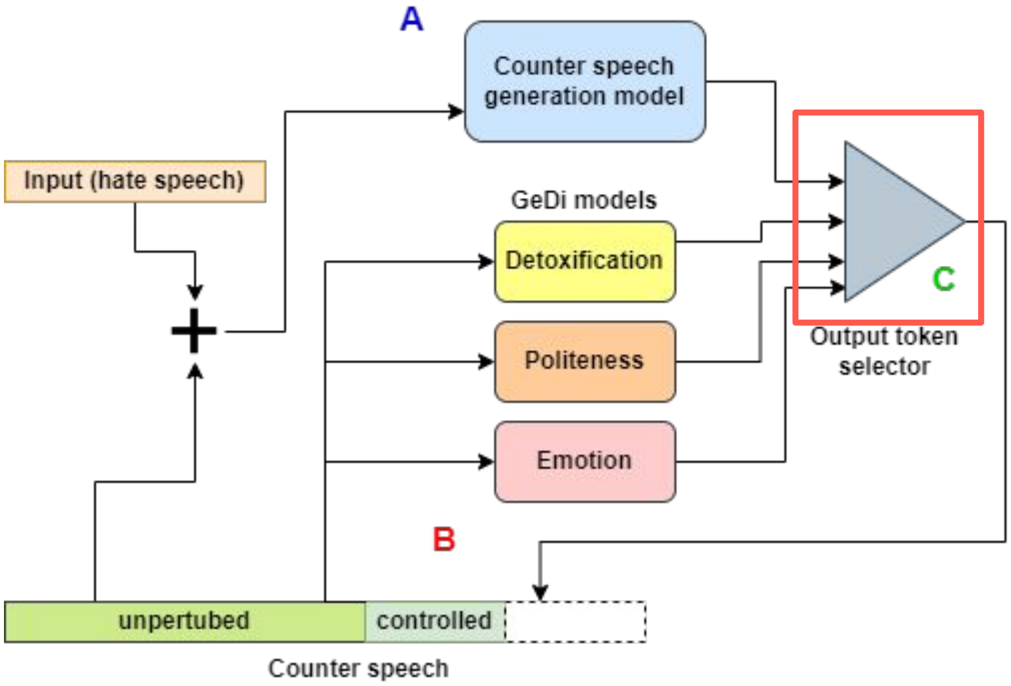
huggingface.co/microsoft



Our proposal - CounterGEDI



Our proposal - CounterGEDI



Performance: Single attribute (Control)

- **Politeness** and **detoxification** score increased by 15-18% and 6-8% respectively across all the datasets
- For the emotion attributes, '**joy**' has the **highest scores** for controlled generation.

Model	D (↑)	P (↑)	J (↑)	A (↑)	S (↑)	F (↑)
CONAN						
GPS	**0.68**	2.01	0.16	**0.12**	0.03	0.01
DialoGPTm	0.64	3.91	0.18	0.09	0.04	0.01
DialoGPTm-c	**0.68**	**4.54**	**0.34**	0.11	**0.08**	**0.05**
Reddit						
GPS	0.82	1.62	0.23	**0.32**	0.04	0.01
DialoGPTm	0.82	5.24	0.63	0.17	0.06	0.00
DialoGPTm-c	**0.87**	**6.05**	**0.72**	0.27	**0.10**	**0.02**
Gab						
GPS	0.79	1.46	0.22	**0.28**	0.04	0.01
DialoGPTm	0.81	5.14	0.66	0.17	0.05	0.00
DialoGPTm-c	**0.85**	**6.11**	**0.77**	0.26	**0.10**	**0.02**

Performance of single attribute setups with the vanilla baseline generate-prune-select (GPS) and DialoGPTm models. Each column name represents the attribute being measured. The attributes measured are politeness (P), detoxification (D), sadness (S), joy (J), anger (A) and fear (F). Politeness (P) is measured in a scale of 0-7 whereas others are measured in the scale [0, 1]. For the last row - controlled DialoGPTm (DialoGPTm-c) the column name also represents the attribute getting controlled. For all the metrics, higher is better and **bold** denotes the best scores.

