Explainable AI in Text Classification

1st Semester of 2025-2026

Vlad Mihai Olăeriu
vlad-mihai.olaeriu@s.unibuc.ro

Abstract

Explainable Artificial Intelligence (XAI) is critical for understanding NLP models used in sensitive domains such as hate speech detection and depression classification. Transformer-based models perform well, but their decision-making procedures are intractable, which is problematic for high-stakes applications like clinical text analysis and content moderation. The reliability and stability of explanation techniques for a hate speech classifier across several target categories are evaluated in this work. Internal attention weights are compared with SHAP and LIME, two post-hoc attribution methods. Entropy, Gini coefficient, and top-k mass concentration are used to examine token importance distributions, while ranking correlations are used for assessing cross-method agreement. Adversarial synonym substitutions are used to further evaluate robustness, and flip rate and confidence change are provided as stability metrics.

1 Introduction

The introduction of the attention mechanism has revolutionized the field of Natural Language Processing (NLP), enabling models to weight the relevance of different parts of an input sequence. By allowing a model to *attend* to specific tokens, attention has significantly improved performance across a wide array of tasks, from neural machine translation to sentiment analysis, and to building large language models.

Despite its performance benefits, a debate has emerged regarding the extent to which attention weights can serve as a reliable *explanation* for a model's decisions. On one hand, the intuitive nature of attention heatmaps leads many researchers to use them as feature importance scores, assuming that a high weight directly corresponds to an influential token in the output. On the other hand, recent critical evaluations have challenged this assumption and come to the conclusion that they are too unstable to be directly used as explanation features.

The stakes for this debate are particularly high in sensitive domains like biomedical NLP and social media monitoring, where the *black-box* nature of models can have damaging consequences. In clinical applications such as depression detection, identifying specific linguistic markers is essential for diagnostic support and clinician trust. Similarly, in hate speech classification of social media posts or comments, models must accurately isolate derogatory tokens to justify content moderation decisions. Thus, it is critical to evaluate the stability and reliability of these explanations.

This work aims at exploring the behavior of a hate speech classifier for different hate speech categories, such as race, religion, or origin, by analyzing the attention weights, LIME, and SHAP scores of each category's tokens, examining their stability, and finding how much they correlate.

2 Related Work

Recent research on explainability methods for NLP tasks has led to similar conclusions that attention weights cannot be purely used as an explainability metric. (Jain and Wallace, 2019) shows that most of the time, learned attention weights are frequently uncorrelated or only weakly correlated with gradient-based and leave-one-out measures of feature importance. Moreover, they demonstrate that by building adversarial attention distributions, entirely different from the learned ones, it can result in the same model prediction. Thus, with weak correlation with more natural features and counterfactual attention distributions, the ability of attention weights to provide meaningful explanations is at least questionable.

In addition to existing research, (Bastings and Filippova, 2020) argues that if the goal is to identify important input tokens, researchers should use dedicated saliency methods rather than attention weights. The authors argue that these input saliency methods are fundamentally better suited for identifying the most relevant tokens because they typically account for the entire computation path from input embeddings to the final prediction. In contrast, attention weights only reflect a single point in the computation and may operate over representations that have already mixed information from other inputs. The saliency methods proposed by them include gradient-based, propagation-based, and occlusion-based methods.

3 Method

3.1 Dataset

The hate speech dataset introduced by (Kennedy et al., 2020) consists of 39,565 comments that were annotated by 7,912 annotators, with the hate speech score being the principal outcome variable, but also including labels for eight identity groups: race, religion, gender, origin, age, sexuality, disability, and politics.

The hate speech score is a continuous value, for which>0.5 is most likely hate speech, <-1 is supportive speech, and everything between is neutral or ambiguous. For this experiment, only the hate speech comments are of interest. To keep them, all the rows of the dataset were aggregated, the mean of the hate speech score was computed and kept as the final score, and the final target category for each comment was represented by the category most present among the annotations.

