

# Benchmarking Individual and Ensemble XAI Methods for Arabic Transformer-Based Text Classification

Youssef Chafiqui

ychafiqui2@gmail.com

Hassan II University

Houda Anoun

Hassan II University

---

## Research Article

**Keywords:** Explainable AI (XAI), Natural language processing, Transformers, Arabic text classification, SHAP, LIME, Integrated Gradients, DeepLIFT

**Posted Date:** October 28th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11093635/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Benchmarking Individual and Ensemble XAI Methods for Arabic Transformer-Based Text Classification

Youssef Chafiqui<sup>1\*</sup> and Houda Anoun<sup>1</sup>

<sup>1</sup>RITM Laboratory, Hassan II University, Casablanca, Morocco.

\*Corresponding author(s). E-mail(s): [ychafiqui2@gmail.com](mailto:yachafiqui2@gmail.com);  
Contributing authors: [houda.anoun@gmail.com](mailto:houda.anoun@gmail.com);

## Abstract

Explainable Artificial Intelligence (XAI) is important for interpreting transformer-based models, yet explanation quality in Arabic language remains underexplored. This work presents a comprehensive benchmark of attribution XAI methods for Arabic sentiment analysis and hate speech detection, evaluating LIME, SHAP, Integrated Gradients, DeepLIFT, and ensemble variants across three Arabic transformer models. We assess explanations using faithfulness and plausibility metrics and complement score-based faithfulness evaluation with rank-based aggregation via Borda count. We show that selective ensembling yields a statistically significant but modest improvement over individual XAI methods in faithfulness, with LIME+SHAP improving the primary Combined score by an average paired difference of approximately 0.012 over LIME in sentiment classification. The same general trend persists under cross-task evaluation on hate-speech detection. In contrast, ensemble configurations provide little improvement in agreement with human-annotated rationales, demonstrating that improved model-centered faithfulness does not necessarily translate to human-centered plausibility. We additionally introduce a parameter-free, automated rationale-selection method based on elbow-detection that achieves competitive faithfulness with lower computational cost. Overall, the study provides a rigorous, reproducible benchmark and practical guidance for efficient and reliable XAI in Arabic NLP. The code to reproduce the benchmark is available at: [https://github.com/yachafiqui/arabic\\_xai\\_benchmark](https://github.com/yachafiqui/arabic_xai_benchmark)

**Keywords:** Explainable AI (XAI), Natural language processing, Transformers, Arabic text classification, SHAP, LIME, Integrated Gradients, DeepLIFT

# 1 Introduction

In recent years, transformer-based models [1] have achieved remarkable performance in Natural Language Processing (NLP). However, their complex architectures often make their predictions difficult to interpret, raising concerns in applications where understanding the rationale behind a model’s decision is essential. Explainable Artificial Intelligence (XAI) addresses this challenge by providing human-understandable explanations of model predictions. Despite growing interest in XAI, its application to Arabic NLP remains relatively underexplored.

Existing XAI studies for Arabic NLP have largely focused on applying individual explanation methods, while systematic evidence and benchmarks regarding ensemble explanations remains limited. Moreover, explanation quality is multidimensional: improvements under model-centered faithfulness criteria do not necessarily imply stronger agreement with human rationales. There is therefore a need for controlled benchmarks that compare individual and ensemble attribution methods under common experimental conditions, across multiple tasks, and from both faithfulness and plausibility perspectives.

Motivated by these gaps, this study addresses the following research questions:

1. How do individual and ensemble attribution methods compare under complementary faithfulness criteria for Arabic transformer-based text classification?
2. Do ensemble effects generalize across Arabic classification tasks, and how do they depend on ensemble composition?
3. Do improvements in model-centered faithfulness translate to agreement with human-annotated rationales?
4. Can we introduce automated rationale selection techniques that achieve competitive performance without relying on manually specified parameters?

To address these questions, we evaluate LIME, SHAP, Integrated Gradients, DeepLIFT, and multiple ensemble configurations across three Arabic transformer classifiers and two classification tasks. The main contributions of this work are:

- A benchmark of individual and ensemble XAI methods for Arabic text classification, covering sentiment analysis and hate-speech detection across three transformer models.
- A multidimensional evaluation framework combining five complementary faithfulness metrics, score- and rank-based aggregation, robustness analysis, and plausibility evaluation against human-annotated rationales.
- A parameter-free residual-based rationale-selection method using elbow detection that provides competitive faithfulness while substantially reducing computational cost compared to existing automated elbow-detection techniques.

The remainder of this paper is organized as follows. Section 2 introduces Explainable AI methods. Section 3 reviews related work. Section 4 presents the benchmark methodology, including ensemble attribution and adaptive hard-rationale selection. Section 5 describes the experimental setup and evaluation framework. Section 6 presents and discusses the results, limitations, and directions for future work. Finally, Section 7 concludes the paper.

## 2 Explainable AI

Explainable Artificial Intelligence (XAI) refers to a set of techniques and methods that aim to make AI models more transparent and understandable. These methods can provide insights into the model’s rationale behind its predictions, which makes prediction validation, error analysis and bias detection much easier. This improves trust in AI systems, especially in critical domains and applications.

The trade-off between model accuracy and interpretability remains a persistent challenge, as simpler models offer greater transparency but underperform on complex tasks. In NLP, explaining transformer-based models remains challenging due to their complex architectures. Transformers process words in relation to their surrounding context, making it difficult to determine the individual contribution of inputs to a prediction. Although attention weights are sometimes used as explanations, their faithfulness as reliable justifications for model predictions remains a subject of ongoing debate in the XAI literature [2] [3].

Feature attribution XAI methods aim to interpret predictions of complex machine learning models by identifying how input features contribute to model outputs. They can be categorized according to their scope as **global** or **local**, and model dependence as **model-specific** or **model-agnostic**. Global methods explain general model behavior across the input space, whereas local methods try to explain the reasoning behind an individual prediction. Model-specific methods exploit the internal structure of a particular architecture, while model-agnostic methods treat the model as a black box and rely only on its inputs and outputs, allowing them to be applied to a wider range of models.

In this benchmark, we focus on **local feature attribution** methods, which assign relevance scores to individual input features to quantify their contribution to a specific prediction. This approach is particularly appropriate for transformer-based text classification, as it enables direct analysis of which tokens drive the model’s class prediction.

The benchmark evaluates four widely used feature attribution methods: LIME [4], SHAP [5], Integrated Gradients [6], and DeepLIFT [7]. These methods differ primarily in how they estimate the contribution of individual input tokens to a model prediction. LIME and SHAP are **model-agnostic** approaches: LIME learns a locally interpretable surrogate model around a given instance, whereas SHAP assigns feature contributions based on Shapley values from cooperative game theory. In contrast, IG and DeepLIFT are **model-specific** attribution methods that exploit gradients or internal activations of deep learning models to quantify feature contributions.

## 3 Literature Review

Ensembling attribution XAI methods aim to reduce method-specific variability and improve robustness. Existing approaches typically normalize attribution scores before combining them through operations such as averaging, median aggregation, or weighted combinations. Such strategies have been explored particularly in computer vision [8], including recent work such as NormEnsembleXAI [9], which systematically examines the influence of normalization and aggregation strategies and shows that

ensemble performance depends on both the constituent explainers and the aggregation procedure. These findings demonstrate that aggregation should not be assumed to improve explanation quality, as combining poorly performing or redundant methods may provide limited benefits and increasing ensemble size does not necessarily yield better explanations. However, while these studies provide evidence for ensemble XAI on image data, the behavior of explanation ensembles in NLP remains less extensively studied.

Evaluating explanation quality remains a central challenge in NLP. ERASER [10] introduced human-annotated rationales and standardized metrics for rationale evaluation, while SIDU-TXT [11] incorporates perturbation-based criteria such as Insertion and Deletion AUC. An important distinction is between *faithfulness*, which assesses whether an explanation reflects the behavior of the underlying model, and *plausibility*, which measures its agreement with human expectations or annotated rationales. FERRET [12] provides a unified framework for benchmarking transformer explanations using multiple faithfulness and plausibility metrics. These properties are related but not interchangeable: explanations that faithfully represent model behavior do not necessarily align with human reasoning [13].

XAI research for Arabic NLP remains comparatively limited. Arabic provides a particularly informative setting because of its morphological richness and dialectal diversity. These properties also interact with transformer tokenization, potentially affecting token-level explanations and their evaluation. Earlier Arabic XAI studies have often applied individual explanation techniques to specific applications. Abdelwahab et al. [14], for example, employed LIME to explain an LSTM-based Arabic sentiment classifier. More recent research has moved toward comparative evaluation. Hadj Azzem et al. [15] evaluated multiple gradient and perturbation-based attribution methods across Arabic sentiment analysis and semantic question similarity using both faithfulness and human-rationale plausibility criteria. However, their study focuses only on individual explainers and employs a more limited set of evaluation criteria, leaving the behavior of explanation ensembles and their robustness across complementary faithfulness metrics unexplored. Moreover, it does not systematically investigate strategies for converting continuous attribution scores into hard rationales or the resulting trade-offs among faithfulness, rationale size, and computational efficiency.