Performance: Single attribute (Quality)

Scores	Detox	Polite	Joy	Anger	Sadness	Fear
CONAN						
BLEU-2	**13.8**	12.1	12.2	11.6	12.0	12.8
COLA	**0.83**	0.72	0.72	0.74	0.76	**0.72**
Reddit						
BLEU-2	**8.1**	7.8	7.7	7.8	7.5	7.3
COLA	0.72	0.77	0.70	0.72	**0.81**	0.70
Gab						
BLEU-2	**8.7**	8.3	8.5	8.3	8.2	8.3
COLA	**0.85**	0.82	0.76	0.76	0.80	0.78

BLEU-2 and COLA performance for single attribute setups for DialoGPTm-c model. Each column name represents the individual attribute model namely politeness (P), detoxification (D), sadness (S), joy (J), anger (A) and fear (F). **Bold** denotes the best scores across the row.

There is slight drop in the **relevancy** and **fluency** metric but overall they are stable when the text is getting controlled.

Performance: Multi-attribute (Control and Quality)

Our experiments with **multi-attributes** further reveals that there are certain complementing attributes for e.g joy + polite + detox which can be used to further increase the single-attribute setups.

Attributes	Detox(↑)	Polite(↑)	Emotion(↑)	B2(↑)	COLA(↑)
CONAN					
Joy(J)+P+D	**0.74**	**4.13**	0.49 (J)	13.4	**0.79**
Anger(A)+P+D	0.67	3.06	0.08 (A)	12.6	0.68
Sad(S)+P+D	0.70	3.56	0.07 (S)	13.2	0.74
Fear(F)+P+D	0.70	4.00	0.06 (F)	**13.6**	0.75
Reddit					
Joy+P+D	**0.89**	**5.79**	0.82 (J)	8.3	**0.81**
Anger+P+D	0.85	4.24	0.19 (A)	**8.3**	0.72
Sad+P+D	0.87	3.56	0.09 (S)	8.2	0.79
Fear+P+D	0.87	4.00	0.01 (F)	7.8	0.79
Gab					
Joy+P+D	**0.87**	5.68	0.85 (J)	**8.8**	**0.85**
Anger+P+D	0.83	4.11	0.19 (A)	8.5	0.75
Sad+P+D	0.85	4.70	0.09 (S)	**8.8**	0.84
Fear+P+D	0.86	**5.82**	0.01 (F)	**8.8**	0.83

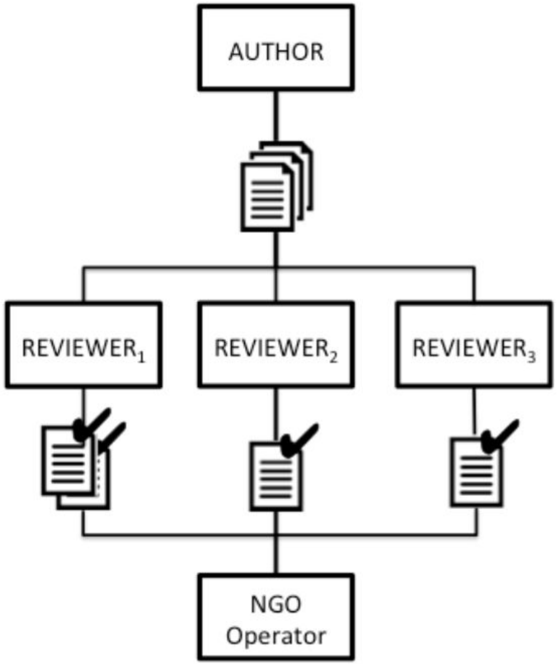
Results of controlling three attributes – politeness, detoxification and one of the emotions in a multi-attribute setting. The columns represent the amount of the attribute present for each setup. The column – *emotion* represents the score of the emotion shown in the parenthesis that is being controlled for that instance. BLEU(B2) and COLA were also reported for different setups. For all metrics, higher is better and **bold** denotes the best scores.

Hybrid strategy of collection

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.



Counterspeech datasets

Dataset	Annotators	Total data	Source of hate	Target Community
Fanton et.al, `21	Expert Annotators	5000 single turn	**SYNTHETIC**	**Multi-target**
Bonaldi et.al, '22	Expert Annotators	3000 multi-turn dialogue	**SYNTHETIC**	**Multi target**

[9] Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. ACL 2021
[10] Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering. EMNLP 2022

Reviewing metrics

For single turn dataset

1) Acceptance Rate
2) HTER
3) Novelty
4) Repetition Rate
5) Vocabulary Expansion
    a) Contribution of author vs reviewer
    b) Cross-fertilization of vocab

Additional metrics for multi-turn

6) Turn deletion
7) Turn swap

Reviewing metrics

For single turn dataset

1) Acceptance Rate
2) HTER
3) Novelty
4) Repetition Rate
5) Vocabulary Expansion
    a) Contribution of author vs reviewer
    b) Cross-fertilization of vocab

