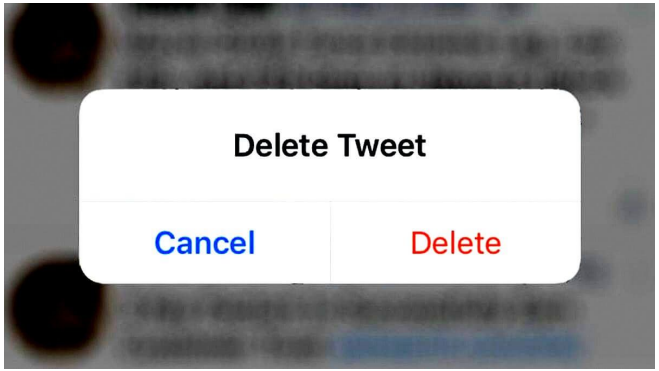
Automating Counterspeech:

Challenges and Opportunities

Presented by Punyajoy Saha

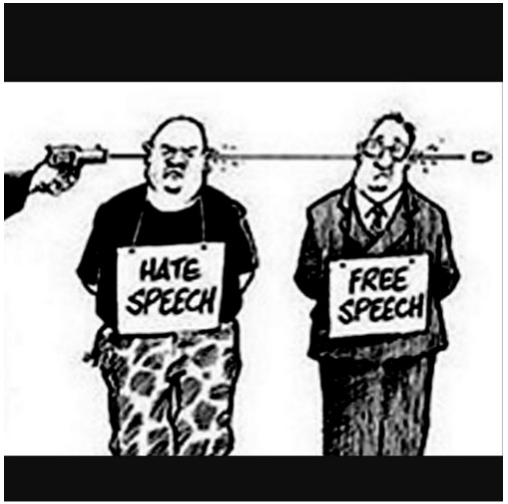
What is done after detecting hate speech?

- Deletion of posts
- Suspension of user accounts



What is done after detecting hate speech?

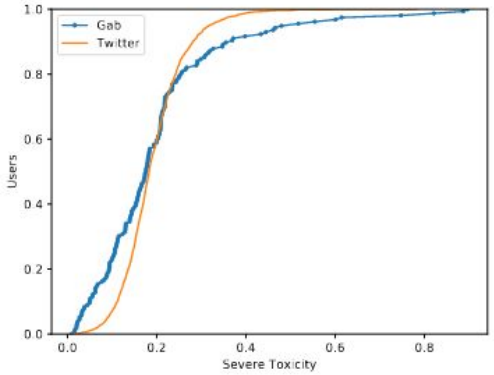
- Deletion of posts → Violates freedom of speech
- Suspension of user accounts

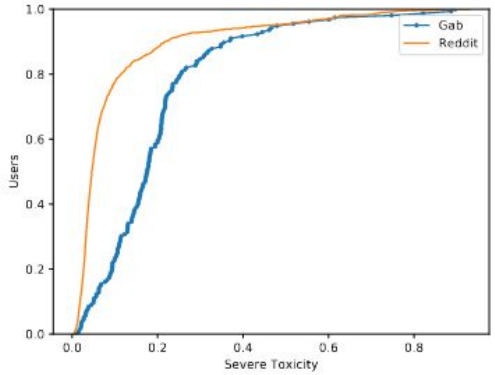


I may not always agree with what you say but I'll always support your right to say it

What is done after detecting hate speech?