Taken together, prior work provides methods for aggregating multiple explainers, general-purpose transformer explanation benchmarks and emerging comparative studies for Arabic NLP. However, the intersection of these directions remains insufficiently investigated. In particular, limited evidence exists on whether selective ensemble attribution provides consistent gains across Arabic classification tasks, whether such gains remain robust across complementary evaluation criteria, and whether improved faithfulness corresponds to stronger human-rationale alignment. The present study addresses these questions through a common multi-model, multi-task, and multi-metric benchmark. Rather than treating attribution aggregation itself as the principal methodological novelty, we focus on empirically characterizing when ensembling is beneficial and additionally examine parameter-free rationale selection in terms of faithfulness, conciseness, and computational efficiency.

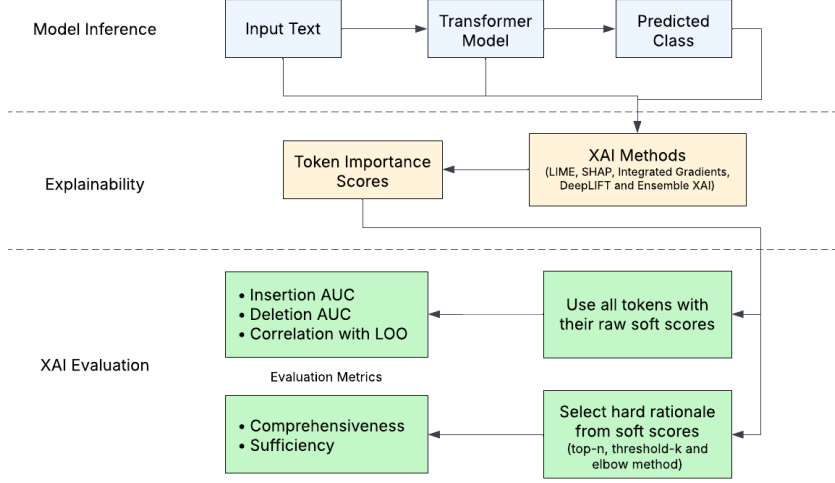


**Fig. 1** Overview of the XAI evaluation workflow.

## 4 Benchmark Methodology

This section describes the methodology used to systematically benchmark individual and ensemble XAI methods for Arabic transformer-based text classification. The framework evaluates multiple attribution methods under common experimental conditions and investigates whether combining attribution-based explainers provides consistent benefits across complementary evaluation criteria. In addition, we examine strategies for converting continuous token-attribution scores into hard rationales, including fixed Top- $n$  tokens, threshold-based selection, parameter-free automated elbow-detection methods, and a newly proposed residual-based approach designed for computationally inexpensive rationale selection.

Figure 1 summarizes the benchmark workflow. Input text is first processed by a transformer classifier to obtain the predicted class. Individual XAI methods (LIME, SHAP, Integrated Gradients, and DeepLIFT) and their ensemble configurations then generate token-level attribution scores for that prediction. Explanation quality is evaluated from complementary perspectives. Continuous attribution scores are assessed using Insertion AUC, Deletion AUC, and correlation with Leave-One-Out (LOO), while discrete hard rationales are evaluated using Comprehensiveness and Sufficiency. Hard rationales are obtained using fixed Top- $n$  and attribution-threshold strategies or automated elbow-detection methods.

### 4.1 Ensemble XAI Approach

To systematically investigate whether aggregating attribution XAI methods improves explanation quality, we evaluate an ensemble framework in which normalized token-level attribution vectors are combined using mean and median aggregation. As shown in Algorithm 1, each method independently produces an importance vector, the

resulting scores are normalized to ensure comparability, and an aggregation function combines them into a single importance score for each token.

---

**Algorithm 1** Ensemble XAI Method for Token Importance Aggregation

---

**Require:**  $x = [t_1, t_2, \dots, t_m]$  {input sequence of  $m$  tokens}  
**Require:**  $f$ : classifier model  
**Require:**  $M = \{M_1, M_2, \dots, M_n\}$  {set of  $n$  selected XAI methods}  
**Require:** `normalize()`: normalization function  
**Require:** `aggregate()`: aggregation function  
**Ensure:**  $v \in \mathbb{R}^m$ : vector of aggregated importance scores for the  $m$  tokens in  $x$

- 1: Initialize score matrix  $S \in \mathbb{R}^{m \times n}$
- 2: Initialize importance vector  $v \in \mathbb{R}^m$
- 3: **for** each  $i \leftarrow 1$  to  $n$  **do**
- 4:    $s_i \leftarrow M_i(x, f)$  {Compute scores with XAI method  $M_i$ }
- 5:    $\hat{s}_i \leftarrow \text{normalize}(s_i)$  {Min-max normalization}
- 6:    $S[:, i] \leftarrow \hat{s}_i$
- 7: **end for**
- 8: **for**  $j \leftarrow 1$  to  $m$  **do**
- 9:    $v[j] \leftarrow \text{aggregate}(S[j, :])$  {Aggregate across methods}
- 10: **end for**
- 11: **return**  $v$

---

Given an input  $x = [t_1, \dots, t_m]$  and a set of XAI methods  $M = M_1, \dots, M_n$ , each method  $M_i$  produces a token importance vector  $s_i$ . To prevent differences in attribution scales from disproportionately influencing the aggregation, each vector is independently min-max normalized to the range  $[-1, 1]$  as follows:

$$\hat{s}_{ij} = 2 \frac{s_{ij} - \min(s_i)}{\max(s_i) - \min(s_i)} - 1, \quad (1)$$

where  $s_{ij}$  denotes the attribution assigned by method  $M_i$  to token  $t_j$ . The normalized score vectors from all the XAI methods are concatenated into one score matrix  $S \in \mathbb{R}^{m \times n}$ , where each column corresponds to one explanation method. The final token importance vector  $v \in \mathbb{R}^m$  is obtained by applying an aggregation function across the normalized scores for each token. We consider two aggregation strategies: **arithmetic mean** and **median**. The mean provides a simple aggregation that can reduce method-specific noise, while the median is more robust to outlying attribution scores.

Rather than assuming that combining more XAI methods necessarily improves explanation quality, we treat ensemble XAI as a design space and empirically evaluate different combinations of constituent methods and aggregation strategies. This enables us to determine whether, and under which configurations, ensembling improves explanation quality. The individual XAI methods are selected to provide diversity in their underlying attribution mechanisms, including perturbation and gradient-based approaches.

## 4.2 Hard Rationale Selection Using the Elbow Method

Several evaluation metrics, including Comprehensiveness and Sufficiency [10], require hard rationales consisting of a discrete subset of the most important tokens. Common selection strategies, such as Top- $k$  and fixed-threshold approaches, require predefined values that may not generalize across inputs, models, or attribution methods. To avoid such manually specified cutoffs, we use elbow-based selection methods to automatically detect important rationales.

Elbow detection is applied in clustering to determine the optimal number of clusters and can identify a point of diminishing returns in a curve. We propose applying elbow detection to dynamically identify the transition between highly important and less important tokens in the ranked importance-score distribution. However, direct elbow identification is typically based on visual inspection and is therefore subjective. Existing automated approaches include Kneedle [16], DFDT/DSDT [17, 18], and the L-Method [19]. L-Method fits two linear models on either side of each candidate point and selects the point minimizing their combined fitting error. While effective, this requires evaluating multiple candidate splits and can become computationally expensive for long sequences.

We propose a simplified residual-based variant of the L-Method that identifies the elbow using a single global linear fit. Given the sorted importance scores  $s = [s_1, \dots, s_n]$ , the proposed method fits a single global linear trend using ordinary least squares (OLS). Since the independent variable consists only of sequential token ranks, the OLS coefficients are computed analytically using the closed-form solution, avoiding the overhead of a general-purpose regression estimator. The residual  $r_i = \hat{s}_i - s_i$  quantifies the downward deviation of each attribution score from the global trend, and the elbow is identified as the non-initial position exhibiting the maximum residual, as formalized in Algorithm 2.

