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

Fear speech in Indian Whatsapp groups

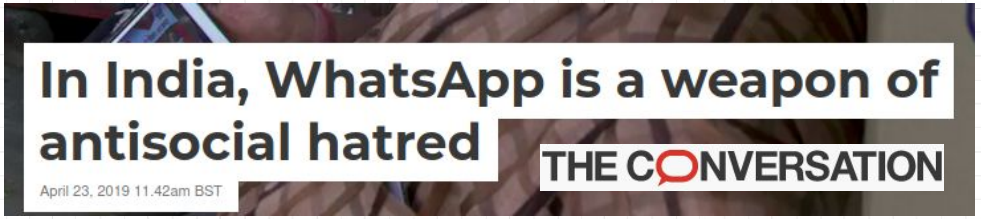
Punyajoy Saha, Binny Mathew,
Kiran Garimella and Animesh Mukherjee

Increasing hate crimes in India

Role of social media

Whatsapp in India

- 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

What we did not find ...

In our initial analysis, we did not find any presence of **direct hate speech!**

BUT ...

What we found ...

In our initial analysis, we did not find any presence of **direct hate speech!**

BUT ...

We found **Fear speech**

"An expression aimed at instilling (existential) fear of a target (ethnic and religious) group."

Target (in our work): Muslims

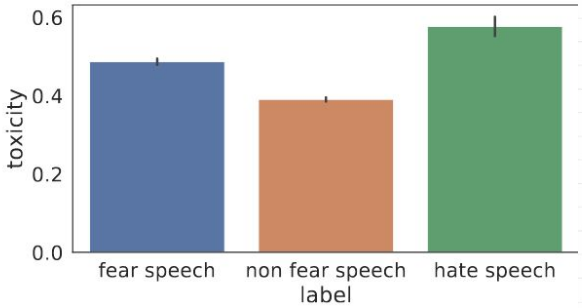
Buyse, Antoine. "Words of violence: Fear speech, or how violent conflict escalation relates to the freedom of expression." *Hum. Rts. Q.* 36 (2014): 779.

Why such camouflaging?

- Absence of direct hate speech may be attributed to
  - Laws against hate speech in India.
  - Political groups have to maintain a public image.
  - We only have access to a subset of public groups.
- **Fear speech possibly specially contrived to bypass the above hindrances.**

Example

Message (original in 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 .. People who do love jihad --- is a Muslim. If you want to give muslims a good answer, please share!!	**Fear speech**
That's why I hate Islam! See how these mu**ahs are celebrating. Seditious traitors!!	**Hate speech**



Toxicity based on perspective api. Hate speech taken from a recent dataset

Argument structure in the Example

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

Table of Contents

01

Data collection

How we collected the data from Whatsapp?

Data collection

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

[1] 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 for around 1 year, from **August 2018 to August 2019**[1].

[1] 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).

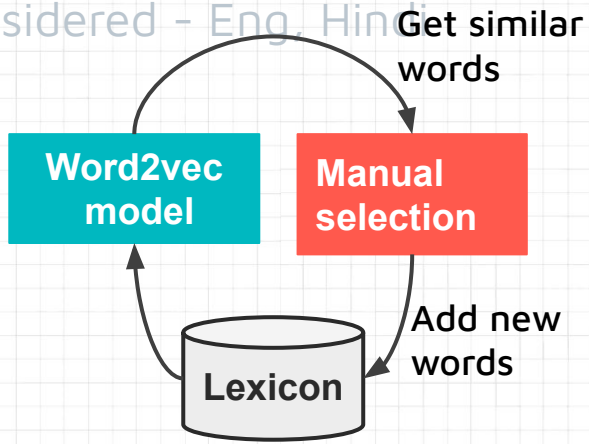
Data filtering

- 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 sampling

- 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 **muslim** community was created using a bootstrapping method



02

Annotating data

How we annotated the fear speech data?