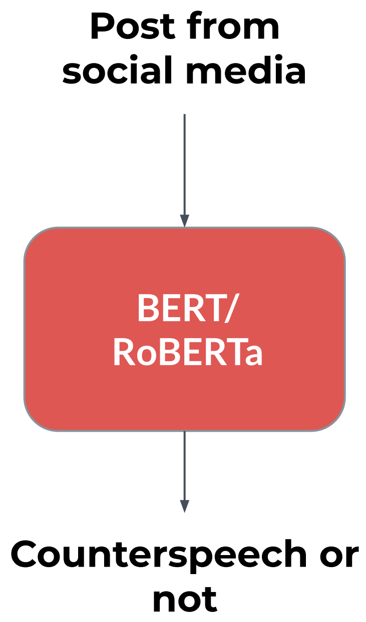
Additional metrics for multi-turn

6) Turn deletion
7) Turn swap

Can we measure the counterspeech quality ?

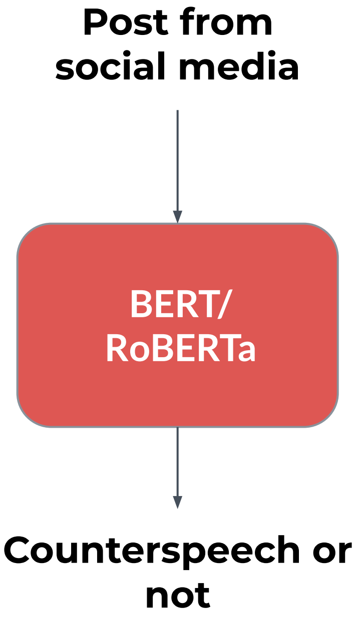
Detection of counterspeech

- **A detection problem**
- **Research challenges**
  - Building a dataset for counterspeech detection
  - Detection framework



Detection of counterspeech

- A **detection** problem
- **Research challenges**
  - **Building a dataset for counterspeech detection**
  - Detection framework



Counterspeech detection datasets

Dataset	Number of counterspeech	Total data	Source	Target Community
Mathew et al., ‘19	6,898	13,922	YouTube	Jews, Blacks, LGBT
He et al. ‘21	517	2,290	Twitter	Asians
Yu et.al ‘21	2,879	6,848(x2)	Reddit	Mixed

[1] Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., ... & Mukherjee, A. (2019, July). Thou shalt not hate: Countering online hate speech. ICWSM 2019
[2] He, Bing, et al. "Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis." ASONAM 2021
[3] Yu, Xinchen, Eduardo Blanco, and Lingzi Hong. "Hate Speech and Counter Speech Detection: Conversational Context Does Matter." NAACL 2022.

Types of counterspeech

Data collected and annotated from comments of youtube videos showing hate towards some communities

	Target community			Total
Type of counterspeech	*Jews*	*Blacks*	*LGBT*	
Presenting facts	308	85	359	752
Pointing out hypocrisy or contradictions	282	230	526	1038
Warning of offline or online consequences	112	417	199	728
Affiliation	206	159	200	565
Denouncing hateful or dangerous speech	376	482	473	1331
Humor	227	255	618	1100
Positive tone	359	237	268	864
Hostile	712	946	1083	2741
Total	2582	2811	3726	9119

Thou Shalt Not Hate: Countering Online Hate Speech [Mathew 2019]

Types of counterspeech

Data collected and annotated from comments of youtube videos showing hate towards some communities

	Target community			Total
Type of counterspeech	*Jews*	*Blacks*	*LGBT*	
Presenting facts	308	85	359	752
Pointing out hypocrisy or contradictions	282	230	526	1038
Warning of offline or online consequences	112	417	199	728
Affiliation	206	159	200	565
Denouncing hateful or dangerous speech	376	482	473	1331
Humor	227	255	618	1100
Positive tone	359	237	268	864
Hostile	712	946	1083	2741
Total	2582	2811	3726	9119

Thou Shalt Not Hate: Countering Online Hate Speech [Mathew 2019]

