

“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



INDIA



India reported 218 hate crimes in 2018, UP tops chart, says Amnesty; cow violence, honour killings most common

Over 200 alleged cases of hate crimes were reported in 2018 against people from marginalised groups, especially Dalits, with Uttar Pradesh recording the highest number of such incidents for the third consecutive year, Amnesty India said in a new report on Tuesday.

WORLD

'This Is It. I'm Going To Die': India's Minorities Are Targeted In Lynchings

August 21, 2019 - 9:35 AM ET



CORONAVIRUS CRISIS

The other virus: Hate crimes against India's Muslims are spreading with Covid-19

On April 7, rumours about Muslims intentionally spitting to spread the virus reportedly led to a riot-like situation in Jharkhand, leaving one person dead.

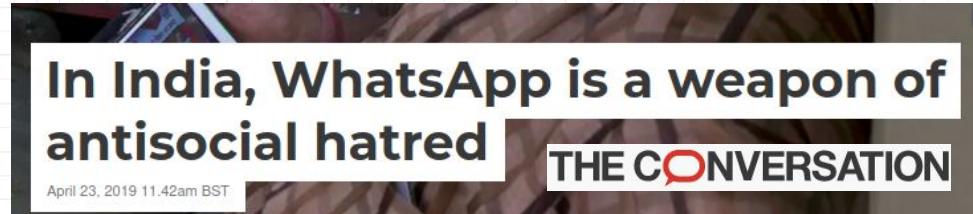
Increasing hate crimes in India



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.



Is there hate in Whatsapp ?



In our initial analysis, we did not find **any hate directly!** This might be due to three reasons:

- Laws against hate speech in India.
- Political groups have to maintain a public image.
- We only have access to a subset of public groups.



Past research

- Works in the past have tried to **identify hate speech** against different target categories^[3].
- But they are mostly capturing **overt** hate speech
- Few works have tried to bridge the gap, by studying weak toxicity against different target community like **muslims**^[1,2].
- Our work tries to operationalise one of such weak toxic speech in a **closed platform** - Whatsapp.

[1] Vidgen, B., & Yasseri, T. (2020). Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1), 66-78.

[2] Sindoni, M. G. (2018). Direct hate speech vs. indirect fear speech. A multimodal critical discourse analysis of the Sun's editorial "1 in 5 Brit Muslims' sympathy for jihadis". *Lingue e Linguaggi*, 28, 267-292

[2] Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018, May). An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

What we found: Fear speech

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

Forms include but are not limited to:

- Harmful things of the past.
- Traditions portrayed as harmful.
- Threat about the future.

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)

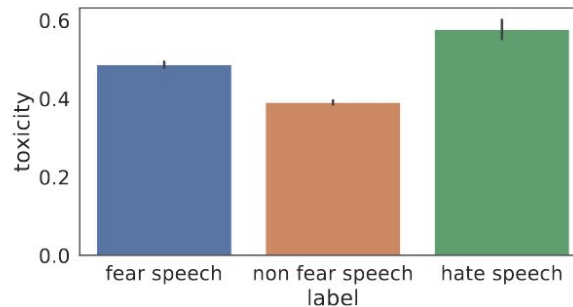
Label

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 <i>love jihad</i> – 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?

02

Annotation

How we annotated the data?

03

Messages

Characteristics of the messages

04

Survey

Survey to understand the users further.

05

Detection

Detection of fear speech

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

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).

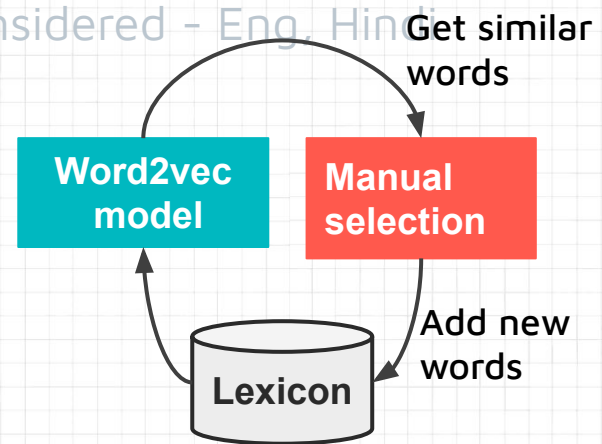
Data filtering

- 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



Data collection

- Data collection tools developed by past works^[1] to gather the WhatsApp data.
- The keyword lists (i.e. **query** terms) cover all **major political parties and politicians**.
- Searched public WhatsApp groups on Google, Facebook and Twitter using “**chat.whatsapp.com +query**”,
- In total **5,000 political groups** having image, videos and text
- Our data collection spans for around 1 year, from **August 2018 to August 2019**.



