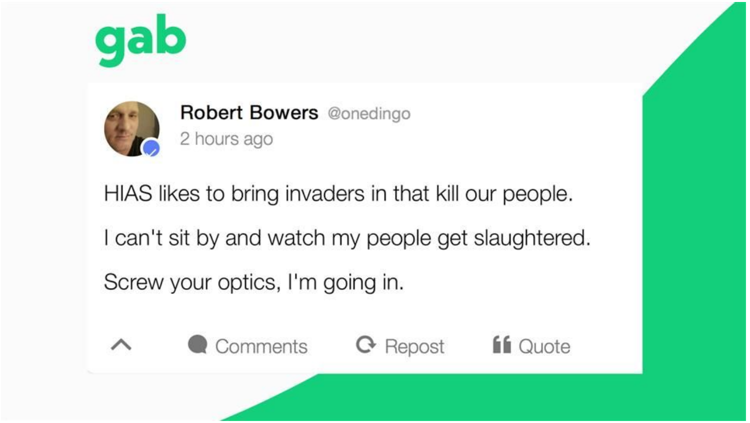
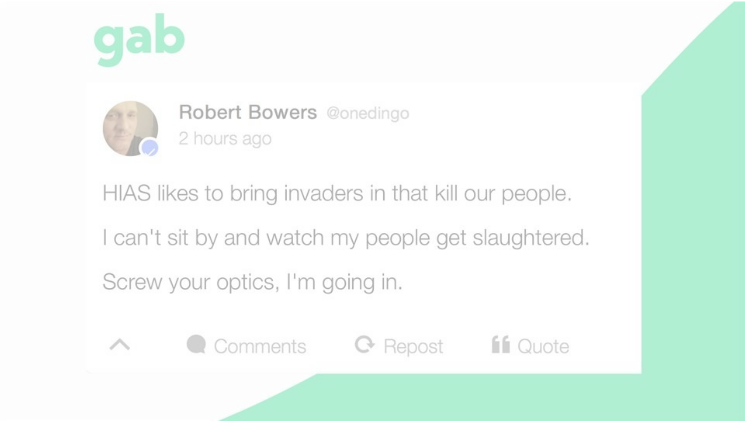
HateXplain:

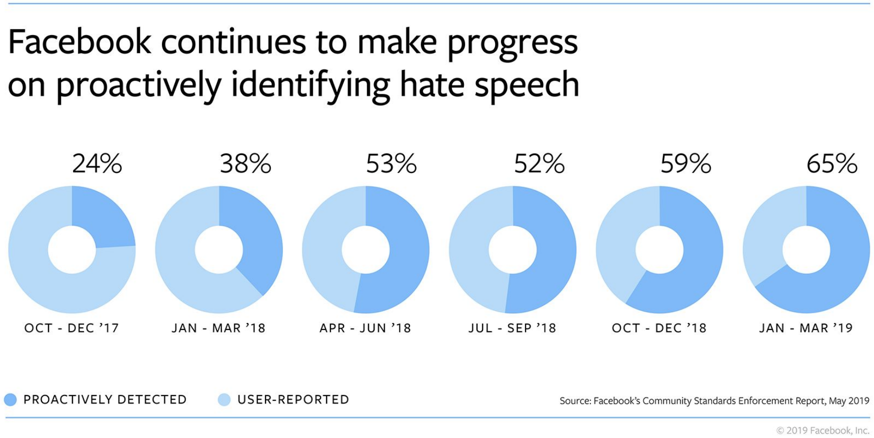
A benchmark dataset for explainable hate speech detection

