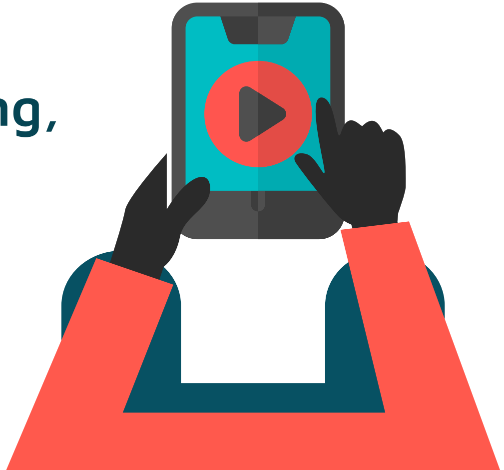
Enabling moderation of harmful content in online social media platforms

by

Punyajoy Saha

Dept. of Computer Science & Engineering,
IIT Kharagpur



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

Table of contents

1. Introduction
2. Detection
3. Explanation
4. Mitigation
5. Conclusion

Harmful speech

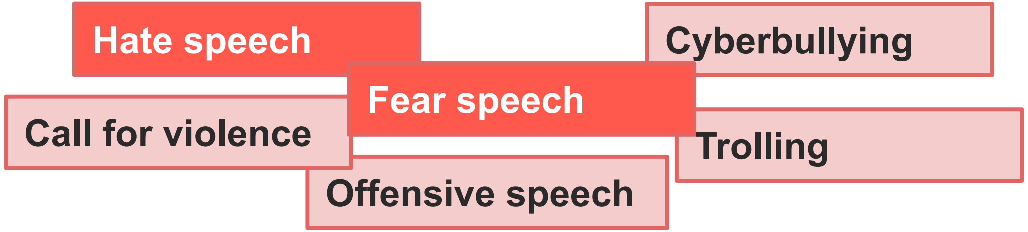
Harmful speech consists of a range of phenomenon that often overlap and intersect, and includes a variety of types of speech that cause different harms.

- 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).

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.



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

Definitions

Hate speech *is a language used to express hatred towards a targeted individual or group or is intended to be derogatory, to humiliate, or to insult the members of the group, based on attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender*[3].

Fear speech *is an expression aimed at instilling (existential) fear of a target group based on attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender*[2].

[2] Buyse, A. (2014). Words of violence:" Fear speech," or how violent conflict escalation relates to the freedom of expression. Hum. Rts. Q., 36, 779.
[3] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In Proceedings of the international AAAI conference on web and social media (Vol. 11, No. 1, pp. 512-515).

Examples

Fear speech	Hate speech
Germany is no longer German. German media celebrates school where 80% of class is non-German	You are a camel piss drinking goat f**king imbecile now get off my timeline you disgusting piece of sh*t.
TILL White people won't protest for their SAFETY. Hell, it's not just Whites. Asian & Middle Eastern shopkeepers are frequent victims.Young Black Males are a DANGER to society. SOME are ok, but we don't know who is who. We need PROTECTION & the RIGHT NOT to race mix!	I hear Botswana is lovely in the spring. All n**gers should go there. And stay.
Jewish poison pouring out of our media and Hollywood is destroying Christianity	Because Jews are lying pigs. I'm really thinking this is a genetic thing..

**Taken from the dataset created in Gab

Harmful speech

Pittsburg Shooting

Srilankan riots

Rohingya Genocide

Psychological effects

Effects of harmful speech

Pittsburg Shooting

Moderation of Harmful speech

Srilankan riots

Rohingya Genocide

Psychological effects

Moderation of Harmful speech

BUT ..


CONSUMER TECH • EDITORS' PICK

Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day

John Koetsier Senior Contributor

John Koetsier is a journalist, analyst, author, and speaker.

Follow

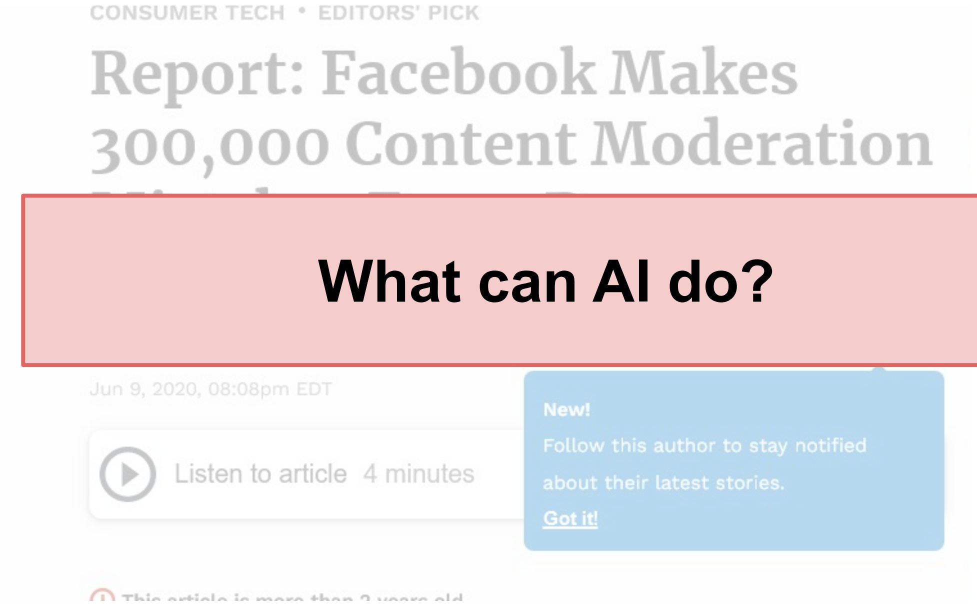
Jun 9, 2020, 08:08pm EDT

Listen to article 4 minutes

New!

Follow this author to stay notified about their latest stories.

Got it!


CONSUMER TECH • EDITORS' PICK

Report: Facebook Makes 300,000 Content Moderation

Jun 9, 2020, 08:08pm EDT

Listen to article 4 minutes

New!

Follow this author to stay notified about their latest stories.

Got it!

What can AI do?

Enablers

Enablers are tools which help in the **moderation pipeline**. We propose the following three enablers

- **Detection -** Identifies harmful content (*fear speech, hate speech .etc*) from the platform using classification systems at scale.

- **Explainability -** Explains the classification system's behaviour to help the moderator understand model behaviour.

- **Mitigation** - Providing mitigation solutions in response to a particular harmful speech.

Table of contents

1 Introduction

2 Detection

3 Explanation

4 Mitigation

5 Conclusion

Enablers

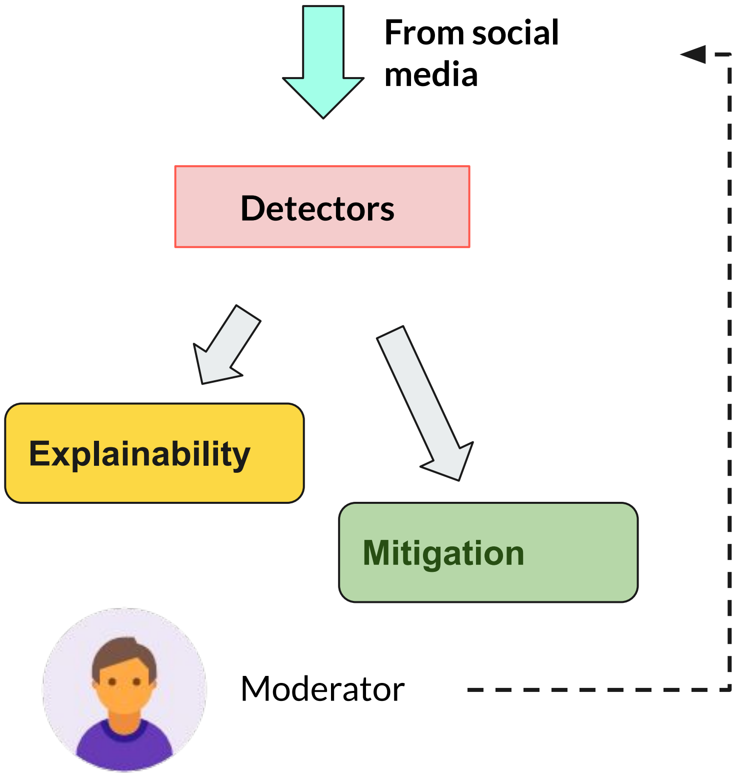


Table of contents

1 Introduction

2 Detection

3 Explanation

4 Mitigation

5 Conclusion

- Building a framework for detection of fear speech

Introduction

In this work, we built a framework for detection and analysis of fear speech (one form of harmful speech) :-

- In this first work, we study prevalence of fear speech in public Whatsapp groups in India.
- In the second work, we extend this analysis to Gab platform and further compare fear speech with hate speech.