Data filtering

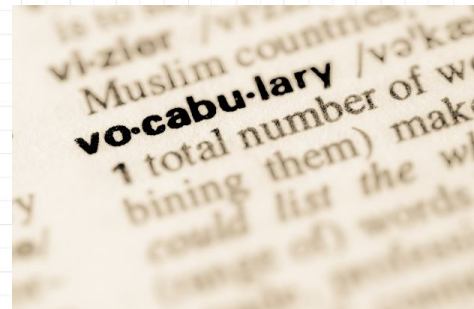
- Filter posts by language - only kept **Hindi** and **English** message (70% of total messages)
- Removed **spam** messages like phishing links, reward points using a high precision lexicon (29% → 3%)

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



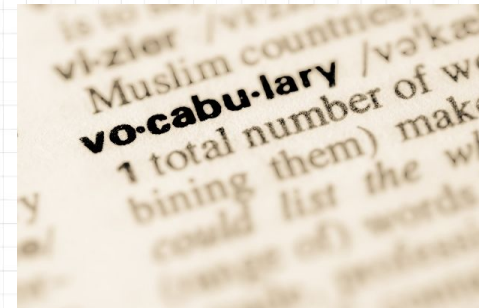
Generating lexicons

1. **Seed lexicon** creation which has word denoting Muslims
2. **Tokenize** each post in the dataset → tokens per post
3. **N grams** (N value 1-3) generated and remove n-grams < 15 frequency
4. Finally, **word2vec** model was created
5. For each word in seed lexicon, we sampled **30 similar words**.
6. Relevant keywords from this set were added to the **seed lexicon**.
7. Step 5 was repeated again until we did not find any more keyword



Generating lexicons

1. **Seed lexicon** creation which has word denoting Muslims
2. **Word2vec model** based on N-gram features
3. For each word in seed lexicon, we sampled **30 similar words** using the word2vec model
4. Relevant keywords were added manually and step 3 was repeated



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

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

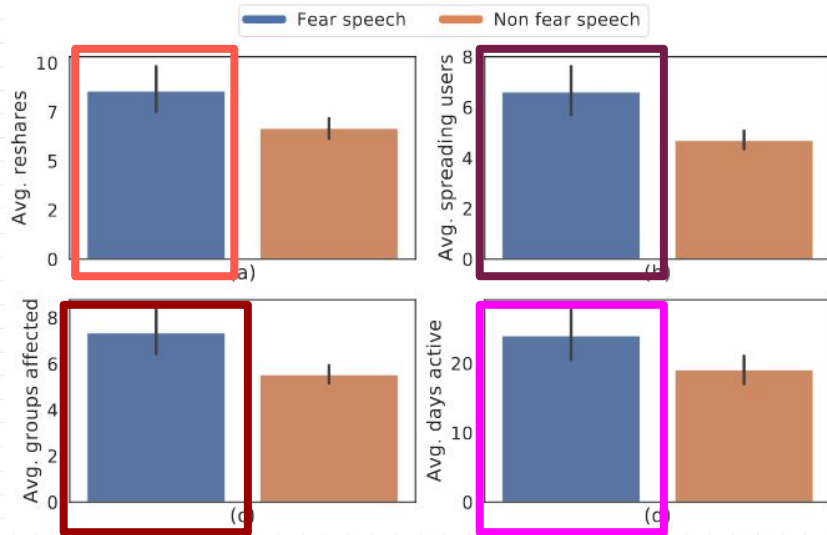
Features	Fear speech	Non fear speech
Number of posts	7,845	19,107
Unique posts (Annotated)	1,142	3,640
Average length of a message (in words)	500	464

03

Messages

Characterisation of messages.