Types of counterspeech

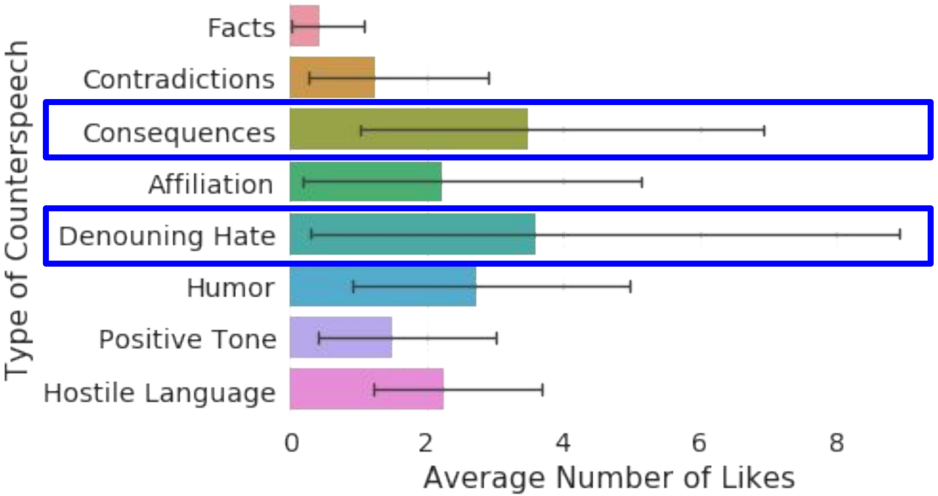
Data collected and annotated from comments of youtube videos showing hate towards some communities

	Target community			Total
Type of counterspeech	*Jews*	*Blacks*	*LGBT*	
Presenting facts	308	85	359	752
Pointing out hypocrisy or contradictions	282	230	$\boxed{526}$	1038
Warning of offline or online consequences	112	417	199	728
Affiliation	206	159	200	565
Denouncing hateful or dangerous speech	376	482	473	1331
Humor	227	255	$\boxed{618}$	1100
Positive tone	359	237	268	864
Hostile	712	946	1083	2741
Total	2582	2811	3726	9119

Thou Shalt Not Hate: Countering Online Hate Speech [Mathew 2019]

Click logo for demo

Types of counterspeech



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

Example:

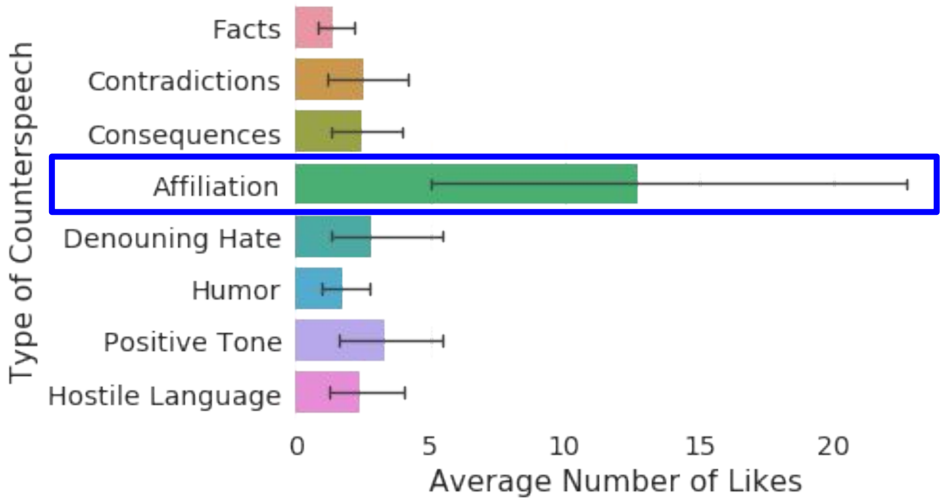
"i hope these cops got fired! this is bullshit"

"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

Types of counterspeech



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.

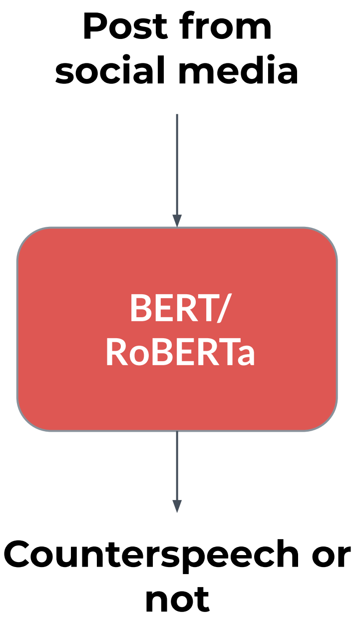
Example:

"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"

