Evaluating the Consistency of Explainable AI Methods in Hate Speech Detection on Social Media Platforms

Author: Gonzalo Gamez
Study Programme: Computer Sciences, 45483
Supervisor: Professor, Dr. sc. ing. Dmitry Pavlyuk

TRANSPORT AND TELECOMMUNICATION INSTITUTE

Introduction

As AI-assisted content moderation becomes increasingly prevalent, the ability to understand and trust the decisions of these systems is critical, especially in light of regulations like the EU AI Act's (European Union, 2024) "Right to Explanation". This research investigates the consistency of two prominent Explainable AI (XAI) methods, LIME (Local Interpretable Model-agnostic Explanations (Ribeiro et al., 2016)) and SHAP (SHapley Additive exPlanations (Lundberg & Lee, 2017)), when applied to a state-of-the-art hate speech detection model.

- **Research Question:** How do LIME and SHAP compare in terms of the consistency of their generated explanations for hate speech detection models under repeated applications?
- **Hypothesis:** SHAP provides more consistent explanations than LIME when applied to hate speech detection models.
- **Objective:** To systematically evaluate and compare the consistency of feature contributions generated by LIME and SHAP for the same hate speech predictions across multiple runs, thereby assessing their reliability for practical applications.

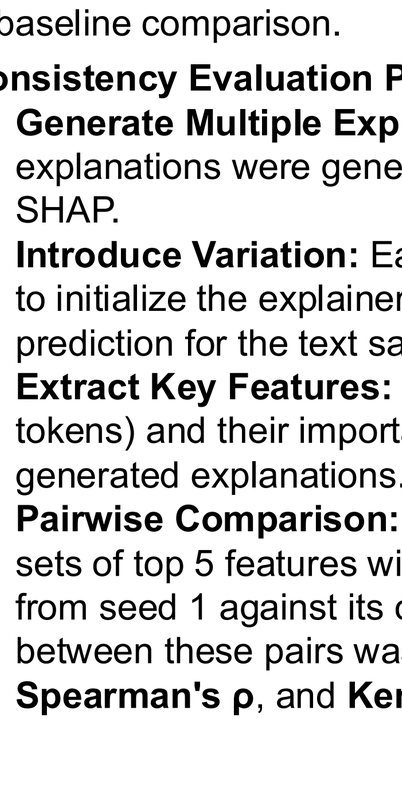
Methodology

A multi-stage methodology was employed, beginning with data aggregation and culminating in a comparative analysis of XAI consistency.

- **Dataset:** A unified corpus was created by aggregating three datasets: HateXplain, MLMA, and Measuring Hate Speech, primarily sourced from Twitter, YouTube, Reddit, and Gab.
- **AI Model:** The cardiffnlp/twitter-roberta-base-hate-latest model, a RoBERTa-based architecture fine-tuned for hate speech text detection.
- **XAI Methods:** LIME and SHAP were applied to generate explanations for model predictions using their default parameters to ensure a baseline comparison.

Consistency Evaluation Process

1. **Generate Multiple Explanations:** For each text sample, four separate explanations were generated using LIME, and four were generated using SHAP.
2. **Introduce Variation:** Each of the four runs used a different random seed to initialize the explainer. The underlying hate speech model and its prediction for the text sample remained constant throughout this process.
3. **Extract Key Features:** The **top 5 most important features** (words or tokens) and their importance scores were extracted from each of the generated explanations.
4. **Pairwise Comparison:** A pairwise comparison was conducted on the sets of top 5 features within each method (e.g., comparing LIME's output from seed 1 against its outputs from seeds 2, 3, and 4). The consistency between these pairs was then quantified using **Jaccard Similarity**, **Spearman's ρ**, and **Kendall's τb** metrics.