Related works

Reference	Contribution
Vidgen, Bertie, and Taha Yasseri. "Detecting weak and strong Islamophobic hate speech on social media." Journal of Information Technology & Politics 17.1 (2020): 66-78.	Studies hate speech against muslims
Klein, Adam. Fanaticism, racism, and rage online: Corrupting the digital sphere. Springer, 2017.	Hints at large presence of fear content in the online communication
Buyse, Antoine. "Words of violence:" Fear speech," or how violent conflict escalation relates to the freedom of expression." Hum. Rts. Q. 36 (2014): 779.	Formal definition of fear speech
Gottschalk, Peter, Gabriel Greenberg, and Gary Greenberg. *Islamophobia: making Muslims the enemy*. Rowman & Littlefield, 2008.	Qualitative analysis of fear against muslims

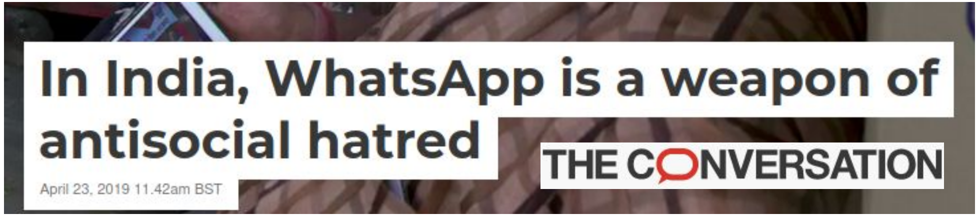
Our work operationalises the *fear speech* definition and performs a quantitative analysis on a social media platform

"Short is the Road that Leads from Fear to Hate"

Fear speech in Indian Whatsapp groups (The Webconference 2021)

Why Whatsapp ?

- Launched in mid 2010s and has reached **500 million users** by 2020
- It is becoming a de facto cheap source for messaging
- Since there is **no moderation**, users are susceptible to misinformation and propaganda.



Delhi riots: WhatsApp group promoted enmity on religion ground, says charge sheet

The Indian EXPRESS

Data collection

- Searched public WhatsApp groups using **“chat.whatsapp.com +keyword”**. **Keyword** represent keywords from different political parties and leaders across India

Data collection

- Searched public WhatsApp groups using **"chat.whatsapp.com +keyword"**. **Keyword** represent keywords from different political parties and leaders across India
- In total **5,000 political groups** having image, videos and text spanning from **August 2018 - 19**[2].

[2] Garimella, K., & Tyson, G. (2018, June). Whatapp doc? a first look at whatsapp public group data. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 12, No. 1).

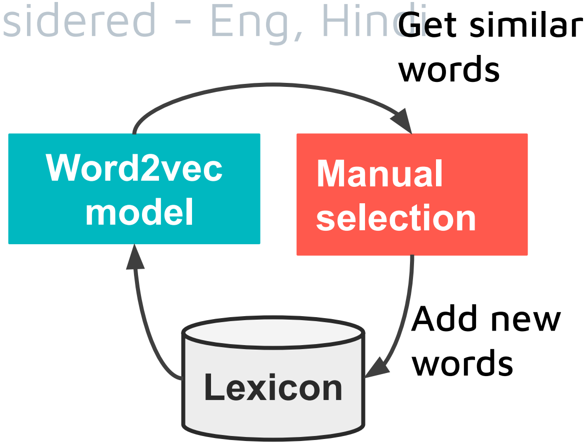
Data collection

- Searched public WhatsApp groups using **"chat.whatsapp.com +keyword"**. **Keyword** represent keywords from different political parties and leaders across India
- In total **5,000 political groups** having image, videos and text spanning from **August 2018 - 19**[1].
- Spam messages were removed, language considered - Eng, Hindi (**70% coverage**)

Features	Count
Number of posts	1,426,482
Number of groups	5,010
Average length of a message (in words)	89

Data collection

- Searched public WhatsApp groups using **"chat.whatsapp.com +keyword"**. **Keyword** represent keywords from different political parties and leaders across India
- In total **5,000 political groups** having image, videos and text spanning from **August 2018 - 19**[1].
- Spam messages were removed, language considered - Eng, Hin
(**70% coverage**)
- To sample data for annotation, **lexicon** about was created using a bootstrapping method



Data Annotation

Initial annotation and training of annotators

- **500** posts was annotated by 2 expert annotators
- Students voluntarily participated using online form and were compensated for the task.
- 7 undergraduate male students aged 19-21 years.
- Training of the annotators was done in 2 rounds of 40 posts.

Main annotation

- Done on docanno annotation platform where each student was provided with a secure account
- Batch size were gradually increased from 100 to 500 posts
- Regular breaks and error analysis were planned

Data Annotation

5k unique posts with Fleiss kappa of **0.36** inter annotator agreement done by **9 annotators**

Challenges

- Message length
- Complex Language

Features	Fear speech	Non fear speech
Number of posts	7,845	19,107
Unique posts (Annotated)	1,142	3,640
Average length of a message (in words)	500	464

Argumentative structure (Qualitative)

Examples of fear speech(FS),hate speech(HS), and non fear speech(NFS).

We show how the 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.**

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

Interesting emojis

Emojis

- Built the co-occurrence network based on emojis.
- Louvain algorithm[4] was used to find emoji communities

Row	Emojis	Interpretation
1	✊, 🚩, 🙏, 🔱, 🏹, 🐚, 🕉, ⛳, 😢, 👆🏻, 🔥, 🌞, 👊🏽, 🐅, 🌹, 🌼, 😊, 🍁, ✅, 🌷, ❗, 🤝, 💐, 🗣, 👍, 〰, ⏺	Hindutva symbols
2	😎, 👺, 📗, ✍, 👉, ❓, ❌, 😈, ⭕, 👿, 🐷, 🐖, 🕋, 🌀, 📌, 🔰	Muslim as demons
3	💣, 🔫, 🗡, ⚔, 🤺, 🦁, 💥, 🐅, ☪	terrorist attacks or riots by Muslims
4	😡, ⚫, 🛑, 🔵, ✋, ⬇	Angry about torture on Hindus

[4] Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." Journal of statistical mechanics: theory and experiment 2008.10 (2008): P10008. APA

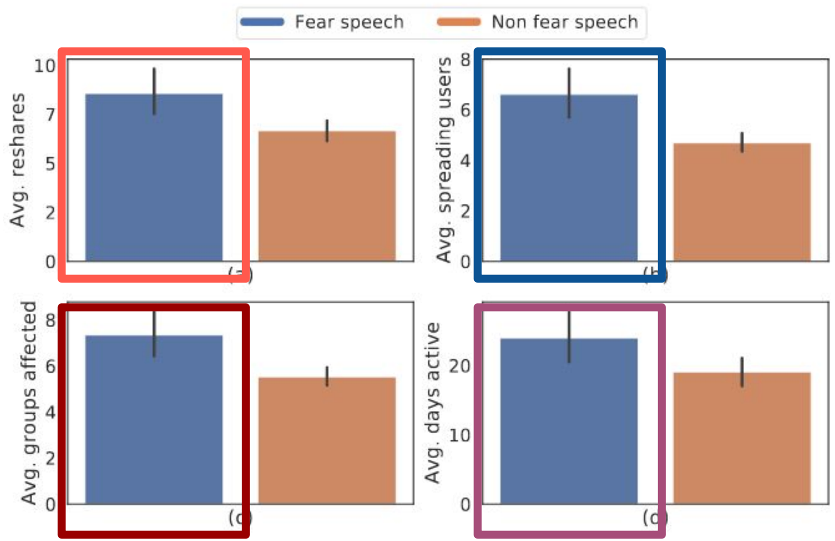
Controversial topics

LDA[5] models to extract topics (number of topics as 10 had highest coherence score)

Topics	Themes of fear speech
Love jihad (Muslim men are forcing hindu women to interfaith marriages)	Painting interfaith marriages in wrong light
Increase in muslim population (Muslim population increasing at an alarming rate)	Using event in the current timeline to spread fear
Kerala riots (Blaming muslims for a past communal riots at Kerala)	Past events used to show how muslims have done harmful things

[5] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John (ed.). "Latent Dirichlet Allocation". Journal of Machine Learning Research.

