DeepXplain Summer School 2025

Francielle Vargas
Ph.D. and M.Sc. in Computer Science and Computational Mathematics
Artificial Intelligence - Natural Language Processing

UNIVERSITY
OF OSLO

Project 1: Description

1. **Interpretable Hate Speech Detection**: Aligning Hate Speech Detection with Human Rationales using Supervised Attention.

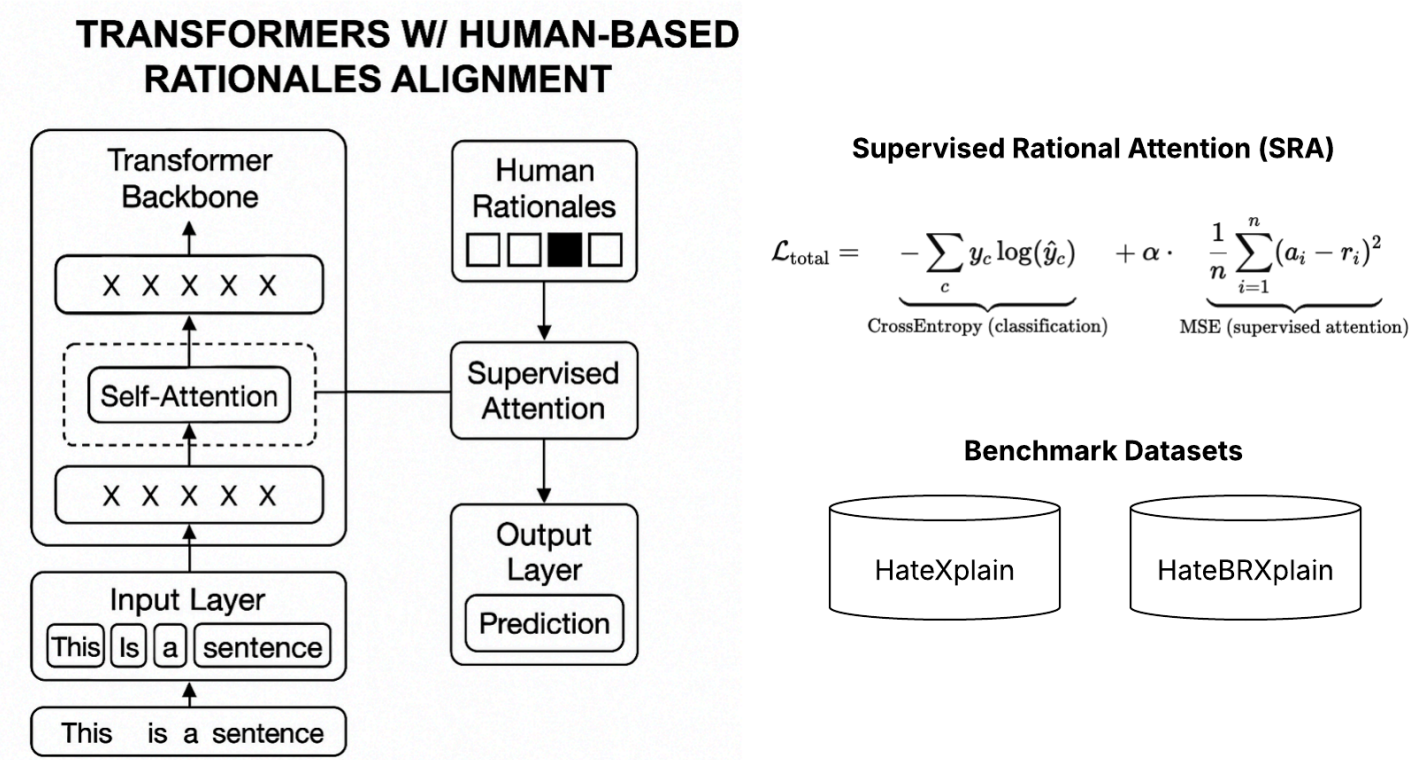
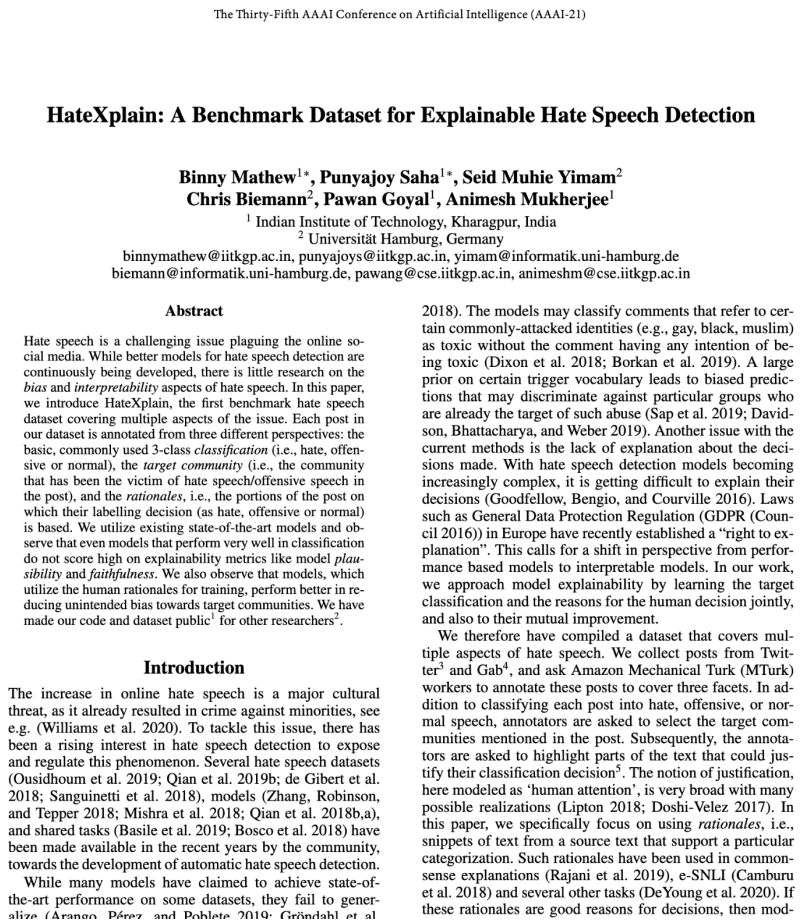


