scientific reports

Check for updates

OPEN

Benchmarking explainable offensive language detection in Somali with human-annotated rationales

Abdisalam Mahamed Badel[1]✉, Ting Zhong[1], Xovee Xu[1], Wenxin Tai[1], Fan Zhou[1] & Abdulahi Abdirashid Dahir[2]

Offensive language presents significant challenges on the internet and requires robust moderation. However, the efficacy of such moderation often depends on providing clear and interpretable justifications for each classification. Unfortunately, many existing datasets lack annotated rationales, and most detection models offer limited interpretability and transparency. These limitations hinder the development of trustworthy systems and the implementation of effective content moderation strategies. In this paper, we introduce SomOffXplain, an interpretable framework for detecting offensive language in Somali, which generates human-understandable explanations for its predictions. SomOffXplain performs span-level rationale extraction at both the word and phrase levels, enabling it to highlight text segments that support its predictions. Given that Somali is a low-resource language, we first construct a new benchmark dataset of 10,175 samples, each annotated with human-provided rationales. We evaluate our method against five fine-tuned pre-trained models using Local Interpretable Model-Agnostic Explanations (LIME). Additionally, we adapt four large language models (LLMs) through few-shot and zero-shot prompting to assess their ability to understand and produce rationales in Somali. Our proposed model demonstrates superiority in terms of explainability and predictive accuracy, exhibiting higher plausibility and faithfulness compared to the baselines. Furthermore, our results reveal that half of the state-of-the-art LLMs evaluated fail to generate high-quality rationales that align with human-annotated ground truth rationales, whereas LIME-based methods also prove to be weak explainers for Somali text. We believe our contributions support online safety, help prevent harassment in under-resourced language communities, enhance the trustworthiness of language models, and promote transparency in artificial intelligence systems.

Nowadays, internet users can anonymously post content, including targeted offensive remarks, with their identities often unknown to victims or readers. Large language models (LLMs) can also generate offensive content[1]. A recent incident involving the Grok AI chatbot exemplifies this issue[2]. Such content may negatively affect those who encounter it[3], especially targeted users. Offensive language, including hate speech, is widely recognized as a significant source of harm[4]. It can result in human rights violations, personal harm, and, in some cases, criminal behavior[5]. It may also pose serious risks, such as suicide attempts[6]. According to Ref[7], offensive language generally refers to any expression that is hurtful, insulting, morally repugnant, or degrading. Hate speech, which is a subset of offensive language, targets people based on characteristics such as sexual orientation, gender, or race[8]. In this study, following[9], we define offensive language as any comment containing a threat, insult, harmful or discriminatory action, swear words, profane language, or improper speech that violates accepted standards of normal discourse, whether targeted or not. Numerous studies, including[10], confirm that hate speech significantly contributes to online abuse and harassment. In contrast[11], notes that some countries, including Canada and France, have implemented laws prohibiting hate speech. A range of resources and datasets has been developed to support research in this field[12–14], along with proposed models[15] to address this issue. While some approaches rely on Transformer-based architectures[3,16], others investigate LLMs' ability for hate speech detection[17,18]. However, many existing datasets do not provide rationales, and most models lack interpretability. Moreover, most existing models fail to offer transparent and trustworthy justifications for their classifications[19,20].

[1]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China. [2]Department of Statistics, University of Kabridahar, 250 Kabridahar, Somali Region, Ethiopia. ✉email: 202214090105@std.uestc.edu.cn

This limitation is significant because platform moderators require explainable tools to understand why specific content violates policy guidelines[21]. A new study on an understudied language[22]argues that a large disparity exists between low-resource and high-resource languages from a performance perspective. On the other hand[23], states that a significant research gap exists in detecting hate speech in low-resource languages, mainly due to linguistic diversity and the limited availability of annotated datasets. Somali, a low-resourced language, lacks many natural language processing (NLP) resources.

Consequently, a language rich in literature and poetry[24,25] remains underrepresented in modern computational linguistics research. To address this gap and contribute to the creation of offensive-free social media and LLMs, we introduce XOLDS, a novel dataset with human-annotated rationales for offensive language detection in Somali. This dataset is not only the largest to date but also the first in the Somali language to include human-annotated rationales. Our dataset comprises 10,175 samples manually collected from TikTok and YouTube, with additional samples gathered from X. Three annotators labeled each sample as offensive or not offensive. Additionally, two of the three annotators selected relevant text segments and provided rationales for their annotations. Using these annotations, we then developed SomOffXplain (Benchmarking Explainable Offensive Language Detection in Somali with Human-Annotated Rationales), an interpretable framework for offensive language detection in Somali. To improve interpretability of SomOffXplain, we adopted class-conditioned contrastive learning and benchmarked the framework's performance against five fine-tuned models using Local Interpretable Model-Agnostic Explanations (LIME)[26], evaluating both explainability and classification accuracy. SomOffXplain implements explainability via extractive span-level rationales, where the model identifies the word or phrase in an input text that is most predictive of the offensive language class. Specifically, it performs span-level rationale extraction and highlights the exact text segments that support its predictions.

We further studied the rationale generation capability of four LLMs using zero-shot and few-shot (2-shot) settings. We evaluated the effectiveness of LLMs in explainability at both the word and phrase levels for rationale generation. Finally, this work proposes the first rationale-based benchmark dataset for the Somali language. We develop a computationally less expensive explainable framework for Somali offensive language detection. By comprehensively evaluating various methods, we create a new benchmark for explainable offensive language detection in Somali that can also be extended to other low-resource languages. Thus, we believe our work makes a significant contribution to the NLP community, particularly by addressing barriers faced by low-resource languages. The main contributions of this paper are as follows:

1. We introduce the first rationale-based and largest dataset for offensive language detection in Somali, consisting of 10,175 samples with human-annotated spans.
2. We propose a simple yet effective interpretable framework, SomOffXplain, for detecting offensive language in Somali, which can also be adapted for other low-resource languages.
3. We investigate the capabilities of four LLMs in rationale generation and offensive language detection for Somali text. We also conduct comprehensive experiments to evaluate and benchmark the performance of SomOffXplain in terms of both explainability and classification accuracy, comparing it against fine-tuned models that use LIME-based explanations and state-of-the-art LLMs.

Related work

In this section, we discuss the Somali language, prior work related to offensive language, and explainability in offensive language detection. Specifically, we first briefly introduce the Somali language, then summarize studies on offensive language, and finally review existing rationale-based work on offensive language detection.

Somali

Somali uses the Latin script and adopts rules from the Roman orthography[25]. It is a member of the Cushitic family and is mostly spoken in the Horn of Africa. The Somali language is the official language of Somalia[25] and one of the five languages designated as future official languages in Ethiopia[27]. It is also widely spoken in some parts of Kenya and is one of the predominant languages in Djibouti[25]. Somali has a system of twenty-two consonants and five vowels[25], and it is morphologically classified as an agglutinative language[24]. In summary, based on our investigation, Somali is a very low-resource language in computational linguistics. The language has limited electronic resources available in this field, which is a strong justification for conducting this project.

Offensive language

The study of offensive language is an ongoing and prominent research field, with numerous studies conducted. We discuss some of these studies in the literature, such as[28], who studied offensive language in Greek and Arabic, while[29]categorized offensive language into distinct subcategories, such as offensive identification, offensive targets, and offensive types. Nevertheless[11], emphasizes the persistent challenge of effectively categorizing offensive instances. Moreover[30], investigated the impact of annotators on hate speech classification, emphasizing the necessity for clear definitions of hate speech. The fact that offensive language detection is an ongoing challenge leads to numerous studies employing diverse methodologies across multiple languages. For example[31], proposed a multi-modal framework for hate speech, while[32] used dual contrastive loss to address semantic information and distributional imbalance in hate speech classification. In addition, the authors in[33]explored implicit hate speech detection using contrastive learning[34], focused on cross-lingual challenges. Furthermore, offensive language detection has been examined from multiple perspectives, including general offensive language[6], sexism[35], hate speech[36,37], abusive language[38], and cyberbullying[39]. Researchers argue that effective moderation requires clearly defined reasoning behind classification decisions[20,21]. This highlights the importance of explainability, which our work seeks to address.

Existing rationale-based works on offensive language

Most prior work in explainable text classification has relied on human-annotated rationales, first introduced by[40], or on automatically generated rationales, later explored by[41]. In recent years, interpretability in offensive language detection has attracted considerable attention and has been studied across multiple languages. For example[42], employed rationales for English hate speech classification, while[43] introduced datasets for code-switched contexts in English-Tamil and English-Kannada.

Similarly[44], examined toxic language detection using span annotations for English. Research on offensive language detection with human-annotated rationales has expanded to other languages, including Vietnamese[45], Persian[46], and Portuguese[20]. Nevertheless, offensive language detection in low-resource languages remains a significant challenge[23]. Moreover, model interpretability often depends on the availability of high-quality, human-annotated rationales[20]. To address this gap in the literature, we introduce a novel rationale-based dataset featuring human-annotated text spans in Somali. Furthermore, we present SomOffXplain, a simple, interpretable model designed for offensive language detection in Somali.

Dataset

In the dataset construction section, we provide a brief overview of the dataset collection process, label annotation methodology, target group identification, rationale generation, and the establishment of ground truth rationales.

Dataset collection

We collected this dataset primarily between April 2022 and August 2023 from the comments sections of YouTube and X/Twitter, with additional data collected from TikTok in 2024. To ensure the integrity of the dataset, we collected the data manually.

The dataset exclusively comprises text data, and we removed all personally identifiable information. We omitted names, links, and other identifying details, or anonymized them using the @user token, following prior work by Ref[42]. We took these measures to protect privacy, prevent potential rights violations, and comply with platform guidelines on ethics.

Dataset annotation

We recruited five data annotators, all native speakers of Somali. All annotators received extensive training and well-defined annotation guidelines. After training, following[42] each annotator completed a pilot annotation on a sample of 100 comments. Annotation was conducted in two stages: *Stage 1:* Three annotators labeled 8,117 samples with labels and targets. *Stage 2:* Three annotators worked together, one from the previous team (who annotated 2,058 samples with labels and targets such as *group*, *women*, *person*, or *other*) and two new annotators (who annotated the full dataset with labels, targets, and rationales/spans explaining why a sample text was offensive or not). In this stage, labels, targets, and rationales were annotated, but rationales were identified only by the two new annotators. In the first phase, our annotation platforms were KoboCollect[47] and Google spreadsheets[48], and in the second phase we used a custom website developed for this task (see Fig. 1), which makes rationale selection easy. On this website, the rationale is selected with a mouse click. Table 1 presents the statistics of our dataset. Please note that we have tried our best to remove non-Somali texts from the dataset. For quality control, we used three annotators per item and applied majority voting. However, to assess voting reliability, we calculated Fleiss kappa (k)[49] for binary label identification among the three annotators, which was 76%. Regarding annotator roles, we assigned one annotator who was more experienced than the others to serve as the disagreement resolver each time. The annotation guidelines were prepared by the first author.