---

### Algorithm 2 Proposed Residual-Based Elbow Method

---

**Require:**  $\mathbf{s} = [s_1, \dots, s_n]$ : token importance scores sorted in descending order

**Ensure:**  $e$ : detected elbow index

```

1: if  $n < 2$  then
2:   return 1
3: end if
4:  $\mathbf{x} \leftarrow [0, 1, \dots, n-1]$ 
5:  $(a, b) \leftarrow \text{OLS}(\mathbf{x}, \mathbf{s})$  {Closed-form linear fit}
6:  $\hat{\mathbf{s}} \leftarrow a\mathbf{x} + b$ 
7:  $\mathbf{r} \leftarrow \hat{\mathbf{s}} - \mathbf{s}$ 
8:  $e \leftarrow \arg \max_{i \in \{2, \dots, n\}} r_i$ 
9: return  $e$ 

```

---

The proposed method assumes that the ranked attribution scores approximately exhibit a two-regime structure, with a steep decline among highly important tokens followed by a flatter tail of less informative tokens. Under this assumption, the transition between the two regimes produces a large deviation from the global linear

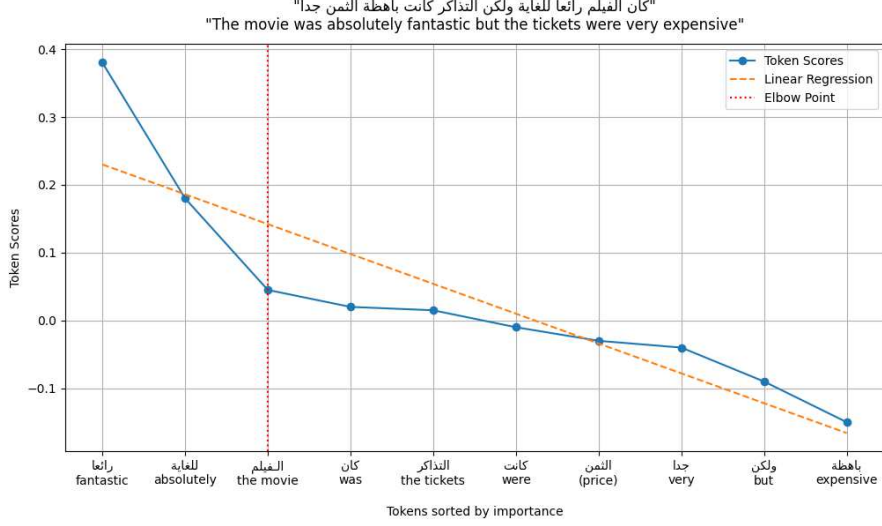


**Fig. 2** Example of the proposed residual-based elbow method applied to token importance scores with a linear fit and residual-based elbow detection. The orange dashed line represents the ordinary least-squares fit across token importance scores, and the elbow point is identified as the token with the largest vertical deviation from the fitted line.

trend, allowing the maximum residual to serve as an estimate of the elbow. All tokens up to the detected elbow are selected as the hard rationale. Unlike Top- $k$  or fixed-threshold selection, the resulting rationale size is determined independently for each input according to its attribution distribution. An example of this approach is shown in Figure 2.

The approach is most suitable when the ranked importance scores exhibit a distinct transition between a high-importance head and a flatter tail. When the scores decrease smoothly without a clear inflection, or when the input contains very few tokens, the maximum residual may not correspond to a meaningful rationale boundary. These limitations are considered when interpreting the resulting explanations. Unless otherwise stated, all experiments use the proposed automated residual-based elbow detection method; the accompanying visual example in Figure 2 is provided only to illustrate its behavior.

## 5 Experimental Setup

This section outlines the experimental settings used to evaluate individual and ensemble XAI methods for Arabic transformer-based text classification. We first introduce the datasets and models used, and human-rationale evaluation settings, followed by implementation details for the attribution methods and the faithfulness and plausibility evaluation protocols. This work constitutes a methodological and technical extension of the XS2A platform [20], with the present study focusing on systematic multi-method, multi-model, and cross-task evaluation.

## 5.1 Datasets & Models

Three Arabic datasets are used to evaluate the proposed framework across complementary experimental settings. Table 1 summarizes their characteristics. For clarity, the datasets are subsequently referred to as D1–D3.

**Table 1** Summary of the datasets used in the experimental evaluation.

| ID | Dataset                           | Task                  | # Instances used       | Class Distribution                            |
|----|-----------------------------------|-----------------------|------------------------|-----------------------------------------------|
| D1 | Tweet Sentiment Multilingual [21] | Sentiment Analysis    | 3,033                  | 1,011 positive, 1,011 negative, 1,011 neutral |
| D2 | HateDay [22]                      | Hate Speech Detection | random subset of 2,000 | 1,000 hate, 1,000 neutral                     |
| D3 | Tripadvisor Reviews [23]          | Sentiment Analysis    | 303                    | 160 positive, 102 negative, 41 neutral        |

D1 serves as the primary benchmark for the comprehensive faithfulness evaluation, while D2 provides cross-task evaluation of the principal findings on hate-speech detection. D3 is used specifically for plausibility evaluation against manually annotated human rationales.

For D3, word-level rationales were collected using our XS2A platform [20]. The first annotator labeled 163 reviews and the second labeled 140 reviews. To assess annotation consistency, an overlap of 55 reviews was annotated by both annotators. Cohen’s  $\kappa$  [24] on these overlapping annotations yielded an average value of 0.57, indicating moderate agreement and reflecting the inherent subjectivity of identifying word-level sentiment rationales.

We also employ three pre-trained and fine-tuned transformer models specifically designed for Arabic NLP from Hugging Face:

- **CAMELBERT-MSA** [25]: Pretrained on Modern Standard Arabic and fine-tuned on sentiment datasets (ASTD [26], ArSAS [27], SemEval [28]).
- **AraBERT-Arabic-SA** [29]: AraBERT fine-tuned on an Arabic tweets dataset and a corpus of Arabic tweets in the Saudi dialect.
- **MarBERT-HateSpeech** [30]: MARBERTv2 fine-tuned for binary text classification (Hate or Neutral) on a sampled version of a custom Egyptian-Arabic hate speech dataset.

All three models are evaluated in this benchmark without additional fine-tuning on the task-specific datasets previously described.

Before model inference, text inputs are preprocessed by removing user mentions (@user), retweet markers (RT), URLs, repeated whitespace, and Arabic diacritics (tashkil). Each classifier is evaluated using its corresponding pretrained tokenizer and explanations are generated at the token level. For each input, XAI methods explain the model’s predicted class rather than the ground-truth class label, since the objective of the benchmark is to evaluate the faithfulness of explanations to the model’s actual prediction.

## 5.2 XAI Implementation Details

All XAI methods are applied under a common token-level evaluation setting. Table 2 summarizes the main implementation parameters used for each explanation method.

**Table 2** Implementation parameters of the evaluated XAI methods.

| XAI Method           | Parameter            | Value                          |
|----------------------|----------------------|--------------------------------|
| LIME                 | Explainer            | <code>LimeTextExplainer</code> |
|                      | Perturbation samples | 100                            |
|                      | Feature selection    | <code>none</code>              |
|                      | Random seed          | 0                              |
| SHAP                 | Explainer            | <code>shap.Explainer</code>    |
|                      | Algorithm            | Partition                      |
|                      | Maximum evaluations  | 100                            |
|                      | Batch size           | 256                            |
|                      | Random seed          | 0                              |
| Integrated Gradients | Attribution level    | Input embeddings               |
|                      | Baseline             | Zero embedding                 |
|                      | Integration steps    | 50                             |
| DeepLIFT             | Attribution level    | Input embeddings               |
|                      | Baseline             | Zero embedding                 |

Parameters not explicitly reported in Table 2 follow the corresponding library defaults. Integrated Gradients and DeepLIFT are deterministic under the fixed model and input configuration used in our experiments.

## 5.3 XAI Evaluation Framework

To evaluate the quality of explanations produced by XAI methods, we use five well-established evaluation metrics: *Comprehensiveness* and *Sufficiency* [10], *Correlation with Leave-One-Out* [12], and *Insertion & Deletion AUC* [11]. Given an input  $x$ , a hard rationale  $R$  containing the selected important tokens, Comprehensiveness measures the model’s confidence drop after removing  $R$  (higher is better), while Sufficiency measures the confidence difference between the full input and  $R$  alone (lower is better). Correlation with LOO measures the agreement between XAI-assigned token importance and the empirical impact of individually removing each token, using Kendall’s rank correlation (higher is better). Insertion and Deletion AUC evaluate the model’s confidence as important tokens are progressively added to or removed from the input, respectively, with higher Insertion AUC and lower Deletion AUC indicating better explanations.

These faithfulness metrics use the predicted-class probability as the model confidence score and perturbations are implemented using the model-specific [MASK] token rather than physically deleting tokens, thereby preserving sequence length

and positional structure. This choice is particularly appropriate for the evaluated CAMeLBER, AraBERT and MarBERT models, as they are BERT-based architectures pretrained using a masked-language-modeling objective and therefore explicitly encounter masked tokens during pretraining.

To provide a unified comparison across the five metrics, we propose the following combined score:

$$\text{Comb.} = \frac{\text{Comp.} + \text{Ins.AUC} + (1 - \text{Suff.}) + (1 - \text{Del.AUC}) + \left(\frac{\text{Cor.LOO}+1}{2}\right)}{5}$$

The score converts all metrics to a common direction, such that higher values consistently indicate better explanations. Comprehensiveness and Insertion AUC are incorporated directly, whereas Sufficiency and Deletion AUC are inverted because lower values are preferable. Correlation with LOO is linearly rescaled from  $[-1, 1]$  to  $[0, 1]$  using  $(\text{Cor.LOO} + 1)/2$ . The resulting score provides a single measure for comparing XAI methods across the evaluation dimensions while preserving the individual metrics for more detailed analysis. Importantly, this combined metric is intended as a pragmatic aggregation tool rather than a theoretically optimal measure of explanation quality.

As a complementary rank-based analysis, we employ the *Borda count* to aggregate the rankings obtained across the five evaluation metrics. For each metric, methods are ranked independently, with rank 1 assigned to the best-performing method. Unlike the combined score, Borda count is insensitive to differences in metric scales and distributions, providing a complementary assessment of the relative ordering of XAI methods. Finally, we compute the Spearman rank correlation between the combined score and Borda rankings to assess the consistency between the two aggregation strategies.