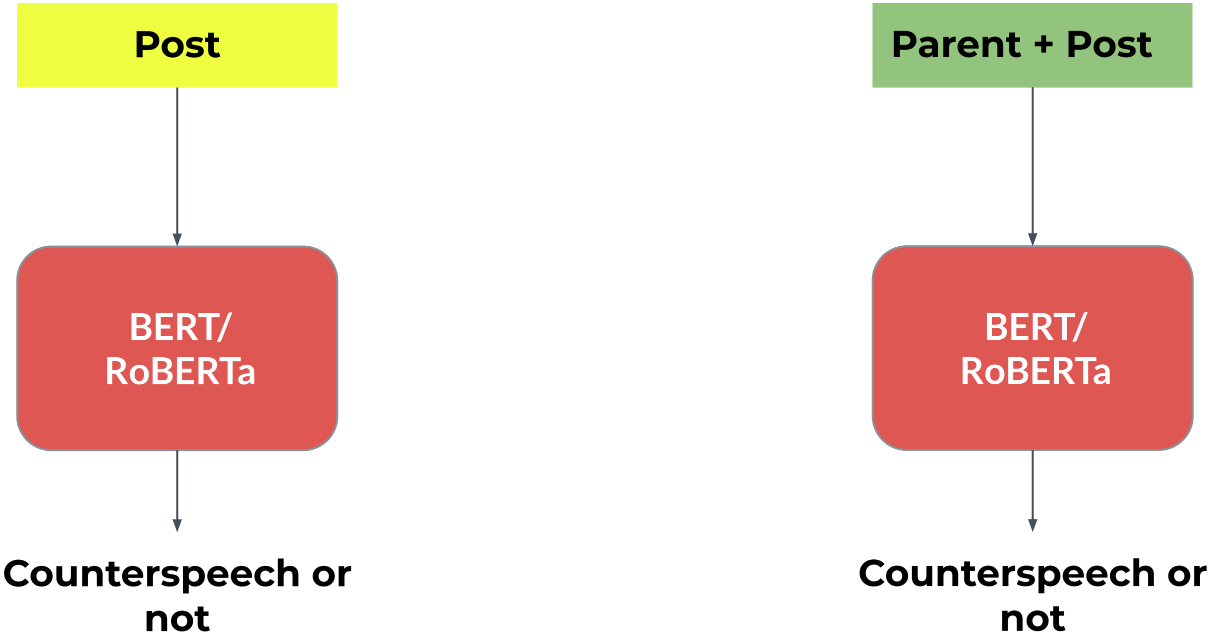
Thou Shalt Not Hate: Countering Online Hate Speech [Mathew 2019]

Detection of counterspeech

- A **detection** problem
- **Research challenges**
  - Building a dataset for counterspeech detection
  - **Detection framework**



Does conversational context matter ?



Does conversational context matter ?

	Hate			Counter-hate			Neutral			Weighted Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Majority Baseline	0.00	0.00	0.00	0.00	0.00	0.00	0.51	1.00	0.67	0.26	0.51	0.34
Trained with Target	0.56	0.55	0.56	0.41	0.36	0.38	0.67	0.71	0.69	0.58	0.59	0.58
+ Silver	0.58	0.55	0.57	0.44	0.42	0.43	0.69	0.72	0.70	0.60	0.61	0.61
+ Related task	0.56	0.55	0.56	0.51	0.41	0.45	0.68	0.74	0.71	0.61	0.61	0.61
+ Silver + Related task	0.55	0.56	0.56	0.49	0.53	0.51	0.67	0.69	0.70	0.61	0.61	0.61
Trained with Parent_Target	0.56	0.62	0.59	0.52	0.38	0.44	0.68	0.72	0.70	0.61	0.62	0.61
+ Silver†	0.58	0.57	0.57	0.49	0.51	0.50	**0.72**	**0.71**	**0.72**	0.63	0.63	0.63
+ Related task†	**0.55**	**0.66**	**0.60**	0.54	0.43	0.48	0.71	0.70	0.71	0.63	0.63	0.63
+ Silver + Related task‡	0.55	0.65	0.60	**0.54**	**0.52**	**0.53**	0.74	0.68	0.71	**0.64**	**0.64**	**0.64**

Is counterspeech effective?

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.