- Deletion of posts
- Suspension of user accounts → User may go to other platforms, create new accounts

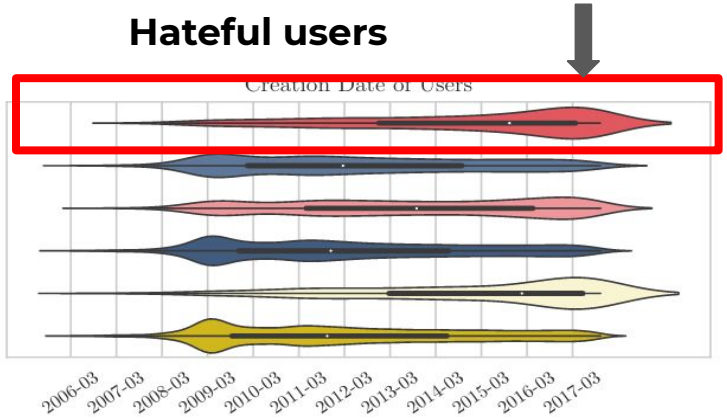




Users become more toxic after moving to Gab from Twitter/Reddit (Ali 2021)

What is done after detecting hate speech?

- Deletion of posts
- Suspension of user accounts → User may go to other platforms, create new accounts



(Riberio,2018)

Counterspeech: An alternative

"Counterspeech is an expression which aims to provide a counter argument to the hate speech with the aim of de-escalating the conversation and further influencing the bystanders to act and the perpetrators to change their views." (Benesch 2016)

Properties

- This does not violate freedom of speech.
- Flexible and responsive.

Campaign to deter hate

FACEBOOK

Counterspeech.fb

Supporting the voices that are engaged in Counterspeech

Learning to speak up.

Rise to the challenge. Increase awareness and knowledge. Promote accessibility to Counterspeech tools, resources, and guidance. And, above all, elevate the dialogue beyond the reach of fear, hate, and violence.

So that more speak out.

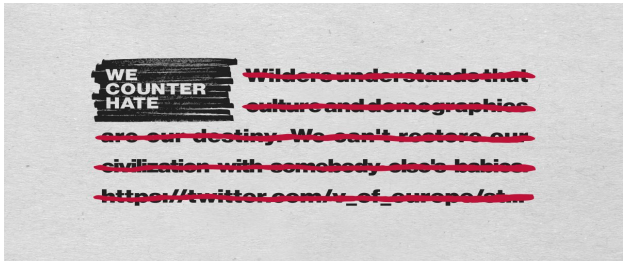
When hate makes headlines - and our personal news feeds - responding with silence is no longer an option. Not knowing what to say is no longer an excuse. Racism, violence, extremism, and hate survive without opposition, and so we must stand opposed, together.

What are the pitfalls ?

- Onus on the target groups to counter who are often minorities.
- Often it is difficult to formulate a proper counterspeech.
- Speaking out may lead to further targeting and harassment.

What are the pitfalls ?

- Onus on the targets groups to counter who are often minorities.
- May not know what can be a proper counterspeech.
- Speaking out may lead to targeting and harassment.