To complement the faithfulness evaluation, we assess explanation plausibility on a subset of 303 sentiment-analysis instances with manually annotated human rationales. The continuous token-attribution scores produced by each XAI method and ensemble are converted into hard rationales using the proposed residual-based elbow method, thereby avoiding a manually specified rationale size or attribution threshold. Since human rationales are annotated at the word level whereas explanations are generated at the model’s native token/subword level, the annotations are aligned with the corresponding model tokens, with all subword tokens belonging to an annotated word assigned to the human rationale. Agreement between the resulting model-generated and human-annotated rationales is evaluated using token-level Precision, Recall, F1-score, and Jaccard similarity, with higher values indicating greater alignment with human rationales.

## 6 Results and Discussion

This section presents and discusses the results obtained from our experimental evaluation and analysis of the different models, XAI methods and hard rationale strategies, including our proposed ensembling framework and residual-based elbow strategy. We first present the performance of the underlying models, then the results of the XAI evaluation metrics for individual and ensemble methods. We then analyze inter-XAI

Method agreement and robustness of the combined metric and evaluate different hard rationale selection methods. Finally, we discuss the implications of our findings and outline potential avenues for future research.

## 6.1 Model evaluation results

Before evaluating the XAI methods, it is crucial to assess the performance of the sentiment analysis and hate speech detection models on each dataset. Table 3 summarizes the average performance of the models, using standard classification metrics. The Camelbert-MSA model achieves the highest overall performance in all evaluation metrics on the sentiment analysis task. For the hate speech detection task, MarBERT demonstrates strong and consistent performance across all metrics (0.841 – 0.852). These results provide a baseline for understanding the reliability of the models whose predictions we aim to explain.

**Table 3** Model Evaluation Results

| Task        | Dataset | Model     | Acc. $\uparrow$ | Prec. $\uparrow$ | Rec. $\uparrow$ | F1 $\uparrow$ |
|-------------|---------|-----------|-----------------|------------------|-----------------|---------------|
| Sentiment   | D1      | CamelBERT | 0.778           | 0.785            | 0.778           | 0.779         |
| Sentiment   | D1      | AraBERT   | 0.601           | 0.650            | 0.601           | 0.605         |
| Hate Speech | D2      | MarBERT   | 0.841           | 0.852            | 0.841           | 0.840         |

## 6.2 XAI evaluation results

### 6.2.1 Faithfulness evaluation for sentiment analysis task

We evaluated several individual XAI methods (LIME, SHAP, Integrated Gradients, and DeepLIFT) and multiple ensembling combinations on the sentiment analysis task. This analysis also serves as an ablation study over ensemble composition and aggregation strategy, allowing us to isolate the contribution of individual explanation methods and aggregation choices.

To convert the soft token importance scores into binary rationales, we use our proposed residual-based elbow method. This binarization strategy is specifically required for evaluation metrics that rely on hard rationales, namely Comprehensiveness and Sufficiency.

Table 4 reports mean performance and standard deviation across CAMELBERT and AraBERT models. Among ensemble configurations, combinations involving LIME and SHAP consistently achieve the highest average Combined scores and lowest Borda ranks, indicating more stable overall rankings across complementary evaluation criteria. Notably, the observed improvements in the combined score are modest in magnitude compared to the best standalone XAI method (LIME) but consistent across all ensembling combinations and aggregation strategies, supporting the interpretation that ensemble explanations provide incremental gains in robustness rather than large performance shifts.

**Table 4** XAI Evaluation Results (average across the two sentiment models). L, S, IG, and DL denote LIME, SHAP, Integrated Gradients, and DeepLIFT, respectively.

| XAI Method         | Comp. $\uparrow$ | Suff. $\downarrow$ | Corr. LOO $\uparrow$ | Ins. AUC $\uparrow$ | Del. AUC $\downarrow$ | Comb. $\uparrow$ | Borda $\downarrow$ |
|--------------------|------------------|--------------------|----------------------|---------------------|-----------------------|------------------|--------------------|
| LIME               | $0.446 \pm 0.30$ | $0.122 \pm 0.26$   | $0.324 \pm 0.22$     | $0.714 \pm 0.15$    | $0.390 \pm 0.21$      | $0.662 \pm 0.10$ | 25                 |
| SHAP               | $0.423 \pm 0.29$ | $0.116 \pm 0.25$   | $0.273 \pm 0.23$     | $0.718 \pm 0.15$    | $0.398 \pm 0.20$      | $0.653 \pm 0.10$ | 29                 |
| IG                 | $0.364 \pm 0.31$ | $0.119 \pm 0.25$   | $0.133 \pm 0.21$     | $0.643 \pm 0.18$    | $0.501 \pm 0.21$      | $0.591 \pm 0.11$ | 39                 |
| DeepLIFT           | $0.329 \pm 0.32$ | $0.119 \pm 0.25$   | $0.012 \pm 0.21$     | $0.583 \pm 0.20$    | $0.573 \pm 0.20$      | $0.545 \pm 0.11$ | 44                 |
| L+S+IG+DL (Mean)   | $0.448 \pm 0.30$ | $0.068 \pm 0.23$   | $0.289 \pm 0.20$     | $0.711 \pm 0.15$    | $0.402 \pm 0.20$      | $0.667 \pm 0.09$ | 22                 |
| L+S+IG+DL (Median) | $0.444 \pm 0.30$ | $0.076 \pm 0.24$   | $0.289 \pm 0.21$     | $0.710 \pm 0.15$    | $0.403 \pm 0.20$      | $0.664 \pm 0.10$ | 27                 |
| L+S+IG (Mean)      | $0.451 \pm 0.29$ | $0.101 \pm 0.25$   | $0.323 \pm 0.21$     | $0.721 \pm 0.15$    | $0.383 \pm 0.20$      | $0.670 \pm 0.09$ | 12                 |
| L+S+IG (Median)    | $0.446 \pm 0.30$ | $0.109 \pm 0.26$   | $0.322 \pm 0.22$     | $0.720 \pm 0.15$    | $0.387 \pm 0.20$      | $0.666 \pm 0.10$ | 19                 |
| L+S (Mean/Median)  | $0.455 \pm 0.29$ | $0.109 \pm 0.25$   | $0.339 \pm 0.22$     | $0.727 \pm 0.14$    | $0.373 \pm 0.20$      | $0.674 \pm 0.09$ | 8                  |

The best-performing ensemble configuration (LIME+SHAP) achieves a statistically significant improvement over LIME in the Combined metric, as confirmed by a paired Wilcoxon signed-rank test ( $p < 10^{-16}$ ). Bootstrap analysis further yields a 95% confidence interval of  $[0.0098, 0.0141]$  for the mean paired performance difference, with an average improvement of  $\Delta \approx 0.012$ . Although the magnitude of the improvement remains modest, the confidence interval lies entirely above zero, indicating that the ensemble provides a statistically reliable positive average gain over LIME rather than a substantial increase in explanation quality.

Despite these improvements, Correlation with Leave-One-Out scores remain relatively low across all evaluated methods. We therefore interpret Cor.LOO as a complementary diagnostic rather than a definitive measure of explanation quality, and include it as part of a multi-metric evaluation framework instead of relying on it in isolation.

### 6.2.2 Cross-Task Generalization

To assess whether the main findings of the benchmark generalize beyond sentiment analysis, we additionally evaluate the individual and ensemble XAI methods on the Arabic hate-speech detection task. The same explanation-generation and faithfulness-evaluation framework is applied, allowing the relative behavior of the XAI configurations to be examined under a different classification task.

Table 5 shows that the main ensemble behavior observed in sentiment analysis largely persists under cross-task evaluation. LIME remains the strongest individual method according to the Combined score (0.676), while the LIME+SHAP ensemble achieves the highest overall Combined score (0.685), corresponding to an absolute improvement of approximately 0.009. The LIME+SHAP+IG mean ensemble achieves a comparable Combined score (0.683) and obtains the strongest Borda ranking (7), indicating that the exact optimal configuration depends on the aggregation criterion.

Importantly, increasing ensemble size does not consistently improve performance. Adding DeepLIFT to the LIME+SHAP+IG combination decreases the Combined

**Table 5** Cross-task faithfulness evaluation on Arabic hate-speech detection. L, S, IG, and DL denote LIME, SHAP, Integrated Gradients, and DeepLIFT, respectively.