Prevalence of fear speech



More reshares, large #users spreading, large #groups affected and a longer lifetime

Fear speech detection: Techniques

Doc2vec + **LR/SVM**

100 dim vectors

SVM with RBF kernel

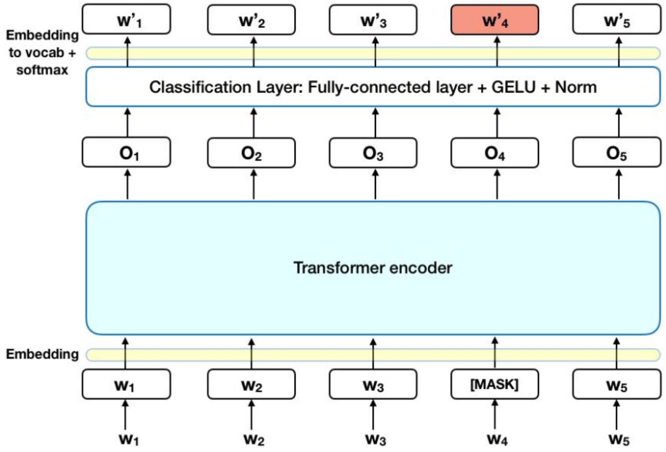
LASER + **LSTM**

1024 dim vectors per sentence

Learning rate - 0.01
Hidden dimensions - 128

Different forms of inputs

(A) n-tokens from the start
(B) n-tokens from the end
(C) n/2-tokens from the start and n/2-tokens from the end append together by a <SEP> token



XLM-Roberta /BERT

Default parameters with token length of 256, learning rate of 2e-5

Fear speech detection : Results

Models	Features	Accuracy	F1-Macro	AUC-ROC
Logistic regression	Doc2vec	0.72	0.65	0.74
SVC (with RBF Kernel)	Doc2vec	0.75	0.69	0.77
LSTM	LASER embeddings	0.66	0.63	0.76
XLM-Roberta +LR	Raw text (c)	**0.76**	**0.71**	**0.83**
mBERT + LR	Raw text (c)	0.72	0.65	0.80

Surveying WhatsApp users

Important to understand the **perception** of people in the WhatsApp groups. Used **facebook's ad** to target **three** types of users (mobile numbers obtained from the WhatsApp public groups analyzed):

- Users posting fear speech message (*UPFG*)- **3000**
- Users present in groups sharing fear speech (*UFSG*) - **9,500**
- Users present in groups not sharing fear speech (*UNFSG*) - **9,500**

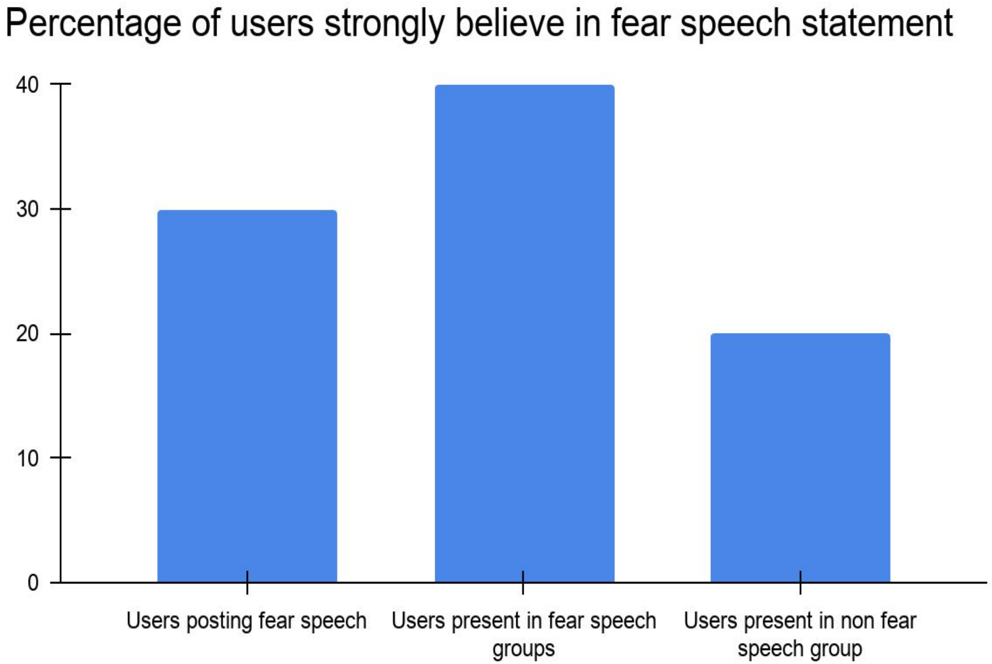
Surveying WhatsApp users

- Important to understand the **perception** of people in the WhatsApp groups. Used **facebook's ad targeting** to **three** types of users selected:
- **3** (user types) X **2** (types of statements). Total **8 statements.**
- With each statement participants were asked about their **belief** and **propensity to share**

Claim in fear speech: In 1761, Afghanistan got separated from India to become an Islamic nation.

Claim in Non fear speech: A Muslim is not a terrorist, and a terrorist is not a Muslim.

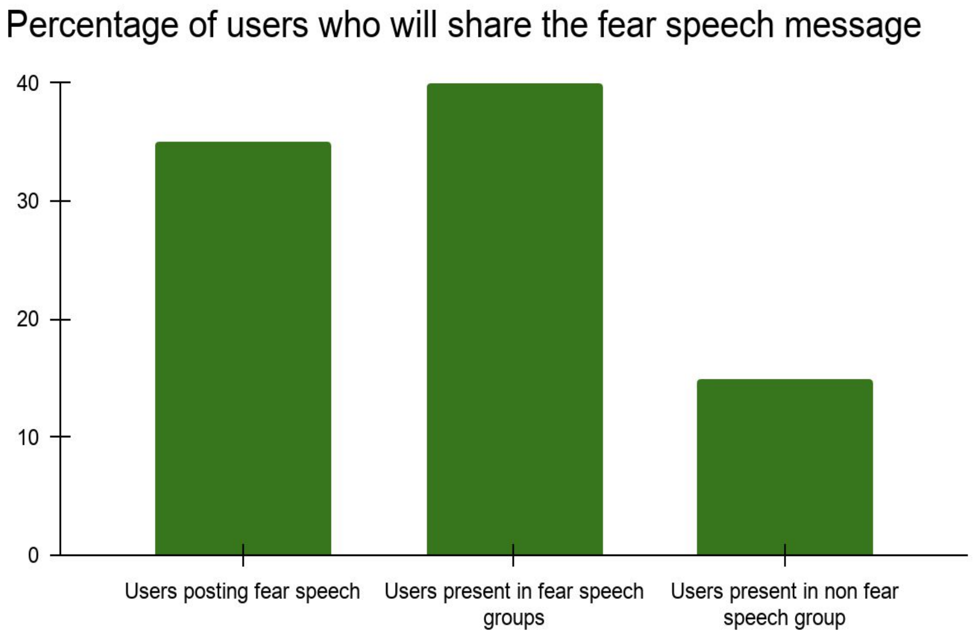
Results from the survey



Users in UPFG and UFSG are more likely to believe in fear speech

Results from the survey

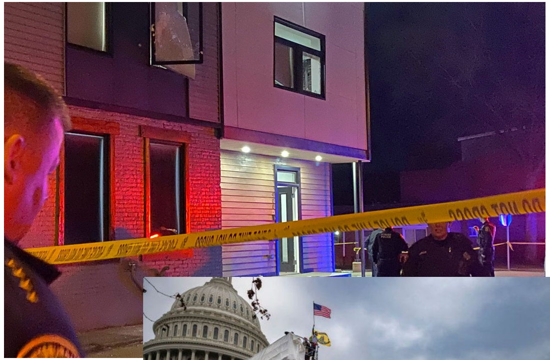
Users in UPFG and UFSG are more likely to share the fear speech



On the rise of fear speech in online social media (PNAS 2022)

Why Gab platform ?

- Promotes itself as “Champion of free speech”.
- Criticised as an echo-chamber for “alt-right users”.
- Gab promotes “free-speech”, allowing users to post hateful content
- **We wanted to further understand if fear speech is also prevalent**



Related works

Reference	Contribution
Kennedy, Brendan, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs et al. "The gab hate corpus: A collection of 27k posts annotated for hate speech." (2018).	Created a large corpus of hate speech in Gab
Mathew, Binny, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. "Spread of hate speech in online social media." In Proceedings of the 10th ACM conference on web science, pp. 173-182. 2019.	Studied diffusion dynamics of users posting hateful posts and their networks
Mathew, Binny, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. "Hate begets hate: A temporal study of hate speech." Proceedings of the ACM on Human-Computer Interaction 4, no. CSCW2 (2020): 1-24.	Characterised the growth of hate speech in Gab and also saw how the hate users affected the community

This work extends the last work to further understand the prevalence of fear speech and its effects.

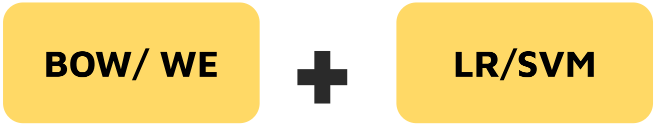
Annotated dataset

- Sampled the posts from a corpus of Gab Data[1] which contains **21 million posts** and their metadata from **October 2016** to **July 2018.**
- **4 expert** annotators and **103 crowd annotators** participated in MTurk platform.
- Total datapoints were ~**10,000**, out of which **1800** were fear speech and **4000** were hate speech.