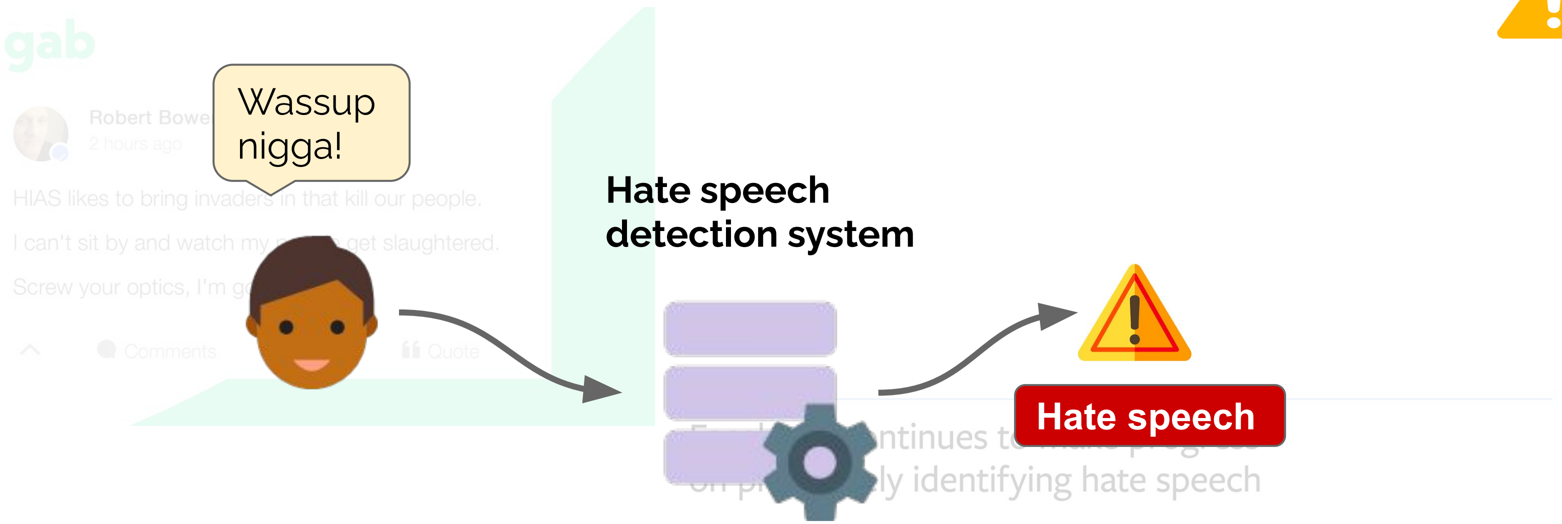
Binny Mathew*, **Punyajoy Saha***, Seid Muhie Yimam,
Chris Biemann, Pawan Goyal, Animesh Mukherjee

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


gab
Robert Bowers @onedingo
2 hours ago
HIAS likes to bring invaders in that kill our people.
I can't sit by and watch my people get slaughtered.
Screw your optics, I'm going in.
Comments
Repost
Quote


gab
Robert Bowers @onedingo
2 hours ago
HIAS likes to bring invaders in that kill our people.
I can't sit by and watch my people get slaughtered.
Screw your optics, I'm going in.
Comments
Repost
Quote


Facebook continues to make progress on proactively identifying hate speech
24%
38%
53%
52%
59%
65%
OCT - DEC ’17
JAN - MAR ’18
APR - JUN ’18
JUL - SEP ’18
OCT - DEC ’18
JAN - MAR ’19
PROACTIVELY DETECTED
USER-REPORTED
Source: Facebook’s Community Standards Enforcement Report, May 2019
© 2019 Facebook, Inc.



But the decisions are not explainable and may be biased.

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	✗	✗
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	✓	✗

1

Data collection

Details of data collection from Twitter and Gab

Collection strategy

- 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

Few finer details

- Did not consider **reposts** and remove **duplicates.**
- Posts do not contain **links**, **pictures** or **videos.**
- The **emojis** are in the text.
- The **usernames** are replaced with **<user>**

2

Annotations

The annotation framework.

Annotation framework

Each post in our dataset contains

- Label
- Target
- Rationales

Final label is selected using **majority**

919 cases did not have clear majority.

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**

A target is selected if it is marked so by **majority** of the annotators

Offensive - Women, African and LGBTQ
Hate speech - African, Islam and Jewish

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.

Data format

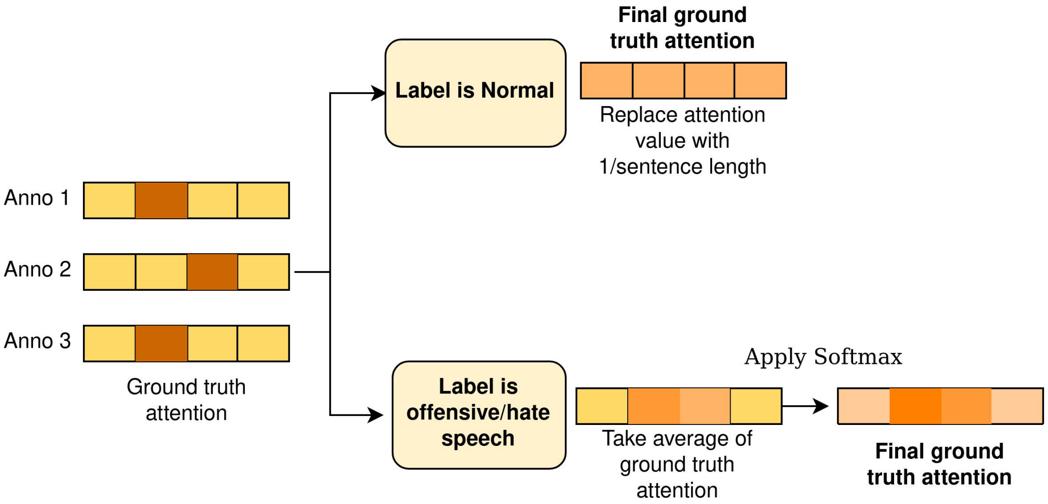
The data is a dictionary having elements in the following **format:**