| XAI Method         | Comp. $\uparrow$ | Suff. $\downarrow$ | Corr. LOO $\uparrow$ | Ins. AUC $\uparrow$ | Del. AUC $\downarrow$ | Comb. $\uparrow$ | Borda $\downarrow$ |
|--------------------|------------------|--------------------|----------------------|---------------------|-----------------------|------------------|--------------------|
| LIME               | 0.572 $\pm$ 0.44 | 0.069 $\pm$ 0.24   | 0.240 $\pm$ 0.34     | 0.837 $\pm$ 0.12    | 0.576 $\pm$ 0.24      | 0.676 $\pm$ 0.14 | 29                 |
| SHAP               | 0.408 $\pm$ 0.44 | 0.068 $\pm$ 0.24   | 0.259 $\pm$ 0.32     | 0.858 $\pm$ 0.11    | 0.538 $\pm$ 0.23      | 0.657 $\pm$ 0.15 | 26                 |
| IG                 | 0.309 $\pm$ 0.43 | 0.106 $\pm$ 0.28   | 0.156 $\pm$ 0.34     | 0.829 $\pm$ 0.15    | 0.599 $\pm$ 0.23      | 0.600 $\pm$ 0.15 | 41                 |
| DeepLIFT           | 0.432 $\pm$ 0.46 | 0.152 $\pm$ 0.32   | -0.024 $\pm$ 0.34    | 0.720 $\pm$ 0.19    | 0.753 $\pm$ 0.18      | 0.543 $\pm$ 0.16 | 43                 |
| L+S+IG+DL (Mean)   | 0.455 $\pm$ 0.45 | 0.057 $\pm$ 0.22   | 0.257 $\pm$ 0.30     | 0.855 $\pm$ 0.12    | 0.549 $\pm$ 0.24      | 0.668 $\pm$ 0.14 | 24                 |
| L+S+IG+DL (Median) | 0.473 $\pm$ 0.45 | 0.060 $\pm$ 0.22   | 0.258 $\pm$ 0.31     | 0.854 $\pm$ 0.12    | 0.549 $\pm$ 0.24      | 0.671 $\pm$ 0.14 | 25                 |
| L+S+IG (Mean)      | 0.483 $\pm$ 0.45 | 0.047 $\pm$ 0.21   | 0.289 $\pm$ 0.32     | 0.863 $\pm$ 0.11    | 0.522 $\pm$ 0.24      | 0.683 $\pm$ 0.14 | 7                  |
| L+S+IG (Median)    | 0.438 $\pm$ 0.45 | 0.062 $\pm$ 0.23   | 0.288 $\pm$ 0.32     | 0.860 $\pm$ 0.11    | 0.527 $\pm$ 0.24      | 0.669 $\pm$ 0.14 | 17                 |
| L+S (Mean/Median)  | 0.508 $\pm$ 0.45 | 0.056 $\pm$ 0.23   | 0.285 $\pm$ 0.32     | 0.858 $\pm$ 0.11    | 0.531 $\pm$ 0.24      | 0.685 $\pm$ 0.14 | 13                 |

score from 0.683 to 0.668 under mean aggregation, while DeepLIFT is also the weakest individual method according to the Combined metric. These results reinforce the finding that ensemble effectiveness depends on the quality and complementarity of the constituent explainers rather than the number of methods being combined. Overall, the cross-task experiment provides evidence that the modest faithfulness advantage of selective XAI ensembling is not restricted to sentiment analysis, while also demonstrating that the relative ranking of ensemble configurations can vary across tasks.

### 6.2.3 Human-Rationale Plausibility Evaluation

While the previous experiments evaluate whether explanations faithfully reflect model behavior, faithfulness does not necessarily imply agreement with human reasoning. We therefore complement the model-centered evaluation with a plausibility analysis using 303 sentiment instances accompanied by manually annotated human rationales. The generated soft attribution scores are converted into hard rationales using the proposed residual-based elbow method and compared with the corresponding human annotations using token-level Precision, Recall, F1-score, and Jaccard similarity.

Table 6 reveals a different pattern from the faithfulness evaluation. SHAP achieves the highest token-level Precision (0.299), whereas DeepLIFT achieves the highest Recall (0.903). The latter is accompanied by comparatively low Precision (0.258), indicating substantial differences in the precision-recall behavior of the generated rationales. The highest F1-score is obtained by the LIME+SHAP mean ensemble (0.4124); however, its advantage over SHAP, which obtained an F1-score of 0.4117 is negligible. The same trend can be observed with Jaccard similarity (0.260 for LIME+SHAP and 0.259 for SHAP).

These findings indicate that the improvements obtained through ensembling in the faithfulness experiments did not translate into improvements in human-rationale alignment. LIME+SHAP, which achieves the strongest Combined faithfulness score in both the primary sentiment benchmark and the cross-task hate speech experiment, obtains

**Table 6** Plausibility evaluation against manually annotated human rationales (average across the two sentiment models).

| XAI Method               | Prec. $\uparrow$ | Rec. $\uparrow$ | F1 $\uparrow$ | Jaccard $\uparrow$ |
|--------------------------|------------------|-----------------|---------------|--------------------|
| LIME                     | 0.263            | 0.739           | 0.388         | 0.240              |
| SHAP                     | 0.299            | 0.659           | 0.4117        | 0.259              |
| IG                       | 0.261            | 0.854           | 0.399         | 0.249              |
| DeepLIFT                 | 0.258            | 0.903           | 0.401         | 0.251              |
| LIME+SHAP+IG+DL (Mean)   | 0.274            | 0.820           | 0.411         | 0.258              |
| LIME+SHAP+IG+DL (Median) | 0.271            | 0.821           | 0.407         | 0.256              |
| LIME+SHAP+IG (Mean)      | 0.282            | 0.743           | 0.408         | 0.257              |
| LIME+SHAP+IG (Median)    | 0.271            | 0.775           | 0.401         | 0.251              |
| LIME+SHAP (Mean/Median)  | 0.287            | 0.734           | 0.4124        | 0.260              |

plausibility F1 and Jaccard scores that are marginal compared with the strongest individual method, SHAP. Consequently, faithfulness and plausibility should be evaluated as complementary rather than interchangeable dimensions of explanation quality.

### 6.3 Inter-XAI Method agreement

To further investigate the behavior of the individual explainers and provide insight into the ensemble results, we analyze the agreement between the token importance scores produced by the different XAI methods on the primary sentiment dataset (D1). For each pair of XAI methods, agreement is measured using the Pearson correlation coefficient between their token-level attribution scores. Inter-method agreement indicates the extent to which different explanation techniques assign similar relative importance to the same input tokens.

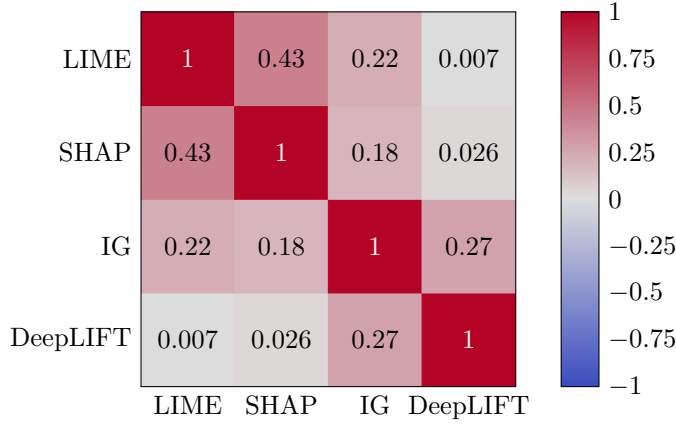


**Fig. 3** Pairwise Pearson correlation between token-level attribution scores produced by the XAI methods on the D1 sentiment benchmark.

Figure 3 presents the pairwise correlations between the token-level attribution scores generated by the four individual explainers. The results reveal substantial differences in agreement across XAI methods. The strongest inter-method correlation is observed between LIME and SHAP (0.43), indicating moderate agreement in the token importance patterns identified by these two model-agnostic perturbation-based explainers. Interestingly, LIME and SHAP also constitute the strongest-performing ensemble in all previous evaluation experiments, suggesting that their combination may benefit from a balance between agreement and complementary attribution information. LIME exhibits lower agreement with IG (0.22) and almost no correlation with DeepLIFT. Similarly, SHAP shows relatively weak agreement with IG and DeepLIFT.

Among the gradient-based methods, IG and DeepLIFT exhibit the second-highest pairwise correlation (0.27). This correlation score remains considerably lower than that observed between LIME and SHAP. Overall, the generally low-to-moderate correlations indicate that the evaluated XAI methods produce substantially different token-level attribution patterns despite explaining predictions from the same models. These findings reinforce the motivation for investigating ensemble explanations, however, the results also show that diversity alone is insufficient to guarantee improved faithfulness. In particular, the superior performance of the LIME-SHAP ensemble compared with larger combinations suggests that effective ensembling depends on the complementarity and individual quality of the explainers rather than simply maximizing disagreement or increasing the number of methods.

#### 6.4 Robustness of the Combined Evaluation Metric

Because the proposed Combined metric assigns equal importance to five complementary faithfulness metrics, we further examine whether the resulting ranking of XAI methods is robust to the metric aggregation procedure and to the contribution of individual evaluation metrics. We conduct two complementary robustness analyses: comparison with the rank-based Borda count and a leave-one-metric-out analysis.

First, we compare the rankings produced by the Combined metric and Borda count using Spearman rank correlation. In the sentiment benchmark (two-model average), the two rankings exhibit a strong negative correlation of -0.97 ( $p < 0.01$ ). Similarly, a negative correlation of -0.77 is observed in the hate-speech benchmark ( $p = 0.016$ ). The negative direction is expected because higher Combined scores indicate better performance, whereas lower Borda scores correspond to better rankings. This substantial agreement indicates that the relative performance of the evaluated XAI methods is largely preserved when replacing the score-based aggregation with a scale-independent rank-based aggregation.

Second, we perform a leave-one-metric-out analysis to determine whether the observed ranking in the sentiment benchmark is disproportionately driven by any single faithfulness metric. The Combined score and Borda count are recomputed five times, each time excluding one of Comprehensiveness, Sufficiency, Correlation with Leave-One-Out, Insertion AUC, or Deletion AUC. The LIME+SHAP ensemble remains the highest-ranked method under both the Combined score and Borda count in every leave-one-metric-out configuration. Detailed results for all five leave-one-metric-out configurations are provided in Appendix A.

