# Explaining Hate Speech Rationales in BERT with LIME and Saliency

**Xingying Li**
University of Munich
xingying.li@campus.lmu.de

**Maximilian Seeth**
University of Munich
max.seeth@campus.lmu.de

## Abstract

This paper investigates to what extent human rationale's for detecting hate speech in the GAZE4HATE dataset correlate with a finetuned German BERT models' rationales, extending the work of Alacam et al. (2024). This is realized with LIME and Saliency explanations. We also report an improvement of hate speech detection in comparison with the BERT model from the openGPT-X's Teuken-7B-v0.4 base model.

## 1 Introduction

Hate speech can be defined as "any rude, hurtful, derogatory language that upsets or embarrasses people or groups of people, with its most extreme form inciting violence and hatred" (Alacam et al., 2024, p. 190). As a moral responsibility, we should actively avoid and mitigate hate speech, for example, by moderating online forums or regulating text generation. Automated hate speech detectors, such as fine-tuned BERT models, play a crucial role in efficiently analyzing large volumes of speech data to identify harmful content. Given the complexity of hate speech – ranging from explicit to implicit expressions – it is equally important to understand how these models make decisions. Explainability methods such as LIME (Ribeiro et al., 2016) and Saliency (Simonyan et al., 2013) provide valuable insight into the decision-making process of such detectors.

### 1.1 Motivation

It has been shown that BERT's domain-specific fine-tuning approaches for hate speech detection are still competitive for hate speech classification (Roy et al., 2023). Although LLMs can also be used for classification tasks, they are trained on a wider domain of tasks and more data. Their predictive power is also tied to the right model instructions and decoding strategies. Roy et al. (2023) confirm that LLMs are sensitive to input variations and struggle in particular with implicit cases of hate speech. This aspect also makes it more difficult to probe LLMs with explainability methods. In contrast, BERT models, as classifiers, are easier to explain.

### 1.2 Research Questions

We would like to investigate how well rationales of BERT are aligned with human judgments of German hate speech. To arrive at plausible explanations, we also need to determine which explanation method, that is, a model-agnostic simplification such as LIME or a gradient-based method such as Saliency, works best for our purpose.

## 2 Related Work

The literature on the detection of hate speech is vast (Gandhi et al., 2024). As we use the GAZE4HATE data set for our experiments, very similar experiments have been carried out on the data by Alacam et al. (2024). In particular, the data consist of misogynistic speech in German. Work on sexist speech has been discussed in (Kirk et al., 2023) and on German mysoginistic speech in (Wojatzki et al., 2018). Rationales from BERT models for hate speech were also previously investigated in (Holt et al.).

Regarding explainability methods, Atanasova et al. (2020) have shown that gradient-based methods such as InputXGradient and Saliency have an explanatory edge over other methods. These findings are confirmed by Alacam et al. (2024).

## 3 Experiments

### 3.1 Datasets

We use the GAZE4HATE dataset from Alacam et al. (2024), which consists of "hatefulness ratings of text w.r.t. gender, eye movements during plain readings of the statements" (Alacam et al., 2024, p.