<post_id>: {
    post_id: <post_id>,
    annotators: <list of annotations>,
    rationales: <2-3 boolean vector length equal to post_tokens>,
    Post_tokens: < list of tokens >
}

The **<list of annotations>** contains annotation from **3 annotators**

- Annotator ID
- Label
- List of targets

Ground truth rationales



Tokens - ["I" , "hate" , "ni**er"]
Rationales - [[0,1,1],[0,0,1],[0,1,1]]
Label - hate speech

Taking average - [0,0.66,1]

Doing softmax - [0.18,0.34,0.48]

3

Models

Deep learning models used in this work

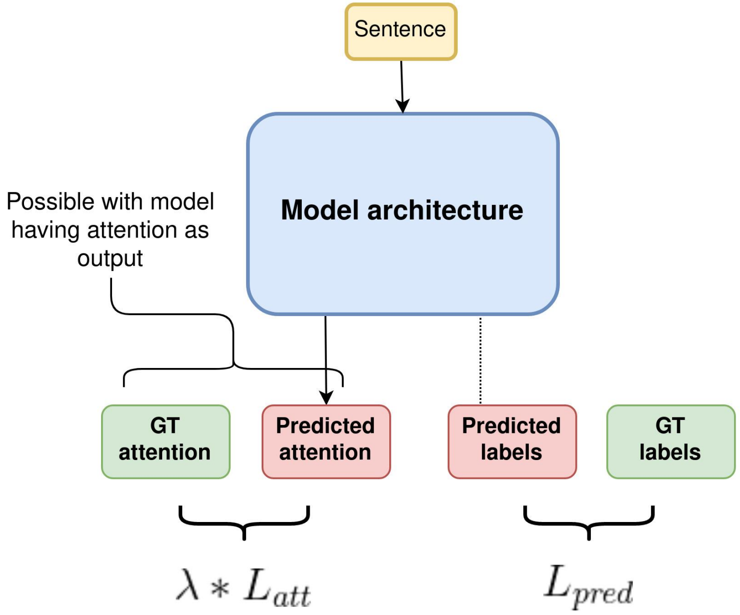
General framework

Models **without** attention supervision

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

Models **with** attention supervision

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



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

Attention supervision

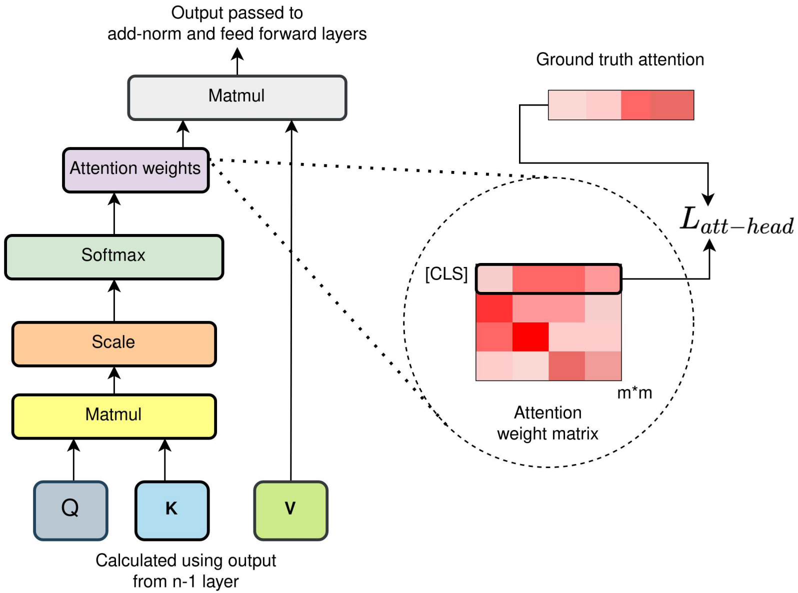
- **BiRNN-HateXplain**

  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



Extracting rationales from models

Attention based (Attn): Here we use the attention weights as final rationales.

