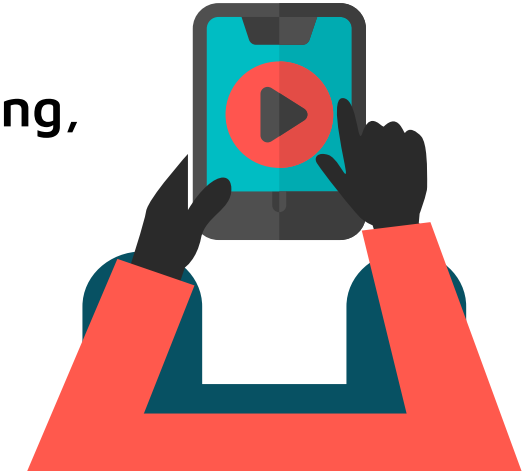
Echoes of Fear: Unraveling the Presence of Fear Speech in Social Media Platforms

by
Punyajoy Saha

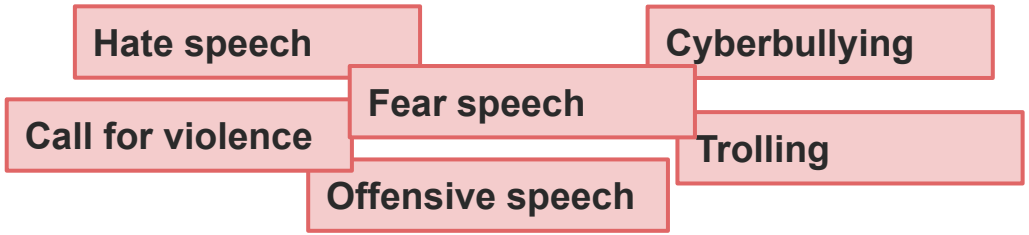
Dept. of Computer Science & Engineering,
IIT Kharagpur



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

Harmful speech

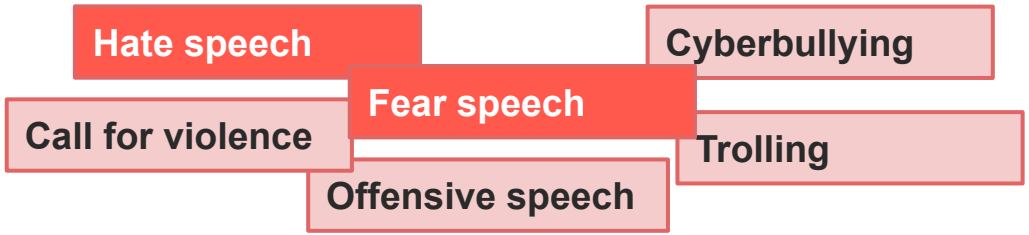
Harmful speech consists of a range of phenomenon that often overlap and intersect, and includes a variety of types of speech that cause different harms.



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

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

Overview

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

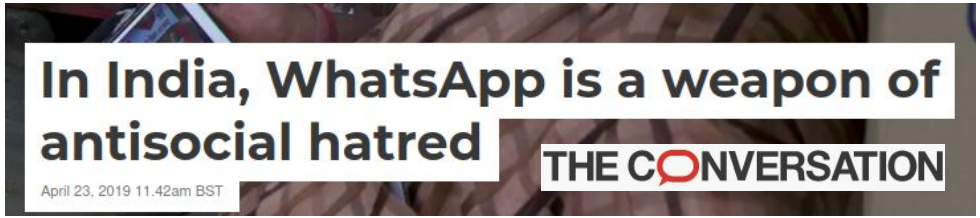
Our work operationalises the *fear speech* definition and performs a quantitative analysis on a social media platform

"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.



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).