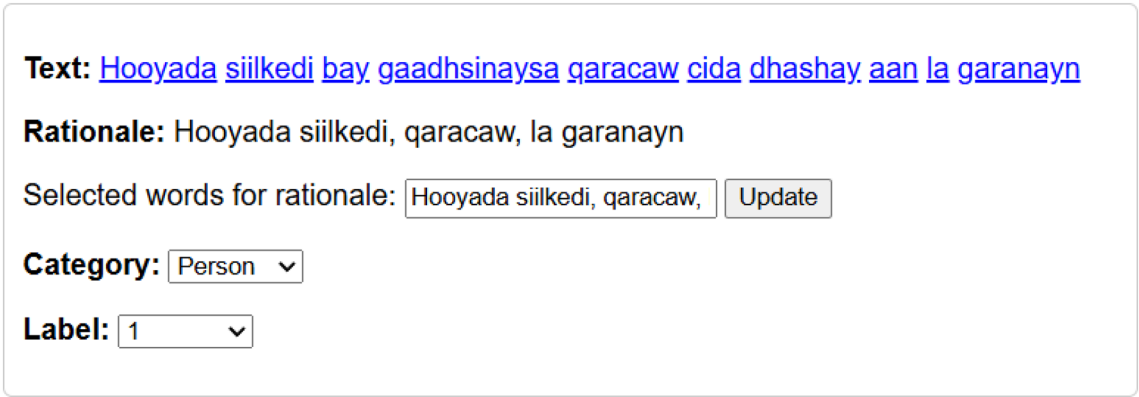


Fig. 1. Screenshot of the custom website designed for annotating the dataset. Rationales are selected by clicking on the words; the words are then automatically inserted into the text field. After the update button is clicked, the category, label, and rationale are stored in our backend, and the rationale is also displayed in front of the rationale label.

Category	Label		Total
	Off	Not-off	Sum
Total samples per label	6123	4052	10,175
Total words per label	67100	33930	101,030
Percentage per label	60.17	39.82	100%
Avg words per sample	10.95	8.37	19.32
Avg word length per sample	5.54	5.57	11.11

Table 1. Statistics of our dataset for offensive (off) vs. not-offensive (not-off) labels.

Target group
In addition to labeling each comment as offensive or non-offensive and providing rationales, we instructed annotators to identify the target of each offensive comment, such as a *group*, *woman*, *person*, or *other*. The *other* category was introduced for instances in which the offensive comment had no explicit target, the target was ambiguous, or the target was difficult to determine. Furthermore, we identified 248 instances in which the annotators failed to reach a consensus. Following[42], we excluded these disputed samples from the dataset. The distribution of targets within the offensive class was 4402 persons, 1140 groups, 124 women, and 457 others. As in prior work by[50], we used majority voting and selected labels based on the majority decision of the three annotators. In cases of disagreement, the third annotator, who had the most experience, served as the adjudicator throughout the annotation process. We resolved all disagreements to ensure that all annotators agreed on a single label. Nevertheless, identifying the target category was challenging. For example, distinguishing between a person and a group, or between a woman and a person, often led to disagreement and required several rounds of resolution.

Ground truth rationale
To create the ground truth rationale, we used the overlapping words (i.e. $rationale = r_1 \cap r_2$) from the two annotators or the combined text spans (i.e. $rationale = r_1 \cup r_2$) selected by them when there was no intersection between the rationales from the two annotators.

Annotator demographics
We worked with two teams, as the dataset was annotated in two rounds. In the first round, the annotation team consisted of three people: one woman and two men. In the second round, the team also consisted of three annotators: one woman and two men. One annotator from the first round also participated in the second round, so we had five annotators in total. Their ages ranged from 24 to 31, and their educational backgrounds included two with master's degrees, two with bachelor's degrees, and one with a high school diploma. All annotators were native Somali speakers.

Proposed framework
In the proposed framework section, we first provide an overview of the entire framework. We then describe the main components of the framework in detail. Specifically, we discuss (1) the backbone, (2) multi-detector learning, which consists of two subcomponents: (a) the offensive classifier and (b) the span extractor, and (**3**) class-conditioned contrastive learning.

Overview
We utilize a BERT encoder[51] as the backbone for processing Somali text. It extracts features and generates embeddings, which are then passed into a multi-detector module that jointly performs (a) offensive language classification and (b) rationale span extraction. The offensive language classifier determines whether the text is offensive or non-offensive, while the span extractor identifies and extracts the relevant spans. The extracted spans and predicted labels are subsequently fed into a class-conditioned contrastive learning module, which uses averaged span embeddings as a semantic explanation and labels as additional information. Finally, at inference time an explainer generates explanations that include the predicted label, the rationale, and a justification for why the text belongs to that label. We apply simple preprocessing steps to the rationale and text, which normalizes them by converting the text to lowercase, removing whitespace, and handling NaN values. Figure 2 provides an overview of the proposed SomOffXplain framework. We now discuss the main components of the proposed framework as follows:

Backbone
We utilize BERT to extract features by tokenizing text $\{w_1,..., w_n\}$ and generating contextualized embeddings $\{e_1,..., e_n\}$ from input sentences. Compared to other embedding generation methods we tried and tested, BERT has shown good performance in our work. For this reason, we chose it.

Multi-detector
This module is a multi-task learning module that jointly performs offensive classification and spans extraction. It comprises two main sub-components: (a) an offensive classifier and (b) a span extractor.

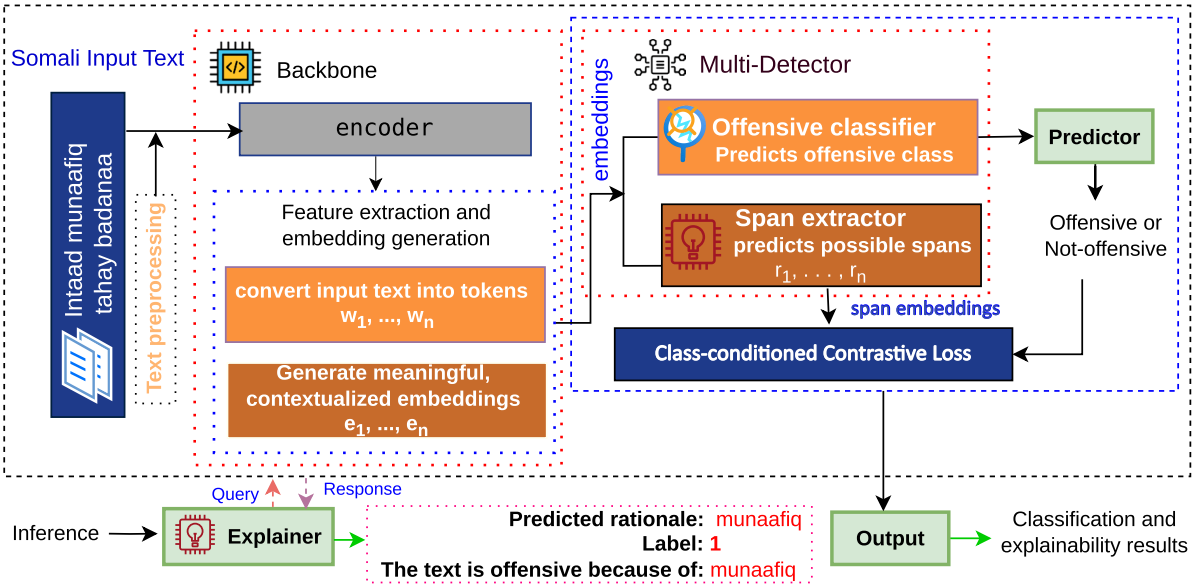


Fig. 2. Overall architecture of the proposed SomOffXplain.

(a) *Offensive classifier*. We apply a linear layer on top of the BERT-generated embeddings to perform binary classification, predicting whether the input is offensive or non-offensive. To define this, let $C = \{w_1, w_2, w_3, \ldots, w_n\}$ denote the input comment, where w_i represents the i-th word in the sentence. The offensive language detection sub-task aims to predict $\hat{y}$, where $y \in \{0, 1\}$, with $y = 1$ indicating offensive content and $y = 0$ indicating non-offensive content. The offensive classifier module is the main component responsible for offensive language detection. The offensive detection component of the proposed framework detects offensiveness using sentence-level prediction. This component uses the hidden representation of the special token `[CLS]`, which is denoted as h_{CLS}, as the aggregated representation of the entire input text. This vector is then passed through a linear classification layer, as defined in Eq. (1):

$$\hat{y} = W_c h_{\text{CLS}} + b_c, \tag{1}$$

where b_c and W_c are trainable parameters.

(b) *Span extractor*. We designed a rationale extractor with two heads: a rationale start head and a rationale end head, which predict the beginning and end of the explanatory span. Let $R = \{r_1, r_2, \ldots, r_n\}$ denote the rationale mask, where $r_i = 1$ if token i belongs to the rationale span and $r_i = 0$ otherwise. We use two linear layers to predict the start and end positions of the explanatory span. Specifically, the start logit for token i is computed as $s_i = W_s h_i + b_s$, and the end logit for token i is computed as $e_i = W_e h_i + b_e$, where h_i is the hidden representation of token i. These predictions are compared with the ground truth start and end positions of the rationale span. In this framework, rationale extraction is formulated as a span prediction problem rather than a token-level classification problem. The rationale extraction component identifies the text span that explains the model's decision, where s_i and e_i denote the start and end logits for token i, respectively. The predicted rationale span is determined from the start and end logits, and the rationale extraction loss is computed from the gold span indices, as defined in Eq. (2):

$$\mathscr{L}_{\text{span}} = \frac{1}{2}\left(\text{CE}(s_i, \hat{s}) + \text{CE}(e_i, \hat{e})\right), \tag{2}$$

where s_i and e_i are the predicted token indices for start and end, and $\hat{s}, \hat{e}$ are the ground truth start and end positions of the rationale, CE is the cross-entropy.

Class-conditioned contrastive learning

Empirically, we observed that most fine-tuned models achieve higher classification performance but fail to deliver good explainability. Based on this observation, we came up with the idea that span embeddings should be separated based on their class, with spans of the same class closer together and spans of dissimilar classes farther apart. This approach allows the contrastive loss to become more discriminative. Thus, we propose a contrastive rationale loss applied to span representations.

Unlike most existing rationale-based approaches like[20,46], our method introduces a class-conditioned contrastive learning loss that uses not only generated span embeddings but also predicted label information. This architecture enables the model to cluster spans more effectively, and enhances explainability. Our class-conditioned contrastive learning can be defined as given span embeddings $v_i \in \mathbb{R}^d$ and their corresponding class labels y_i, we formulate the contrastive loss to encourage spans within the same class to have similar embeddings,

while ensuring that spans from different classes are well-separated, and discriminated we define this in Eq. (3). By adopting label information, the contrastive loss not only improves the interpretability of the model but also contributes to the overall performance of the offensive classification. Following the approach of Ref[42], we set the temperature variable (τ) to 0.1 to encourage the model to focus on the rationale.

$$\mathscr{L}_{\text{contrastive}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|P(i)|}\sum_{j\in P(i)}\log\frac{\exp\left(\frac{z_i^\top z_j}{\tau}\right)}{\sum_{k\neq i}\exp\left(\frac{z_i^\top z_k}{\tau}\right)} \tag{3}$$

where N denotes the total number of samples in each batch. $z_i^\top z_j$ denotes the cosine similarity between normalized embeddings z_i and z_j. $P(i)$ is the set of positive samples that share the same label as anchor i, excluding the anchor itself, and $|P(i)|$ denotes the number of positive samples for anchor i. τ is the temperature variable that scales the similarity scores, which we set to $\tau = 0.1$.