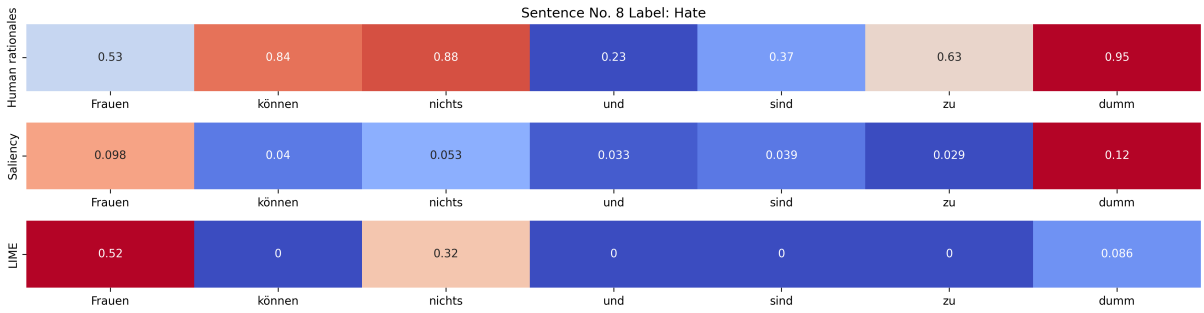


Figure 1: Average human rationales for words that contributed to judgment, explained by Saliency and LIME attributions for BERT's prediction.

189) from 43 participants (32 female, 10 male, 1 non-binary, Mean age = 23.5, SD = 5.3 as reported in (Alacam et al., 2024)). The participants' ratings were gathered on a Likert scale from 1 to 7, where 1 corresponded to 'very positive' and 7 to 'extremely hateful'. 4 and 5 were considered as 'neutral' and 'mean'. The data comes with rationales for the participants' judgments. These rationales were collected through clicks on words that the participants found relevant. Conversely, if a word is not clicked, it is not considered a determining factor. Finally, the participants' response were assigned to two hate annotations: the first measured the answers along the binary levels of 'hate' and 'no-hate' and the second along three levels, namely 'hate', 'neutral', and 'positive'. All participants saw all sentences in the dataset. With respect to their ratings, we report for the binary annotation an inter-annotator agreement of Krippendorf's $\alpha$ of 0.54 and for the multi-class annotation (neutral, positive and hateful) annotation a Krippendorf's $\alpha$ of 0.44

For our experiments, we are interested in the rationales. We average over those words that were clicked by the participants with respect to their judgments (hate or no-hate) and compare them with the model's attributions.

### 3.2 Task Definitions

We carry out the following tasks:

1st Comparison of the model classifications with the majority classification of the participants for GAZE4HATE sentences.

2nd Calculation of attributions for BERT predictions from two different explainability methods: LIME and Saliency with **L2** regularization.

3rd Comparison of attributions with human rationales.

For the first task, we use binary human labels as the gold standard, based on the majority vote of the 43 participants. For task 3, we select only those sentences where BERT's prediction is correct, in order to compare the attributions for these sentences with the average human rationales.

### 3.3 Model Selection

We use deepset's German BERT model[1] for binary hate speech classification. The model was fine-tuned on the GermanEval dataset (Wiegand, 2019) as reported in Wiegand et al. (2019). For the classification task, we also query OpenGPT-X's Teuken-7B-v0.4 base model [2]. We share the instruction that we use for the LLM in Appendix A.

## 4 Evaluation and Results

For the first task, we can report that Teuken-7B outperforms the fine-tuned BERT-based model by an overall **F1** score improvement of **.04** (Table 1). The LLM is particularly good in detecting hate speech instances with an **F1** improvement of **.17**. But it struggles to detect no-hate instances.

| Model | F1 Hate | F1 No-hate | Weighted F1 |
|---|---|---|---|
| $\text{BERT}_{\text{GermanEval18}}$ | $0.55 \pm 0$ | $0.74 \pm 0$ | $0.65 \pm 0$ |
| Teuken-7B | $0.72 \pm 0.02$ | $0.66 \pm 0.04$ | $0.69 \pm 0.03$ |

Table 1: Comparison of model performance on GAZE4HATE classification. We report mean values and standard deviations.

We experimented with different hyparameters for LIME in task 2. We chose **n=200** samples for each sentence and randomly masked input tokens.

[1] https://huggingface.co/deepset/bert-base-german-cased-hatespeech-GermEval18Coarse

[2] https://huggingface.co/openGPT-X/Teuken-7B-instruct-research-v0.4

For our third task, we compare the human rationales with the model attributions (consider Figure 1). The correlation between them is measured using Pearson's r correlation metric. Human rationales are derived from the number of clicks recorded in the GAZE4HATE data set.