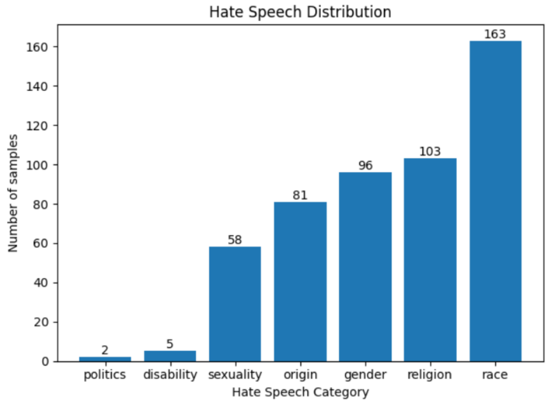

Figure 1: Hate Speech Target Group Distribution

As can be observed in Figure 1, after the aggregation of the rows, one target category disappeared completely, *age*, and two of them have only a few samples: *politics* and *disability*. The less present target groups were also removed to avoid imbalances among the explanability scores.

3.2 Classifier

The model used for classification of the hate speech comments was the fine-tuned RoBERTa-large model **facebook/roberta-hate-speech-dynabench-r4-target** optimized for detecting hate speech. The data that was trained on was collected through four rounds of human-and-model-in-the-loop training, in which human annotators tried to fool previous versions of the model into making mistakes. Because of this training method, the model is performant in identifying implicit hate speech and distinguishing it from non-hateful but noisy text.

3.3 Explanation Features

3.3.1 Attention Weights

The default RoBERTa model has 24 layers and 16 attention heads. We are interested in the attention weights of the last L layers and of all attention heads, the total number of heads being denoted with H. In the case of the hate speech classifier, the query token will be the $[CLS]$ token, and the key will be the token for each word we want to compute the attention weight.

$$\text{attention score}_k = \sum_{i}^{L} \sum_{j}^{H} \frac{1}{H} a_{ijk}$$

In the above formula, a_{ij} represents the attention at layer i at head j for the k^th token. The attention from a single layer will be the mean attention from all of its heads, and afterwards, these means will be summed. In our experiments, we set $L = 4$ and $H = 12$, to use the last four layers with all of their attention heads.

3.3.2 LIME

LIME (Ribeiro et al., 2016), Local Interpretable Model-agnostic Explanations, provides an explanation of the model's behavior by fitting a linear surrogate model around perturbed versions of an input instance. For each text sample, the `LimeTextExplainer` is used to perturb the input by randomly masking tokens and observing the resulting changes in the model's class probabilities.

In the implementation, the number of features is set to the number of non-special tokens in the current input sequence, ensuring that the explainer focuses only on meaningful words. A total of $N = 1000$ samples were generated for the local linear regression to ensure a stable approximation of the decision boundary. The resulting feature attribution for a token k is the coefficient of the linear model, representing that token's contribution to the predicted class $y = \text{hate speech}$:

$$\text{LIME}_k = w_{yk}$$

where w represents the learned weights of the surrogate model. These weights provide a local measure of importance, where a positive score indicates that the token's presence increases the model's confidence in the predicted hate speech category.

3.3.3 SHAP

SHAP (Lundberg and Lee, 2017), SHapley Additive exPlanations0, which uses a game-theory-based framework to assign an importance score to each token, representing its contribution to the model's prediction. In this experiment, we extracted the logits of the RoBERTa model to calculate the log-odds of the hate label versus the non-hate label.

To compute the score for a token k, SHAP analyzes how the model is behaving in its absence by calculating the average marginal contribution of that token across all possible subsets of other input tokens:

$$\phi_k = \sum_{S \subseteq N \setminus \{k\}} \frac{|S|!(n - |S| - 1)!}{n!} [f(S \cup \{k\}) - f(S)]$$

where N is the total number of tokens of the input, $f(S)$ are the logits of the model when the tokens in S are present. Using logit units ensures that the SHAP values are additive; the sum of the attributions plus a base value equals the model's current prediction.

3.3.4 Comparison

The primary objective is to investigate how the model's reasoning logic differs when identifying different hate speech categories, and to compare the alignment of internal mechanisms, attention weights, with external evidence of feature importance, as well as LIME and SHAP scores.