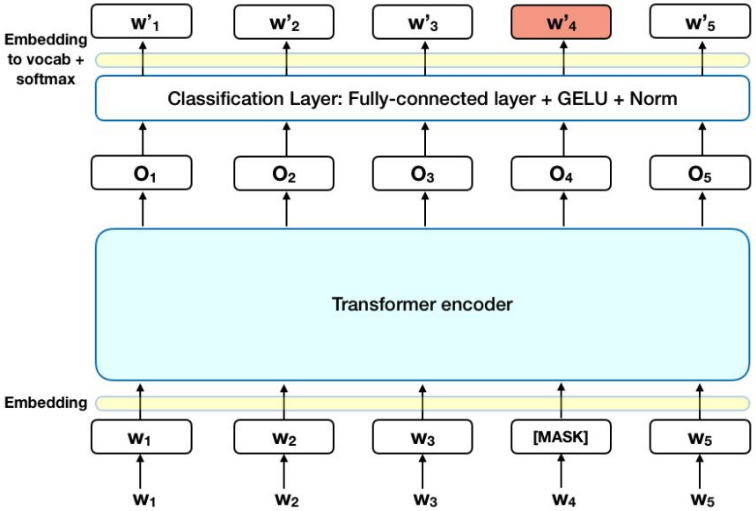
[1] Mathew, Binny, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. "Hate begets hate: A temporal study of hate speech." Proceedings of the ACM on Human-Computer Interaction 4, no. CSCW2 (2020): 1-24.

Fear speech detection

- **Baseline models**
  - Features - BOW, WE and TFIDF
  - Models - LR, SVM, XGBoost
- **Transformers**
  - Pretrained for e.g. BERT
  - Finetuned for e.g Hatexplain
  - MLM-Pre Trained for e.g GabBERT
- **Additional features**
  - Emotion vector



SVM with RBF kernel

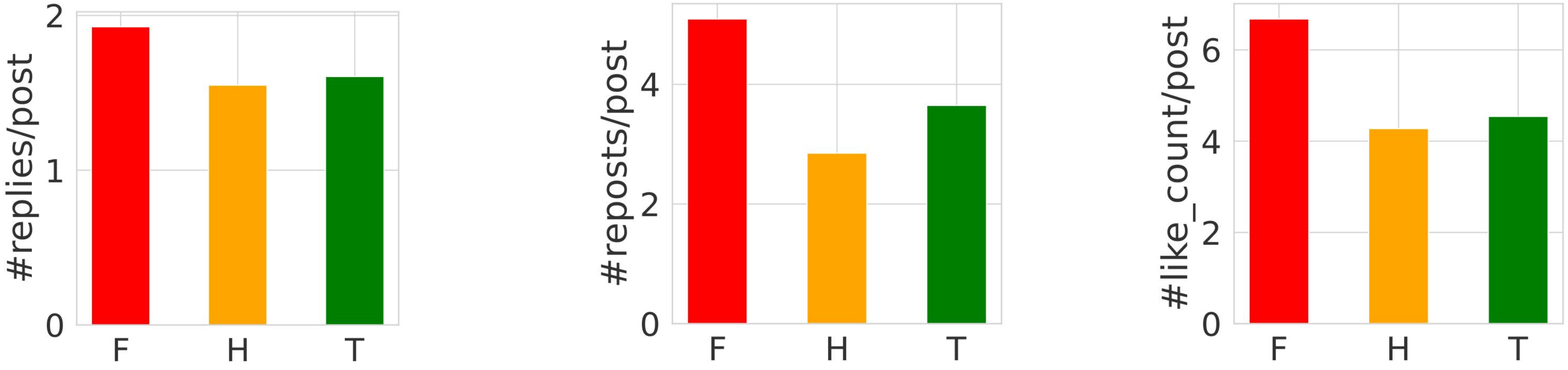


Scaled up dataset

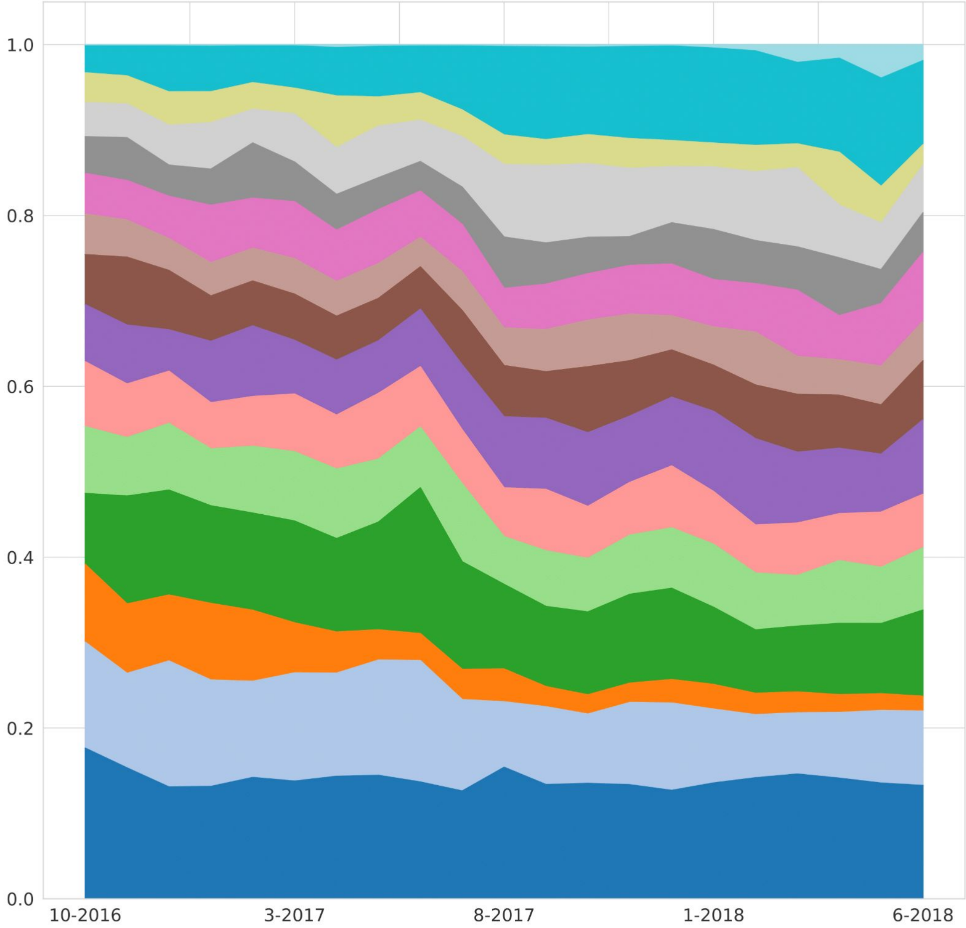
- We got the best performance by GabBERT and emotion vector of **0.63 f1 score**
- Applied this model on the whole dataset (**21M**) and got **400k** fear speech and **700k** hate speech
- We also selected **ExHate** and **ExFear** users (~500) based on the top 10 percentile of posting fear/hate speech.