Finally, the overall training of the proposed framework combines three objectives: classification loss using cross-entropy, span loss using cross-entropy, and contrastive loss using a supervised contrastive loss function. These components detect offensive language, predict the start and end token positions, and form a rationale span embedding, respectively. Therefore, the final loss of SomOffXplain is computed as in Eq. (4):

$$\mathscr{L} = \alpha\mathscr{L}_{\text{contrastive}} + \mathscr{L}_{\text{span}} + \mathscr{L}_{\text{cls}}, \tag{4}$$

where $\mathscr{L}_{\text{cls}}$ is the classification loss, in this experiment, we used the standard cross-entropy classification loss, and α controls the contribution of the contrastive term.

Data ethics

To annotate this dataset, we manually collected publicly available textual data. We explained the purpose of the dataset to our annotators, assured them that their privacy would be protected, and obtained their full consent to participate, and we also paid them properly. In addition, we followed responsible NLP guidelines to preserve data privacy and strictly applied data anonymization. Our aim is to make the Somali language more resourced and to assist future researchers in creating offensive-free social media. We do not intend to harm anyone; therefore, we have properly anonymized the dataset and used only comments written in text, ensuring that identifying information has been removed or replaced with the @user token to the best of our ability. Overall, our methods were performed in accordance with relevant regulations and guidelines. However, we acknowledge the possibility of incorrect annotations resulting from human error. in addition, it should be noted that the text may include comments that some readers may find disturbing and that do not align with our own views. It is important to note that the creation of this dataset was exclusively for academic research purposes, and its utilization for any other purpose necessitates additional scrutiny. Finally, the work in this paper did not involve experiments on animals, humans, or human tissue. However, the data annotation protocol received ethical approval from the Research Ethics Committee of the University of Kabridahar with reference number UoK/RPED//051/17. We therefore confirm that the experiments in this work were conducted in accordance with relevant guidelines and international and national regulations.

Experiment

In this section, we provide a comprehensive overview of the dataset used, training details, baselines, and evaluation metrics.

Dataset

We used a dataset containing 10,175 samples and adopted a splitting strategy by allocating 70% for training, 15% for validation, and 15% for testing. We selected a dataset that fit the addressed problem well, followed a standard random splitting approach, and ensured splits were representative of the whole dataset. Inspired by many offensive language detection studies, including[42], we report only the results obtained from the test dataset.

Training details

We trained SomOffXplain using the BERT-base-uncased encoder with a hidden size of 768. We used the AdamW optimizer with a learning rate of 3e-5, a batch size of 8, and a maximum sequence length of 128 for 3 epochs. We set random seeds to 42 for reproducibility. All experiments were conducted on a Windows 11 machine configured with an NVIDIA GeForce RTX 3080 Ti Laptop GPU.

Baselines

To evaluate the effectiveness of SomOffXplain, we use two types of baselines: (1) fine-tuned models with LIME as the explainability baseline, and (2) four LLMs and the fine-tuned models as classification baselines. In addition to these two baselines, we further compare the rationale generation capabilities of the four LLMs with human-generated ground-truth rationales. Finally, the baselines we used are:

1. *LIME-based baselines.* Explainability Baseline Using Fine-Tuned Models: We apply LIME, which is proposed by Ref[26], to five fine-tuned pretrained models: RoBERTa[52], mBERT[51], DistilBERT[53], XLNet[54], and ALBERT[55]. We also use these fine-tuned models themselves as classification baselines. During fine-tuning, we tested two well-known explainability frameworks, LIME and SHAP, and selected LIME over SHAP for memory efficiency. For example, when experimenting with SHAP, we faced challenges during model inference and training

with medium batch sizes greater than 4. Both frameworks have their advantages and disadvantages, but in this specific experiment, our main justification for selecting LIME was memory efficiency, along with lower complexity and reduced latency.

2. *LLM baselines.* To establish a strong future baseline, we further examine the performance of LLMs on Somali offensive text explainability and classification tasks. Specifically, we evaluate four state-of-the-art LLMs: DeepSeek-V3[56], GPT-3.5, GPT-4o, and GPT-4o-mini[57], under few-shot and zero-shot prompting settings to compare their rationale-generation capabilities with human-generated ground-truth rationales. We employ two evaluation strategies. First, we evaluate the models' rationale-generation performance on the full test set (15% of the total data, 1,527 samples) by comparing model-generated rationales with human annotations, on both phrase-level and word-level evaluations. Second, we evaluate a smaller subset (10% of the test set) to evaluate the quality of LLM-generated rationales against inter-annotator agreement, based on word-level rationale alignment between two human annotators. Finally, we compare the binary classification performance of the LLMs with that of our proposed model.

Evaluation metrics

We briefly discuss two types of evaluation metrics in this section: explainability metrics and classification metrics. For explainability, we employed IOU F1, Token F1, AUPRC, comprehensiveness, and sufficiency to evaluate our model and the LIME-based models. For the LLMs, we use F1 score, precision, recall, and exact match. For classification, we used metrics such as macro F1, macro precision, macro recall, F1 score, precision, accuracy, AUROC, and recall. We discuss the explainability metrics and the classification metrics as follows:

*Explainability metrics.*To measure explainability, we follow the ERASER framework of Ref[58] and prior works such as[20,42], which provide definitions and methods for evaluating model explainability. We evaluate explainability using two key measures: plausibility and faithfulness. Faithfulness assesses the correctness of the model's reasoning[59], while plausibility evaluates how persuasive the model's interpretations are to humans. To measure plausibility, we apply both soft selection and discrete metrics. For discrete metrics, we use Intersection-over-Union (IOU) F1 score, Area Under the Precision-Recall Curve (AUPRC), and token F1-score. IOU F1 score measures the overlap between two rationale spans divided by their combined token length, while token F1 is calculated based on token-level precision and recall[58]. We assess faithfulness using sufficiency and comprehensiveness[58]. Sufficiency evaluates whether the model can make accurate predictions using only the extracted rationales.

We calculate sufficiency as Sufficiency = $P(y_l \mid x) - P(y_l \mid r)$, where x denotes the original input text, r is the generated rationale, y_l is the class which is predicted, $P(y_l \mid x)$ is the model's probability using the full text, and $P(y_l \mid r)$ is the probability using only the rationale. Comprehensiveness examines whether the extracted tokens include all important information for accurate predictions. We calculate it as: Comprehensiveness = $P(y_l \mid x) - P(y_l \mid x \setminus r)$, where $x \setminus r$ is the text with the rationale removed, measuring how important the rationale is to the prediction. For LLM interpretability evaluation, we use both phrase-level and word-level evaluation. Phrase-level evaluation captures the overlap of full rationale phrases or semantic units, while word-level evaluation measures the overlap between individual words in the human-annotated ground truth rationales and those generated by the LLMs. For this, we use standard metrics such as F1 score, precision, and recall. Additionally, we evaluate LLMs using exact match rate, which focuses on rigorous equality between the predicted and ground-truth rationales.

Classification metrics. Our evaluation metrics follow established practices in offensive language detection evaluation and prior works such as[20,42,46]. We assess the model's binary classification performance using macro F1, macro precision, macro recall, F1 score, precision, recall, accuracy, and AUROC.

Results and discussion

In the results and discussion section, we report the explainability performance results of our proposed framework in comparison with other LIME-based models. We also present the alignment between the LLM-generated rationales and the human-generated ground-truth rationales. Finally, we report the binary classification results of our proposed framework, the fine-tuned methods, and the LLMs. Additionally, we present the results of the ablation study, error analysis, and robustness analysis.

Comparison of SomOffXplain and LIME-based models

The explainability results show a significant difference between our proposed model and the baseline models, as presented in Table 2. In the presented results, the proposed SomOffXplain model achieves outstandingly higher explainability performance compared to other models that use LIME-based explanations. This indicates that SomOffXplain consistently outperforms all baseline models in both plausibility and faithfulness, achieving a Token F1 of 70.77%, an IOU F1-Score of 70.32%, and an AUPRC of 79.01%. The results demonstrate substantial alignment between the model-generated rationales and the ground truth annotations. In terms of plausibility, SomOffXplain exhibits a significant performance advantage over the best-performing baseline, DistilBERT [LIME], which achieves 40.63% for Token F1, 40.63% for IOU F1-Score, and 34.59% for AUPRC.

These findings suggest that SomOffXplain produces rationales faithfully aligned with those selected by humans. Moreover, SomOffXplain attains a higher Comprehensiveness score of 43.79%, showing that its extracted rationales exert a stronger influence on its predictions. However, the LIME-based XLNet [LIME] reports a lower Sufficiency score of 0.34%, which may indicate that its rationales are more faithful than those of SomOffXplain, though this result should be interpreted carefully.

Among the baseline models, DistilBERT [LIME] demonstrates the highest plausibility, likely due to its lightweight and efficient architecture. In Fig. 3a and b, we visualize saliency maps for selected text samples from the test dataset, comparing results from binary classification and target-wise saliency maps. This visualization

Model	Plausibility			Faithfulness	
	Token F1↑	IOU F1↑	AUPRC↑	Comp.↑	Suff.↓
RoBERTa [LIME]	0.3367	0.3367	0.2505	0.1385	0.0700
DistilBERT [LIME]	*0.4063*	*0.4063*	*0.3459*	*0.2217*	0.0624
mBERT [LIME]	0.3933	0.3933	0.3325	0.1774	0.1493
ALBERT [LIME]	0.3222	0.3222	0.2343	0.0543	0.0682
XLNet [LIME]	0.3326	0.3326	0.2495	0.0013	**0.0034**
SomOffXplain	**0.7077**	**0.7032**	**0.7901**	**0.4379**	*0.0173*

Table 2. Comparison of explainability results between SomOffXplain and the LIME-based baseline methods on the test dataset. Bolded values indicate the best scores, while italic values indicate the second-best scores. ↑ indicates that a higher value is better, while ↓ indicates that a lower value is better. Comp = comprehensiveness, and Suff = Sufficiency.