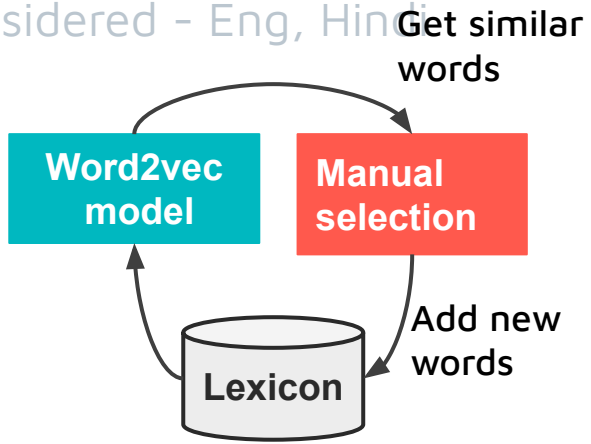
Data collection

- 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, Hindi (**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

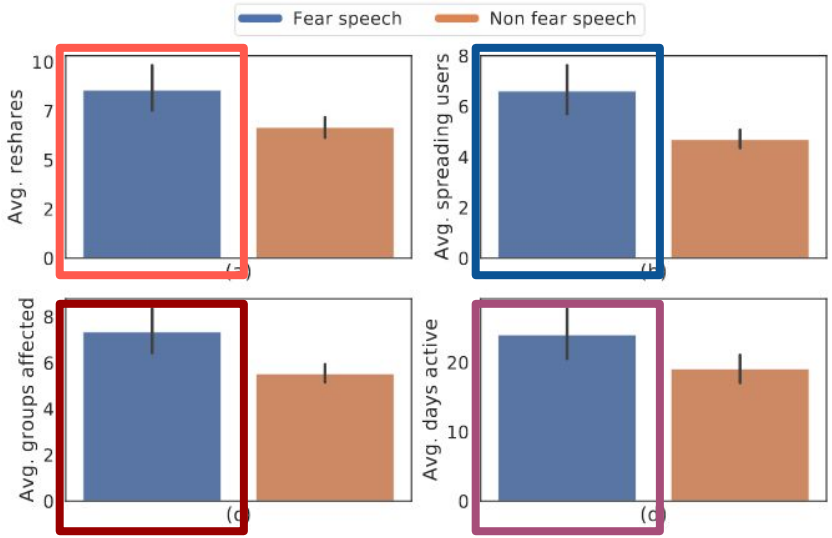
Controversial topics

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

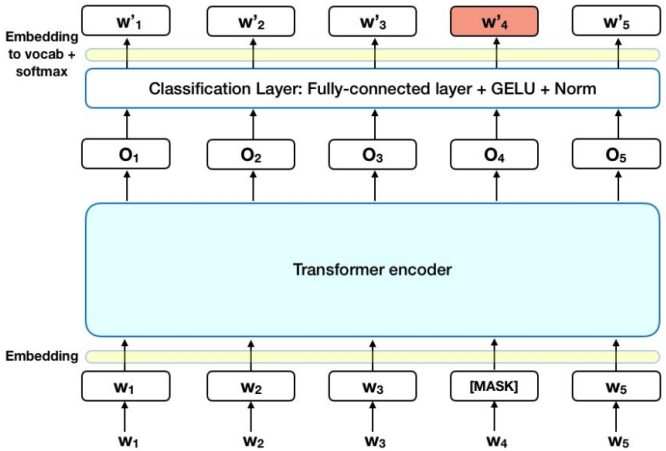
LASER + **LSTM**

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**

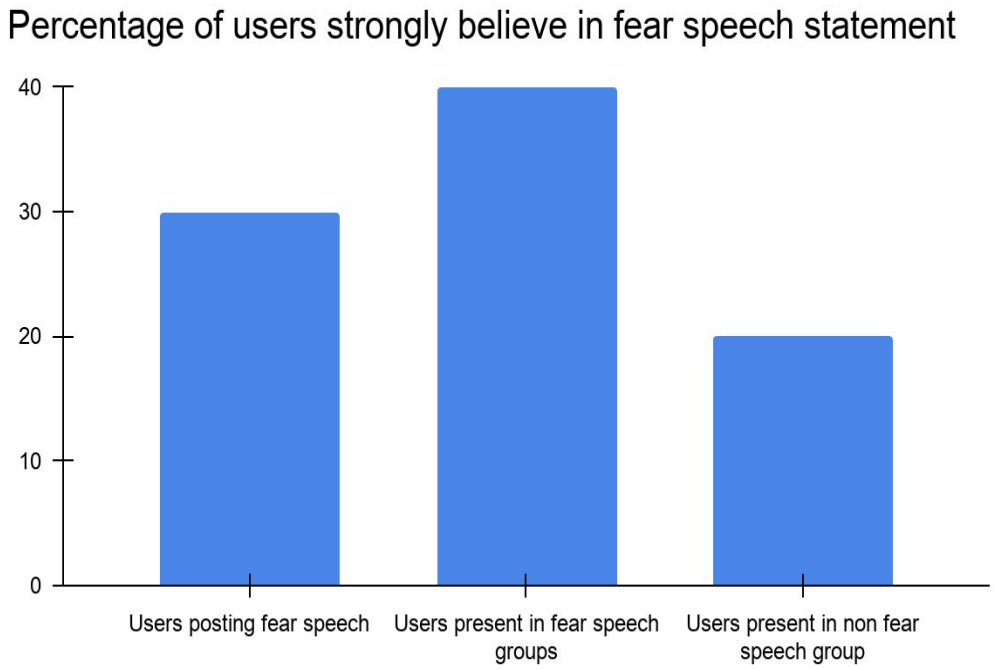
Surveying WhatsApp users

- 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.