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

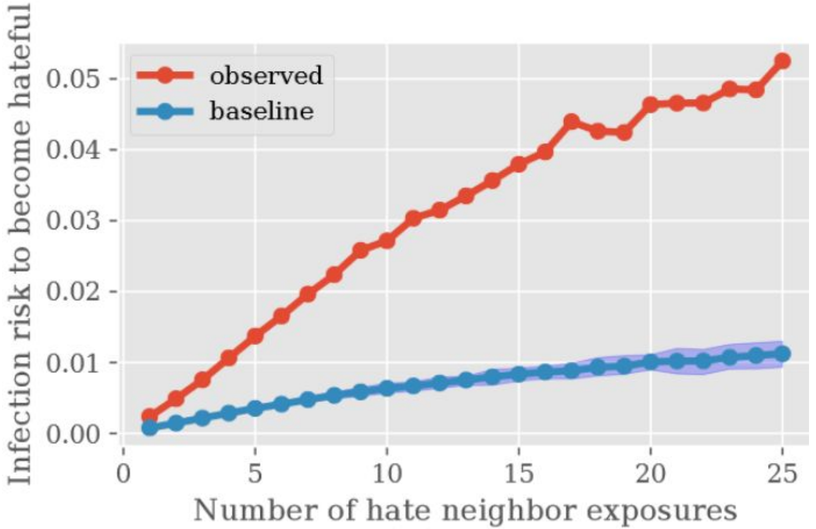
- Warning of Consequences
- Shaming/Labeling
- Empathy and Affiliation
- Humor
- Images

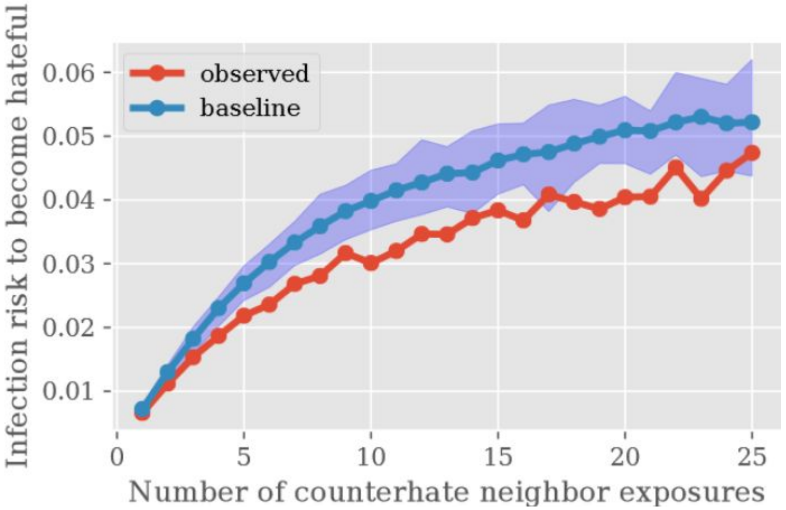
Discouraged Strategies

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

Evidence from social media platforms

Analysis reveals that **counterhate messages can discourage users from turning hateful** in the first place. [Ziem 2020]





Evidence from social media platforms

Their findings suggest that organized hate speech is associated with changes in public discourse and that counter speech—**especially when organized**—may help **curb hateful rhetoric** in online discourse [Garland 2020]

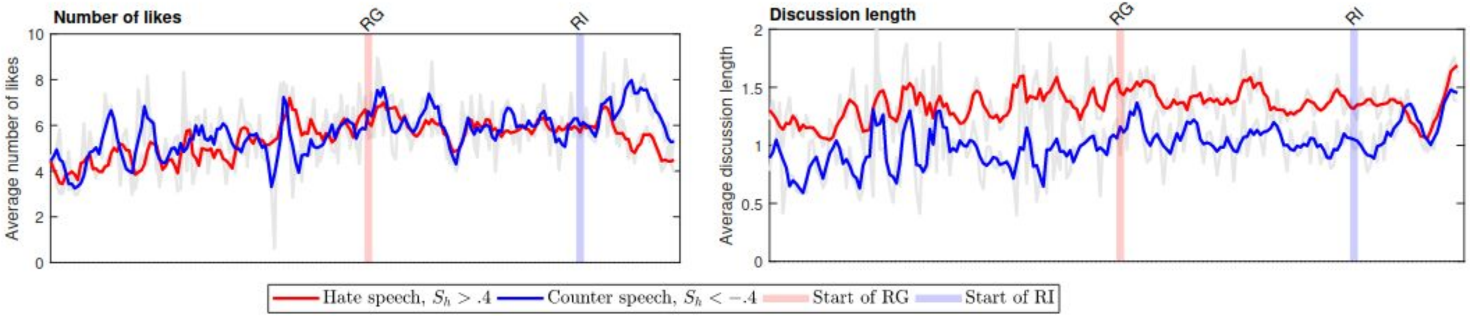


Figure 5: **Impressions of hate and counter speech.**Impact of hate and counter speech messages over time as quantified by the average number of likes and length of conversation they initiate. The emergence of organized counter speech (RI, blue vertical line). Results are for 181,370 reply trees from January 2015 to December 2018. Each data point is a week average and trends are smoothed over a month-long window. The timeline on the x-axis is the same as in other figures but was omitted for space, except for markers of the emergence of RG and RI.