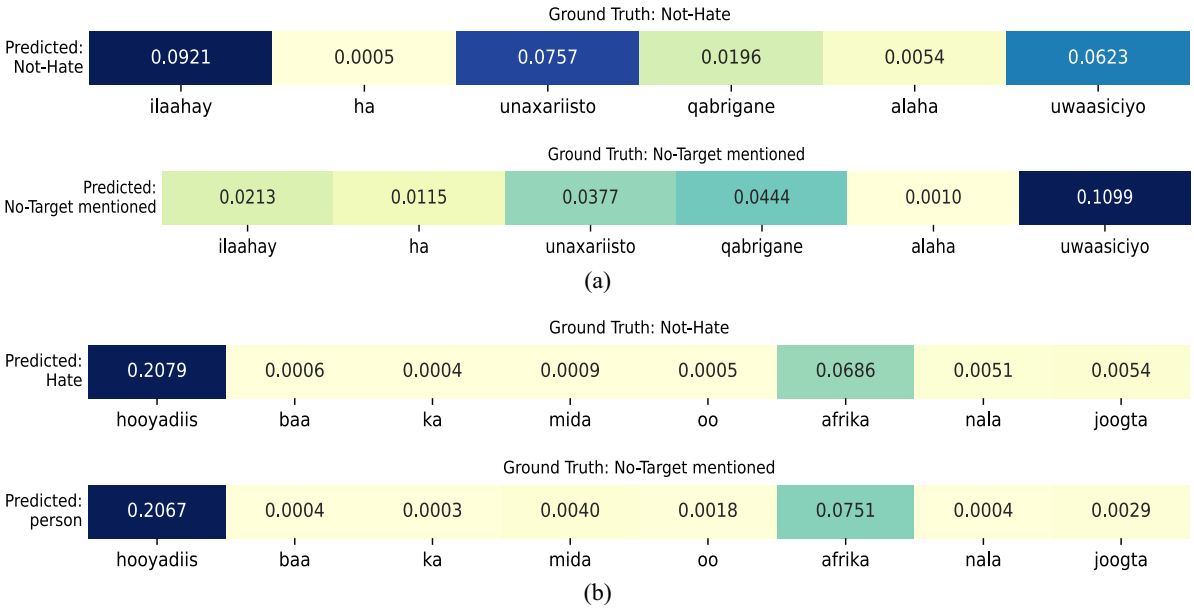


Fig. 3. Visualization of word importance using saliency maps to analyze the influence of words in binary classification compared to target-wise classification for sample texts. Each pair of figures represents the same sample.

emphasizes how model predictions change across different samples and highlights the important words that influence model predictions. For instance, in Fig. 3b, the model wrongly predicts the target as *person*, even though no *person* target is present in the ground truth. Despite such misclassifications, SomOffXplain's predictions generally align with human rationales across most saliency maps, as further shown in Table 3 for the tested sample sentence. Overall, SomOffXplain delivers robust and reliable explainability, underscoring its suitability for offensive language detection tasks. Furthermore, DistilBERT [LIME], as displayed in Table 3, shows a good selection of rationales that are close to the human-identified rationales for the sample sentence investigated.

Rationale generation using LLMs

To assess the effectiveness of LLMs in providing explainability for Somali text, we evaluate how well they generate rationales that align with human-provided ground-truth rationales. First, we use the entire dataset to evaluate the LLMs under zero-shot and few-shot prompting settings. In this evaluation, we perform both word-level and phrase-level analyses. The word-level evaluation measures the semantic overlap between human-generated and LLM-generated rationales, while the phrase-level evaluation measures how many of the rationale phrases produced by the LLMs truly match the human rationales. To this end, we summarize the findings in Table 4. In the few-shot setting, GPT-4o-mini achieves the highest phrase-level F1-score of 51%, although GPT-4o achieves comparable performance, indicating that their reasoning process at the phrase level is better than other models and aligns with human-generated rationales. In contrast, DeepSeek-v3 shows the lowest phrase-level performance of 33% in F1-score, suggesting that its rationale generation diverges more from human reasoning. In the word-level evaluation, all LLMs achieve higher F1-scores: 52%, 49%, 55%, and 56% for DeepSeek-v3, GPT-3.5, GPT-4o, and GPT-4o-mini, respectively, indicating a higher degree of overlap between

Method	Text
Human Annotator	Hooyada siilkedi bay gaadhsinaysa qaracaw cida dhashay aan la garanayn
RoBERTa [LIME]	Hooyada siilkedi bay gaadhsinaysa qaracaw cida dhashay aan la garanayn
DistilBERT [LIME]	Hooyada siilkedi bay gaadhsinaysa qaracaw cida dhashay aan la garanayn
mBERT [LIME]	Hooyada siilkedi bay gaadhsinaysa qaracaw cida dhashay aan la garanayn
ALBERT [LIME]	Hooyada siilkedi bay gaadhsinaysa qaracaw cida dhashay aan la garanayn
XLNet [LIME]	Hooyada siilkedi bay gaadhsinaysa qaracaw cida dhashay aan la garanayn
SomOffXplain	Hooyada siilkedi bay gaadhsinaysa qaracaw cida dhashay aan la garanayn

Table 3. Comparison of sample human-annotated and model-predicted rationale. We highlight rationales identified by both humans and models in orange, those identified only by models in green, and those identified only by humans in blue. Using direct translation we translate the sentence: *Hooyada siilkedi bay gaadhsinaysa qaracaw cida dhashay aan la garanayn*, as "It reaches out to your mother's p*ssy, you illegally born a child of unknown parentage.".

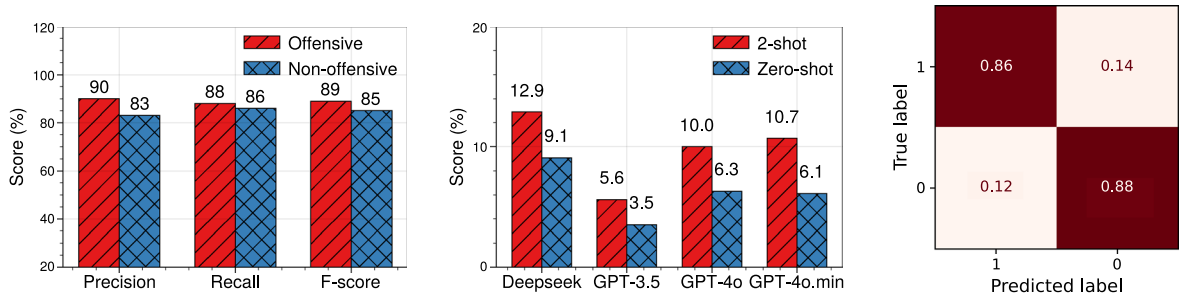
Setting	Model	Phrase-level			Word-level		
		F1-score	Precision	Recall	F1-score	Precision	Recall
2-shot	Deepseek-v3	0.3263	0.3356	0.3636	0.5227	0.4929	0.6867
	GPT-3.5	0.3486	0.2965	0.5172	0.4877	0.3905	0.8346
	GPT-4o	*0.5083*	0.4245	0.7707	*0.5516*	0.4561	0.8576
	GPT-4o.min	**0.5096**	0.4477	0.7015	**0.5560**	0.4818	0.7905
Zero-shot	Deepseek-v3	*0.2305*	0.2355	0.2586	**0.5151**	0.4576	0.7496
	GPT-3.5	0.1946	0.1733	0.2687	0.4548	0.3637	0.7760
	GPT-4o	**0.2464**	0.2207	0.3364	*0.5094*	0.4116	0.8541
	GPT-4o.min	0.21906	0.2010	0.2838	0.4367	0.3685	0.6710

Table 4. Comparison of zero-shot and few-shot performance in generating rationales by LLMs against human-provided ground-truth rationales on the test dataset, evaluated at the word and phrase levels.

Method	Text
Human Annotator	Hooyada siilkedi bay gaadhsinaysa qaracaw cida dhashay aan la garanayn
Deepseek-v3	Hooyada siilkedi bay gaadhsinaysa qaracaw cida dhashay aan la garanayn
GPT-3.5	Hooyada siilkedi bay gaadhsinaysa qaracaw cida dhashay aan la garanayn
GPT-4o	Hooyada siilkedi bay gaadhsinaysa qaracaw cida dhashay aan la garanayn
GPT-4o.min	Hooyada siilkedi bay gaadhsinaysa qaracaw cida dhashay aan la garanayn

Table 5. Comparing zero-shot LLMs identified rationales and human-identified rationales, we highlight rationales identified by both humans and models in orange, those identified only by models in green, and those identified only by humans in blue.

LLM-generated and human-generated words compared to phrase-level evaluations. Our findings reveal that most LLMs can detect explicit offensive words more effectively than implicit ones. This is evident in Table 5, where most LLMs select explicit offensive words. This suggests that LLMs rely more on dictionary-based words and direct translations but struggle to understand out-of-dictionary Somali offensive terms. Figure 4b presents the exact match rate between LLM-generated and human ground-truth rationales. DeepSeek-v3 achieves the highest exact match rate, scoring 12.9% in the few-shot setting. Table 6 shows the evaluation of the LLMs on 10% of the dataset. In these results, we compare human inter-annotator agreement with the LLMs' few-shot performance on the ground-truth rationales. The two human annotators achieved a 65% F1-score in word-level agreement, while GPT-4o-mini performed closest to humans, with an F1-score of 48%. As presented in Table 4, in the zero-shot setting, GPT-3.5 exhibits greater conceptual disparity from humans, scoring an F1-score of 19% in the phrase-level evaluation, whereas DeepSeek-v3 achieves 52% in the word-level evaluation, indicating that its rationale generation process is more closely aligned with humans. Moreover, other models score moderately: GPT-3.5 scores 45%, GPT-4o scores 51%, and GPT-4o-mini scores 44% in the zero-shot setting. In the 10% subset of the test dataset, GPT-4o shows the highest exact match rate of 7.8% among the LLMs, as shown in Table 6, while GPT-4o-mini achieves the highest recall of 74%, indicating that it selects more rationales from the human-selected rationales compared to the other LLMs. We conclude that existing state-of-the-art LLMs can recognize and select well-known offensive or non-offensive words as rationales, but they mostly fail to detect words containing implicit offensive language. We believe that enhancing data quality and creating more benchmarks, such as ours, may help prevent both online and LLM-generated offensive content.



(a) SomOffXplain. (b) LLM exact-match rate. (c) Confusion matrix.

Fig. 4. (**a**) The performance of SomOffXplain on the individual labels, offensive and non-offensive; (**b**) the exact-match rate performance comparison of the LLMs in zero-shot and few-shot rationale generation; and (**c**) the confusion matrix of SomOffXplain for binary classification performance.

Model	F1-score	Precision	Recall	Exact match (%)
Deepseek-v3	0.4624	0.4016	0.6973	3.2
GPT-3.5	0.4424	0.3685	0.7036	1.3
GPT-4o	0.4784	0.4134	0.7211	*7.8*
GPT-4o.min	*0.4814*	0.4049	0.7423	4.6
Human evaluation	**0.6515**	**0.6550**	**0.7191**	**25**

Table 6. Few-shot LLM performance compared to human performance in rationale generation, evaluated on 10% of the test dataset at the word level.