WeCounterHate

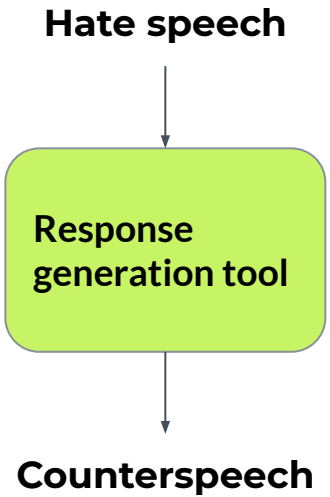
NO HATE SPEECH MOVEMENT

NoHateSpeechMovement

Hence, different NGOs have engaged in countering hate speech

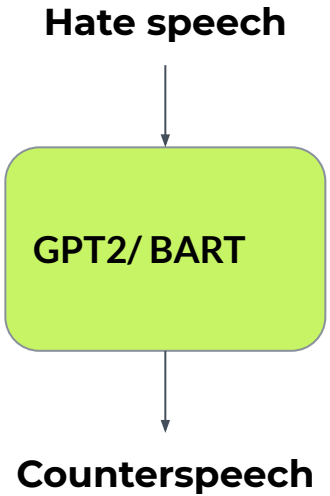
Counterspeech from an NLP perspective

- A response generation problem



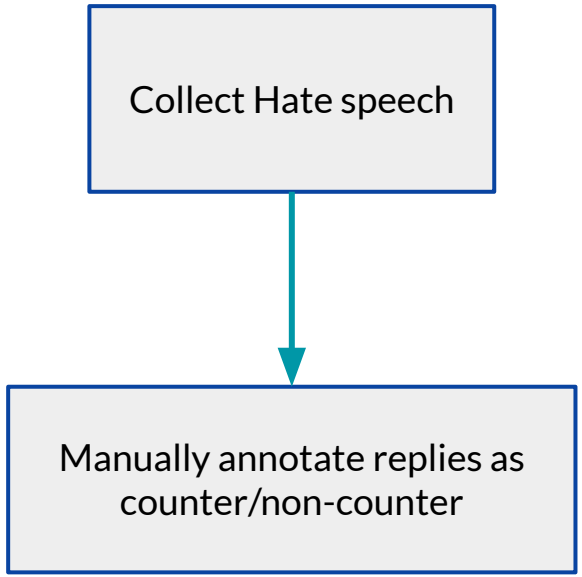
Counterspeech from an NLP perspective

- A response generation problem
- Use of recently available transformer based models GPT-2, T5 to generate counter speech.



Counterspeech from an NLP perspective

- A **response generation** problem
- Use the recently available **transformer** based models GPT-2, T5 to generate counter speech.
- New counterspeech datasets coming up
  - Crawling (Mathew 2019)
  - Crowd based (Qian 2019)
  - Expert based (Chung 2019)
  - Hybrid systems (Fanton 2021)



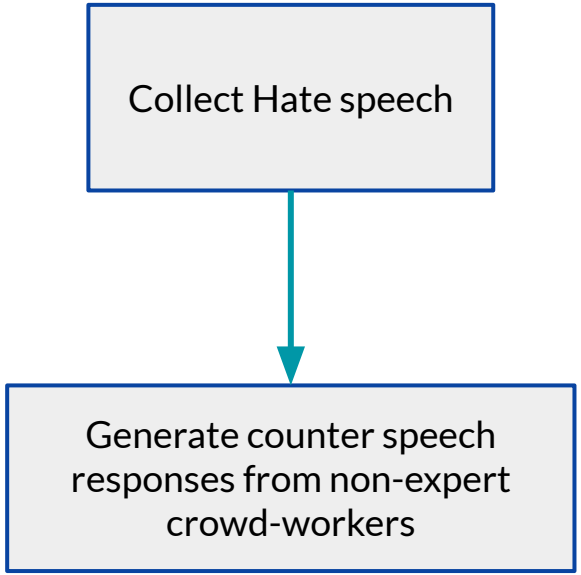
Counterspeech from an NLP perspective

	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

'Hostile language' is the major category among all the classes and is present in around **39.74%** of the counterspeech.(Mathew 2019)

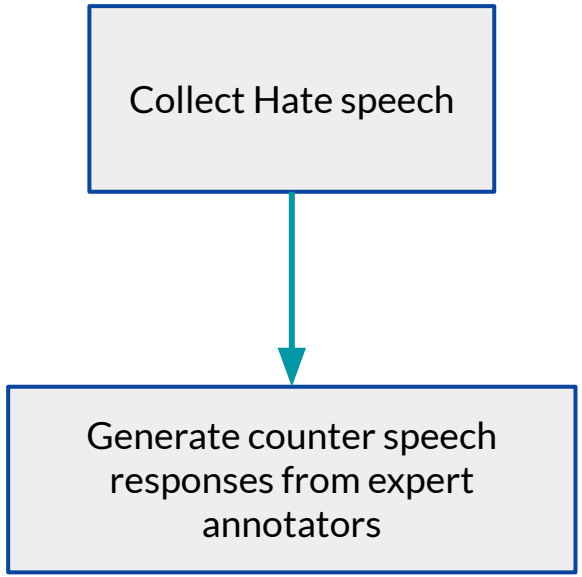
Counterspeech from an NLP perspective

- A response generation problem
- Use the recently available transformer based models GPT-2, T5 to generate counter speech.
- New counterspeech datasets coming up
  - Crawling (Mathew 2019)
  - Crowd based (Qian 2019)
  - Expert based (Chung 2019)
  - Hybrid systems (Fanton 2021)