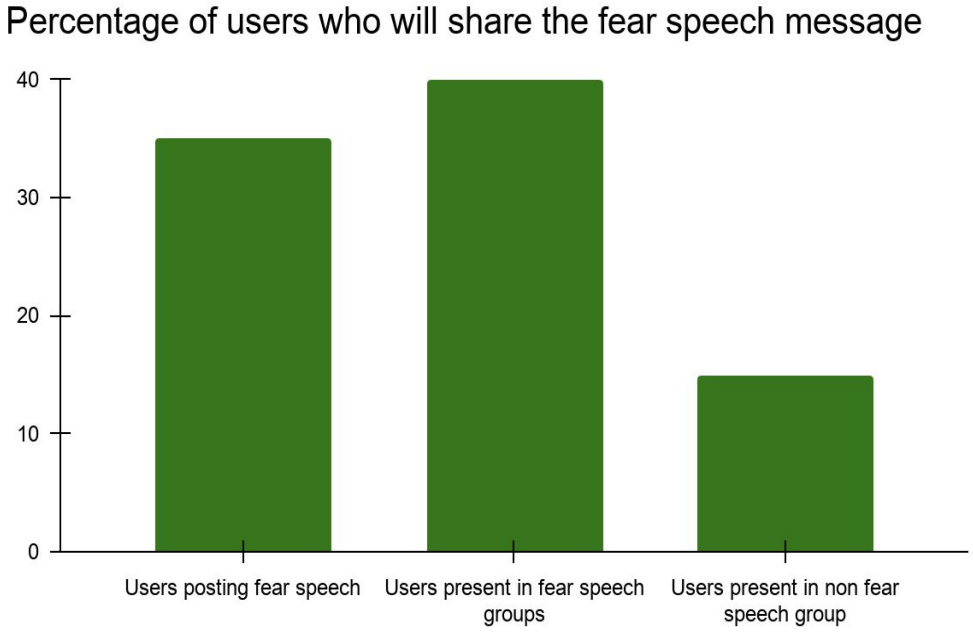
Results from the survey



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

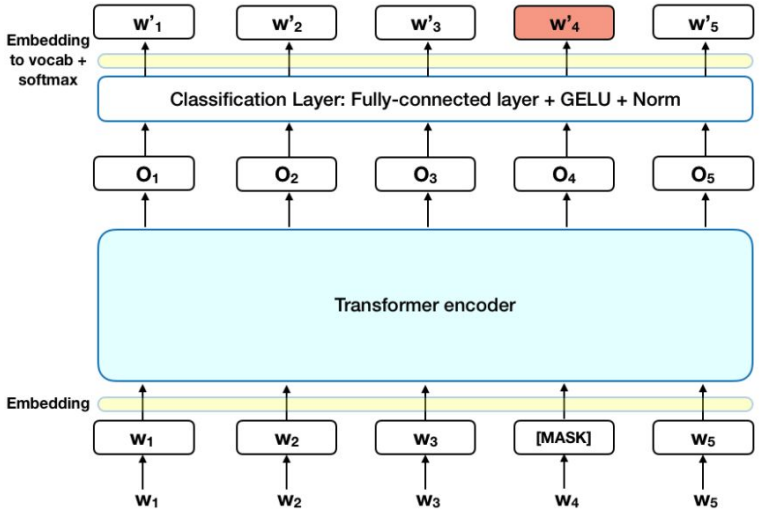
- Sampled the posts from a corpus of Gab Data[1] which contains **21 million posts** and their metadata from **October 2016** to **July 2018.**
- **4 expert** annotators and **103 crowd annotators** participated in MTurk platform.
- Total datapoints were ~**10,000**, out of which **1800** were fear speech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

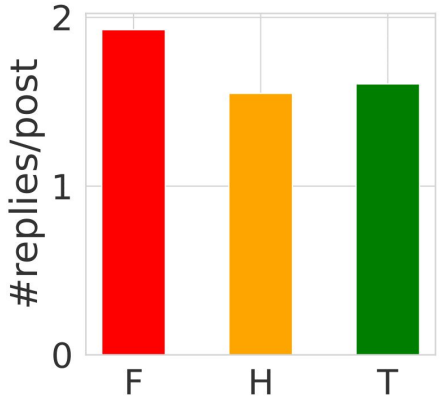


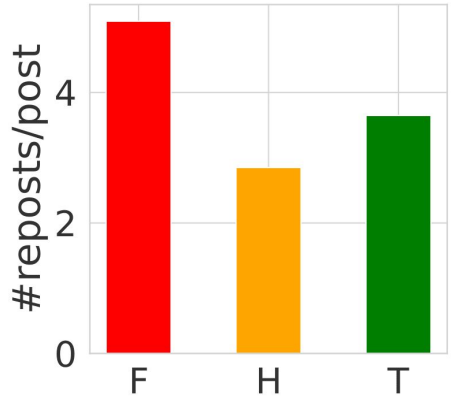
Scaled up dataset

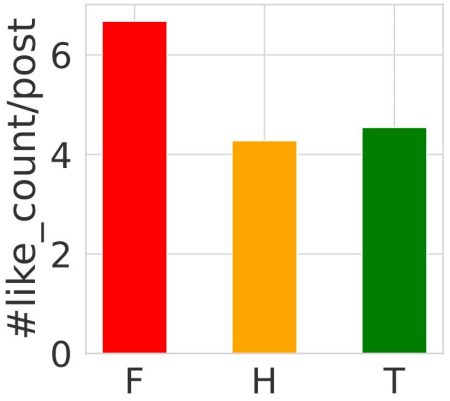
- We got the best performance by GabBERT and emotion vector of **0.63 f1 score**
- Applied this model on the whole dataset (**21M**) and got **400k** fear speech and **700k** hate speech
- We also selected **ExHate** and **ExFear** users (~500) based on the top 10 percentile of posting fear/hate speech.