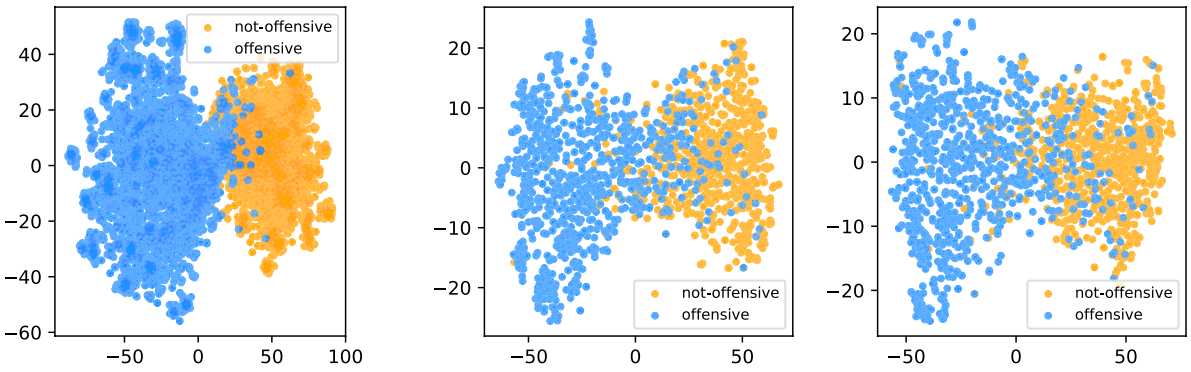
Setting	Model	Accuracy	Macro F1	Macro precision	Macro recall
2-shot	Deepseek-v3	0.7308	0.7682	0.6849	0.6858
	GPT-3.5	0.6929	0.6938	0.7002	0.6906
	GPT-4o	*0.7852*	0.7786	0.7843	0.7805
	GPT-4o.min	*0.7852*	0.7786	0.7843	0.7805
Zero-shot	Deepseek-v3	0.7380	0.6951	0.7767	0.6929
	GPT-3.5	0.6346	0.6328	0.6821	0.6689
	GPT-4o	*0.7852*	0.7786	0.7843	0.7805
	GPT-4o.min	0.7826	0.7757	0.7809	0.7776
-	SomOffExplian	**0.8710**	**0.8674**	**0.8657**	**0.8697**

Table 7. Classification performance of LLMs in zero-shot and few-shot settings compared to SomOffExplain on the test dataset. We bold the best-performing scores and italic the the second best performing scores.

Classification results

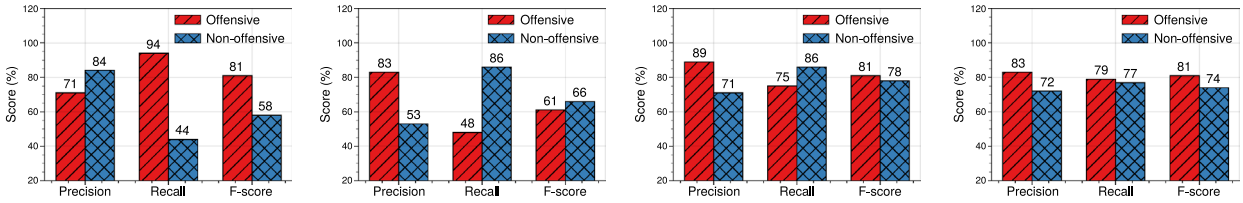
We report the binary classification results of SomOffXplain, compared with the LLMs, in Table 7. SomOffXplain achieves a Macro-F1 of 87% and an accuracy of 87%, making it the highest-scoring model. This strong performance is likely due to the combination of span extraction with the offensive classifier, which may indirectly aid in classification. In the few-shot setting, GPT-4o and GPT-4o-mini achieve better classification performance than the other LLMs, both reaching an accuracy of 79% and a Macro-F1 of 78%. Similarly, in the zero-shot setting, GPT-4o and GPT-4o-mini again outperform the other LLMs in binary classification, with scores close to those obtained in the few-shot scenario: accuracy of 79% and 78%, respectively.

In Fig. 4a, we show the performance of our model on the individual labels offensive and non-offensive. SomOffXplain achieves the highest F1-score on the offensive label at 89%, which may be because the offensive label has a larger number of samples in the dataset. In Fig. 4c, we present the confusion matrix of SomOffXplain for binary classification performance. Figure 5a–c also visualizes the t-SNE of the training, development, and test sets for SomOffXplain, showing that the model separates classes reasonably well. We also present and analyze the performance of the LLMs on the individual offensive and non-offensive classes. Through these evaluations, the zero-shot results in Fig. 6a–d show that among all four LLMs, DeepSeek-V3 achieves the lowest F1-score of 58% in the non-offensive class. However, DeepSeek-V3, GPT-4o, and GPT-4o-mini display similar performance in the offensive class, each obtaining an F1-score of 81%. GPT-3.5 performs moderately, while DeepSeek-V3 attains the highest recall score of 94% in the offensive class. In the few-shot setting, GPT-4o-mini achieves an F1-score



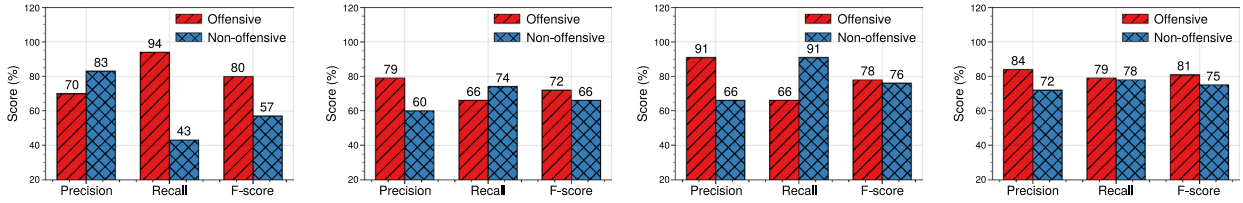
(a) Train. (b) Dev. (c) Test.

Fig. 5. SomOffXplain t-SNE visualization of training, validation, and testing sets.



(a) Deepseek-v3. (b) GPT-3.5. (c) GPT-4o. (d) GPT-4o-mini.

Fig. 6. Zero-shot classification performance of the LLMs on the individual labels offensive and non-offensive in the test dataset.



(a) Deepseek-v3. (b) GPT-3.5. (c) GPT-4o. (d) GPT-4o-mini.

Fig. 7. Few-shot classification performance of the LLMs on the individual labels offensive and non-offensive in the test dataset.

of 81% in the offensive class, outperforming the other models, with DeepSeek-V3 ranked second, as shown in Fig. 7a–d. However, in the non-offensive class, DeepSeek-V3 shows the lowest score at 57% F1, whereas GPT-4o is the best-performing model with an F1-score of 76% in the non-offensive class. In summary, although the LLMs show some consistency in their individual class performances, they still lag significantly behind our proposed model. This suggests that there is still room for improvement in LLMs for low-resource languages. In this context, our contribution of a high-quality labeled dataset stands out as a significant addition to this area of research. In Table 8, we report the performance of fine-tuned models in comparison with SomOffXplain.

In this comparison, DistilBERT performs best among the baselines with a Macro-F1 of 85%, while SomOffXplain outperforms all models. We also ran five experiments using different random seeds and conducted a comprehensive statistical significance analysis. As shown in Table 9, the model achieved a mean F1 score of 88.45% ± 0.85%, indicating high stability of the performance of the proposed framework.

Human-understandable explanations

We measure the overlap between model predicted rationale spans and human-annotated gold rationale spans. This allows us to quantify how closely the extracted rationales correspond to explanations selected by human annotators. Specifically, we report Token-F1, IOU-F1, and AUPRC. As shown in Table 2, SomOffXplain achieves 70.77% Token-F1, 70.32% IOU-F1, and 79.01% AUPRC. It also achieves a token-level precision of 70.41% and recall of 73.00%. These results indicate that the rationales generated by SomOffXplain substantially

Model	Accuracy	Macro F1	AUROC
RoBERTa	0.8153	0.8013	0.9075
DistilBERT	*0.8559*	*0.8540*	**0.9424**
mBERT	0.8507	0.8401	0.9302
XLNet	0.7937	0.7812	0.8933
ALBERT	0.7367	0.7059	0.8059
SomOffXplain	**0.8710**	**0.8674**	*0.9408*

Table 8. Classification performance of SomOffXplain compared to fine-tuned baseline methods on the test dataset.

Metric	Per-Run Results					Mean ± Std
	Run 1	Run 2	Run 3	Run 4	Run 5	
F1-score	0.8719	0.8905	0.8794	0.8906	0.8899	0.8845 ± 0.0085
Accuracy	0.8585	0.8651	0.8422	0.8710	0.8625	0.8599 ± 0.0109
AUROC	0.9426	0.9438	0.9446	0.9407	0.9424	0.9428 ± 0.0015
Precision	0.9533	0.8666	0.8109	0.9052	0.8550	0.8782 ± 0.0538
Recall	0.8033	0.9158	0.9607	0.8765	0.9279	0.8968 ± 0.0603

Table 9. SomOffXplain performance over five runs using five different random seeds for the binary classification.

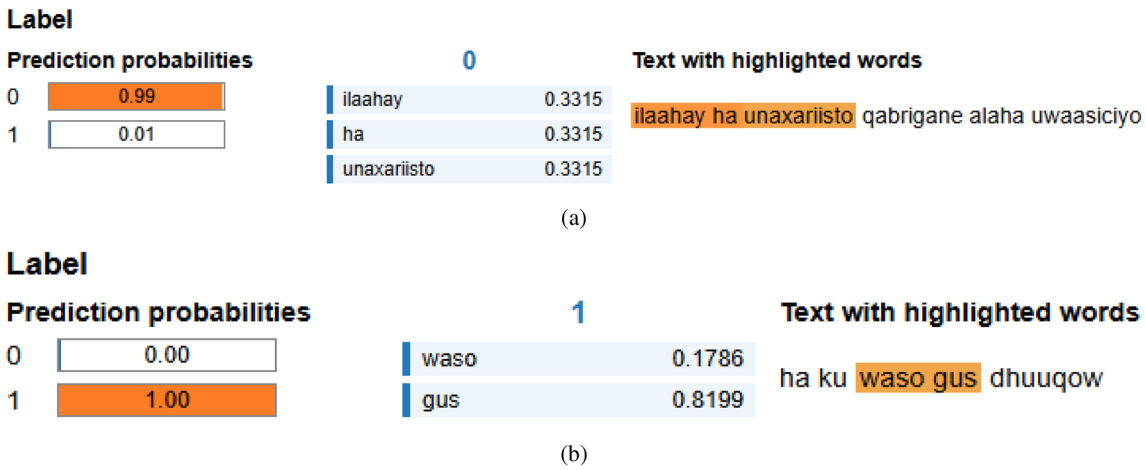


Fig. 8. Screenshot of two qualitative examples of rationale visualization, demonstrating how SomOffXplain produces human-understandable rationales.

overlap with human-annotated rationales, supporting their plausibility and interpretability. In addition to the quantitative evaluation, we provide qualitative examples to further illustrate the human interpretability of the generated rationales. For example, in the offensive sentence *sharmutoyinka dhan isku dooq waye*, meaning "all whores have the same hobby," the ground-truth rationale is *sharmutoyinka*, meaning "whore." The proposed framework correctly selects *sharmutoyinka*, the offensive term responsible for the classification. Similarly, for the non-offensive sentence *horay u soco walalkey qaali*, meaning "keep going my dear brother," the ground-truth rationale is *walalkey qaali*, meaning "my dear brother," and the proposed framework selects the same rationale while correctly classifying the sentence as non-offensive.