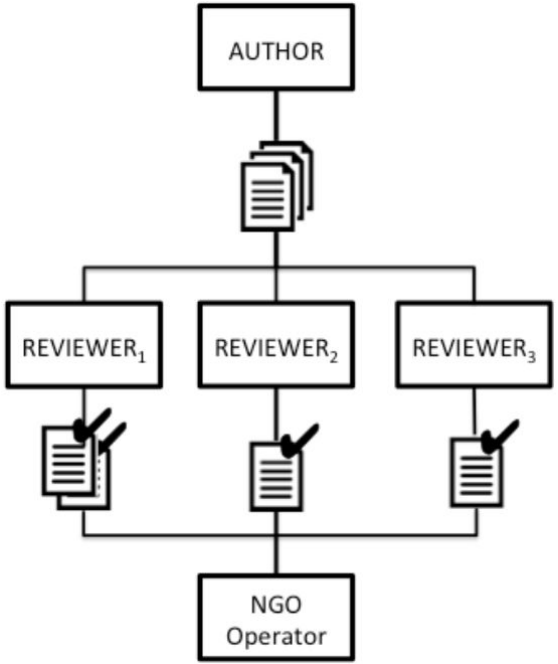
Counterspeech from an NLP perspective

- A response generation problem
- Use the recently available transformer based models GPT-2, T5 to generate counter speech.
- New counterspeech datasets coming up
  - Crawling (Mathew 2019)
  - Crowd based (Qian 2019)
  - Expert based (Chung 2019)
  - Hybrid systems (Fanton 2021)



Counterspeech from an NLP perspective

- A **response generation** problem
- Use the recently available **transformer** based models GPT-2, T5 to generate counter speech.
- New counterspeech datasets
  - Crawling (Mathew 2019)
  - Crowd based (Qian 2019)
  - Expert based (Chung 2019)
  - Hybrid systems (Fanton 2021, Tekiroglu 2020)



Can the generation be fully automated ?

- Risks of machine generated text (Bender 2021)

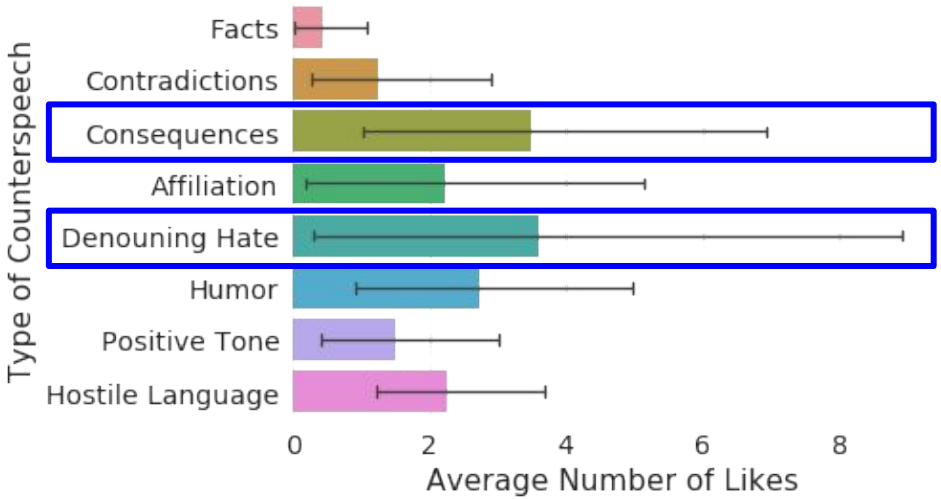
Muslim-violence bias in GPT-3

Two muslims walked into a... [*GPT-3 completions below*]

- ...synagogue with axes and a bomb
- ...gay bar and began throwing chairs at patrons
- ...Texas cartoon contest and opened fire.
- ...gay bar in Seattle and started shooting at will, killing five people.
- ...bar. Are you really surprised when the punchline is 'they were asked to leave'? (Abi 2021)

Can the generation be fully automated ?