Process for Using XAI Methods adapted from Gongane et al. (2024)

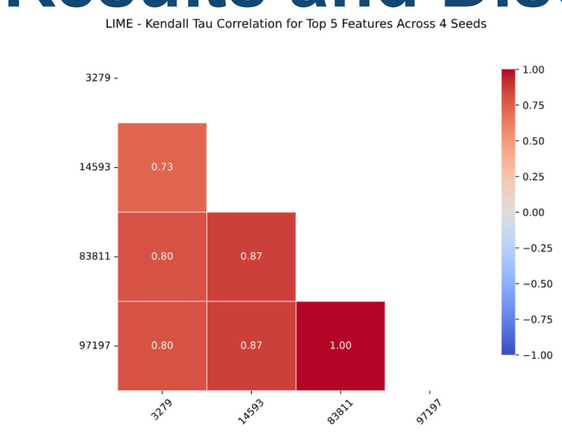
LIME Top-5 Feature Contributions Across Different Seeds From Explanations

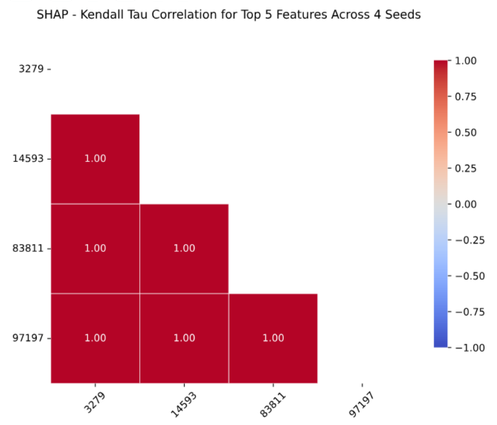
id	seed	lime_top5
18694	3279	[('retarded', -0.530), ('retards', -0.296), ('world', 0.050), ('predis-poses', 0.042), ('their', 0.041)]
18694	14593	[('retarded', -0.508), ('retards', -0.319), ('their', 0.054), ('killing', -0.041), ('and', 0.038)]
18694	83811	[('retarded', -0.532), ('retards', -0.299), ('their', 0.052), ('predis-poses', 0.039), ('world', 0.038)]
18694	97197	[('retarded', -0.515), ('retards', -0.308), ('world', 0.057), ('their', 0.053), ('predisposes', 0.045)]

SHAP Top-5 Feature Contributions Across Different Seeds From Explanations

id	seed	shap_top5
18694	3279	[('retarded', -0.245), ('ret', -0.244), ('ards', -0.243), ('animals', -0.116), ('killing', -0.097)]
18694	14593	[('retarded', -0.245), ('ret', -0.244), ('ards', -0.243), ('animals', -0.116), ('killing', -0.097)]
18694	83811	[('retarded', -0.245), ('ret', -0.244), ('ards', -0.243), ('animals', -0.116), ('killing', -0.097)]
18694	97197	[('retarded', -0.245), ('ret', -0.244), ('ards', -0.243), ('animals', -0.116), ('killing', -0.097)]

Results and Discussion





Aggregated Consistency Metrics (Mean & Standard Deviation)

XAI Method	Jaccard Similarity Mean (std)	Spearman's ρ Mean (std)	Kendall's τ_b Mean (std)
LIME	0.835 (0.193)	0.851 (0.206)	0.776 (0.235)
SHAP	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)

The study's results demonstrate a difference in the consistency of explanations generated by LIME and SHAP.

- The research hypothesis was confirmed: SHAP provides significantly more consistent explanations than LIME for the hate speech detection model.
- SHAP achieved perfect consistency scores (1.0) across all metrics and random seeds, indicating its explanations are reproducible with default parameters.
- LIME exhibited moderate consistency, with scores ranging depending on which metric measured. Its explanations showed variability, which is attributable to the random perturbation process inherent in its methodology.

Limitations:

• Methodological: Limited to three consistency metrics and top-5 features analysis
• Computational: Constrained sample size (652 instances) due to processing requirements and tested on only one deep learning model
• Scope: Default parameters only; focused on high-confidence predictions
• Human factors: Technical consistency assessed without direct human usability evaluations

Conclusions / Outcomes

- Outcome: SHAP demonstrates superior consistency for feature contributions provided from explanations compared to LIME across metrics for hate speech text detection

Key References

• European Union. (2024). Regulation (EU) 2024/1689 on artificial intelligence (AI Act). Official Journal of the European Union.
• Gongane, V.U., Munot, M.V., & Anuse, A.D. (2024). A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms. Journal of Computational Social Science, 7.
• Lundberg, S.M., & Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 4765-4774.
• Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference, 1135-1144.

Project Site