As shown in Fig. 8a, the sentence *ilaahay ha unaxariisto qabrigane alaha uwaaciyo*, meaning "may God brighten his grave and have mercy on him," has the ground-truth rationale *ilaahay unaxariisto*, meaning "may God have mercy." The model generates *ilaahay ha unaxariisto* as the rationale, which is semantically consistent with the human annotation. These examples show that SomOffXplain selects meaningful text spans that correspond closely to human-provided explanations. To demonstrate how the proposed framework produces such human-understandable rationales, we applied simple HTML and CSS to the explanation module. This display is similar to LIME's visualization style, showing rationale selection by assigning probabilities to words. As shown in Fig. 8a and b, the visualization highlights the selected rationale within the input text, making it easier

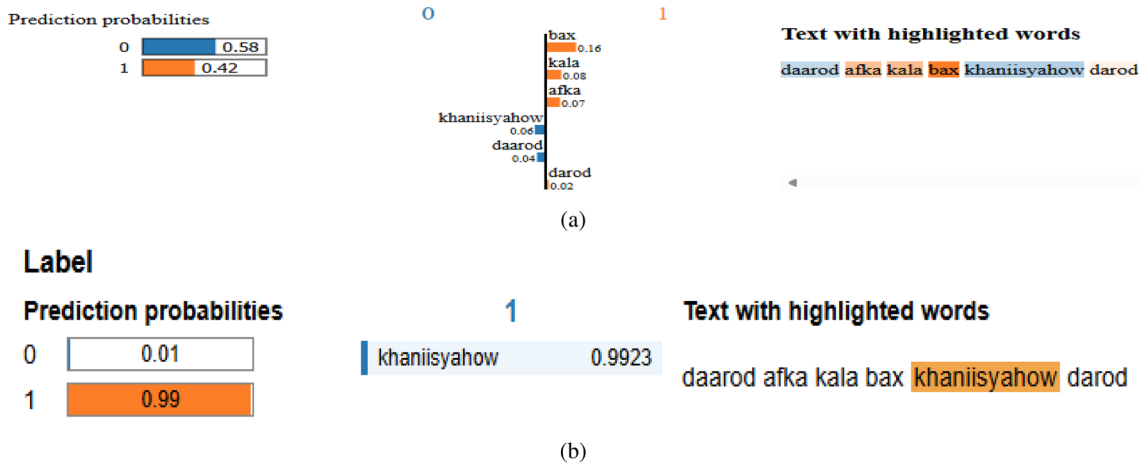
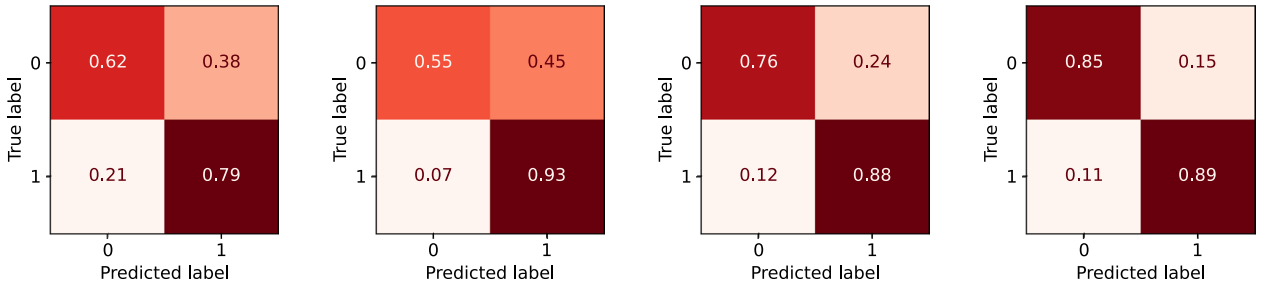


Fig. 9. Screenshot of an example of rationale generation comparison between SomOffXplain (**b**) and the best-performing baseline DistilBERT [LIME] (**a**).



(a) 10% of training data. (b) 25% of training data. (c) 50% of training data. (d) 100% of training data.

Fig. 10. Generalization performance of SomOffXplain on binary classification using different training data subsets and the same test set (1 = offensive, 0 = non-offensive).

for users to understand which words contributed to the model prediction. Additional examples are provided in Fig. 9.

Figure 9a shows rationales produced by LIME-based baseline models, whereas Fig. 9b shows rationales generated by SomOffXplain. In Fig. 9b, SomOffXplain correctly identifies the offensive expression *khaniisyahow*, meaning "you gay," while the best-performing LIME-based model, DistilBERT [LIME], fails to identify the offensive rationale and incorrectly classifies the sentence as non-offensive, as shown in Fig. 9a. Overall, the quantitative overlap with human-annotated rationales, together with the qualitative examples, provides stronger evidence that the explanations generated by SomOffXplain are plausible, semantically meaningful, and interpretable to humans.

Real-world applicability and generalizability of SomOffXplain

In this subsection we discuss the generalizability and possible real-world applicability of the proposed framework.

Real-world applicability. The contributions of this work can be applied to areas such as political discourse monitoring, social media moderation, online news comment filtering, and moderation assistance for human content moderators. The proposed framework has several practical advantages, including fast extractive explanations and the ability for a single model to predict both the label and the evidence span. Compared to LIME-based approaches, SomOffXplain has a simpler architecture, which suggests it may be more suitable for low-resource environments. It is computationally less expensive and can be integrated into production systems with minimal latency. Furthermore, the benchmark dataset may support applications such as fine-tuning models for Somali language offensive and hate speech content detection, and could contribute to broader efforts in social media platforms or chatbots to mitigate offensive and toxic content.

Generalizability. To test the generalizability and robustness of SomOffXplain under low-resource conditions, we conducted a reduced-data experiment using 10%, 25%, 50%, and 100% of the available labeled training data. The proposed framework achieved accuracies of 72%, 78%, 84%, and 87% on the respective data subsets in the binary classification task. Figure 10a–d present the confusion matrices for SomOffXplain's performance across these subsets. The performance trends indicate how well the framework scales with data availability and also show that SomOffXplain remains effective when supervision is scarce or training data is limited.

No.	Text
1	*howlgal lagu dilay dagaalamayaal ka tirsan alshabaab oo ka dhacay gobolka jubbada hoose.* **Translation:** An operation in which fighters belonging to Al-Shabaab were killed took place in the Lower Juba zone.
2	*weerar qaraxyo culus ku bilaawday oo ka dhacay deegaanka janaay cabdalle ee gobolka jubbada hoose.* **Translation:** An attack that began with heavy explosions occurred in the Janaay Cabdalle area of the Lower Juba zone.
3	*raysal wasaare oo maanta muqdisho ka daah furay mashruuc muhiim ah.* **Translation:** The Prime Minister today inaugurated an important project in Mogadishu.
4	*shaxda ugu fiican ilaa haatan horyaalka la liga xilli ciyaareedkan oo la soo saaray.* **Translation:** The best table so far this season of the La Liga championship has been released.

Table 10. Examples of rationales selected by SomOffXplain when testing generalization performance on four news headlines. Green words or phrases are the model predicted rationales, their English translations are also highlighted in green.

Model	Method	Plausibility			Faithfulness	
		Token F1↑	IOU F1↑	AUPRC↑	Comp.↑	Suff.↓
SomOffXplain	With label	0.7077	0.7032	0.7901	0.4379	0.0173
	Without label	0.6853	0.6781	0.7703	0.3643	0.0000

Table 11. Impact of the class-conditioned on explainability performance. ↑ indicates that a higher value is better, while ↓ indicates that a lower value is better.

We also performed inference to evaluate the generalizability of SomOffXplain's rationale generation on four news headline samples drawn from a completely different platform than the training data (Table 10). The table provides English translations and the selected rationales. Most of the main polarity-bearing words are selected; however, some important words are not. For example, *weerar* which means "attack" in the second sentence could be considered a rationale but was not selected. Overall, the rationale selection of SomOffXplain is promising and reliable. Nevertheless, further validation on cross-lingual and cross-domain settings across diverse low-resource languages is needed to establish broad generalizability, leaving room for future work.

Error analysis

To identify the weaknesses of the proposed framework and the factors affecting its performance, we conducted an error analysis.

The analysis revealed that the model's reasoning is generally correct; however, word-mismatch errors still arise during inference. In particular, rationale misalignment is mainly caused by lexical variation. For instance, *dilay* "killed" may not match the lexical form of *dilaan* "to kill" in the sentence *garxajis waa inay noocan ay iska dilaan intanay somaliland madaxtinimo ka doonin*, translated as "garxajis should kill this type before they go to find leadership in Somaliland." In this case, the model classified the sentence as non-offensive while simultaneously producing a mixed rationale that highlighted both offensive and non-offensive spans. Furthermore, words such as *naagahan* "these women", *ninkan* "this man", and *hooyadiis* "his mother" appear to affect both rationale generation and classification, possibly because they occur frequently in offensive sentences (see Fig. 13). These findings suggest that lexical variation and the prominence of certain words may bias the model's predictions. Overall, the error analysis indicates that the proposed framework is sensitive to specific lexical forms: although it can correctly identify most offensive expressions, it may fail to recognize morphologically different variants, thereby limiting its generalization across syntactic constructions. We also found that the LLMs sometimes generate incorrect rationales and misclassify sentences that do not contain explicit offensive words. For example, the sentence *walal ha u gafin waa qalad gafku e* was labeled as offensive by at least two LLMs, which also highlighted incorrect spans, even though the sentence is non-offensive. Therefore, we believe that addressing these issues remains an important direction for future work.

Ablation study

In this section, we present three ablation studies. To examine the influence of labels on explainability and the effect of hyperparameters, and dataset size on SomOffXplain, we conducted three ablation studies: (1) we tested explainability without using the class-conditioned, (2) we evaluated hyperparameter sensitivity by applying different hyperparameter settings to SomOffXplain, and (3) We divided the dataset into partitions and tested the effect of their size on performance.

Explainability without the class-conditioned

As shown in Table 11, we removed the labels from the contrastive learning module and conducted an ablation study. We observed a moderate performance decline in several metrics: Token F1 decreased from 70.77% to 68.53%, IOU F1-Score from 70.32% to 67.81%, and AUPRC from 79.01% to 77.03%. Comprehensiveness also dropped by 7.36% points. The Sufficiency score in the label-free experiment fell to 0.0%. In summary, these results indicate that labels significantly enhance the model's explainability performance, likely because contrastive learning becomes more discriminative when labels are used.

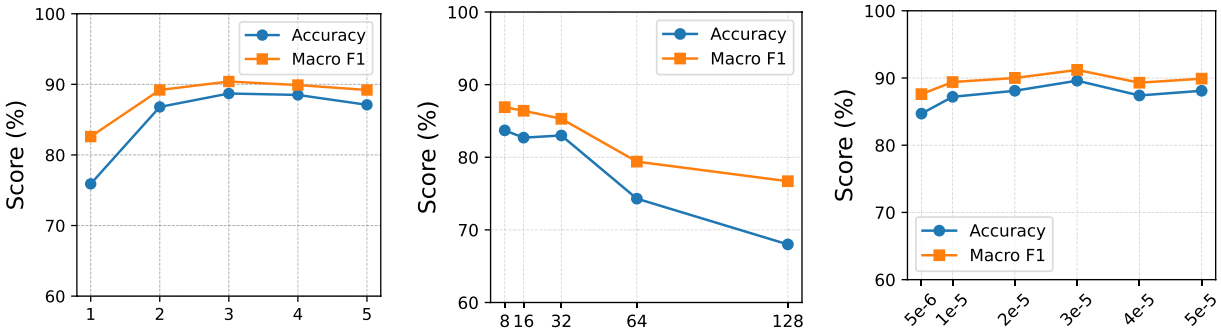


Fig. 11. Sensitivity analysis on SomOffXplain on the impact of epochs, batch size, and learning rate during training.