Figure: In the supervised rational attention method, we are incorporating human rationale into models via attentions.

Project 1: Reference



The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection

Binny Mathew[1]*, Punyajoy Saha[1]*, Seid Muhie Yimam[2]
Chris Biemann[2], Pawan Goyal[1], Animesh Mukherjee[1]
[1] Indian Institute of Technology, Kharagpur, India
[2] Universität Hamburg, Germany
binnymathew@iitkgp.ac.in, punyajoys@iitkgp.ac.in, yimam@informatik.uni-hamburg.de
biemann@informatik.uni-hamburg.de, pawang@cse.iitkgp.ac.in, animeshm@cse.iitkgp.ac.in

Abstract

Hate speech is a challenging issue plaguing the online social media. While better models for hate speech detection are continuously being developed, there is little research on the *bias* and *interpretability* aspects of hate speech. In this paper, we introduce HateXplain, the first benchmark hate speech dataset covering multiple aspects of the issue. Each post in our dataset is annotated from three different perspectives: the basic, commonly used 3-class *classification* (i.e., hate, offensive or normal), the *target community* (i.e., the community that has been the victim of hate speech/offensive speech in the post), and the *rationales*, i.e., the portions of the post on which their labelling decision (as hate, offensive or normal) is based. We utilize existing state-of-the-art models and observe that even models that perform very well in classification do not score high on explainability metrics like model *plausibility* and *faithfulness*. We also observe that models, which utilize the human rationales for training, perform better in reducing unintended bias towards target communities. We have made our code and dataset public[1] for other researchers[2].

Introduction

The increase in online hate speech is a major cultural threat, as it already resulted in crime against minorities, see e.g. (Williams et al. 2020). To tackle this issue, there has been a rising interest in hate speech detection to expose and regulate this phenomenon. Several hate speech datasets (Ousidhoum et al. 2019; Qian et al. 2019b; de Gibert et al. 2018; Sanguinetti et al. 2018), models (Zhang, Robinson, and Tepper 2018; Mishra et al. 2018; Qian et al. 2018b,a), and shared tasks (Basile et al. 2019; Bosco et al. 2018) have been made available in the recent years by the community, towards the development of automatic hate speech detection.

While many models have claimed to achieve state-of-the-art performance on some datasets, they fail to generalize (Arango, Pérez, and Poblete 2019; Gröndahl et al. 2018). The models may classify comments that refer to certain commonly-attacked identities (e.g., gay, black, muslim) as toxic without the comment having any intention of being toxic (Dixon et al. 2018; Borkan et al. 2019). A large prior on certain trigger vocabulary leads to biased predictions that may discriminate against particular groups who are already the target of such abuse (Sap et al. 2019; Davidson, Bhattacharya, and Weber 2019). Another issue with the current methods is the lack of explanation about the decisions made. With hate speech detection models becoming increasingly complex, it is getting difficult to explain their decisions (Goodfellow, Bengio, and Courville 2016). Laws such as General Data Protection Regulation (GDPR (Council 2016)) in Europe have recently established a "right to explanation". This calls for a shift in perspective from performance based models to interpretable models. In our work, we approach model explainability by learning the target classification and the reasons for the human decision jointly, and also to their mutual improvement.