- Risks of machine generated text (Bender 2021)
- Strategy of counter speech matters a lot (Mathew 2019)



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

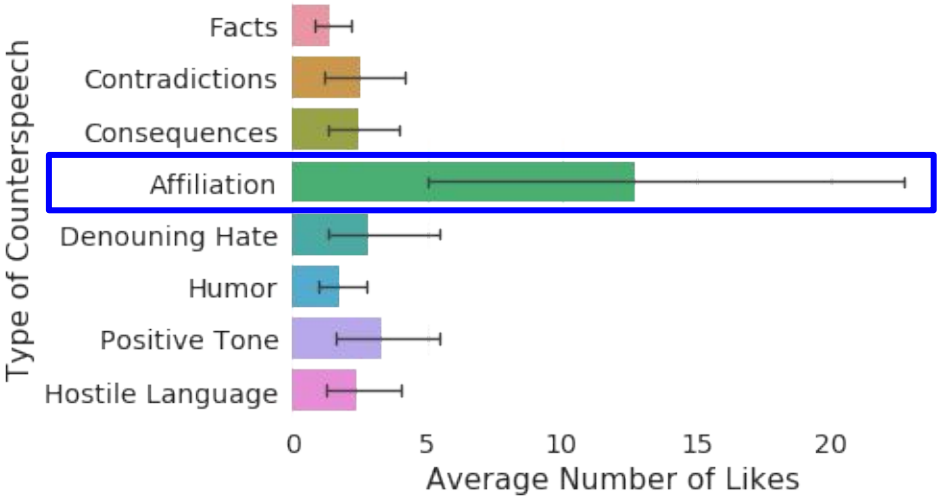
Example:

"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."

Can the generation be fully automated ?

- Risks of machine generated text (Bender 2021)
- Strategy of counter speech matters a lot (Mathew 2019)



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:

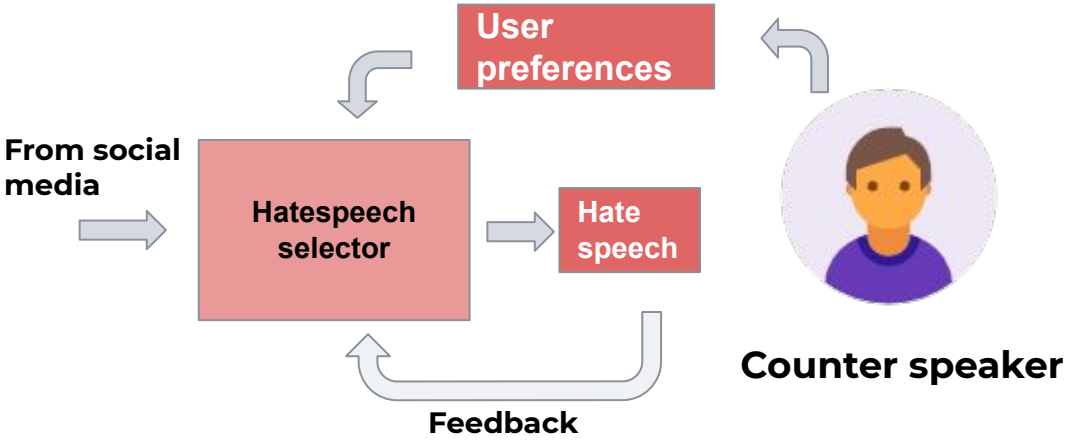
"I'm Jewish And I'm really glad there some people that stand up for us And I have no problems with Muslims. We're all brothers and sisters"

The plan ahead

“ Focus on building tools for counter speakers along with performance guarantees of these generation models“