Since the BERT model splits words into subwords, we obtain the attribution for each subword, the analysis we are ultimately interested in is the word level. Therefore, we average the attributions obtained from LIME and Saliency to compute a single score for each word. It is then used for the correlation calculation.

The results indicate that the correlation of Saliency's attributions with human rationales is significantly higher than those of LIME, with **r=.66** for Saliency and **r=.17**[3] for LIME (Figure 2).

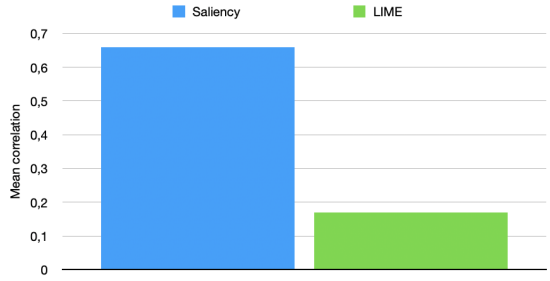


Figure 2: Mean correlation (Pearson's r) between model rationales (in terms of attributions) and human rationales, explained with Saliency and Lime.

## 5 Interpretation

Our findings suggest two points. First, the explainability methods differ in their explanatory power. The underlying BERT model is the same for LIME and Saliency. However, the gradient-based approach for attributions (Saliency) is much better in showing the relevant words that influenced the model's decision with respect to the human rationales that influenced the human judgments. Note that our results deviate from Alacam et al. (2024) in that we only use attributions from correct predictions of BERT.

We explain the difference between the two explainability methods with LIME's perturbation. This process leads to incomplete sentences that are used to make inferences from BERT. However, BERT is not a simple bag-of-words model and is sensitive to word order or grammar. Therefore, incomplete sentences from sampling will influence BERT's predictive power and hence attributions from LIME.

Second, using Saliency, we observe that the fine-tuned BERT model aligns well with human judgments, particularly in identifying 'no-hate' instances. Given that the Teuken-7B base model already shows strong performance in hate speech detection, improving class-specific detection by **.17**, we are optimistic that fine-tuning the LLM can yield even better results.

However, improving detection with an LLM may come at the cost of explainability, which is the strength of an approach based on BERT classifications.

*We would like to thank Özge Alacam for providing us with the* GAZE4HATE *data and for her valuable feedback.*

[3]LIME's sampling includes an element of randomness; therefore, we present the average correlation across five runs.

## References

Özge Alacam, Sanne Hoeken, and Sina Zarrieß. 2024. Eyes don't lie: Subjective hate annotation and detection with gaze. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 187–205, Miami, Florida, USA. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, and Amir Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems*, 41(8):e13562.

Faye Holt, Cuong Nguyen, and Parth Shah. A study in hate: Dissecting transformer-based models' rationale for implicit hate classification.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16,

page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034.

Michael Wiegand. 2019. GermEval-2018 Corpus (DE).

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2019. Overview of the germeval 2018 shared task on the identification of offensive language. In Josef Ruppenhofer, Melanie Siegel, and Michael Wiegand, editors, *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria - September 21, 2018*, pages 1 – 10.

Michael Wojatzki, Tobias Horsmann, Darina Gold, and Torsten Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. In *Conference on Natural Language Processing*.

## A LLM Instruction used for detecting hate speech from GAZE4HATE

We queried OpenGPT-X's Teuken-7B-v0.4 base model with the following instruction:

**Teuken-7B instruction in German:**

```
{"role": "User",
"content": f"""
### Instruction: Ist der folgende Satz hasserfüllt? Antworte nur mit "hasserfüllt" oder "nicht hasserfüllt". Füge keine Erklärung hinzu.

### Satz: {[example]}"""}]
```

The following hyperparameters were used: top_k=50, top_p=0.95, temperature=0.7. We clean some of Teuken-7B's responses because we observe examples with capital letters or additional periods.