Can Hate Be Explained ?

Group 20

An overview

PROBLEM

Direct Consequences of Online Hate Speech
With riots in India, Myanmar, Anti-Semitic Attacks, in addition to cultural distortion and mental health consequences

Are there Efforts

- Yes, Facebook alone removed 31.5 million posts in Q2 2021
- Various Automatic Hate Speech Detection models claim state-of-the art performance

Is 90% Enough?

- Those are tall claims
- The research Suggests otherwise
- Most models fail to generalize on unseen data

Research

- “State-of-the-art” well, the data is more important!
- Annotated data for better “explain-ability”

The Hypothesis and the Experiments

Does a better Annotated Data really Help ?

We Pick **Hate-Xplain** a deeply annotated dataset, with a rigorous check on how the annotation was done along with an effort to **“explain”** hatred through “**rationales”**

Krippendorff’s alpha of 0.46

A reliability coefficient to measure agreeableness, higher than other datasets

And what about Augmentation?

- We test the claims from previous research if data augmentation does really impact generalization.
- We test sensitivity to different sentence lengths, and randomly delete words from our data for training and test in another dataset

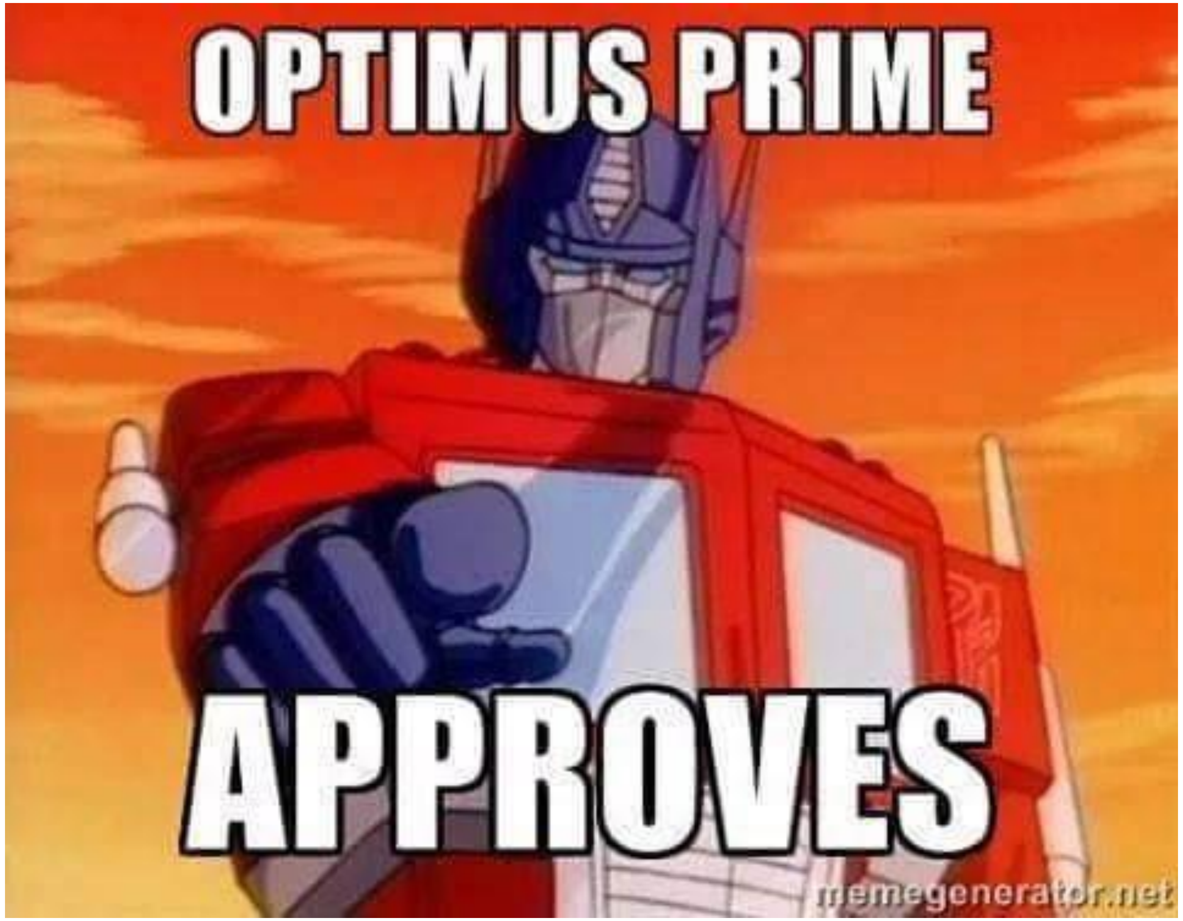
How do we test ?