Reactions on posts

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



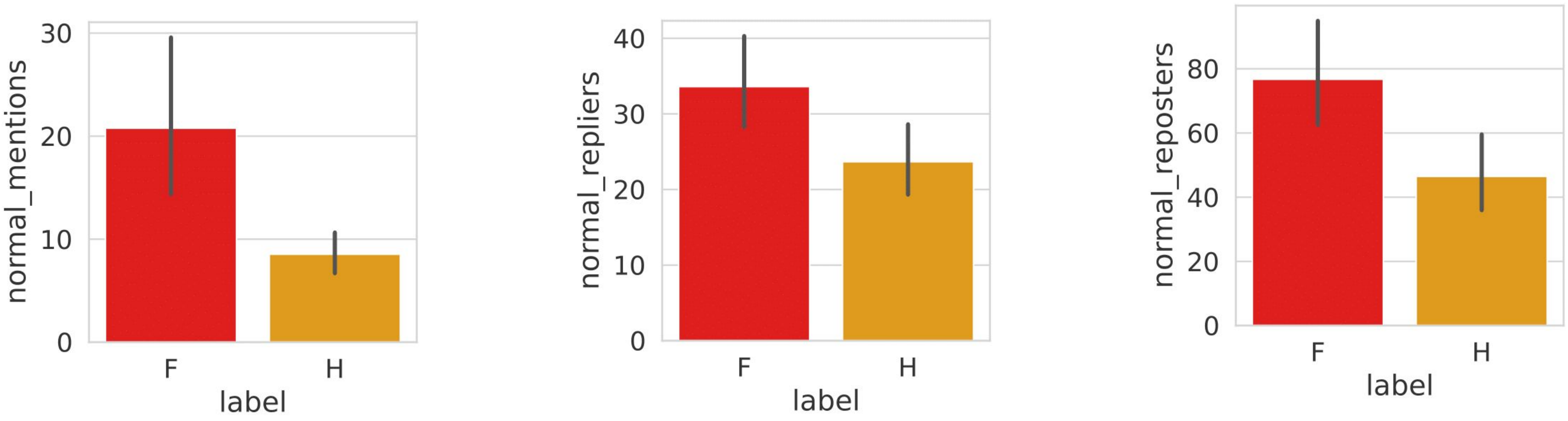
Temporal topics



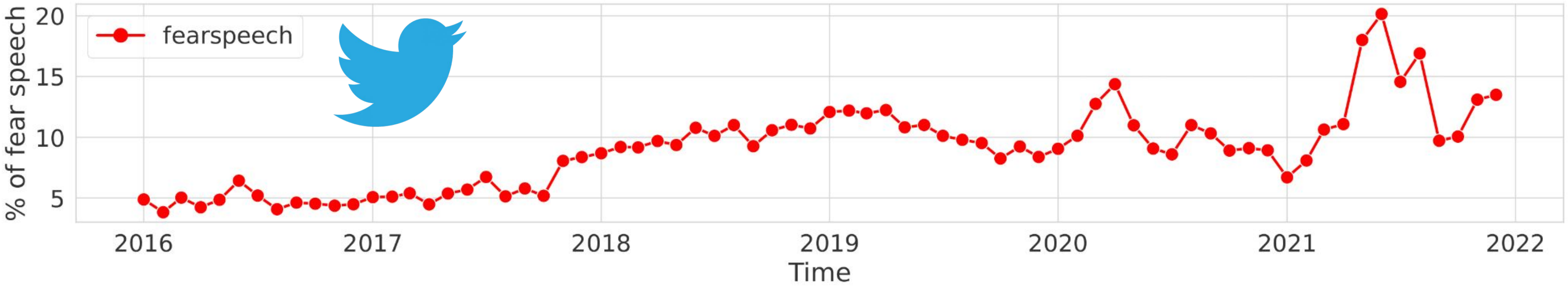
Topics in the fear speech mostly portrayed other communities as perpetrators in a subtle and argumentative style

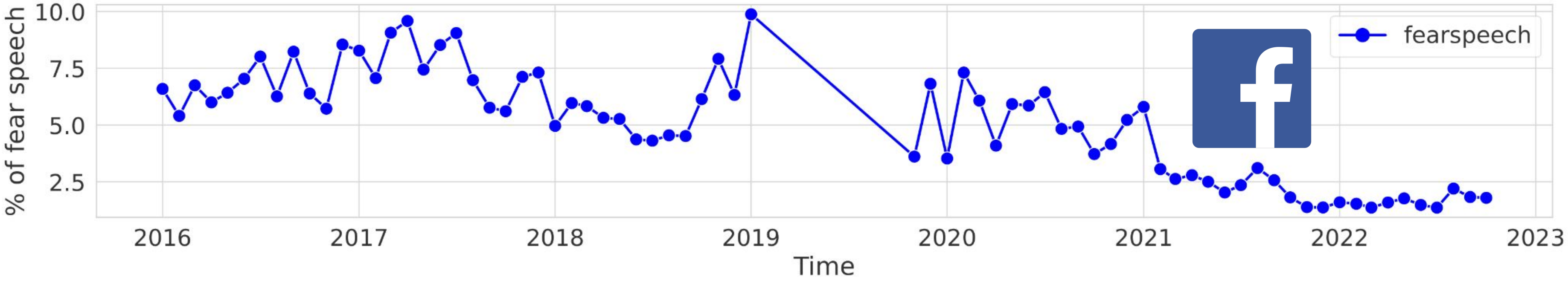
Effect on normal users?

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



What about other platforms?





In the wild users

- Task was to mark the more believable one.
- Created **100 pairs** of fear speech and hate speech from the dataset
- Each of them was judged by **9** annotators. **246** unique annotators took part in the task
- In 69% of the cases fear speech was more believable

What can be done?

- Need cross-disciplinary dialogue
  - Policy
  - Media
  - Technology
- Possible joint activities
  - **Educating the users to moderate content (making them socially responsible)**
  - **Laying out tangible policies of moderation**
  - **Improving existing technologies to implement such policies**

Summary

- We studied the idea of one form of harmful speech - in both US and Indian context
  - Content wise - subtle argumentative structure, emojis
  - User wise - affecting normal users more

Future plans

- Study more fine-grained structure in fear speech
- Study other forms of harmful speech like dangerous speech

Dataset and Code: https://github.com/hate-alert/Fear-speech-analysis
Paper: https://dl.acm.org/doi/10.1145/3442381.3450137

Table of contents

1 Introduction

2 Detection

3 Explanation

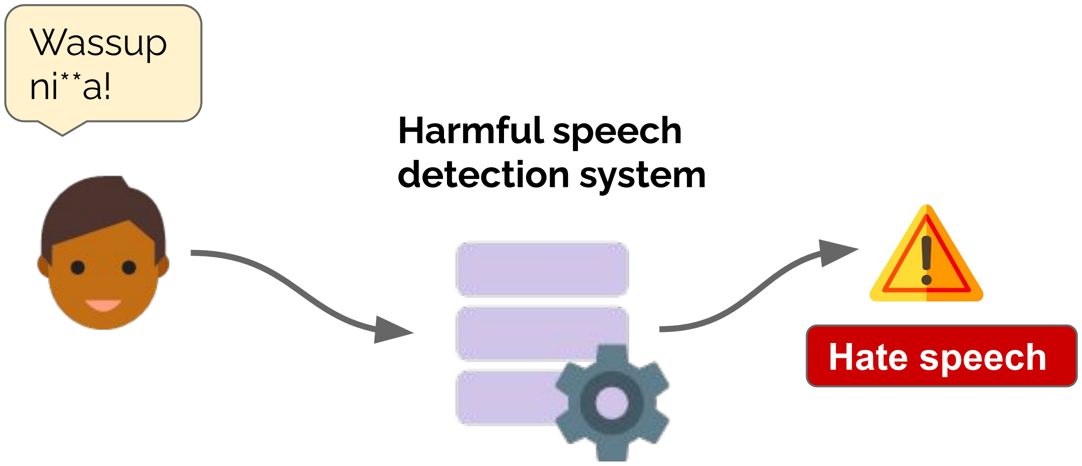
- Need for explanation?
- Hatexplain

4 Mitigation

5 Conclusion

HateXplain: A benchmark dataset for explainable hate speech detection (AAAI 2021)

Need for explanation



But the decisions are not explainable, hence might be difficult to rely on these machines

Research in hate speech

Dataset	Labels	Total size	Language	Target Labels ?	Rationales?
Waseem & Hovy '16	Racist, Sexist, Normal	16,914	English	❌	❌
Davidson et al. '17	Hate speech, Offensive, Normal	24,802	English	❌	❌
Founta et al. '18	Hate speech, Abusive, Normal, Spam	80,000	English, French Arabic	❌	❌
Ousidhoum et al. '19	five different aspects	13,000	English	✔️	❌

Research in hate speech

Dataset	Labels	Total size	Language	Target Labels ?	Rationales?
Waseem & Hovy '16	Racist, Sexist, Normal	16,914	English	✗	✗
Davidson et al. '17	Hate speech, Offensive, Normal	24,802	English	✗	✗
HateXplain '20	Hate speech, Offensive, Normal	20,148	English	✓	✓
Founta et al. '18	Hate speech, Abusive, Normal, Spam	80,000	English, French, Arabic	✗	✗
Ousidhoum et al. '19	five different aspects	13,000	English	✓	✗

Dataset sampling

- Collected data from **gab** and **twitter** using a lexicons
- Lexicon was created from three previous works.
- Gab - dataset created by previous work[1]
- Twitter - 1% random sample from January '19 to June '20.

[1] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of Hate Speech in Online Social Media. WebSci'19

Annotation framework

Each post in our dataset contains -

- Label
- Target
- Rationales

Text	Label
guess the ni**er have been to busy to kill off this mudsh**k.	**Hatespeech**
y is big baby davis a fa**ot on shameless doe.	**Offensive**
People act as if you can not say the same about the states obviously not all americans are pro guns not.	**Normal**

Annotation framework

Each post in our dataset contains -

- Label
- Target
- Rationales

Group	Categories
Race	**African**, **Arabs**, **Asians**, **Caucasian**, **Hispanic**
Religion	Buddhism, Christian, Hindu, **Islam**, **Jewish**
Gender	Men, **Woman**
Sexual Orientation	Heterosexual, **LGBTQ**
Miscellaneous	**Refugee**, Indigenous

*more than 100 posts

Annotation framework

Each post in our dataset contains -

- Label
- Target
- Rationales

Text: I guess the **ni**er** have been to busy to **kill off this mudsh**k**.