Annotation guidelines

Definitions of fear speech and **flowchart** to identify fear speech

Forms of fear speech with **examples**:

A. Fear induced by using examples of past events,

B. Fear induced by referring to present events,

C. Fear induced by cultural references,

D. Fear induced by speculation of dominance by the target group.

A post was marked as fear speech, even if it contained some fear elements in it

Annotating the data

Initial annotation and training of annotators

- **500** posts was annotated by expert annotators
- Students voluntarily participated using online form and were compensated for the task
- 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

https://github.com/doccano/doccano

Final dataset

5k unique posts with Fleiss kappa of **0.36** inter annotator agreement.

Challenges

- Length of the message
- Some of non fear speech message contain quotes from Quran

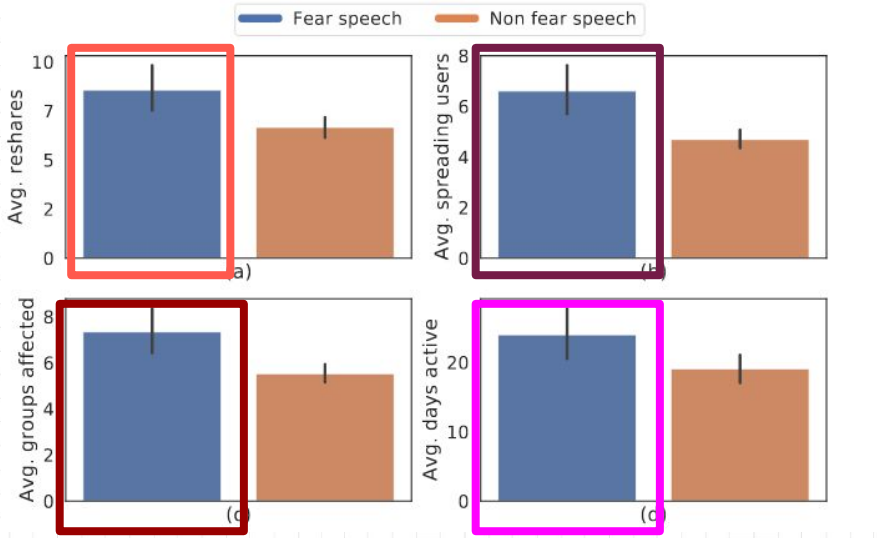
Features	Fear speech	Non fear speech
Number of posts	7,845	19,107
Unique posts (Annotated)	1,142	3,640
Average length of a message (in words)	500	464

03

Messages

Characterisation of messages.

Fear speech characteristics: Counts

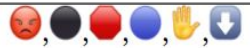


More reshares, large #users spreading, large #groups affected and a longer lifetime

Fear speech characteristics: Emojis

Emojis

- Built the co-occurrence network based on emojis.
- Louvain algorithm[1] 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

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

Fear speech characteristics: Topics

LDA[1] 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

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

04

Survey

Understanding perspective of the users associated with such groups

Surveying WhatsApp users

- Important to understand the **perception** of people in the WhatsApp groups. Used **facebook's ad** to target **three** types of users:
    - 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**