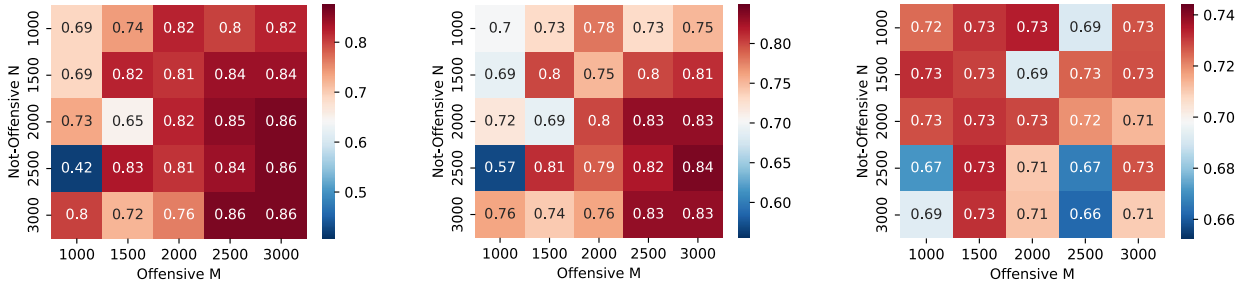


Fig. 12. Sensitivity analysis on dataset size, the effects on intersection-over-union, accuracy, and macro-F1 in the training dataset.

Hyperparameter sensitivity

We analyzed the sensitivity of hyperparameters by experimenting with different numbers of epochs (Fig. 11a), batch sizes (Fig. 11b), and learning rates (Fig. 11c). In most cases both a large number of epochs (e.g. 5) and a very small number (e.g. 1) reduced performance. A batch size of 8, which we selected for our experiment, consistently maintained higher performance, while other batch sizes led to declines in Macro-F1 and accuracy, with batch size 128 performing the worst.

Most learning rates produced comparable results, with no substantial differences in most cases. The learning rate 5e-5 yielded Macro-F1 and accuracy scores close to those of 3e-5; however, 3e-5 demonstrated superior explainability performance in our experiment.

Dataset sensitivity analysis

For the explainability, we selected 1500 offensive samples as they achieved a better and more consistent IOU F1-score of 73%. With regard to the classification performance, increasing the dataset size substantially impacts our model's performance. As the dataset size grows, performance metrics such as Macro-F1 and accuracy show noticeable improvements. We identified 3000 samples as the optimal dataset size for achieving the best scores across both Macro-F1 and accuracy metrics. This is illustrated in Fig. 12a–c, where Macro-F1 achieves a higher score compared to the other metrics. This difference may be influenced by class imbalance, as the 'non-offensive' class contains fewer samples than the 'offensive' class.

Case study

We examined the presence of bias in both binary and target-wise classifications, with the objective of identifying which categories were frequently incorrectly classified. To accomplish this objective, we visualized the dataset using saliency maps.

Our findings show that the model misunderstands some contextual meanings, classifying words such as *hooyadiis/hooyadaa* "his mother/your mother" and *aabahaa/aabe* "your father/father" as offensive, despite their positive use in context. This misclassification likely results from the high frequency of these words in offensive sentences, where they are commonly used in insults. For instance, as shown in Fig. 3b, the sentence *hooyadiis baa ka mida oo Afrika nala joogta*, which translates to "his mother is a member and stays in Africa with us," was classified as offensive, although it is not. The model's decision seems to have been influenced by the presence of *hooyadiis*. When we removed this word, the model correctly classified the sentence as non offensive (see

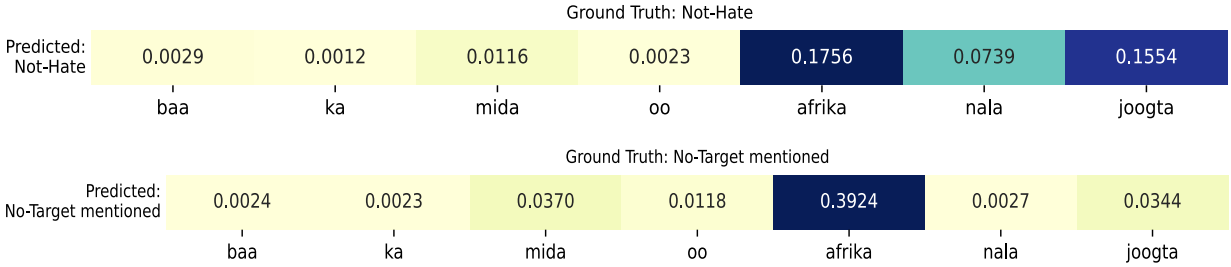


Fig. 13. A case study on why some samples are wrongly classified in the test dataset.

Fig. 13). In addition, understanding context remains a significant challenge for most models. In Somali, the lack of essential NLP tools, such as a part-of-speech tagger (POS-tagger), makes it difficult to handle morphological variations. Words with multiple forms can easily mislead the model. These findings suggest that the development of fundamental NLP tools for Somali considerably helps to improve text preprocessing and overall model performance in this under-resourced language.

Limitations and future work

In this section of the proposed framework, we briefly discuss the limitations of SomOffXplain and future research directions.

Limitations of SomOffXplain. First, we manually collected the dataset over a long period, which made the manual collection process challenging, time-consuming, and costly. Moreover, Somali is a low-resource language, and LLMs may overlook some contextual information, which can sometimes render them unable to detect implicit offensive content. Additionally, our model sometimes incorrectly classifies text when certain words, often used in conjunction with offensive terms, are present, even if these words have positive meanings. Finally, the proposed model faces challenges from different lexical forms and syntactic constructions of the same words.

Future work. for future work, we recommend linguistically informed resource development, such as a well-curated Somali POS-tagged corpus, to improve contextual understanding and enhance LLM-based offensive language detection in Somali. We also recommend using stemmers to resolve lexical ambiguity. Finally, we recommend new studies on targeted sexism in Somali.

Conclusion

In this study, we developed an interpretable framework for offensive language detection in the Somali language. To this end, we created a dataset of 10,175 samples, each annotated with human-provided rationales. We then proposed a framework that performs offensive language classification while concurrently extracting relevant rationales. These rationales provide clear explanations for the model's predictions, thereby enhancing its interpretability and trustworthiness. To establish a benchmark, we evaluated our framework against several baseline models, including LIME-based models and LLMs. Our model demonstrated superior performance in explainability, achieving higher plausibility and faithfulness compared to the baselines.

It also outperformed them in binary classification, attaining a Macro-F1 score and an accuracy of 87%. Our findings show that most LLMs evaluated were unable to detect some implicit offensive text in Somali. Furthermore, LIME-based models achieved lower scores than our model in terms of explainability for Somali text. These results underscore the importance of our work and the dataset we created, in advancing computational linguistics, particularly for Somali. Given that Somali is a low-resource language, we consider both our dataset and framework to be significant contributions to Somali NLP and the broader field of low-resource language research. Moreover, our framework is adaptable to other low-resource languages, and our dataset provides a generalizable approach to data collection and annotation. In summary, our work advances the field of NLP, especially in offensive language detection, where interpretability and transparency are essential for responsible content moderation. This contribution is particularly relevant in today's digital era, where the widespread use of LLMs and social media necessitates robust methods for identifying and mitigating offensive content.

Data availability

We release the code associated with this paper at https://github.com/Abdisalam-Badel/SomOffXplain-XOLDS . The dataset used in this study is available from the corresponding author upon reasonable request.

Received: 30 October 2025; Accepted: 14 May 2026
Published online: 28 May 2026

References

1. Hinduja, S. Generative AI as a Vector for Harassment and Harm. https://cyberbullying.org/generative-ai-as-a-vector-for-harassment-and-harm (Accessed 2 May 2025) (2023).
2. Hagen, L., Jingnan, H. & Nguyen, A. Elon musk's ai chatbot, grok, started calling itself 'mechahitler'. https://www.npr.org/2025/07/09/nx-s1-5462609/grok-elon-musk-antisemitic-racist-content (Accessed 8 September 2025) (2025).