- On **ETHOS,** an annotated dataset thought without rationales
- We test how well the models generalize with and without the augmentation

The architectural choice?

- Well “Transformers”

Well, thanks Optimus but not those



The Models

Attention, and more of it

- Multi-headed attention is a better, more optimized choice to learn representations
- Faster, better performance in literature for similar tasks

BERT

- Bidirectional Encoder Representations from Transformers (BERT)
- “State of the art” performance in related work to capture representations effectively

345 mil Parameters

Less Parameters

-40 %

Distilbert

- A distilled version of the BERT
- 40% less parameters
- Preserves 90% of the performance
- pre-trained on the same corpora, so can it generalize as well or just learns some trigger words?

So Does “Explaining” Hate Work?

Well, partly

- The larger “BERT” saw drops, but not as aberrant as we saw in the literature not even for other classes
- Distilbert generalized poorly after training from HateXplain and had a rather opposite trend to BERT
- But it was only true for training on Hate-Xplain the performance did drop badly for training on ETHOS

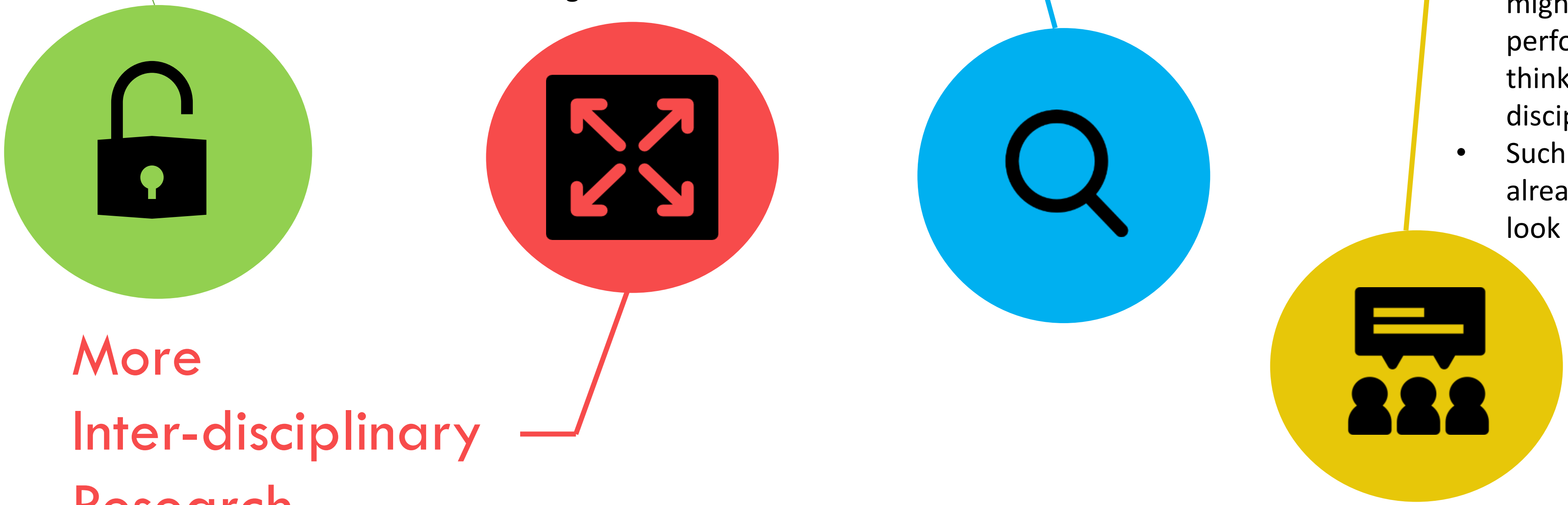
Did the augmentation help?

- Only to Bert and perhaps only aggravated the problem of “trigger” vocabulary
- That is learning keywords in some contexts and shooting conclusions over them

What about ETHOS?

- Training on ETHOS saw massive drops for BERT and rather consistent performance around 0.50 for all metrics for Distilbert
- Distilbert fared better but performance on Hate-Xplain seemed much better

Is Data Preparation the Key?



The hate problem

- We believe the answer is more inter-disciplinary, especially in the context of how and what hate is explained, understood, and agreed upon.

More Inter-disciplinary Research

- While, for other tasks perhaps we would not see linguistics, and convergence from representation but we did look at research from Dr. Chomsky and Dr. Sapolsky and wondered if interdisciplinary research in understanding how hatred is understood, processed, practiced, and experienced can inspire different approaches or better datasets

Datasets then?

- Verily we believe more annotated datasets with stringent quality crowdsourcing for annotations like Hate-Xplain can perhaps give results that generalize better

Multi-Task Learning

- But we also believe we perhaps multi-task learning with related tasks might enhance the performance given we do think this is more inter-disciplinary
- Such approaches have already been tried which look promising