Fear speech characteristics: **Counts**





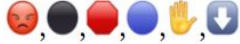


More reshares, large #users spreading, large #groups affected and a longer lifetime

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

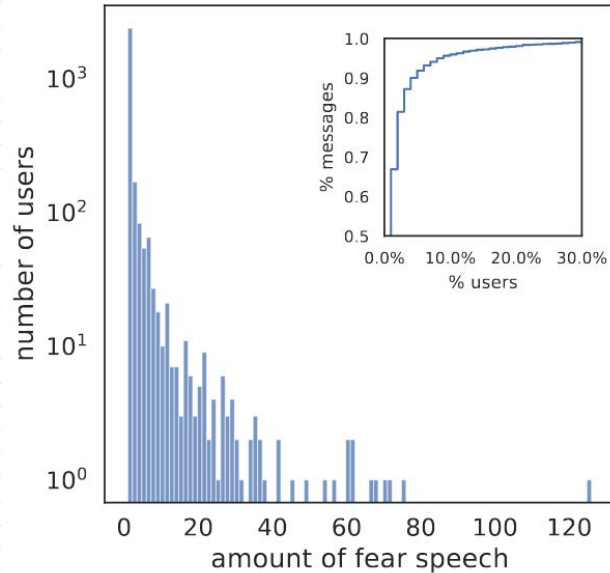
Fear speech characteristics: **Topics**

LDA^[1] models to extract topics (number of topics as 10 had highest coherence score)

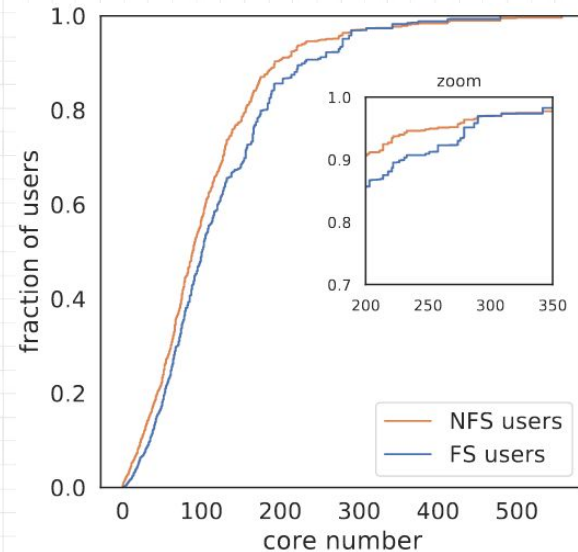
Topics	Themes of fear speech
Love jihad (Muslim men are forcing hindu women to interfaith marriages)	Painting interfaith marriages in wrong light
Increase in muslim population (Muslim population increasing at an alarming rate)	Using event in the current timeline to spread fear
Kerala riots (Blaming muslims 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.

User characterisation



10% of the users post 90% content



Some of the fear speech users constitute a hub like structure using k -core analysis^[1]

[1] Shin, Kijung, Tina Eliassi-Rad, and Christos Faloutsos. "Corescope: Graph mining using k -core analysis—patterns, anomalies and algorithms." 2016 IEEE 16th international conference on data mining (ICDM). IEEE, 2016.

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.

Surveying Whatsapp users

- Used the Custom Audience targeting feature provided by Facebook, targeted users based on lists of phone number.
- Three types of users selected:
 - 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**
- Around **50 %** of the users had an active Facebook account

Survey Design

- Short survey (<3 min) with no monetary benefit
- Used a generic template to avoid priming



Survey questions

Survey questions

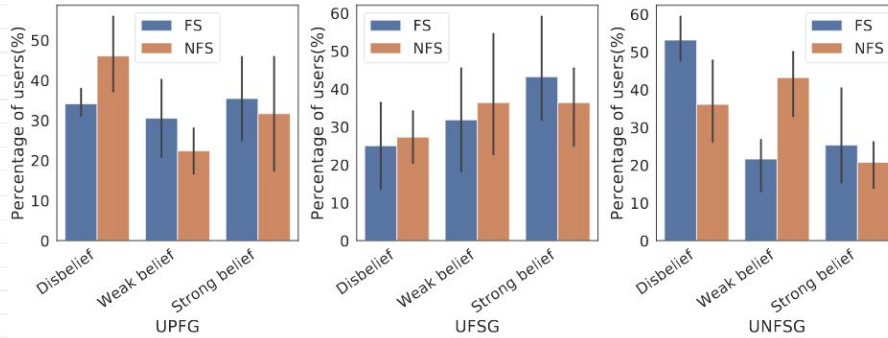
- 3 (user types) X 2 (types of statements)
- 8 statements (in the form of claims)
- With each statement participants were asked about their belief and propensity to share

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.