Reactions on posts

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

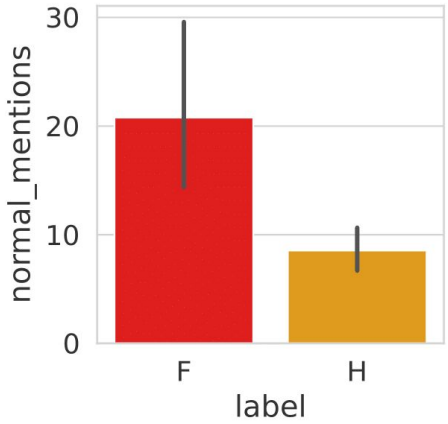


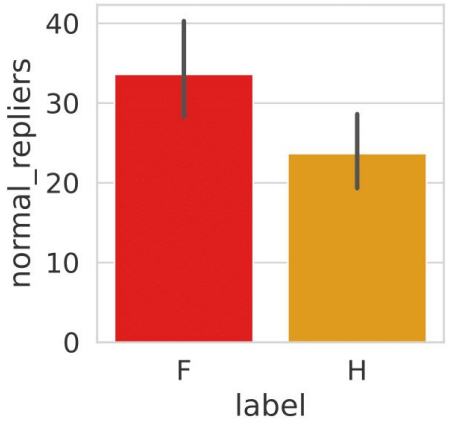


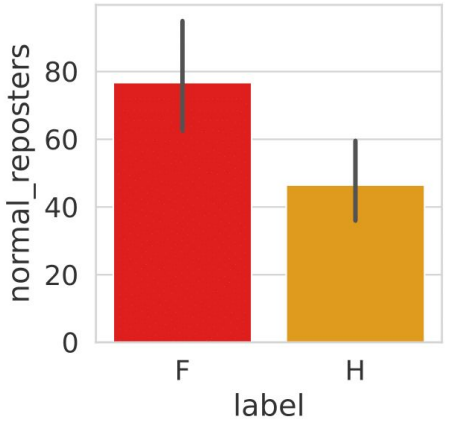