For hate speech category differentiation, the attention weights, LIME, and SHAP scores were normalized per sample by applying softmax over the outputs to obtain a probability distribution of token importance. Afterwards, multiple metrics were computed for each sample's score distribution: normalized entropy, Gini coefficient, and top-k mass concentration with $k = 5$. Afterwards, these metrics were averaged per hate speech category to identify which category is best represented by the classifier.

To compare the explainers' scores, different **ranking metrics** for tokens were used: Spearman correlation, Kendall Tau's coefficient, and top-k Jaccard overlap, with $k = 5$. Each explainer was compared to the other two, and the average ranking score was computed for each hate speech category to see how well they correlate with each other.

3.4 Adversarial Examples

To further examine the stability of each explainer, it was used by adversarial examples that would substitute tokens with their synonyms for tokens with a high score, respectively, with a high attention weight. The question that needs to be answered is *how stable are the model's predictions when high score tokens are replaced by semantically similar synonyms?*. For this experiment, the top-5 tokens with the highest score, separately for each explainer, were substituted with synonyms.

Flip rate (FR) represents the number of samples that were labeled as not hate speech after the synonyms substitution was applied to the samples. The average flip rate for each hate speech category was outlined.

$$\text{FR} = \frac{\text{\# non hate speech prediction}}{\text{\# perturbed examples}}$$

Confidence distribution (CD) represents how much the probability of being hate speech was changed after the synonym substitution. For this distribution, the mean and standard deviation were reported per hate speech category.

$$\text{CD} = p_{\text{hate speech}}(\text{original}) - p_{\text{hate speech}}(\text{perturbed})$$

4 Results

4.1 Explanation Features

Tables 1, 2, and 3 show the distribution metrics for attention weights, LIME, and SHAP respectively. A notable observation across all three explainers is the very high entropy values (ranging from 0.93 to 0.99), which indicates that the importance scores are distributed relatively uniformly across tokens.

This means that for most samples, no single token or small set of tokens dominates the explanation. In other words, the explainers often assign similar importance to many different tokens rather than identifying a few critical ones.

The Gini coefficient, which measures inequality in the distribution of scores, is consistently low across all explainers (0.06 to 0.23). This further confirms the uniform nature of the distributions. The top-5 mass concentration provides an additional perspective on how much importance is concentrated in the top-5 tokens. SHAP shows the highest top-5 mass values (0.40 to 0.68), meaning it concentrates more importance on a small number of tokens.

When looking at specific hate speech categories, sexuality shows notable differences. For top-5 mass, sexuality has the highest values across all three explainers (0.61 to 0.75), suggesting that hate speech targeting sexuality is characterized by a smaller set of discriminative tokens. In contrast, origin shows the lowest top-5 mass values (0.36 to 0.42), indicating that hate speech targeting origin often uses more varied language patterns and tokens.

Category	**Entropy**	**Gini**	**Top-5 Mass**
Gender	0.9874	0.0922	0.4490
Origin	0.9885	**0.0936**	0.3581
Race	0.9891	0.0851	0.4068
Religion	**0.9884**	0.0850	0.3625
Sexuality	0.9926	0.0786	**0.6108**

Table 1: Distribution Metrics for Attention Weights

Category	**Entropy**	**Gini**	**Top-5 Mass**
Gender	0.9936	0.0636	0.5131
Origin	0.9961	0.0565	0.4223
Race	0.9945	0.0575	0.5068
Religion	0.9939	0.0617	0.4575
Sexuality	**0.9900**	**0.0671**	**0.7538**

Table 2: Distribution Metrics for LIME

Category	**Entropy**	**Gini**	**Top-5 Mass**
Gender	0.9449	0.2181	0.5193
Origin	0.9643	0.1820	0.4036
Race	0.9583	0.1945	0.4659
Religion	0.9569	0.1922	0.4169
Sexuality	**0.9348**	**0.2284**	**0.6840**

Table 3: Distribution Metrics for SHAP

Table 4 shows the ranking metrics comparing attention weights and SHAP. The Spearman correlation values are very low (ranging from -0.18 to 0.05), with most categories showing near-zero or negative correlation. The Kendall Tau values are similarly low, ranging from -0.16 to 0.02. These results indicate poor agreement between attention weights and SHAP in terms of token ranking.