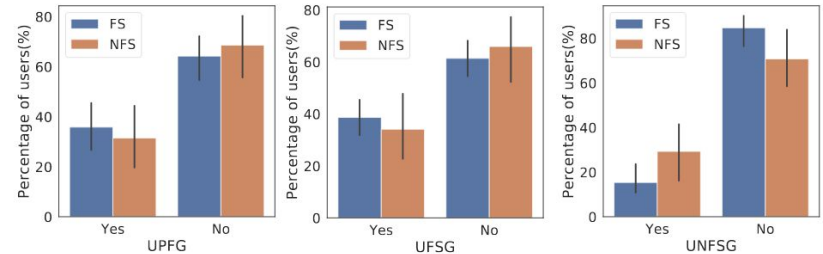
Results

From the 119 responses we found that →



Users in UPFG and UFG are more likely to share in fear speech

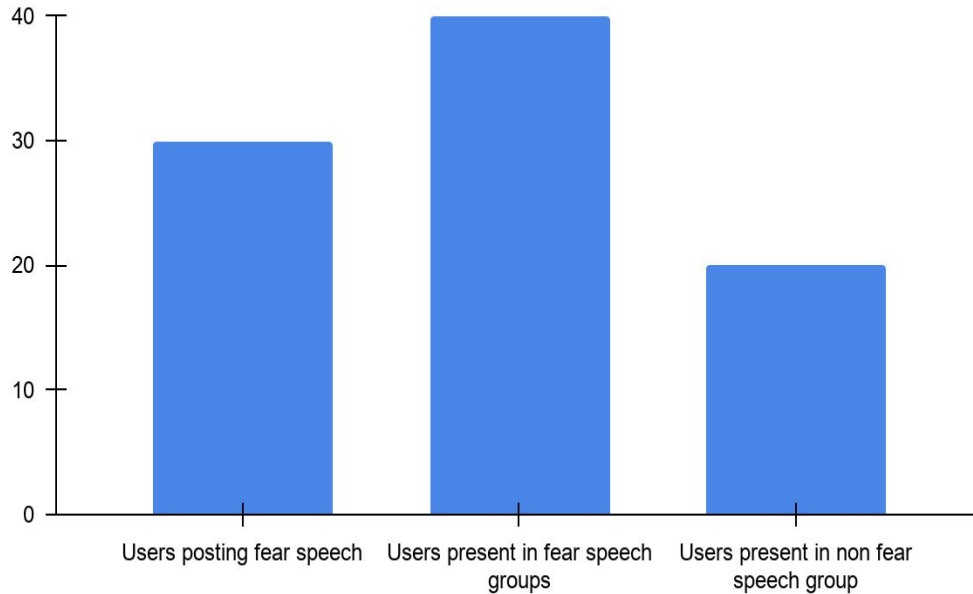
Users in UPFG and UFG are more likely to believe in fear speech



From other questions we also found that UPFG and UFG users support BNP, Support CAB and blame Muslims for COVID-19 hotspot