Average number of tokens is **~5** in rationales **out of ~23** in a post.

Top content words

Offensive - retarded, bitch and white.

Hate speech - ni**er, k*ke and m**lems.

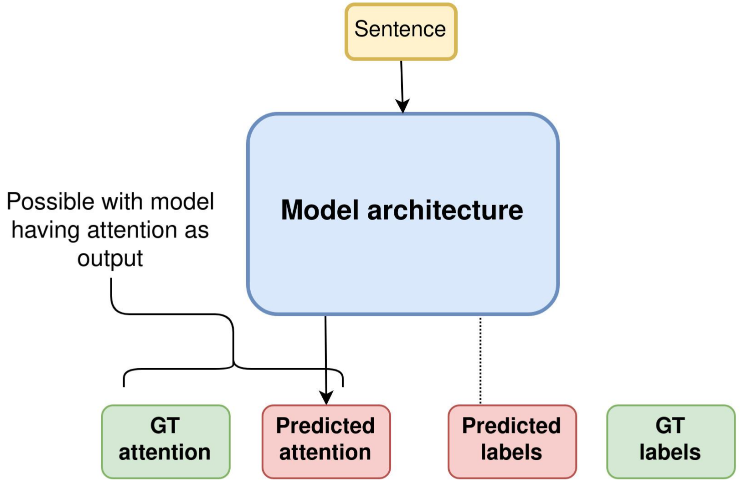
General framework

Models **without** attention supervision

- CNN-GRU
- BiRNN
- BiRNN-Attention
- BERT

Models **with** attention supervision

- **BiRNN-HateXplain**
- **BERT-HateXplain**

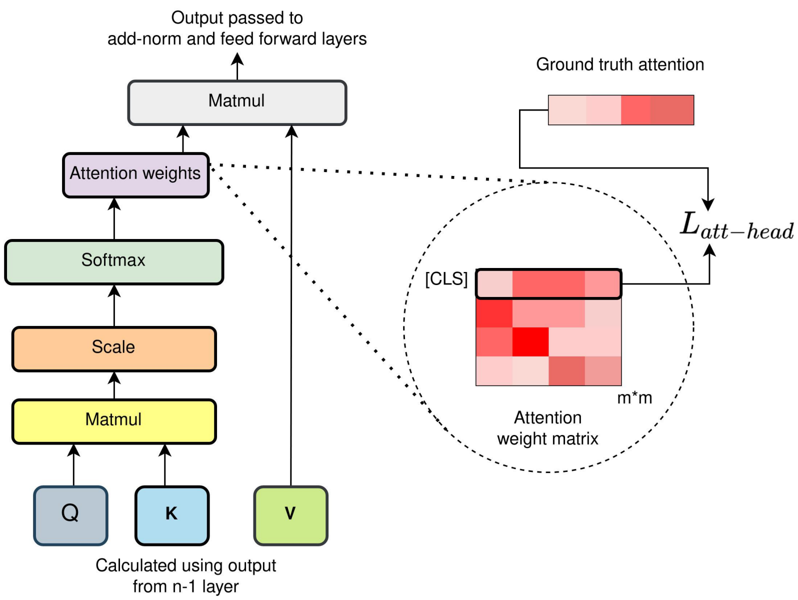


$$\underbrace{\text{GT attention} \quad \text{Predicted attention}}_{\lambda * L_{att}} \qquad \underbrace{\text{Predicted labels} \quad \text{GT labels}}_{L_{pred}}$$

$$L_{total} = L_{pred} + \lambda * L_{att}$$

Attention supervision

- **BiRNN-HateXplain**
  Cross entropy of attention weights and ground truth rationales.
- **BERT-HateXplain**
  12 layers, each having 12 heads.
  We can control which layer and how many heads to supervise



Performance Results

Models	Accuracy	F1 Score	AUROC
CNN-GRU	0.627	0.606	0.793
BiRNN	0.595	0.575	0.767
BiRNN-Attn	0.621	0.614	0.795
BiRNN-HateXplain	0.629	0.629	0.805
BERT	0.690	0.674	0.843
BERT-HateXplain	**0.698**	**0.687**	**0.851**

Bias Results

Models	GMB-Sub	GMB-BPSN	GMB-BNSP
CNN-GRU	0.654	0.623	0.659
BiRNN	0.640	0.604	0.671
BiRNN-Attn	0.653	0.662	0.668
BiRNN-HateXplain	0.691	0.636	0.674
BERT	0.762	0.709	0.757
BERT-HateXplain	**0.807**	**0.745**	**0.763**

Models which use rationales while training **perform better** and has **less unintended bias**

Explainability results

Models	Plausibility			Faithfulness	
	IOU F1	Token F1	AUPRC	Comp.	Suff.
CNN-GRU [LIME]	0.167	0.385	0.648	0.316	**-0.082**
BiRNN [LIME]	0.162	0.361	0.605	0.421	-0.051
BiRNN-Attn [Attn]	0.167	0.369	0.643	0.278	0.001
BiRNN-Attn [LIME]	0.162	0.386	0.650	0.308	-0.075
BiRNN-HateXplain [Attn]	**0.222**	**0.506**	**0.841**	0.281	0.039
BiRNN-HateXplain [LIME]	0.174	0.407	0.685	0.343	-0.075
BERT [Attn]	0.130	0.497	0.778	0.447	0.057
BERT [LIME]	0.118	0.468	0.747	0.436	0.008
BERT-HateXplain[Attn]	0.120	0.411	0.626	0.424	0.160
BERT-HateXplain[LIME]	0.112	0.452	0.722	**0.500**	0.004

Models do not fare well in terms of **model interpretability**

Model	Text	Label
Human Annotator	The jews are again using holohoax as an excuse to spread their agenda .Hilter should have eradicated them	**Hatespeech**
BiRNN-Attn	The jews are again using holohoax as an excuse to spread their agenda .Hilter should have eradicated them	**Hatespeech**
BiRNN-HateXplain	The jews are again using holohoax as an excuse to spread their agenda .Hilter should have eradicated them	**Hatespeech**
BERT	The jews are again using holohoax as an excuse to spread their agenda .Hilter should have eradicated them	**Offensive**
BERT-HateXplain	The jews are again using holohoax as an excuse to spread their agenda .Hilter should have eradicated them	**Offensive**

Human

Only model found important

Both model and human found important

Summary

- Discussed why explainability is important
- We created a new dataset for benchmarking explainable hate speech detection

Future works

- We need to look towards other forms of explanation - free form explanations.
- How can we use these rationales and improve on other datasets?

Data & Code repository : https://github.com/hate-alert/HateXplain

Table of contents

1 Introduction

2 Detection

3 Explanation

4 Mitigation

5 Conclusion

- Counterspeech as a response
- Counterspeech generation task - Better and diverse generation

Mitigation of harmful speech

Inaction: By not responding to the harmful speech.
Deletion: Deleting or Suspending the user account is the most common way used by online platforms such as Facebook and Twitter.
Counterspeech: Directly intervening with textual response that counter the harmful-content.

Mitigation of harmful speech

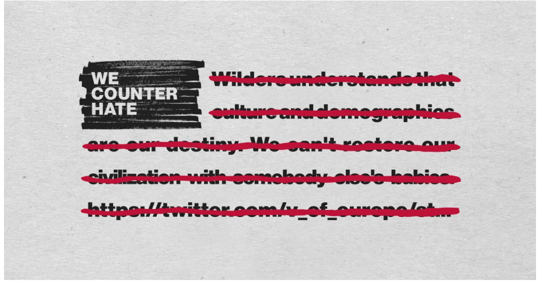
Counterspeech: Directly intervening with textual response that counter the hate-content.

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.



[WeCounterHate]

[NoHateSpeechMovement]

Mitigation of harmful speech

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

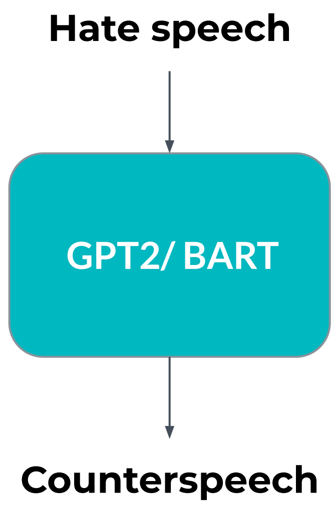
Why counter speech?

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

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



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**?