3. Thanh Nguyen, L. Vihatet5: Enhancing hate speech detection in vietnamese with a unified text-to-text transformer model. In Ku, L.-W., Martins, A. & Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024, 5948–5961, https://doi.org/10.18653/v1/2024.findings-acl.355 (Association for Computational Linguistics, 2024).
4. Fortuna, P. & Nunes, S. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.* https://doi.org/10.1145/3232676 (2018).
5. Wardle, C. A conceptual analysis of the overlaps and differences between hate speech, misinformation and disinformation. https://peacekeeping.un.org/sites/default/files/report_-_a_conceptual_analysis_of_the_overlaps_and_differences_between_hate_speech_misinformation_and_disinformation_june_2024_qrupdate.pdf (Accessed 5 May 2025) (2024).
6. Ranasinghe, T. et al. SOLD: Sinhala offensive language dataset. *Lang. Resour. Eval.* **59**, 297–337. https://doi.org/10.1007/s10579-024-09723-1 (2025).
7. Law-insider. Offensive language definition. https://www.lawinsider.com/dictionary/offensive-language (Accessed 18 May 2025) (2023).
8. United, N. What is hate speech? https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech (Accessed 12 May 2025) (2023).
9. Zampieri, M. et al. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In May, J. et al. (eds.) *Proc. 13th International Workshop on Semantic Evaluation*, 75–86, https://doi.org/10.18653/v1/S19-2010 (Association for Computational Linguistics, 2019).
10. Vogels, E. A. The state of online harassment. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/ (Accessed 13 September 2024) (2021).
11. Davidson, T., Warmsley, D., Macy, M. & Weber, I. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media* **11**(1), 512–515. https://doi.org/10.1609/ICWSM.V11I1.14955 (2017).
12. İhtiyar, M., Özdemir, Ö., Erengül, M. & Özgür, A. A dataset for investigating the impact of context for offensive language detection in tweets. In Bouamor, H., Pino, J. & Bali, K. (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1543–1549, https://doi.org/10.18653/v1/2023.findings-emnlp.103 (Association for Computational Linguistics, 2023).
13. Beyhan, F. A turkish hate speech dataset and detection system. In *Proc. Thirteenth Language Resources and Evaluation Conference* (ed. Calzolari, N.), 4177–4185 (European Language Resources Association, 2022).
14. Nghiem, H. & Daumé Iii, H. Hatecot: An explanation-enhanced dataset for generalizable offensive speech detection via large language models. In Al-Onaizan, Y., Bansal, M. & Chen, Y.-N. (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2024*, 5938–5956, https://doi.org/10.18653/v1/2024.findings-emnlp.343 (Association for Computational Linguistics, 2024).
15. Lang, J. et al. Biting off more than you can detect: Retrieval-augmented multimodal experts for short video hate detection. In *Proc. ACM on Web Conference 2025, WWW '25*, 2763–2774, https://doi.org/10.1145/3696410.3714560 (Association for Computing Machinery, 2025).
16. Mnassri, K., Rajapaksha, P., Farahbakhsh, R. & Crespi, N. Bert-based ensemble approaches for hate speech detection. In *GLOBECOM 2022–2022 IEEE Global Communications Conference*, 4649–4654, https://doi.org/10.1109/GLOBECOM48099.2022.10001325 (2022).
17. Jin, Y., Wanner, L. & Shvets, A. Gpt-hatecheck: Can llms write better functional tests for hate speech detection? In Calzolari, N. et al. (eds.) *Proc. 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 7867–7885, https://aclanthology.org/2024.lrec-main.694/ (ELRA and ICCL, 2024).
18. Wang, Y., Li, H. & Wei, N. Sahsd: Enhancing hate speech detection in llm-powered web applications via sentiment analysis and few-shot learning. In *Proc. ACM on Web Conference 2025, WWW '25*, 3014–3025, https://doi.org/10.1145/3696410.3714644 (Association for Computing Machinery, 2025).
19. Nirmal, A., Bhattacharjee, A., Sheth, P. & Liu, H. Towards interpretable hate speech detection using large language model-extracted rationales. In Chung, Y.-L. et al. (eds.) *Proc. 8th Workshop on Online Abuse and Harms (WOAH 2024)*, 223–233, https://doi.org/10.18653/v1/2024.woah-1.17 (Association for Computational Linguistics, 2024).
20. Salles, I., Vargas, F. & Benevenuto, F. HateBRXplain: A benchmark dataset with human-annotated rationales for explainable hate speech detection in Brazilian Portuguese. In Rambow, O. et al. (eds.) *Proc. 31st International Conference on Computational Linguistics*, 6659–6669, https://aclanthology.org/2025.coling-main.446/ (Association for Computational Linguistics, 2025).
21. Calabrese, A., Ross, B. & Lapata, M. Explainable abuse detection as intent classification and slot filling. *Trans. Assoc. Comput. Linguist.* **10**, 1440–1454. https://doi.org/10.1162/tacl_a_00527 (2022).
22. Haturusinghe, S., Weerasooriya, T. C., Homan, C. M., Zampieri, M. & Liyanage, S. R. Subasa - adapting language models for low-resourced offensive language detection in Sinhala. In brahimi, A. et al. (eds.) *Proc. 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 4: Student Research Workshop), 260–270, https://doi.org/10.18653/v1/2025.naacl-srw.26 (Association for Computational Linguistics, 2025).
23. Singh, A. & Thakur, R. Generalizable multilingual hate speech detection on low resource indian languages using fair selection in federated learning. In Duh, K., Gomez, H. & Bethard, S. (eds.) *Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1: Long Papers), 7211–7221, https://doi.org/10.18653/v1/2024.naacl-long.400 (Association for Computational Linguistics, 2024).
24. Mohamed, A. An Overview of the Interface Between Aspects of Somali Phonology and Morphology: A Resource Guide for ESL Teachers, Master's thesis, St. Cloud State University, St. Cloud, Minnesota (2013).
25. Saeed, J. I. *London Oriental and African Language Library* (John Benjamins Publishing, 1999).
26. Ribeiro, M., Singh, S. & Guestrin, C. "why should I trust you?": Explaining the predictions of any classifier. In DeNero, J., Finlayson, M. & Reddy, S. (eds.) *Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 97–101, https://doi.org/10.18653/v1/N16-3020 (Association for Computational Linguistics, 2016).
27. Demilie, W. B. & Salau, A. O. Detection of fake news and hate speech for Ethiopian languages: A systematic review of the approaches. *J. Big Data* https://doi.org/10.1186/s40537-022-00619-x (2022).
28. Saroj, A., Chanda, S. & Pal, S. IRlab@IITV at SemEval-2020 task 12: Multilingual offensive language identification in social media using SVM. In Herbelot, A. et al. (eds.) *Proc. Fourteenth Workshop on Semantic Evaluation*, 2012–2016, https://doi.org/10.18653/v1/2020.semeval-1.265 (International Committee for Computational Linguistics, 2020).
29. Zampieri, M. et al. Predicting the type and target of offensive posts in social media. In Burstein, J., Doran, C. & Solorio, T. (eds.) *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), 1415–1420, https://doi.org/10.18653/v1/N19-1144 (Association for Computational Linguistics, 2019).
30. Waseem, Z. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In Bamman, D. et al. (eds.) *Proc. First Workshop on NLP and Computational Social Science*, 138–142, https://doi.org/10.18653/v1/W16-5618 (Association for Computational Linguistics, 2016).
31. Prabhu, R. & Seethalakshmi, V. A comprehensive framework for multi-modal hate speech detection in social media using deep learning. *Sci. Rep.* **15**, 13020. https://doi.org/10.1038/s41598-025-94069-z (2025).
32. Lu, J. et al. Hate speech detection via dual contrastive learning. *IEEE/ACM Trans. Audio Speech Lang. Process.* **31**, 2787–2795. https://doi.org/10.1109/TASLP.2023.3294715 (2023).

33. Kim, Y., Park, S. & Han, Y.-S. Generalizable implicit hate speech detection using contrastive learning. In Calzolari, N. et al. (eds.) *Proc. 29th International Conference on Computational Linguistics*, 6667–6679 (International Committee on Computational Linguistics, 2022).
34. Almahdi, A. J., Mohades, A., Akbari, M. & Heidary, S. Enhancing cross-lingual hate speech detection through contrastive and adversarial learning. *Eng. Appl. Artif. Intell.* **147**, 110296110296. https://doi.org/10.1016/j.engappai.2025.110296 (2025).
35. Luo, X. et al. Beyondgender: A multifaceted bilingual dataset for practical sexism detection. *Proc. AAAI Conference on Artificial Intelligence* **39**, 24750–24758. https://doi.org/10.1609/aaai.v39i23.34656 (2025).
36. Cinelli, M. et al. Dynamics of online hate and misinformation. *Sci. Rep.* **11**, 22083. https://doi.org/10.1038/s41598-021-01487-w (2021).
37. Yoo, S., Jeon, E., Hyeon, J. & Cho, J. Adaptive ensemble techniques leveraging BERT based models for multilingual hate speech detection in Korean and English. *Sci. Rep.* **15**, 19844. https://doi.org/10.1038/s41598-025-88960-y (2025).
38. Tuarob, S., Satravisut, M., Sangtunchai, P., Nunthavanich, S. & Noraset, T. Falcon: Detecting and classifying abusive language in social networks using context features and unlabeled data. *Inf. Process. Manag.* **60**, 103381. https://doi.org/10.1016/j.ipm.2023.103381 (2023).
39. Yang, Q. *et al.* SCCD: A session-based dataset for Chinese cyberbullying detection. In *Proc. 31st International Conference on Computational Linguistics* (ed. Rambow, O.) 9533–9545 (Association for Computational Linguistics, 2025).
40. Zaidan, O., Eisner, J. & Piatko, C. Using, "annotator rationales" to improve machine learning for text categorization. In Sidner, C., Schultz, T., Stone, M. & Zhai, C. (eds.) *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 260–267 (Association for Computational Linguistics, 2007).
41. Yessenalina, A., Choi, Y. & Cardie, C. Automatically generating annotator rationales to improve sentiment classification. In Hajič, J., Carberry, S., Clark, S. & Nivre, J. (eds.) *Proc. ACL 2010 Conference Short Papers*, 336–341 (Association for Computational Linguistics, 2010).
42. Mathew, B. et al. HateXplain: A benchmark dataset for explainable hate speech detection. *Proc. AAAI Conference on Artificial Intelligence* **35**, 14867–14875. https://doi.org/10.1609/aaai.v35i17.17745 (2021).
43. Ravikiran, M. & Annamalai, S. DOSA: Dravidian code-mixed offensive span identification dataset. In Chakravarthi, B. R., Priyadharshini, R., Kumar M, A., Krishnamurthy, P. & Sherly, E. (eds.) *Proc. First Workshop on Speech and Language Technologies for Dravidian Languages*, 10–17 (Association for Computational Linguistics, 2021).
44. Pavlopoulos, J., Sorensen, J., Laugier, L. & Androutsopoulos, I. SemEval-2021 task 5: Toxic spans detection. In *Proc. 15th International Workshop on Semantic Evaluation (SemEval-2021)* (ed. Palmer, A.) 59–69 https://doi.org/10.18653/v1/2021.semeval-1.6. (Association for Computational Linguistics, 2021).
45. Hoang, P. G., Luu, C. D., Tran, K. Q., Nguyen, K. V. & Nguyen, N.-T. ViHOS: Hate speech spans detection for Vietnamese. In *Proc. 17th Conference of the European Chapter of the Association for Computational Linguistics* (eds Vlachos, A. & Augenstein, I.) 652–669 https://doi.org/10.18653/v1/2023.eacl-main.47. (Association for Computational Linguistics, 2023).
46. Delbari, Z., Moosavi, N. S. & Pilehvar, M. T. Spanning the spectrum of hatred detection: A Persian multi-label hate speech dataset with annotator rationales. *Proc. AAAI Conference on Artificial Intelligence* **38**, 17889–17897. https://doi.org/10.1609/aaai.v38i16.29743 (2024).
47. KoboToolbox. Kobotoolbox. https://www.kobotoolbox.org/ (Accessed 2 September 2023).
48. Google. Google sheets. https://docs.google.com/spreadsheets (Accessed 15 September 2023).
49. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**, 378–382. https://doi.org/10.1037/h0031619 (1971).
50. Rizwan, H., Shakeel, M. H. & Karim, A. Hate-speech and offensive language detection in Roman Urdu. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (eds Webber, B. et al.) 2512–2522 https://doi.org/10.18653/v1/2020.emnlp-main.197. (Association for Computational Linguistics, 2020).
51. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint at arXiv:1810.04805 (2019).
52. Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. Preprint at arXiv:1907.11692 (2019).
53. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. Preprint at arXiv:1910.01108 (2020).
54. Yang, Z. *et al.* Xlnet: Generalized autoregressive pretraining for language understanding. Preprint at arXiv:1906.08237 (2020).
55. Lan, Z. *et al.* Albert: A lite bert for self-supervised learning of language representations. Preprint at arXiv:1909.11942 (2020).
56. DeepSeek, Inc., Your first api call | deepseek api docs. https://api-docs.deepseek.com/ (Accessed 04 May 2025) (2025).
57. OpenAI. Openai api. https://openai.com/api/ (Accessed 02 May 2025).
58. DeYoung, J. et al. ERASER: A benchmark to evaluate rationalized NLP models. In Jurafsky, D., Chai, J., Schluter, N. & Tetreault, J. (eds.) *Proc. 58th Annual Meeting of the Association for Computational Linguistics,* 4443–4458, https://doi.org/10.18653/v1/2020.acl-main.408 (Association for Computational Linguistics, 2020).
59. Jacovi, A. & Goldberg, Y. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Jurafsky, D., Chai, J., Schluter, N. & Tetreault, J. (eds.) *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 4198–4205, https://doi.org/10.18653/v1/2020.acl-main.386 (Association for Computational Linguistics, 2020).

Author contributions

Abdisalam Mahamed Badel designed the methods, collected the data, wrote the original manuscript, and conducted the experiments. Ting Zhong provided supervision. Xovee Xu edited and reviewed the paper and contributed to the methods. Wenxin Tai reviewed the paper and proofread it. Fan Zhou provided supervision and funding. Abdulahi Abdirashid Dahir contributed to the dataset, and proofread the paper.

Funding

This work was supported in part by the Ethiopian Somali Regional State (SRS) Finance Office and in part by the National Natural Science Foundation of China under Grant Nos. 62072077 and 62176043.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.M.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026