Effect on normal users?

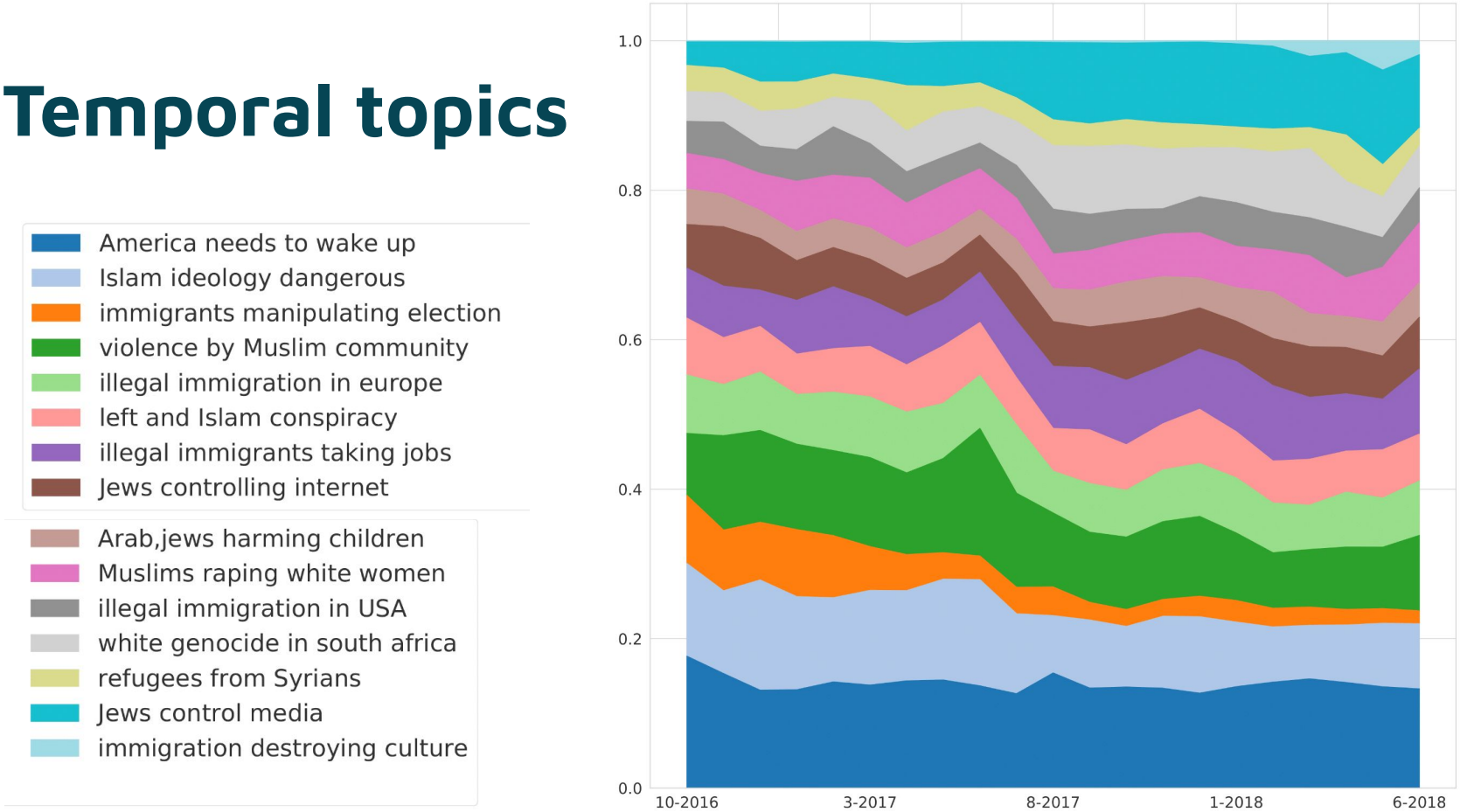
Normal users get mentioned more, reply more and repost more to fear speech than hate speech