Together, these analyses indicate that the superior overall ranking of the LIME+SHAP ensemble is not an artifact of the proposed Combined-score formulation or the influence of a single evaluation metric. Rather, the main performance trend remains stable across score-based and rank-based aggregation and under systematic removal of individual evaluation criteria.

## 6.5 Hard-Rationale Selection Analysis

Comprehensiveness and Sufficiency depend not only on the underlying attribution scores but also on the strategy used to convert these continuous scores into a discrete set of rationale tokens. We therefore compare our proposed residual-based elbow method, used throughout the preceding experiments, with three categories of rationale-selection strategies: fixed Top- $n$  selection, attribution-threshold selection, and established automated elbow-detection methods. The latter include Kneedle, DFDT, and the classical L-Method, evaluated using their implementations provided by the `kneeliverse` [31] Python library. All strategies are evaluated using the LIME+SHAP ensemble configuration. In addition to Comprehensiveness and Sufficiency, we report the average number and proportion of selected tokens to account for differences in rationale size. This is important because selecting larger rationales can systematically improve perturbation-based faithfulness scores, while producing less concise explanations. We additionally report the average selection runtime to compare the computational efficiency of the different hard-rationale strategies. Runtime measurements exclude attribution generation and model inference and measure only the hard-rationale selection procedure. Each strategy is evaluated repeatedly 30 times on the same attribution inputs, and the execution order is randomized across repetitions to reduce systematic timing-order effects.

**Table 7** Evaluation results for different token selection strategies, including average selection runtime across both sentiment transformer models.

| Selection Method      | Comp. $\uparrow$                  | Suff. $\downarrow$                | Comb. $\uparrow$                  | Borda $\downarrow$ | # Tokens $\downarrow$               | Tokens (%) $\downarrow$               | Time ( $\mu$ s) $\downarrow$      |
|-----------------------|-----------------------------------|-----------------------------------|-----------------------------------|--------------------|-------------------------------------|---------------------------------------|-----------------------------------|
| Top- $n$ ( $n = 3$ )  | $0.30 \pm 0.27$                   | $0.21 \pm 0.26$                   | $0.55 \pm 0.18$                   | 20                 | $2.99 \pm 0.05$                     | $18.0\% \pm 12.6\%$                   | $9.7 \pm 4.2$                     |
| Top- $n$ ( $n = 5$ )  | $0.39 \pm 0.28$                   | $0.16 \pm 0.25$                   | $0.61 \pm 0.18$                   | 14                 | $4.98 \pm 0.19$                     | $29.5\% \pm 18.3\%$                   | $10.3 \pm 3.9$                    |
| Top- $n$ ( $n = 10$ ) | $0.46 \pm 0.26$                   | $0.08 \pm 0.23$                   | $0.69 \pm 0.15$                   | 6                  | $9.69 \pm 1.10$                     | $54.1\% \pm 24.1\%$                   | $10 \pm 4.3$                      |
| Top- $n$ ( $n = 20$ ) | $0.52 \pm 0.27$                   | $0.01 \pm 0.15$                   | $0.76 \pm 0.14$                   | 2                  | $16.90 \pm 4.46$                    | $84.4\% \pm 17.6\%$                   | $9.7 \pm 4.4$                     |
| Thresh. ( $k = 0.1$ ) | $0.42 \pm 0.26$                   | $0.16 \pm 0.26$                   | $0.63 \pm 0.16$                   | 12                 | $7.23 \pm 5.68$                     | $33.6\% \pm 20.1\%$                   | $10.4 \pm 4.5$                    |
| Thresh. ( $k = 0.3$ ) | $0.34 \pm 0.26$                   | $0.20 \pm 0.26$                   | $0.57 \pm 0.18$                   | 18                 | $4.65 \pm 3.87$                     | $22.5\% \pm 15.4\%$                   | $10.8 \pm 4.2$                    |
| Thresh. ( $k = 0.5$ ) | $0.26 \pm 0.25$                   | $0.25 \pm 0.28$                   | $0.50 \pm 0.19$                   | 22                 | $2.69 \pm 2.26$                     | $13.8\% \pm 10.6\%$                   | $10.9 \pm 4.2$                    |
| Thresh. ( $k = 0.7$ ) | $0.18 \pm 0.22$                   | $0.31 \pm 0.31$                   | $0.44 \pm 0.20$                   | 24                 | $1.39 \pm 1.08$                     | $7.9\% \pm 7.0\%$                     | $10.4 \pm 4.2$                    |
| Elbow (Kneedle)       | $0.45 \pm 0.28$                   | $0.12 \pm 0.26$                   | $0.66 \pm 0.18$                   | 10                 | $9.25 \pm 7.49$                     | $42.5\% \pm 25.0\%$                   | $194.5 \pm 93.8$                  |
| Elbow (DFDT)          | $0.48 \pm 0.29$                   | $0.05 \pm 0.24$                   | $0.71 \pm 0.18$                   | 4                  | $12.85 \pm 9.59$                    | $58.1\% \pm 29.6\%$                   | $423 \pm 177.1$                   |
| Elbow (L-Method)      | $0.35 \pm 0.28$                   | $0.19 \pm 0.26$                   | $0.57 \pm 0.18$                   | 16                 | $3.92 \pm 1.99$                     | $21.7\% \pm 13.8\%$                   | $1268.6 \pm 569.4$                |
| <b>Elbow (Ours)</b>   | <b><math>0.45 \pm 0.29</math></b> | <b><math>0.11 \pm 0.25</math></b> | <b><math>0.67 \pm 0.18</math></b> | <b>8</b>           | <b><math>11.90 \pm 10.50</math></b> | <b><math>51.9\% \pm 34.5\%</math></b> | <b><math>42.4 \pm 19.9</math></b> |

Table 7 presents the results of this comparison. Among the fixed-size strategies, increasing  $n$  consistently improves both Comprehensiveness and Sufficiency. Top-20

achieves the highest Comprehensiveness (0.520), lowest Sufficiency (0.002), highest Combined score (0.759), and lowest Borda count. However, it selects on average 16.91 tokens, corresponding to approximately 84.4% of the input. Its strong faithfulness scores should therefore be interpreted in conjunction with its substantially larger rationales. In contrast, Top-3 selects only approximately 18.0% of the input but exhibits considerably lower Comprehensiveness (0.303) and higher Sufficiency (0.208). These results illustrate the inherent trade-off between rationale faithfulness and conciseness and demonstrate why comparisons of rationale-selection strategies should account for the number of selected tokens.

A similar pattern is observed for fixed-threshold selection. Increasing the attribution threshold progressively reduces rationale size, from approximately 33.6% of the input at  $k = 0.1$  to only 7.9% at  $k = 0.7$ , but is accompanied by decreasing Comprehensiveness and increasing Sufficiency. For example, the  $k = 0.7$  configuration produces the most concise rationales in the benchmark, selecting only 1.40 tokens on average, but obtains the lowest Combined score (0.436). Consequently, neither fixed Top- $n$  nor fixed-threshold results identify a universally preferable cutoff; their behavior is strongly dependent on the manually specified selection parameter.

The automated elbow-based approaches avoid this predefined rationale-size parameter by adapting the number of selected tokens to the attribution distribution of each input. The proposed residual-based method selects an average of 11.87 tokens (51.9% of the input) and achieves a Combined score of 0.673. This substantially outperforms the classical L-Method, which selects considerably shorter rationales (21.7%) but obtains a Combined score of only 0.575, and slightly outperforms Kneedle (0.665). Among the adaptive methods, DFDT obtains the strongest faithfulness results, with a Combined score of 0.713 and Borda score of 4, while selecting approximately 58.1% of the input.

In terms of computational efficiency, the proposed method exhibits the lowest runtime among the automated adaptive strategies, requiring an average of only 42.41  $\mu$ s per input. In comparison, Kneedle, DFDT, and the classical L-Method require 194.50  $\mu$ s, 423.03  $\mu$ s, and 1268.64  $\mu$ s, respectively. The proposed approach is therefore approximately 4.6 times faster than Kneedle, 10 times faster than DFDT, and 29.9 times faster than the classical L-Method. This computational advantage results from the simplicity of the proposed procedure, which requires only a single closed-form OLS fit followed by residual computation and identification of the maximum deviation.

Overall, the results demonstrate that rationale-selection performance cannot be assessed independently of rationale size and computational cost. Strategies selecting a large fraction of the input, such as Top-20, naturally preserve more information and consequently obtain favorable Comprehensiveness and Sufficiency scores, but at the expense of rationale conciseness. The proposed residual-based approach does not dominate all competing strategies in terms of faithfulness; rather, it provides a parameter-free alternative that achieves competitive faithfulness while exhibiting substantially lower computational cost than the other automated elbow-detection approaches. In particular, DFDT provides the strongest faithfulness among the adaptive strategies, whereas the proposed method offers a more favorable computational-efficiency trade-off. These findings indicate that no single rationale-selection strategy universally dominates across faithfulness, conciseness, and computational efficiency.