Does type of counterspeech matter?

Affiliation - Control accounts ("bots") to sanction the harassers. The author found that subjects who were countered by a **high-follower white male** significantly **reduced** their use of a racist slur.

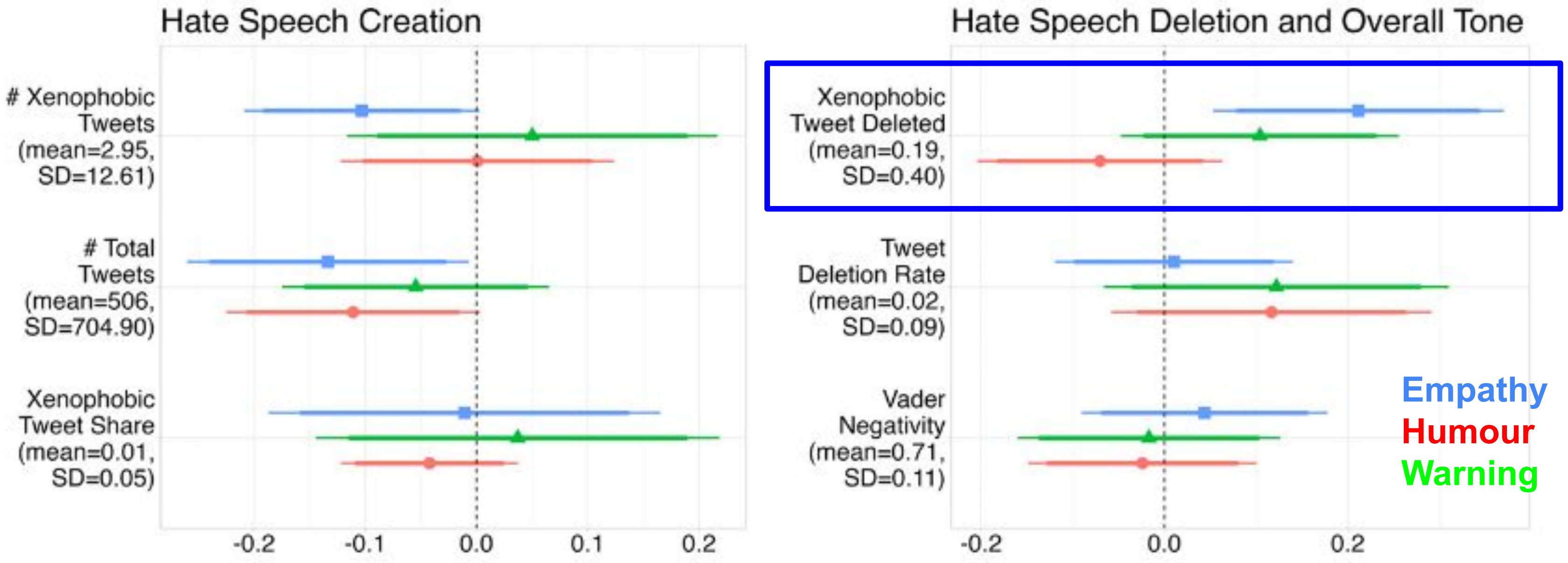
Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment [Munger 2016]

Does type of counterspeech matter?

- The authors compared different types of counter speech - Warning of consequences, Humour and Empathy [Hangartnera,2021]