Results from the survey

Percentage of users strongly believe in fear speech statement

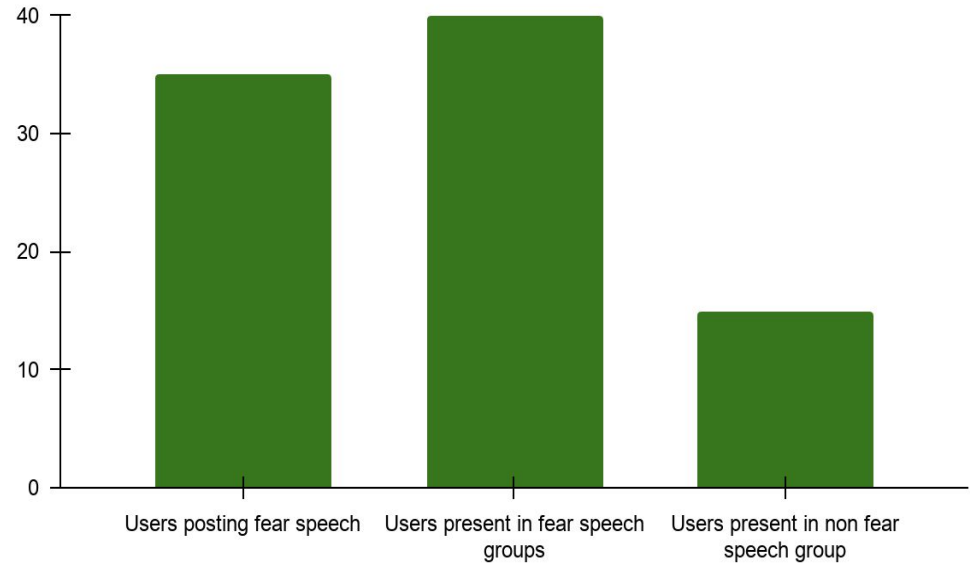


Users in UPFG and UFG are more likely to believe in fear speech

Results from the survey

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

Percentage of users who will share the fear speech message



05

Detection

Automatic detection of fear speech

Detection methods

Range of classification models for detecting fear speech

- **Classical** - Doc2Vec embeddings with LR or SVM classifier
- **LASER + LSTM** - Paragraph was divided into sentences, which were reported using LASER embeddings. Now LSTM was used as a classifier.
- **Transformers** - Stack of self attention blocks, state of the art in many tasks. We fix the number of tokens per sentence to 256
 - a. 256 tokens from start
 - b. 256 tokens from end
 - c. 128 from start and end

Results

Models	Features	Accuracy	F1-Macro	AUC-RO C	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

Best model is XLM-Roberta with 128 tokens from start and end



Fear speech detection

Models	Features	Accuracy	F1-Macro	AUC-RO C	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?

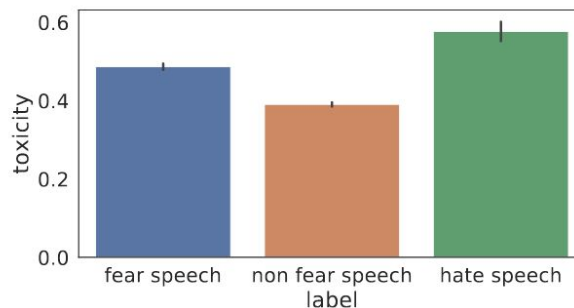


- Need cross-disciplinary dialogue
 - Policy
 - Media
 - Technology
- Possible joint activities
 - **Educating the users to moderate content (making them socially responsible)**
 - **Laying out tangible policies of moderation**
 - **Improving existing technologies to implement such policies**

Error analysis

The best model (XLM-Roberta + LR) was further passed through LIME(model explanation toolkit) to understand the problem

- Confounding factors
- Target identified but fear emotion not detected



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

Text (translated from Hindi)	GT and TE
Big breaking: Dharmendra Shinde, a marriageist Hindu brother, was murdered by Muslims for taking out the procession of Dalit in front of the mosque ... Mr. Manoj Parmar reached Piplarawa and showed communal harmony and prevented the atmosphere from deteriorating as well as taking appropriate action against the accused ... There is still time for all Hindus to stay organized, otherwise, India will not take much time to become Pakistan. Share it as much as the media is not showing it "	FS and CF
Increase brotherhood, in the last journey, on uttering the name of Ram, the procession was attacked mercilessly by the Muslims . This was the only thing remaining to happen to Hindus , now that has also happened	FS and TNE





What can be done?

- We analysed a subset of Whatsapp group, which is only the “tip of the iceberg”.
- Whatsapp is an end-to-end encrypted platform, content moderation completely left out to users. **Hence educating the users is one of the important step.**
- One possible solution might be a **client side classifier** but the accuracy and size of the model are still not at par.
- Discussion about how to moderate such subtle and indirect message is another point for the **policy makers.**

Acknowledgements

- We want to thank the reviewers for their valuable feedback.
- We want to thank the annotators for performing this difficult task of annotation.
- Finally, we want to thank the group at IDSS, MIT for collecting the WhatsApp public group dataset.

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

 [@punyajoy_saha](https://twitter.com/punyajoy_saha)



Binny Mathew

 [@_BinnyM](https://twitter.com/BinnyM)



Kiran Garimella

 [@gvrkiran](https://twitter.com/gvrkiran)



Animesh Mukherjee

 [@Animesh43061078](https://twitter.com/Animesh43061078)

Send your questions at punyajoy@iitkgp.ac.in



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