## 6.6 Discussion, Limitations, and Future Work

The experimental results provide a systematic comparison of XAI methods for Arabic sentiment analysis and hate speech detection and lead to four main observations. First, selective ensembling provides small but reproducible faithfulness gains. LIME+SHAP achieves the strongest Combined score in the primary sentiment benchmark. This advantage also persists under alternative metric aggregation and cross-task evaluation, suggesting that selective ensembling provides incremental robustness rather than substantially stronger explanations.

Second, adding more explainers does not necessarily improve performance. The LIME+SHAP ensemble performs comparably to or better than larger configurations, and adding DeepLIFT to LIME+SHAP+IG decreases faithfulness in both sentiment analysis and hate speech tasks, indicating that ensemble effectiveness depends more on the quality and complementarity of its constituents than on ensemble size.

Third, faithfulness and plausibility exhibit different patterns: although LIME+SHAP achieves the strongest Combined faithfulness score across both tasks and also obtains the highest human-rationale F1-score and Jaccard similarity, its plausibility advantage over the strongest individual method SHAP is negligible. Thus, the more consistent faithfulness advantage observed for selective ensembling does not translate into a comparably meaningful improvement in human-rationale alignment, reinforcing the need to evaluate faithfulness and plausibility as complementary rather than interchangeable objectives.

Finally, hard-rationale evaluation reveals an important rationale-size trade-off. Comprehensiveness and Sufficiency can favor larger, less concise rationales. These metrics should therefore be interpreted alongside rationale size rather than in isolation. Within this trade-off, the proposed residual-based elbow method provides competitive faithfulness while offering parameter-free rationale selection at substantially lower computational cost than the established automated elbow-detection methods evaluated in this study.

Several limitations should nevertheless be considered when interpreting these results. First, the perturbation-based evaluation metrics replace tokens with the model-specific [MASK] token. Although all three evaluated models encounter masked tokens during masked-language-model pretraining and this strategy preserves sequence length and positional structure, heavily masked inputs may still differ from the natural-text distribution encountered during downstream fine-tuning. The resulting distribution shift may therefore influence metric results. Second, Integrated Gradients and DeepLIFT use a zero-embedding reference. While this provides a consistent reference representing the absence of input-embedding signal, it does not correspond to a linguistically meaningful neutral input, and the resulting attributions may be sensitive to the choice of baseline. Third, computational constraints require finite approximation budgets for perturbation-based explanation methods. In particular, LIME is evaluated using 100 perturbation samples, which may potentially affect both its individual performance and ensembles containing LIME.

Future work will address these limitations in several directions. The sensitivity of LIME and other approximation-based methods to their computational budgets should be investigated through systematic convergence analyses. Alternative perturbation

mechanisms and reference baselines for Integrated Gradients and DeepLIFT could be evaluated to quantify the influence of these methodological choices on the resulting rankings. Regarding ensemble construction, adaptive or instance-dependent aggregation could be investigated to account for variations in the reliability of individual XAI methods across models and inputs without relying on evaluation information from the test set. Finally, the proposed hard-rationale selection approach could be further refined for attribution distributions that do not exhibit a clearly defined elbow.

## 7 Conclusion

This study establishes a systematic empirical benchmark of attribution-based XAI methods for Arabic transformer-based text classification, comparing LIME, SHAP, Integrated Gradients, DeepLIFT, and ensemble configurations across sentiment analysis and hate-speech detection. The results show that selective ensembling, particularly LIME+SHAP, provides small but statistically reliable faithfulness improvements over individual methods, with the main benefit lying in the consistency and robustness of these gains rather than their absolute magnitude. However, increasing ensemble size does not guarantee better explanations, indicating that ensemble composition is more important than simply combining additional explainers. The human-rationale evaluation further shows that improvements in model-centered faithfulness do not necessarily translate into stronger plausibility, emphasizing the need to assess these dimensions separately. In addition, the hard-rationale experiments demonstrate that Comprehensiveness and Sufficiency are sensitive to rationale size, highlighting the importance of reporting explanation conciseness alongside faithfulness. Within this trade-off, the proposed residual-based selector provides a parameter-free and computationally inexpensive alternative to established automated elbow-detection methods, while achieving competitive but not dominant faithfulness. Overall, these findings provide evidence-based guidance for evaluating and selecting XAI methods in Arabic NLP and highlight the importance of multi-metric, cross-task, and human-centered evaluation when assessing explanation quality.

## Declarations

### 7.1 Funding

No funding was received for conducting this study.

### 7.2 Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

### 7.3 Data Availability Statement

The data and source code supporting the findings of this study are publicly available at: [https://github.com/ychafiqui/arabic\\_xai\\_benchmark](https://github.com/ychafiqui/arabic_xai_benchmark)

## 7.4 Ethical Approval

This study involved manual annotation of a publicly available dataset by two adult researchers. The study did not involve patients, vulnerable populations, or the collection of personal or sensitive data. No formal ethics committee approval was sought for this research.

## 7.5 Clinical Trial Number

Not applicable.

## 7.6 Consent to Participate

All annotators participated voluntarily in the annotation process.

## 7.7 Consent to Publish

Not applicable.

# Appendix A Leave-One-Metric-Out Analysis

The following tables report the leave-one-metric-out robustness analysis for the sentiment-analysis benchmark, averaged across CAMELBERT-MSA and AraBERT-Arabic-SA. In each analysis, one faithfulness metric is excluded and the Combined score and Borda count are recomputed using the remaining four metrics.

**Table A1** Leave-one-metric-out analysis excluding Comprehensiveness.

| XAI Method         | Suff. ↓ | Corr. LOO ↑ | Ins. AUC ↑ | Del. AUC ↓ | Comb. ↑      | Borda ↓  |
|--------------------|---------|-------------|------------|------------|--------------|----------|
| LIME               | 0.122   | 0.324       | 0.714      | 0.390      | 0.716        | 20       |
| SHAP               | 0.116   | 0.273       | 0.718      | 0.398      | 0.710        | 22       |
| IG                 | 0.119   | 0.133       | 0.643      | 0.501      | 0.647        | 31       |
| DeepLIFT           | 0.119   | 0.012       | 0.583      | 0.573      | 0.599        | 35       |
| L+S+IG+DL (Mean)   | 0.068   | 0.289       | 0.711      | 0.402      | 0.721        | 19       |
| L+S+IG+DL (Median) | 0.076   | 0.289       | 0.710      | 0.403      | 0.719        | 21       |
| L+S+IG (Mean)      | 0.101   | 0.323       | 0.721      | 0.383      | 0.725        | 10       |
| L+S+IG (Median)    | 0.109   | 0.322       | 0.720      | 0.387      | 0.721        | 15       |
| L+S (Mean/Median)  | 0.109   | 0.339       | 0.727      | 0.373      | <b>0.728</b> | <b>7</b> |

## References

- [1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

**Table A2** Leave-one-metric-out analysis excluding Sufficiency.

| XAI Method         | Comp. $\uparrow$ | Corr. LOO $\uparrow$ | Ins. AUC $\uparrow$ | Del. AUC $\downarrow$ | Comb. $\uparrow$ | Borda $\downarrow$ |
|--------------------|------------------|----------------------|---------------------|-----------------------|------------------|--------------------|
| LIME               | 0.446            | 0.324                | 0.714               | 0.390                 | 0.608            | 16                 |
| SHAP               | 0.423            | 0.273                | 0.718               | 0.398                 | 0.595            | 23                 |
| IG                 | 0.364            | 0.133                | 0.643               | 0.501                 | 0.518            | 32                 |
| DeepLIFT           | 0.329            | 0.012                | 0.583               | 0.573                 | 0.461            | 36                 |
| L+S+IG+DL (Mean)   | 0.448            | 0.289                | 0.711               | 0.402                 | 0.600            | 21                 |
| L+S+IG+DL (Median) | 0.444            | 0.289                | 0.710               | 0.403                 | 0.599            | 25                 |
| L+S+IG (Mean)      | 0.451            | 0.323                | 0.721               | 0.383                 | 0.613            | 9                  |
| L+S+IG (Median)    | 0.446            | 0.322                | 0.720               | 0.387                 | 0.610            | 14                 |
| L+S (Mean/Median)  | 0.455            | 0.339                | 0.727               | 0.373                 | <b>0.619</b>     | <b>4</b>           |

**Table A3** Leave-one-metric-out analysis excluding Correlation with Leave-One-Out.

| XAI Method         | Comp. $\uparrow$ | Suff. $\downarrow$ | Ins. AUC $\uparrow$ | Del. AUC $\downarrow$ | Comb. $\uparrow$ | Borda $\downarrow$ |
|--------------------|------------------|--------------------|---------------------|-----------------------|------------------|--------------------|
| LIME               | 0.446            | 0.122              | 0.714               | 0.390                 | 0.662            | 23                 |
| SHAP               | 0.423            | 0.116              | 0.718               | 0.398                 | 0.657            | 22                 |
| IG                 | 0.364            | 0.119              | 0.643               | 0.501                 | 0.597            | 31                 |
| DeepLIFT           | 0.329            | 0.119              | 0.583               | 0.573                 | 0.555            | 35                 |
| L+S+IG+DL (Mean)   | 0.448            | 0.068              | 0.711               | 0.402                 | 0.672            | 16                 |
| L+S+IG+DL (Median) | 0.444            | 0.076              | 0.710               | 0.403                 | 0.669            | 22                 |
| L+S+IG (Mean)      | 0.451            | 0.101              | 0.721               | 0.383                 | 0.672            | 9                  |
| L+S+IG (Median)    | 0.446            | 0.109              | 0.720               | 0.387                 | 0.667            | 15                 |
| L+S (Mean/Median)  | 0.455            | 0.109              | 0.727               | 0.373                 | <b>0.675</b>     | <b>7</b>           |