- **BiRNN -** attention weights corresponding to the single head
- **BERT -** attention weights from 12 heads averaged.

Lime based (LIME): Here we pass the model outputs through LIME and then consider the top *K* words.

4

Evaluation

Evaluation metrics employed in this work

Metrics used for evaluation

- **Performance**

  Accuracy, **F1-score** and **AUROC** of final classification label

- **Bias**

  Subgroup AUC, BPSN , BNSP to understand target level bias

- **Explainability**

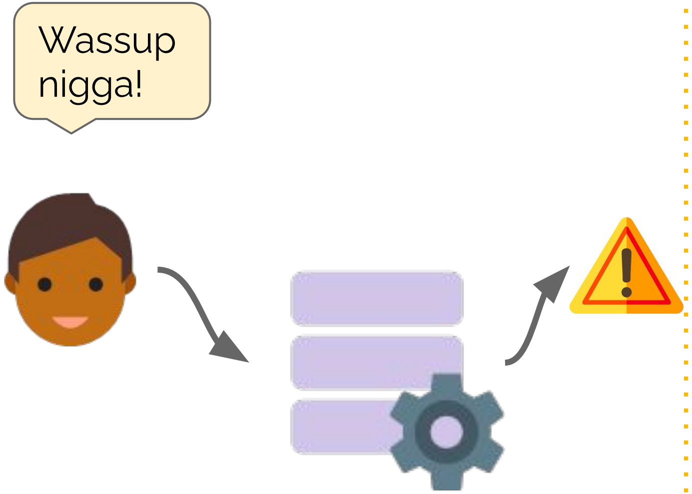
  Plausibility (IOU F1-score & token F1 score) and **Faithfulness** (comprehensiveness & sufficiency) to understand explainability aspect

Bias metrics

Classify between **toxic** (hate speech, offensive) and **non-toxic** (normal)

Measure the unintended bias of the models using

- **Subgroup AUC**
- **BPSN**
- **BNSP**



Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification

Bias metrics

Classify between **toxic** (hate speech, offensive) and **non-toxic** (normal).

Measure the unintended bias of the models using

- **Subgroup AUC**
- **BPSN**
- **BNSP**

Sub group AUC

1. Collect all the posts in **test data** belonging to a **community**
2. Measure the **AUC-ROC** score
3. **Higher score** means the model is able to distinguish toxic vs non toxic posts.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification

Bias metrics

Classify between **toxic** (hate speech, offensive) and **non-toxic** (normal).

Measure the unintended bias of the models using

- **Subgroup AUC**
- **BPSN**
- **BNSP**

Background positive, sub group negative

1. Collect **normal posts** that **mention** target community and **toxic posts** that **do not mention** target community
2. Measure the **AUC-ROC** score
3. **Higher score** means the model is less likely to confuse.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification

Bias metrics

Classify between **toxic** (hate speech, offensive) and **non-toxic** (normal).

Measure the unintended bias of the models using

- **Subgroup AUC**
- **BPSN**
- **BNSP**

Background negative, sub group positive

1. Collect **toxic posts** that **mention** target community and **normal posts** that **do not mention** target community
2. Measure the **AUC-ROC** score
3. **Higher score** means the model is less likely to confuse.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification

Bias metrics

Generalized Mean of Bias(GMB) AUC: This metric was used in the "Jigsaw Unintended Bias in Toxicity Classification"