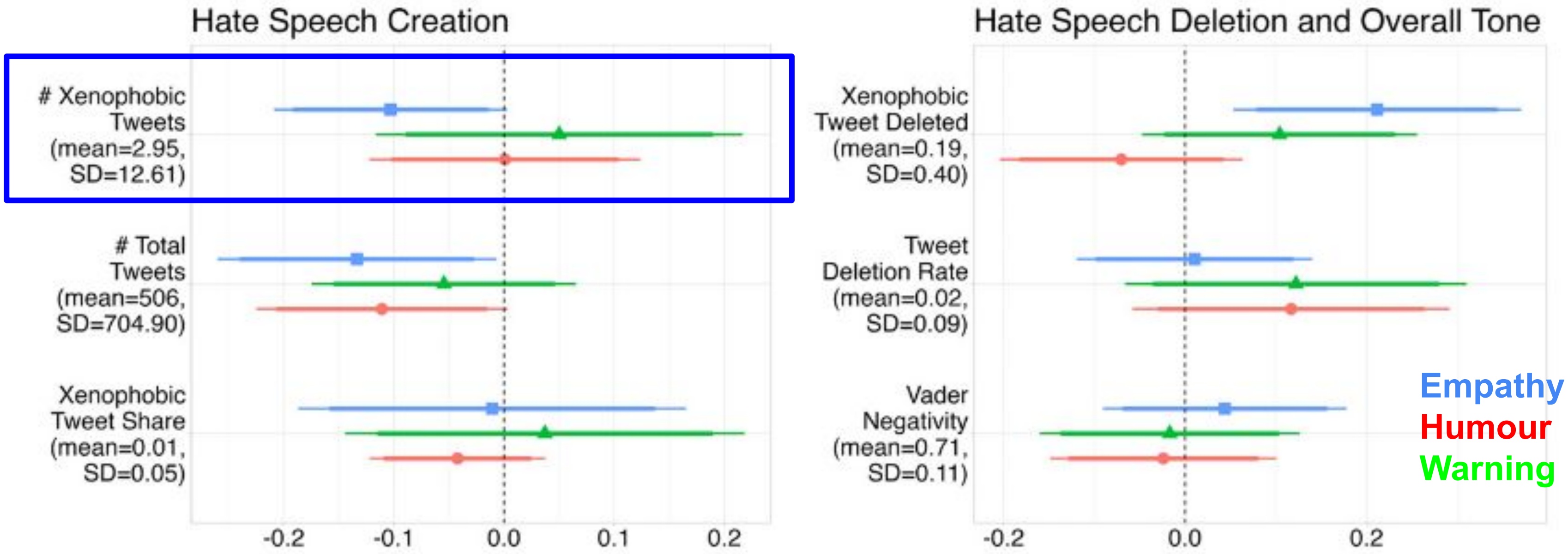
Does type of counterspeech matter?

Empathy based counter speech increase the **retrospective deletion of xenophobic hate speech**(0.2 SD) and reduce the prospective creation of xenophobic hate speech over a 4-wk follow-up period by 0.1 SD. [Hangartnera,2021]



Does type of counterspeech matter ?

Empathy based counter speech increase the retrospective deletion of xenophobic hate speech(0.2 SD) and reduce the **prospective creation of xenophobic hate speech** over a 4-wk follow-up period by 0.1 SD [Hangartnera,2021].



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

Weakness

Opportunity

Threat

Strengths

- Advancement in NLP i.e. Transformers
- Multilinguality
- NGO Initiatives
- Multiple datasets
- Theme, Research grants etc.

Weakness

Opportunity

Threat

Strengths

Weakness

- Inconsistent annotations
- Diverse tasks
- Lack of generalisability
- Bias in data as well as in models
- Lack of explainability

Opportunity

Threat

Strengths

Weakness

Opportunity

- Multimodal datasets
- User as an important aspect
- New variants coming up - Fearspeech, Dangerous speech
- Counter speech as mitigation

Threat

Strengths

Weakness

Opportunity

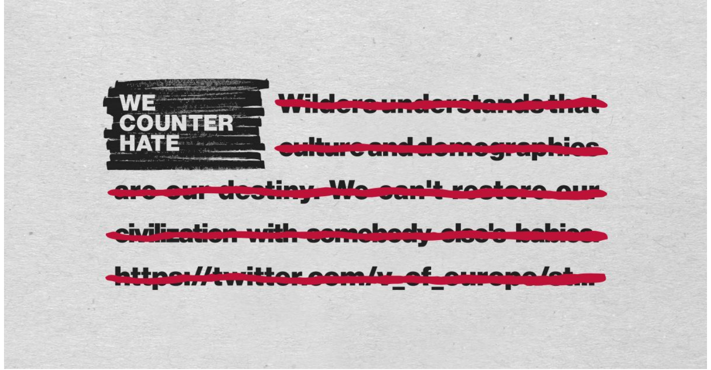
Threat

- Users vs detection
- Alternative (echo chamber) platforms - Gab
- Govt agencies weaponizing hate
- Laws used to silence dissent

Campaigns to deter hate

Counterspeech.fb

ADL



WeCounterHate

NoHateSpeechMovement

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