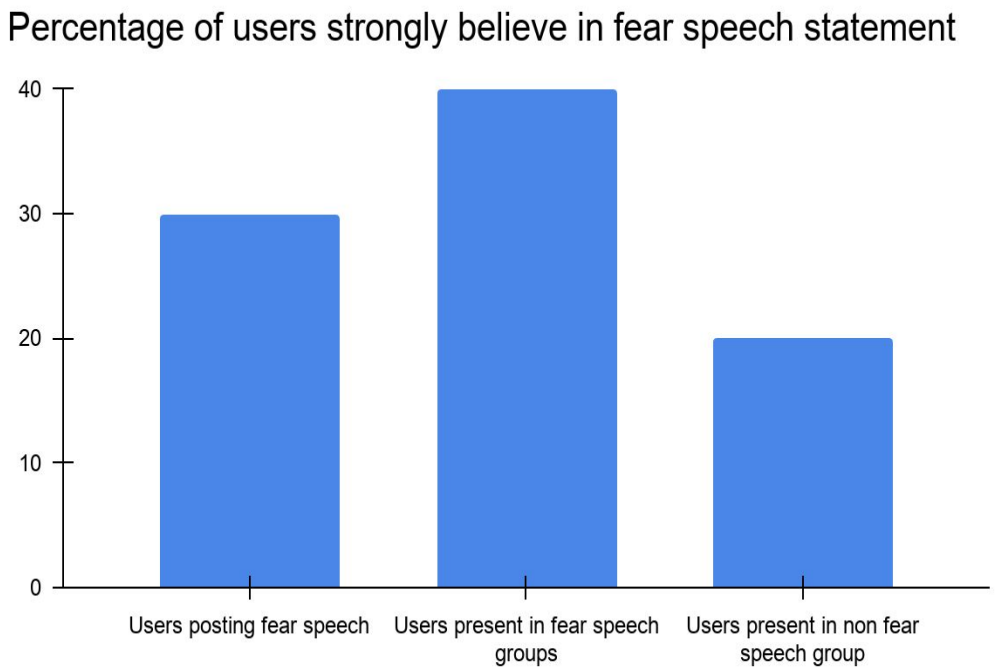
Surveying WhatsApp users

- 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. These double faces must be exposed.

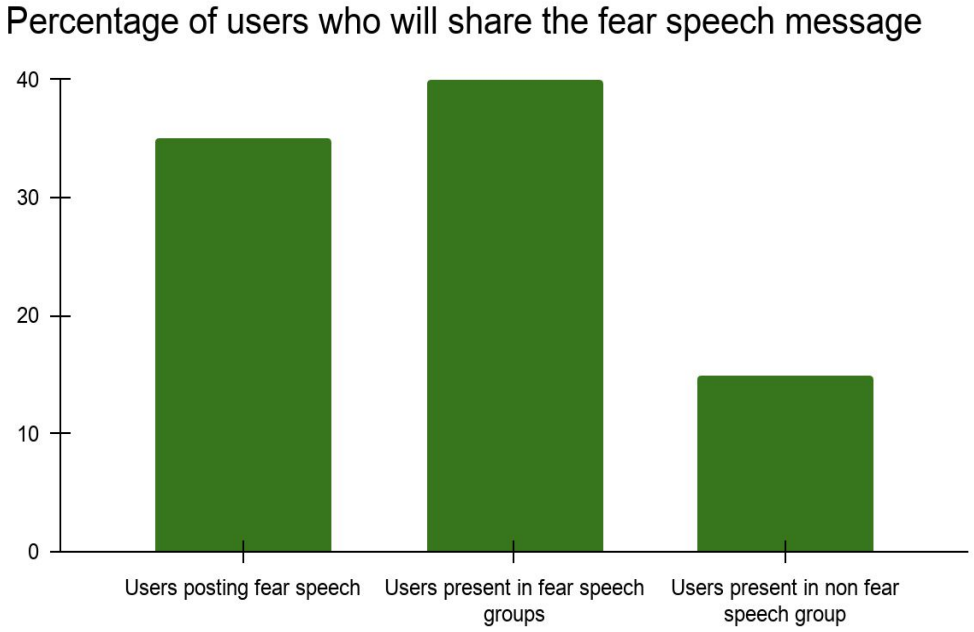
Results from the survey



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 in fear speech



05

Detection

Automatic detection of fear speech

Fear speech detection

Models	Features	Accuracy	F1-Macro	AUC-ROC	Precision(FS)
Logistic regression	Doc2vec	0.72	0.65	0.74	**0.44**
SVC (with RBF Kernel)	Doc2vec	0.75	0.69	0.77	**0.45**
LSTM	LASER embeddings	0.66	0.63	0.76	**0.39**
XLM-Roberta +LR	Raw text (b)	0.76	0.71	0.83	**0.51**
mBERT + LR	Raw text (b)	0.72	0.65	0.80	**0.48**

None of the current models are precise, such that we can deploy them to detect fear speech at a scale

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**

Takeaways

- We curate one of the **first dataset** about fear speech in India, whose timeline is co-located with 2019 Elections.
- We identify **topics** and **emojis** which indicate the different ways to vilify Muslims
- State of the art detection models fail to identify fear speech with **high precision**
- Our **survey** further identifies anti-muslim attitudes of the users present in the fear speech group

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

Thanks!

Punyajoy Saha
@punyajoysaha

Binny Mathew
@_BinnyM

Kiran Garimella
@gvrkiran

Animesh Mukherjee
@Animesh43061078

Send your questions at punyajoys@iitkgp.ac.in

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