$$M_p(m_s) = \left(\frac{1}{N}\sum_{s=1}^{N} m_s^p\right)^{\frac{1}{p}}$$

M_p = the pth power-mean function

m_s = the bias metric m calulated for subgroup s

N = number of identity subgroups

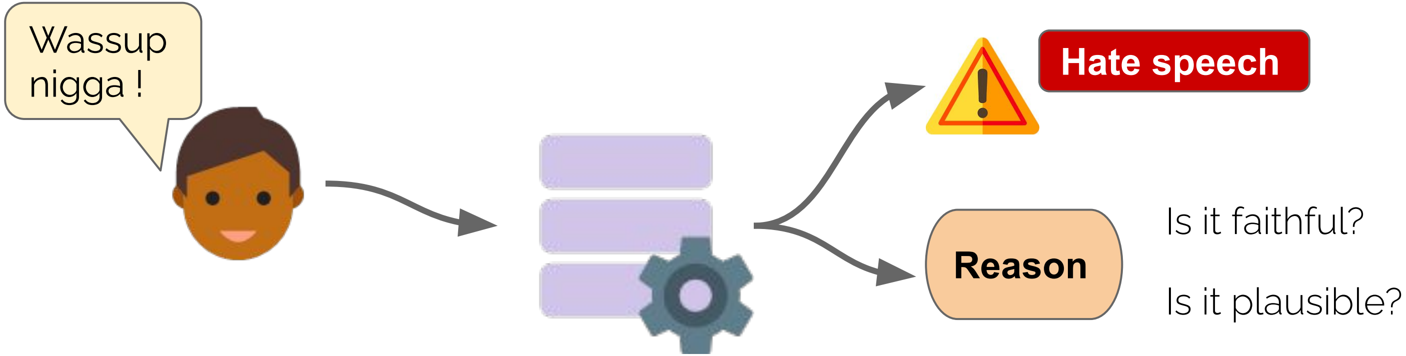
p is **-5** and number of sub groups are **10.**

https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview/evaluation

Explainability metrics

Plausibility: .Is the explanation correct or something we can believe is true, given our current knowledge of the problem?

Faithfulness: how to provide explanations that accurately represent the true reasoning behind the model's final decision



Explainability metrics

Plausibility is measured using ground truth and predicted rationales

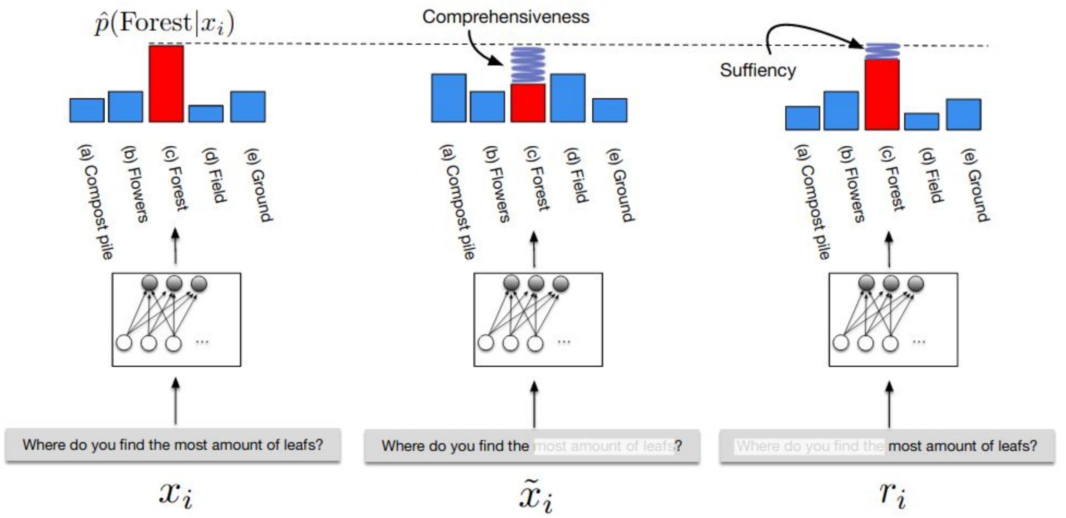
- **IOU F1 score** (Hard)
- **Token F1 score** (Hard)
- **AUPRC score** (Soft)

DeYoung, Jay, et al. "Eraser: A benchmark to evaluate rationalized nlp models." *arXiv preprint arXiv:1911.03429* (2019).

Explainability metrics

Faithfulness is measured using the predicted rationales

- **Comprehensiveness**
- **Sufficiency**



DeYoung, Jay, et al. "Eraser: A benchmark to evaluate rationalized nlp models." *arXiv preprint arXiv:1911.03429* (2019).

5

Results

Results and observations

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**

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

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

Conclusion

- We curate a dataset of **20k posts** from Twitter and Gab having **label**, **target** and **rationale**
- Models show **good performance**, do not fare well in terms of **model interpretability**.
- Models which use rationales while training **perform better** and has **less unintended bias**

Data & Code repository : https://github.com/punyajoy/HateXplain

Thanks!

Binny Mathew

Punyajoy Saha*

Seid Muhie Yimam

Chris Biemann

Pawan Goyal

Animesh Mukherjee

Any questions?

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