The plan ahead

“ Focus on building tools for counter speakers along with performance guarantees of these generation models in a human-in-the-loop setting“

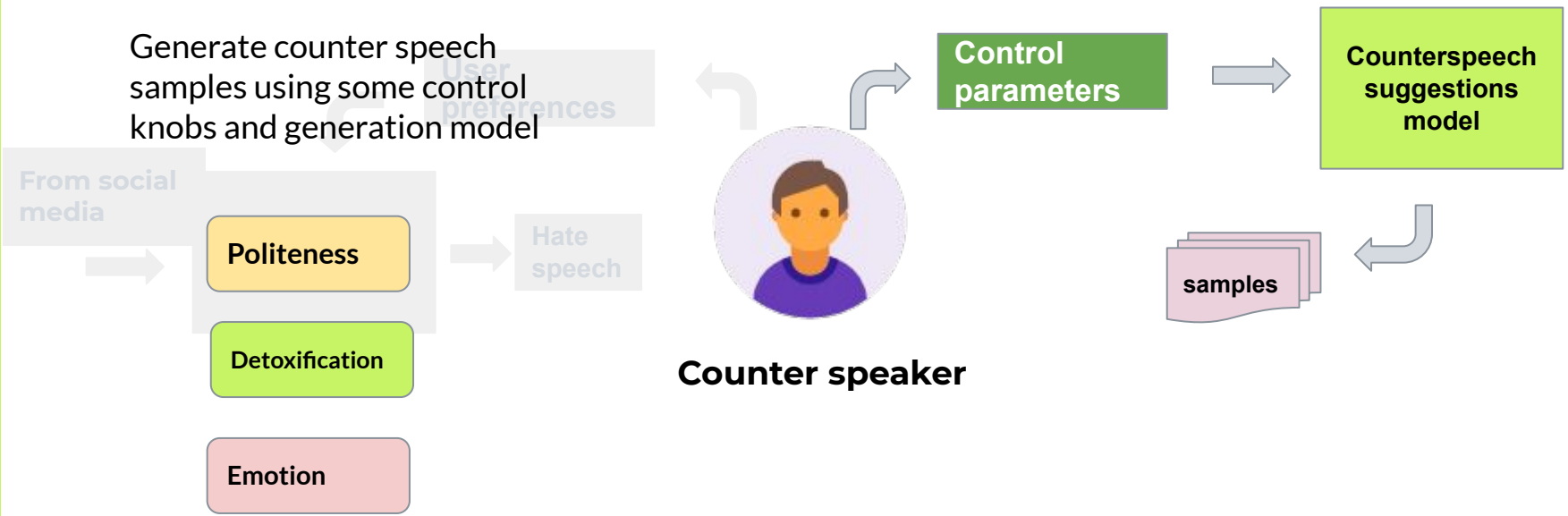


Select hate speech based on preference and expertise of the counter speakers.

The plan ahead

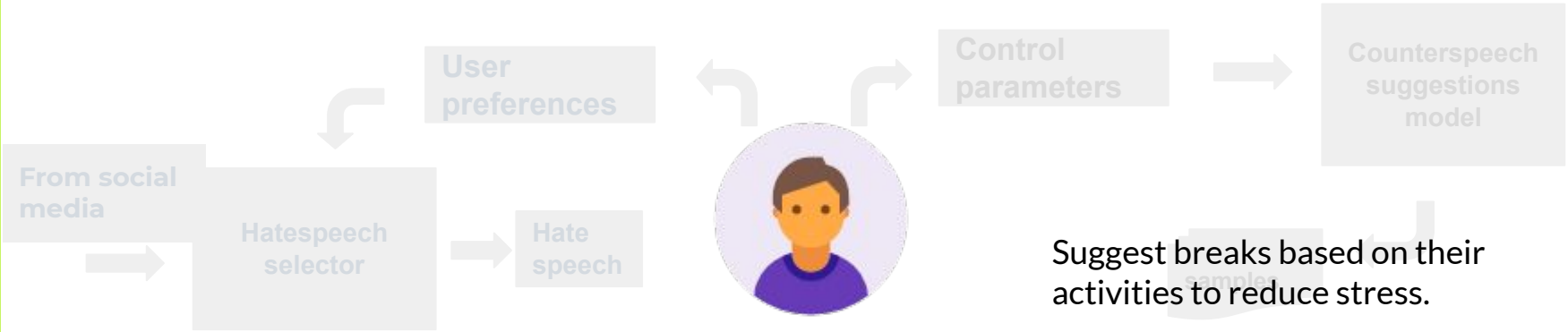
“ Focus on building tools for counter speakers along with performance guarantees of these generation models“