The Jaccard overlap for top-5 tokens is also limited, ranging from 0.09 to 0.18. This means that attention weights and SHAP rarely identify the same set of top-5 important tokens. The sexuality category shows the highest Jaccard overlap (0.18), which aligns with the observation that sexuality has more concentrated importance scores.

These low correlation values suggest that attention weights and SHAP are capturing different aspects of model behavior. This finding supports the argument from recent research that attention weights alone should not be used as the sole explanation for model decisions, and that multiple explanation methods should be used in combination for more robust interpretability.

Category	**S-ρ**	**K-τ**	**Jaccard**
Gender	0.0331	0.02323	0.0859
Origin	0.0477	0.0388	0.0753
Race	0.0152	0.0075	0.0947
Religion	0.0148	0.0103	0.0810
Sexuality	**-0.1779**	**-0.1574**	**0.1784**

Table 4: Ranking Metrics for Attention Weights - SHAP

4.2 Adversarial Examples

The adversarial experiments tested the stability of model predictions when top-5 important tokens, as identified by each explainer, were replaced with their WordNet synonyms. Tables 5, 6, and 7 show the flip rate and confidence distribution metrics for each explainer and hate speech category.

Overall, the flip rates are relatively low across all explainers and categories, ranging from 1.04% to 12.35%. However, notable differences exist between explainers and categories. SHAP shows the highest overall flip rates, suggesting that tokens identified as important by SHAP are more critical to the model's hate speech predictions.LIME shows intermediate flip rates, while attention weights show the lowest flip rates. This finding indicates that SHAP-identified tokens have a stronger influence on model predictions compared to tokens identified by other explainers.

The confidence distribution (CD) measures how much the probability of the hate speech class changed after token replacement. The mean CD values are generally small (0.01 to 0.08), indicating that even when tokens are replaced, the model's confidence levels do not change dramatically. The standard deviation of CD is higher (0.22 to 0.37), showing significant variation in how different samples respond to perturbation.

Category	FR	CD-μ	CD-σ
Gender	**1.04%**	-0.0785	**0.2790**
Origin	7.41%	-0.0404	0.3505
Race	3.68%	-0.0693	0.3023
Religion	6.86%	0.0102	0.3457
Sexuality	5.17%	**0.0027**	0.3110

Table 5: Adversarial Metrics for Attention Weights

Category	FR	CD-μ	CD-σ
Gender	**2.08%**	-0.0585	0.2894
Origin	9.88%	-0.0216	0.3627
Race	2.45%	-0.0498	0.2749
Religion	4.90%	0.0239	0.2416
Sexuality	3.45%	**0.0160**	**0.2179**

Table 6: Adversarial Metrics for LIME

Category	FR	CD-μ	CD-σ
Gender	9.38%	-0.0638	0.3018
Origin	12.35%	-0.0222	0.3673
Race	**6.13%**	-0.0537	0.2799
Religion	6.86%	0.0253	0.2451
Sexuality	6.90%	**0.0166**	**0.2218**

Table 7: Adversarial Metrics for SHAP

5 Conclusions

By contrasting internal attention mechanisms with exterior saliency techniques like LIME and SHAP, this study investigated the interpretability of hate speech classification. Although our approach to assessing stability via adversarial perturbations and cross-method correlation offers a strong foundation for XAI research, the findings show a notable discrepancy between the factors that models consider and the factors that influence their choices. The weak association between attention weights and SHAP scores, in particular, raises the possibility that conventional attention mechanisms might not be adequate stand-alone explanations for model behavior in delicate domains.

The adversarial experiments demonstrated that while SHAP identified more influential tokens, model confidence remained relatively stable even after synonym substitution. These results highlight the difficulty of detecting hate speech and indicate that more reliable, naturally interpretable attention-based architectures should be the focus of future studies.Despite the current lack of alignment between methods, this work contributes to the vital goal of building trustworthy NLP systems, providing a positive foundation for developing more transparent models that can be effectively justified to clinicians and content moderators alike.

References

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? *arXiv preprint arXiv:2010.05607*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

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*, pages 1135–1144.