- [2] Jain, S., Wallace, B.C.: Attention is not explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, pp. 3543–3556 (2019)
- [3] Wiegrefe, S., Pinter, Y.: Attention is not not explanation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, pp. 11–20 (2019)
- [4] Ribeiro, M.T., Singh, S., Guestrin, C.: 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, San Francisco, CA, pp. 1135–1144 (2016)
- [5] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, pp. 4768–4777 (2017)

**Table A4** Leave-one-metric-out analysis excluding Insertion AUC.

| XAI Method         | Comp. $\uparrow$ | Suff. $\downarrow$ | Corr. LOO $\uparrow$ | Del. AUC $\downarrow$ | Comb. $\uparrow$ | Borda $\downarrow$ |
|--------------------|------------------|--------------------|----------------------|-----------------------|------------------|--------------------|
| LIME               | 0.446            | 0.122              | 0.324                | 0.390                 | 0.649            | 20                 |
| SHAP               | 0.423            | 0.116              | 0.273                | 0.398                 | 0.636            | 25                 |
| IG                 | 0.364            | 0.119              | 0.133                | 0.501                 | 0.578            | 31                 |
| DeepLIFT           | 0.329            | 0.119              | 0.012                | 0.573                 | 0.536            | 35                 |
| L+S+IG+DL (Mean)   | 0.448            | 0.068              | 0.289                | 0.402                 | 0.656            | 16                 |
| L+S+IG+DL (Median) | 0.444            | 0.076              | 0.289                | 0.403                 | 0.652            | 20                 |
| L+S+IG (Mean)      | 0.451            | 0.101              | 0.323                | 0.383                 | 0.657            | 10                 |
| L+S+IG (Median)    | 0.446            | 0.109              | 0.322                | 0.387                 | 0.653            | 16                 |
| L+S (Mean/Median)  | 0.455            | 0.109              | 0.339                | 0.373                 | <b>0.661</b>     | <b>7</b>           |

**Table A5** Leave-one-metric-out analysis excluding Deletion AUC.

| XAI Method         | Comp. $\uparrow$ | Suff. $\downarrow$ | Corr. LOO $\uparrow$ | Ins. AUC $\uparrow$ | Comb. $\uparrow$ | Borda $\downarrow$ |
|--------------------|------------------|--------------------|----------------------|---------------------|------------------|--------------------|
| LIME               | 0.446            | 0.122              | 0.324                | 0.714               | 0.675            | 21                 |
| SHAP               | 0.423            | 0.116              | 0.273                | 0.718               | 0.665            | 24                 |
| IG                 | 0.364            | 0.119              | 0.133                | 0.643               | 0.613            | 31                 |
| DeepLIFT           | 0.329            | 0.119              | 0.012                | 0.583               | 0.575            | 35                 |
| L+S+IG+DL (Mean)   | 0.448            | 0.068              | 0.289                | 0.711               | 0.684            | 16                 |
| L+S+IG+DL (Median) | 0.444            | 0.076              | 0.289                | 0.710               | 0.681            | 20                 |
| L+S+IG (Mean)      | 0.451            | 0.101              | 0.323                | 0.721               | 0.683            | 10                 |
| L+S+IG (Median)    | 0.446            | 0.109              | 0.322                | 0.720               | 0.679            | 16                 |
| L+S (Mean/Median)  | 0.455            | 0.109              | 0.339                | 0.727               | <b>0.686</b>     | <b>7</b>           |

- [6] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70. Sydney, Australia, pp. 3319–3328 (2017)
- [7] Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70. Sydney, Australia, pp. 3145–3153 (2017)
- [8] Zou, L., Goh, H., Liew, C., Quah, J., Gu, G., Chew, J., Kumar, M., Ang, C., Ta, A.: Ensemble image explainable AI (XAI) algorithm for severe community-acquired pneumonia and COVID-19 respiratory infections. IEEE Transactions on Artificial Intelligence (2022)
- [9] Hryniewska-Guzik, W., Sawicki, B., Biecek, P.: NormEnsembleXAI: Unveiling the strengths and weaknesses of XAI ensemble techniques. CoRR **abs/2401.17200** (2024)

- [10] DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C.: ERASER: A benchmark to evaluate rationalized NLP models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online (2020)
- [11] Jahromi, M.N.S., Muddamsetty, S.M., Jarlner, A.S.S., Høgenhaug, A.M., Gammeltoft-Hansen, T., Moeslund, T.B.: SIDU-TXT: An XAI algorithm for NLP with a holistic assessment approach. *Natural Language Processing Journal* **7**, 100078 (2024)
- [12] Attanasio, G., Pastor, E., Di Bonaventura, C., Nozza, D.: ferret: A framework for benchmarking explainers on transformers. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (2023)
- [13] Lyu, Q., Apidianaki, M., Callison-Burch, C.: Towards faithful model explanation in NLP: A survey. *Computational Linguistics* **50**(2), 657–723 (2024) [https://doi.org/10.1162/coli\\_a.00511](https://doi.org/10.1162/coli_a.00511)
- [14] Abdelwahab, Y., Kholief, M., Sedky, A.A.H.: Justifying arabic text sentiment analysis using explainable AI (XAI): LASIK surgeries case study. *Information* **13**(11), 536 (2022)
- [15] AZZEM, Y.C.H., Harrag, F., Bellatreche, L.: Exploring explainability in arabic language models: An empirical analysis of techniques. *Procedia Computer Science* **244**, 212–219 (2024) <https://doi.org/10.1016/j.procs.2024.10.194> . 6th International Conference on AI in Computational Linguistics
- [16] Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B.: Finding a ‘kneedle’ in a haystack: Detecting knee points in system behavior. In: 2011 31st International Conference on Distributed Computing Systems Workshops, pp. 166–171 (2011)
- [17] Antunes, M., Gomes, D., Aguiar, R.L.: Knee/elbow estimation based on first derivative threshold. In: 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService), pp. 237–240 (2018)
- [18] Antunes, M., Ribeiro, J., Gomes, D., Aguiar, R.L.: Knee/elbow point estimation through thresholding. In: 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud), pp. 413–419 (2018)
- [19] Salvador, S., Chan, P.: Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In: 16th IEEE International Conference on Tools with Artificial Intelligence, pp. 576–584 (2004)
- [20] Chafiqui, Y., Anoun, H., Aloiradi, L., Nazih, F.: XS2A: A platform for explainable sentiment analysis in arabic comments. (2024)

- [21] Barbieri, F., Espinosa Anke, L., Camacho-Collados, J.: XLM-T: Multilingual language models in twitter for sentiment analysis and beyond. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, pp. 258–266 (2022)
- [22] Tonneau, M., Liu, D., Malhotra, N., Hale, S.A., Fraiberger, S., Orozco-Olvera, V., Röttger, P.: HateDay: Insights from a global hate speech dataset representative of a day on twitter. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Vienna, Austria, pp. 2297–2321 (2025)
- [23] ElSahar, H., El-Beltagy, S.R.: Building large arabic multi-domain resources for sentiment analysis. In: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing, pp. 23–34. Springer, Cham (2015)
- [24] Cohen, J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**(1), 37–46 (1960) <https://doi.org/10.1177/001316446002000104>
- [25] Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., Habash, N.: The interplay of variant, size, and task type in arabic pre-trained language models. In: Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kyiv, Ukraine (Online) (2021)
- [26] Nabil, M., Aly, M., Atiya, A.: ASTD: Arabic sentiment tweets dataset. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp. 2515–2519 (2015)
- [27] Elmadany, A., Mubarak, H., Magdy, W.: ArSAS: An arabic speech-act and sentiment corpus of tweets (2018)
- [28] Rosenthal, S., Farra, N., Nakov, P.: SemEval-2017 Task 4: Sentiment analysis in twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, Canada, pp. 502–518 (2017)
- [29] Ragab, A.: PRAlI22/AraBert-Arabic-Sentiment-Analysis. Hugging Face. [Online]. Available: <https://huggingface.co/PRAlI22/AraBert-Arabic-Sentiment-Analysis> (2024). <https://huggingface.co/PRAlI22/AraBert-Arabic-Sentiment-Analysis>
- [30] Ahmed, I., Abbas, M., Hatem, R., Ihab, A., Fahkr, M.W.: Fine-tuning arabic pre-trained transformer models for egyptian-arabic dialect offensive language and hate speech detection and classification. In: 2022 20th International Conference on Language Engineering (ESOLEC), vol. 20, pp. 170–174 (2022). <https://doi.org/10.1109/ESOLEC54569.2022.10009167>
- [31] Antunes, M., Estro, T., Bhandari, P., Gandhi, A., Kuenning, G., Liu, Y., Waldspurger, C., Wildani, A., Zadok, E.: Kneeliverse: A universal knee-detection

library for performance curves. SoftwareX **30**, 102161 (2025)