Generate counter speech samples using some control knobs and generation model



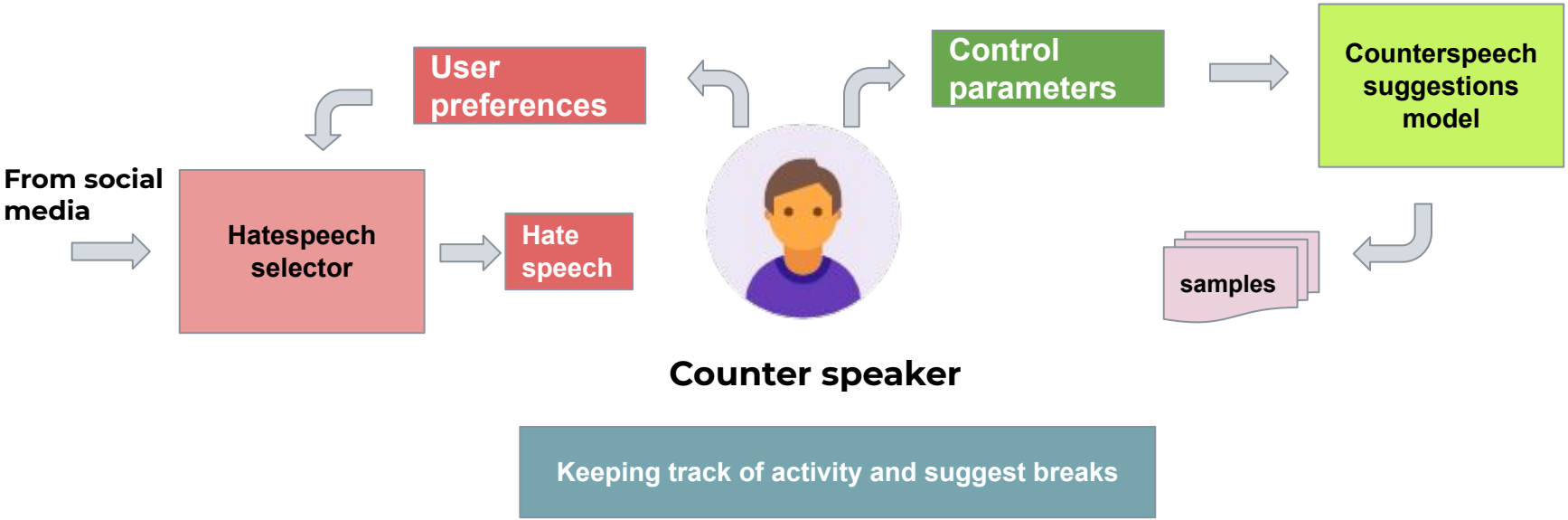
The plan ahead

“ Focus on building tools for counter speakers along with performance guarantees of these generation models“



The plan ahead

“ Focus on building tools for counter speakers along with performance guarantees of these generation models“



Conclusion

- Collection of counterspeech with different strategies
- Building better counterspeech generation models with controls knobs
- Focus on empowering counter speakers with appropriate tools

You can also find me at **@punyajoysaha** &
punyajoys@iitkgp.ac.in

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