Counterspeech generation

Reference	Contribution
de los Riscos, Agustín Manuel, and Luis Fernando D'Haro. "Toxicbot: A conversational agent to fight online hate speech." In Conversational dialogue systems for the next decade, pp. 15-30. Springer, Singapore, 2021.	Generates counter speech after detecting hate speech
Pranesh, Raj Ratn, Ambesh Shekhar, and Anish Kumar. "Towards automatic online hate speech intervention generation using pretrained language model." (2021).	Finetuning on counter speech generation dataset
Zhu, Wanzheng, and Suma Bhat. "Generate, Prune, Select: A Pipeline for Counterspeech Generation against Online Hate Speech." In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 134-149. 2021.	Additional classifiers with finetuning to select more relevant examples

Counterspeech generation

Fine tuning the generation models with **hatespeech-counterspeech** pairs.

But ..

- Writing counterspeech is hard.
- We don’t have any control over what we generate.

Counterspeech generation

Fine tuning the generation models with **hatespeech-counterspeech** pairs.

But ..

- Writing counterspeech is hard.
- We don’t have any control over what we generate. **Can we add additional control to make the framing of counterspeech better?**

CounterGeDi: A controllable approach to generate polite, detoxified and emotional counterspeech (IJCAI 2022, AI for social good)

Add control to counterspeech datasets ?

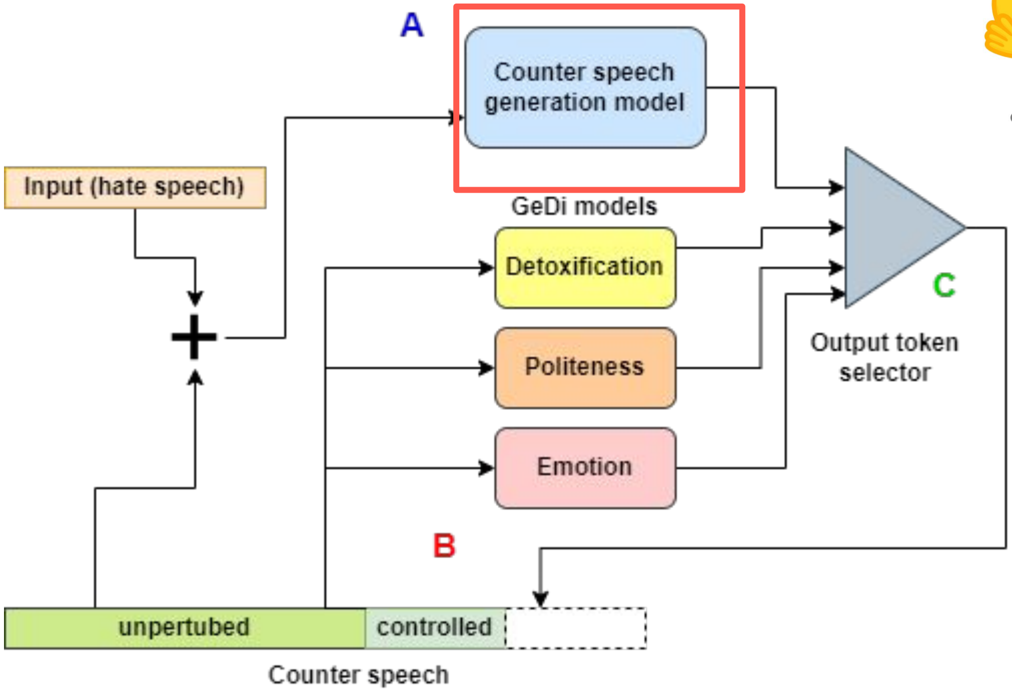
Dataset	Annotators	Unique hate speech	Source of hate	Target Community
Qian et al., ‘19	Crowdworkers	3,847	**REDDIT**	**Mixed**
		11,169	**GAB**	**Mixed**
Chung et al. ‘19	Expert Annotators	408	**SYNTHETIC**	**Muslims**