We therefore have compiled a dataset that covers multiple aspects of hate speech. We collect posts from Twitter[3] and Gab[4], and ask Amazon Mechanical Turk (MTurk) workers to annotate these posts to cover three facets. In addition to classifying each post into hate, offensive, or normal speech, annotators are asked to select the target communities mentioned in the post. Subsequently, the annotators are asked to highlight parts of the text that could justify their classification decision[5]. The notion of justification, here modeled as 'human attention', is very broad with many possible realizations (Lipton 2018; Doshi-Velez 2017). In this paper, we specifically focus on using *rationales*, i.e., snippets of text from a source text that support a particular categorization. Such rationales have been used in commonsense explanations (Rajani et al. 2019), e-SNLI (Camburu et al. 2018) and several other tasks (DeYoung et al. 2020). If these rationales are good reasons for decisions, then mod-

Figure: https://cdn.aaai.org/ojs/17745/17745-13-21239-1-2-20210518.pdf

Project 1: Reference

HateBRXplain: A Benchmark Dataset with Human-Annotated Rationales for Explainable Hate Speech Detection in Brazilian Portuguese

Isadora Salles
Federal University of Minas Gerais
isadorasalles@dcc.ufmg.br

Francielle Vargas
University of São Paulo
francielleavargas@usp.br

Fabrício Benevenuto
Federal University of Minas Gerais
fabricio@dcc.ufmg.br

Abstract

Nowadays, hate speech technologies are surely relevant in Brazil. Nevertheless, the inability of these technologies to provide reasons (rationales) for their decisions is the limiting factor to their adoption since they comprise bias, which may perpetuate social inequalities when propagated at scale. This scenario highlights the urgency of proposing explainable technologies to address hate speech. However, explainable models heavily depend on data availability with human-annotated rationales, which are scarce, especially for low-resource languages. To fill this relevant gap, we introduce HateBRXplain[1], the first benchmark dataset for hate speech detection in Portuguese, with text span annotations capturing rationales. We evaluated our corpus using mBERT, BERTimbau, DistilBERTimbau, and PTT5 models, which outperformed the current baselines. We further assessed these models' explainability using model-agnostic explanation methods (LIME and SHAP). Results demonstrate plausible post-hoc explanations when compared to human annotations. However, the best-performing hate speech detection models failed to provide faithful rationales.

To mitigate this issue, hate speech detection systems have been developed as effective countermeasures to inhibit offensive and hateful language from being published or spread on the Web and social media. Nonetheless, while there was significant progress in the hate speech research area, for instance, new expert and comprehensive datasets (Vargas et al., 2024a; Guest et al., 2021; Fortuna et al., 2019b; Vargas et al., 2024b), the high performance of deep learning models (Zimmerman et al., 2018; Gambäck and Sikdar, 2017) and transformer architectures (Caselli et al., 2021), these recent models are becoming less interpretable (Tsvetkov et al., 2019) highlighting a lack of transparency posing unwanted risks, such as unintended biases that has recently been identified as a major concern in the area (May et al., 2019).

In modern Natural Language Processing (NLP), a prevalent approach to building hate speech classifiers consists of training on hate speech datasets using fine-tunning Large-Scale Language Models (LLMs), which, according to (Davani et al., 2023), leads to representational biases, such as preferring European American names over African American names (Caliskan et al., 2017), associating words with more negative sentiment against persons with

Figure: https://aclanthology.org/2025.coling-main.446/

Project 1: Reference

Inferring Which Medical Treatments Work from Reports of Clinical Trials

Eric Lehman
Northeastern University
lehman.e@northeastern.edu

Jay B. DeYoung
Northeastern University
deyoung.j@northeastern.edu

Regina Barzilay
MIT
regina@csail.mit.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Abstract

How do we know if a particular medical treatment actually works? Ideally one would consult all available evidence from relevant clinical trials. Unfortunately, such results are primarily disseminated in natural language scientific articles, imposing substantial burden on those trying to make sense of them. In this paper, we present a new task and corpus for making this unstructured evidence actionable. The task entails inferring reported findings from a full-text article describing a randomized controlled trial (RCT) with respect to a given intervention, comparator, and outcome of interest, e.g., inferring if an article provides evidence supporting the use of *aspirin* to reduce *risk of stroke*, as compared to *placebo*.

We present a new corpus for this task comprising 10,000+ prompts coupled with full-text articles describing RCTs. Results using a suite of models — ranging from heuristic (rule-based) approaches to attentive neural architectures — demonstrate the difficulty of the task, which we believe largely owes to the lengthy, technical input texts. To facilitate further work on this important, challenging problem we make the corpus, documentation, a website and leaderboard, and code for baselines and evaluation available at http://evidence-inference.ebm-nlp.com/.

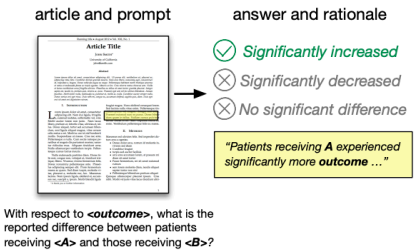


Figure 1: The task. Given a treatment **A**, a comparator **B**, and an **outcome**, infer the reported relationship between **A** and **B** with respect to **outcome**, and provide evidence supporting this from the text.

as: *Does infliximab reduce dysmenorrhea (pain) scores, relative to placebo*?

Given the critical role published reports of trials play in informing evidence-based care, organizations such as the Cochrane collaboration and groups at evidence-based practice centers (EPCs) are dedicated to manually synthesizing findings, but struggle to keep up with the literature (Tsafnat et al., 2013). NLP can play a key role in automating this process, thereby mitigating costs and keeping treatment recommendations up-to-date with the evidence as it is published

Figure: https://aclanthology.org/N19-1371.pdf

Thank you.

https://franciellevargas.github.io/