13TH INTERNATIONAL SYMPOSIUM ON APPLIED ENGINEERING AND SCIENCES
(SAES2025)
10th – 11th NOVEMBER 2025
UNIVERSITI PUTRA MALAYSIA

Presentation Type: Oral
Presentation Topic: AI and Big Data in Scientific Discovery
Presenter Name: Roesman Ridwan Raja

Ensemble Validation of Automated Annotations for Sexual Harassment Detection in Low-Resource Languages

Roesman Ridwan Raja,[1*] Kazutaka Shimada,[1]
[1]Graduate School of Informatics, Kyushu Institute of Technology, Iizuka, Japan
*Corresponding author's e-mail: raja.roesman-ridwan757@mail.kyutech.jp

Keywords – Large Language Models, Automated Annotation, Sexual Harassment Detection.

1. INTRODUCTION

The creation of high-quality, annotated datasets is a well-established bottleneck in developing effective, task-specific AI across domains. This challenge is further amplified in low-resource languages such as Bahasa Indonesia, where the scarcity of labeled data constrains both model training and evaluation. The conventional approach of manual annotation is notoriously inefficient, demanding significant time and financial investment, which renders it impractical for a scalable application. To address this, Large Language Models (LLMs) have emerged as a powerful alternative for automating the process. However, a critical problem arises: LLM outputs are not inherently reliable and require structured validation [1], which creates a new dilemma, as using human validators reintroduces the very costs and time delays that automation was meant to solve. This paper introduces and analyzes a fully automated, multi-stage pipeline designed to balance efficiency and quality. A large model (GPT-OSS 120b) is employed for initial annotation, while a comparatively smaller model (Qwen 3 32b) is used for validation. Using a lighter model reduces computational demand, and employing a different model helps avoid self-validation bias. This validation involves generating multiple judgments at different temperatures, which are then consolidated using robust aggregation methods, namely Mean Rating and Majority Voting. The primary objective of this study is to critically analyze the performance of such a framework, with particular attention to its applicability in low-resource contexts and its effectiveness for the sensitive task of sexual harassment detection.

2. METHODOLOGY

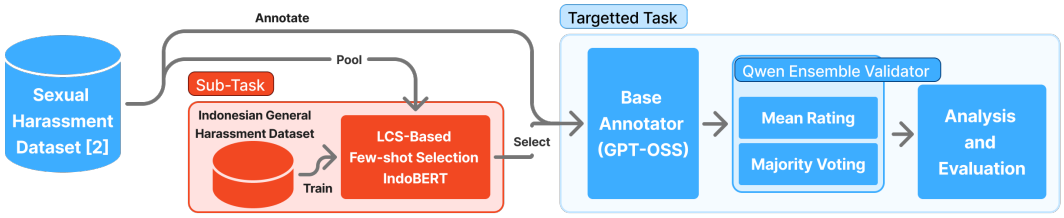

Figure 1. Proposed Method

The overall workflow of the proposed method is illustrated in Figure 1. Our methodology begins with the Nikmah et al. [2] dataset, an Indonesian corpus of 2,990 tweets manually labeled for sexual harassment (32.2% harassment, 67.8% non-harassment). The pipeline consists of two parts: a preparatory sub-task for example selection and a targeted annotation task. In the sub-task, an Indonesian general harassment fine-tuned IndoBERT model is used to identify the eight most informative examples based on the Least Confidence Score (LCS). This approach assumes that the

examples most difficult for a smaller model provide the strongest guidance for a larger model's in-context learning [3]. These curated examples are then supplied to the GPT-OSS 120b as a base annotator, which generates initial labels for the entire dataset. In the next stage, an ensemble validation step is performed using Qwen 3 32b, a LLM that assigns correctness scores ranging from 0 to 100 for each annotated by GPT-OSS instance under three temperature settings (T=0, T=0.5, T=1). Finally, the outputs from these three individual validators are aggregated into refined labels through two methods: Mean Rating (MR), which computes the average of the three confidence scores ($r1, r2, r3$) provided by the validator:

$$MR = \frac{(r1 + r2 + r3)}{3}$$

If the mean score falls at or below a threshold of 50, the label from the base annotator is flipped. In other words, the initial decision of GPT-OSS is changed to its opposite class (e.g., from Harassment to Non-harassment or vice versa). In contrast, if the score remains above 50, the label is held. Alternatively, the Majority Voting (MV) method converts each of the three ratings into a discrete flip or hold judgment. Each rating will represent a flip vote (<50) and a hold vote (>50). If two or more ratings vote to flip, the label from the base annotator is changed; otherwise, the label is held. This mechanism enables the validator to refine annotations by preserving reliable labels or correcting potential errors. To provide a comprehensive assessment, we evaluate each approach using four standard metrics: accuracy, precision, recall, and F1-score, with the latter serving as the primary indicator of balance between precision and recall.

3. RESULTS AND DISCUSSION

As shown in Table 1, the Mean Rating (MR) method achieved the highest F1-score (0.80). Although the gain is small, it is notable that the approaches did not reduce accuracy compared to other approaches. The analysis of individual validator temperatures shows a slight declining performance as temperature increases, with T=1 performing the worst (F1=0.78). The effectiveness of the proposed method was limited. The reason for this result is the simplicity of the combination method. To achieve higher validation accuracy, exploring alternative ensemble methods, such as scoring schemes with multiple parameters, remains an important direction for future work.

Table 1. Classification Report for each approach.

Approach	Accuracy	Precision	Recall	F1-Score
GPT-OSS Base (Unvalidated)	0.80	0.79	0.80	0.79
Individual Qwen T = 0	0.80	0.80	0.80	0.79
Individual Qwen T = 0.5	0.80	0.79	0.80	0.79
Individual Qwen T = 1	0.78	0.78	0.78	0.78
Majority Vote (MV)	0.80	0.80	0.80	0.79
Mean Rating (MR)	0.81	0.80	0.81	0.80

4. CONCLUSION

We proposed ensemble validation methods based on MR and MV. However, the improvement was slight because these approaches were simple. Future work will investigate more sophisticated validation techniques.

References

[1] N. Pangakis, S. Wolken, and N. Fasching, "Automated annotation with generative ai requires validation," *arXiv preprint arXiv:2306.00176*, 2023.

[2] T. L. Nikmah, M. Z. Ammar, Y. R. Allatif, R. M. P. Husna, P. A. Kurniasari, and A. S. Bahri, "Comparison of LSTM, SVM, and naive bayes for classifying sexual harassment tweets," *Journal of Soft Computing Exploration*, vol. 3, no. 2, pp. 131–137, 2022.

[3] K. Imazato and K. Shimada, "Automatic Few-shot Selection on In-Context Learning for Aspect Term Extraction," in *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2024, pp. 15–20.