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.**

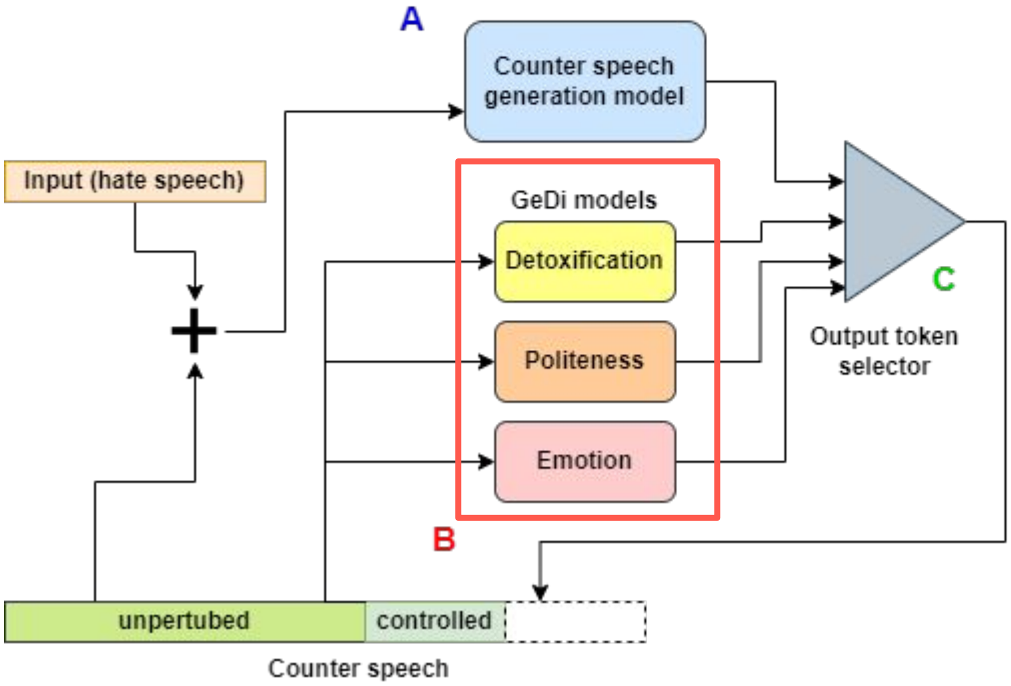
Our proposal - COUNTERGEDI

DialoGPT

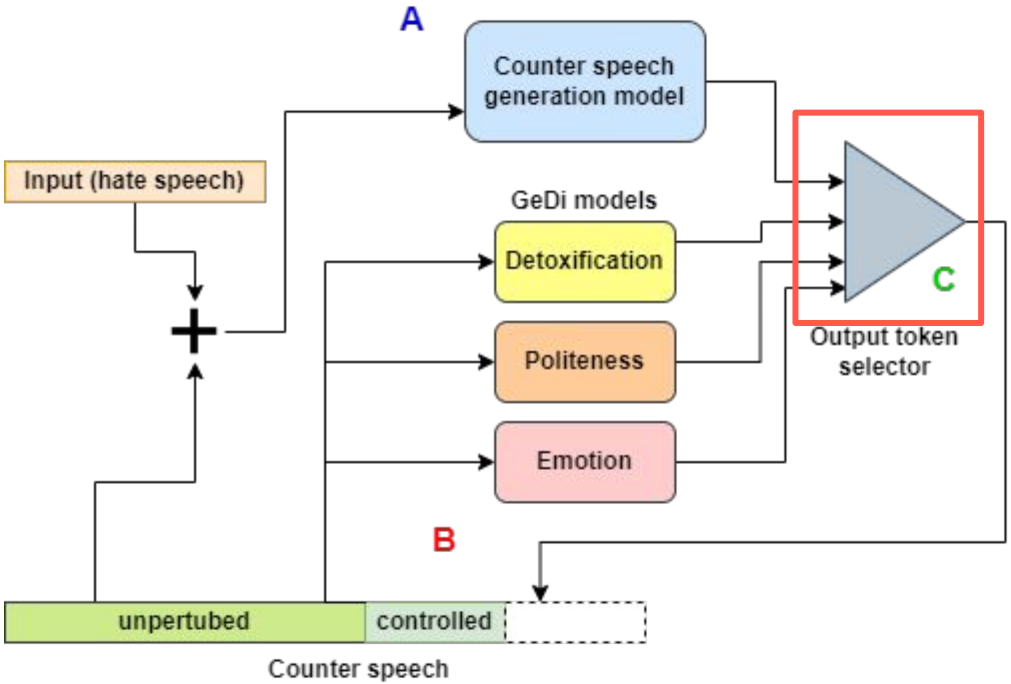
huggingface.co/**microsoft**



Our proposal - COUNTERGEDI



Our proposal - COUNTERGEDI



Controllable text generation

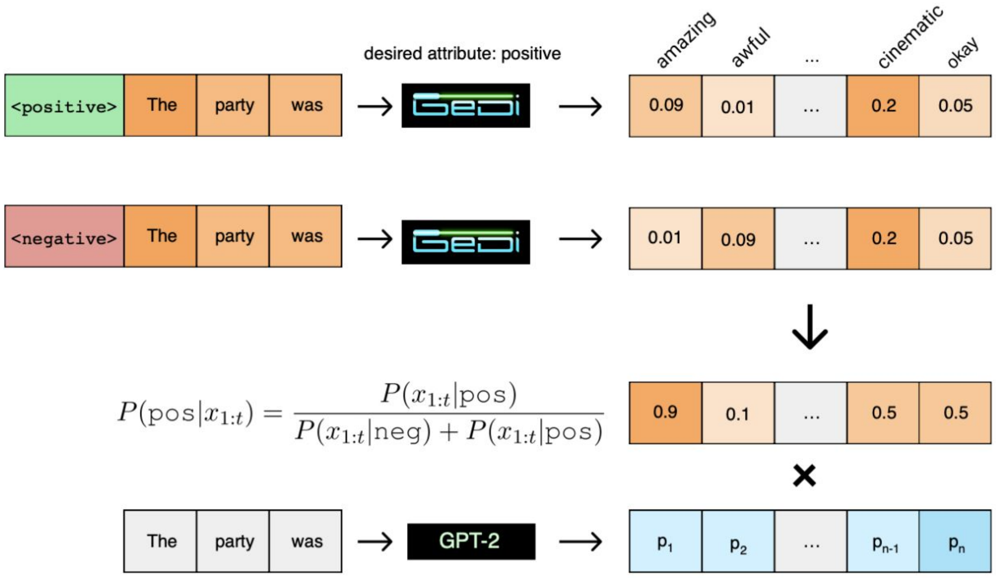
- We steer the generation model to contain certain quality attributes such as:
  - **Emotion** - Generating more diverse responses catering to large number of communities.
    - **Sadness** - Show affiliation with the targeted communities.
    - **Joy** - Convey positivity in the counterspeech.
    - **Anger** - Express disagreement with the speaker.
  - **Politeness** - Toward more empathetic counterspeech.
  - **Detoxification** - Minimize hostile behaviour (slur words) in generated responses.

Attribute datasets

Dataset	+ve	-ve	T_r (%+ve)	V (%+ve)	T_e (%+ve)
Polite	p	n-p	1.12M (20%)	137k (20%)	137k (20%)
Toxic	t	n-t	143k (10%)	16k (10%)	153k (4%)
Emotion	j	o	333k (34%)	42k (34%)	42k (34%)
Emotion	f	o	333k (11%)	42k (11%)	42k (11%)
Emotion	s	o	333k (29%)	42k (29%)	42k (29%)
Emotion	a	o	333k (14%)	42k (14%)	42k (14%)

This table shows the attribute datasets, positive and negative classes and data present in train, validation and test part for each. T_r: Train, V: Validation, T_e: Test, p: polite, n-p: non-polite, t: toxic, n-t: non-toxic, s: sadness, j: joy, a: anger, f: fear, o: others. The % associated with the T_r, V and T_e are the % of positive labels.

GEDI: Generative Discriminator Guided Sequence Generation



Attribute control

Trained separate GEDI Models for each controllable parameter

Single-attribute control

- Combined single attribute GEDI Model with Fine-tuned base Dialo-GPT model.

Multi-attribute control

- Combined of several single attribute GEDI Models with Fine-tuned base Dialo-GPT model.
- **Equal Weights** provided for each attribute while combining probabilities at the time of generation

Experimental setup

Models considered for Experiments:

- **Baseline 1: Generate, prune, select (GPS)**
  - A three stage pipelined approach for counterspeech generation, Zhu and Bhat [2021]
- **Baseline 2: Dialo-GPT fine-tuned base model**
  - Used a variant of the GPT model - Dialo-GPT Zhang et al. [2020]
  - Fine-tuned on respective datasets: CONAN, Reddit and Gab
- **CounterGEDI: Single attribute and multi-attribute**
  - Our model with GEDI models trained for different controlling attributes
  - Generation performed with single and multi-attribute combination

Metrics

- **Generation metrics**
  - Novelty, Diversity, BLEU (relevancy) and COLA (Fluency)
- **Controller metrics :** We use third-party classifiers for evaluating each attribute.
  - **Politeness :** Trained a bert-base-uncased model for politeness detection[1].
  - **Emotion :** Used the Ekman version of Go-Emotions Model[2] .
  - **Detoxification :** Evaluated using HateXplain[3] Model's confidence for the toxic class.

[1]https://github.com/AlafateABULIMITI/politeness-detection
[2]https://huggingface.co/monologg/bert-base-cased-goemotions-ekman
[3]https://huggingface.co/Hate-speech-CNERG/bert-base-uncased-hatexplain-rationale-two

Baselines

Model	B2 (↑)	COLA (↑)	M (↑)	N (↑)	D (↑)
CONAN					
GPS	**41.5**	**0.82**	0.14	0.18	0.60
DialoGPTm	12.7	0.78	**0.18**	**0.84**	**0.80**
Reddit					
GPS	**14.1**	**0.82**	0.11	0.30	0.47
DialoGPTm	6.9	0.75	**0.17**	**0.82**	**0.74**
Gab					
GPS	**13.9**	**0.82**	0.12	0.15	0.41
DialoGPTm	7.7	0.80	**0.17**	**0.80**	**0.72**

Evaluation results for the three datasets. We report BLEU-2 (B2), COLA, METEOR (M), novelty (N) and diversity (D) to compare the two baselines: generate-prune-select (GPS) framework and DialoGPTm. For all metrics, higher is better and **bold** denotes the best scores.

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	<u>0.70</u>	3.56	0.07 (S)	13.2	0.74
Fear(F)+P+D	<u>0.70</u>	<u>4.00</u>	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	<u>4.24</u>	0.19 (A)	**8.3**	0.72
Sad+P+D	<u>0.87</u>	3.56	0.09 (S)	8.2	0.79
Fear+P+D	<u>0.87</u>	4.00	0.01 (F)	7.8	0.79
Gab					
Joy+P+D	**0.87**	<u>5.68</u>	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	<u>0.86</u>	**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.

Human evaluation

Human evaluation: Judgement on a scale of 0 to 5 for a particular control attribute. We observe an improvement in most of the attribute scores for the controlled model over the two baselines.

Model	Polite (↑)	Joy (↑)	Anger (↑)	Sad (↑)	Fear (↑)
CONAN					
GPS	0.50	1.30	2.50	**1.00**	0.00
DGPTm	0.59	**2.50**	3.00	0.75	0.75
DGPTm-c	**2.00**	1.00	**4.00**	**1.00**	**2.00**
Reddit					
GPS	1.83	0.93	1.50	0.33	0.36
DGPTm	2.66	2.50	1.50	0.66	**1.33**
DGPTm-c	**3.50**	**3.33**	**2.00**	**2.00**	1.25
Gab					
GPS	1.56	1.28	0.81	0.4	0.17
DGPTm	2.17	2.50	1.66	1.11	0.89
DGPTm-c	**3.21**	**2.92**	**1.90**	**2.03**	**1.00**

Average human judgement scores (scale 0-5) for each of the models – GPS, DialoGPTm and controlled DialoGPTm (DGPTm). Each column represents the attribute that DialoGPTm-c (DGPTm-c) is controlled for. For all the metrics, higher is better and **bold** indicates best scores.

Controlled examples

Model	Text
(Instance)	wats wrong with being a commie jew? don't you want that high-iq baby inside your nazi womb?.
Dialo-GPT (joy)	there is no need to degrade an entire religion based on the actions of a few people here, please moderate your speech outside of this space and think about how using such language could be hurtful or even unfunny. **thank you! happy new year**!
Dialo-GPT (politeness)	its inappropriate to use this word when referring specifically to muslims and can be viewed as racist. **please do not** use these terms or consider them hurtful. thank you for using better language in the future.
Dialo-GPT (sadness)	making fun of nationality, religion and sexual preference is not kosher in todays world. try rereading them both times just in case they are harmful for you or anyone! first time using words used here makes those **hurt even more**.

Summary

- We discussed how counter speech can be an important mitigation strategy
- Discussed challenges of the counter speech dataset and generation

Future work

- Creation of the diverse counterspeech dataset while maintaining higher quality.
- How to improve counterspeech generation without using a lot of annotated data ?

Table of contents

1 Introduction
2 Detection
3 Explanation
4 Mitigation
5 Conclusion

Conclusion

- Discussed the problem of harmful speech and our plan to improve the moderation pipeline using detection, explanation and mitigation.
- **Detection** - Building the framework for detection of fear speech and studying it in the context of Whatsapp and Gab.
- **Explainability** - Building a benchmark dataset for explainable hate speech detection.
- **Mitigation -** Building a framework for controlling counterspeech generation.

Thanks !

Do you have any questions?

@punyajoysaha

punyajoys@iitkgp.ac.in

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