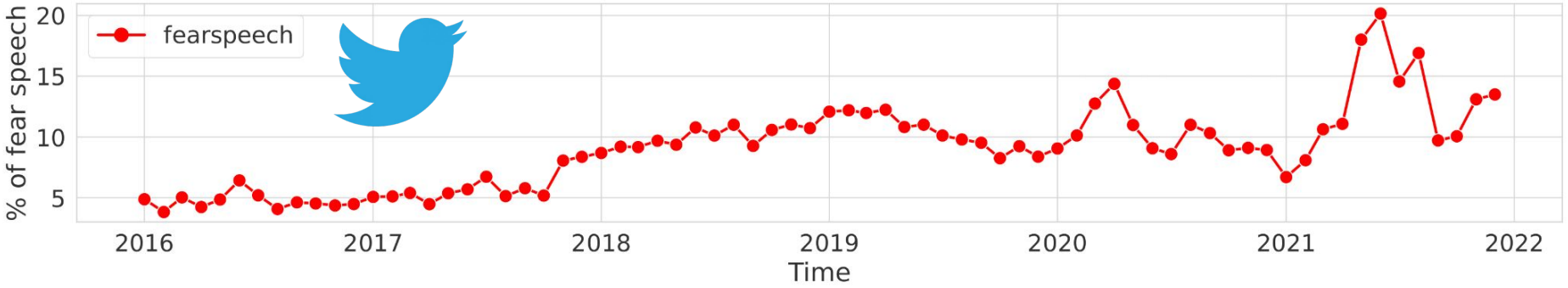


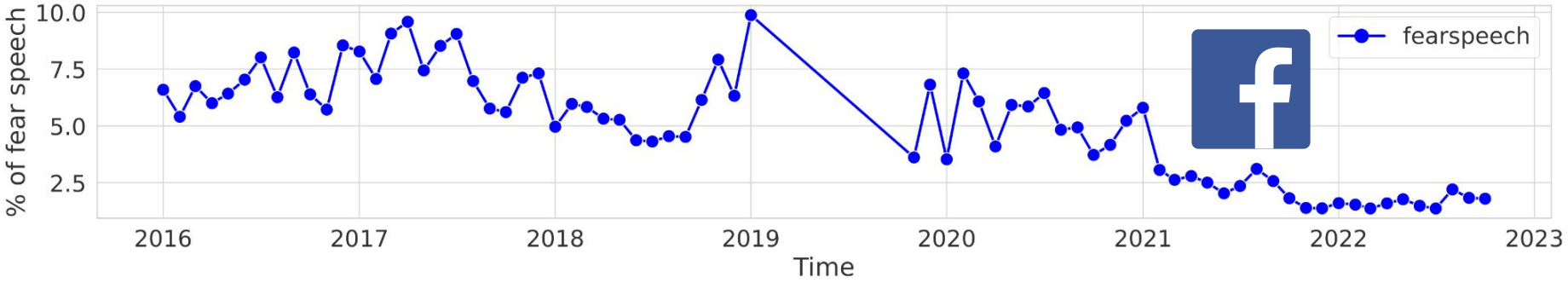
Temporal topics



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

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**

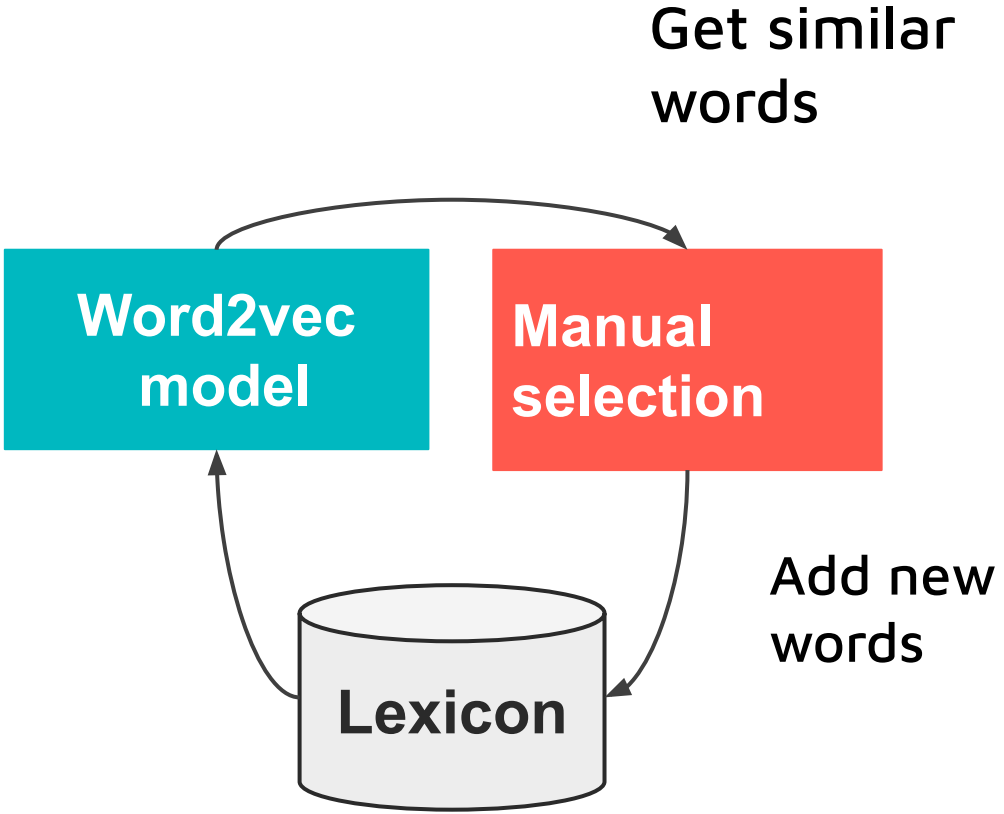
Summary

- 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


Get similar words
Word2vec model
Manual selection
Add new words
Lexicon

Thanks !

Do you have any questions?

@punyajoysaha

punyajoys@iitkgp.ac.in

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