Static or Dynamic? Exploring Cross-Lingual Few-Shot Transfer for Indonesian Misogynistic Meme Detection

Endang Wahyu Pamungkas[1,5,*,†], Maryam[1], Rona Rizkhy Bunga Chasana[2,5], Dade Nurjanah[3] and Ikhlasul Amal[4,†]

[1]Informatics Department, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia
[2]Communication Science Department, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia
[3]School of Computing, Telkom University, Bandung, Indonesia
[4]Computer Science and Electronics Department, Universitas Gadjah Mada, Yogyakarta, Indonesia
[5]Social Informatics Research Center, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia

Abstract

Online misogyny is often conveyed through memes that combine images and text to express stereotypes, objectification, and demeaning portrayals of women. Detecting such content is especially difficult in low-resource languages, where cultural references, local slang, and implicit meanings are poorly represented in available resources. This study compares three cross-lingual few-shot strategies for Indonesian misogynistic meme detection: static selection, dynamic textual retrieval, and dynamic multimodal retrieval. Demonstrations from six source languages are used to prompt a Multimodal Large Language Model without translating the Indonesian memes. Overall, 16 of 18 few-shot configurations outperform the zero-shot baseline. Static selection is the most consistent, achieving the best result in four of six language settings and the highest macro-F1 of 0.7549 with Spanish examples. Dynamic textual and multimodal retrieval perform best for Tamil and Chinese, respectively. These results show that cross-lingual few-shot transfer is generally effective, although no single demonstration-selection strategy is universally optimal.

Keywords

misogynistic meme detection, cross-lingual transfer, few-shot prompting, multimodal large language model, low-resource language

Content Warning: This paper includes examples of misogynistic memes containing offensive language and imagery that some readers may find uncomfortable.

1. Introduction

Internet memes have become a prominent form of digital communication, serving as vehicles for humor, social commentary, and political expression [2, 3]. At the same time, their viral and multimodal nature enables the dissemination of harmful content, including online misogyny [4, 5]. Misogynistic memes may communicate stereotypes, objectification, and demeaning portrayals of women through the interaction between visual elements and overlaid text. Their meaning is often conveyed implicitly through sarcasm, metaphor, and culturally specific references, making them difficult to identify using conventional text-based moderation systems [6, 7]. Effective detection therefore requires models that can interpret both modalities and understand how their interaction contributes to the intended message [8, 9, 10].

Substantial progress has been made in multimodal misogyny detection, although existing research remains concentrated on high-resource languages, particularly English. The SemEval-2022 Task 5 on Multimedia Automatic Misogyny Identification (MAMI) [11] established an influential benchmark and motivated approaches based on transformer-based multimodal fusion [12, 13], ensemble learning [14], and variational representation learning [15]. More recent studies have investigated context-aware

NL4AI 2026: Ninth Workshop on Natural Language for Artificial Intelligence, October 6-9th, 2026, Perugia, Italy [1]
*Corresponding author.
†These authors contributed equally.
ewp123@ums.ac.id (E. W. Pamungkas); maryam@ums.ac.id (Maryam); rrb608@ums.ac.id (R. R. B. Chasana); dadenurjanah@telkomuniversity.ac.id (D. Nurjanah); ikhlasulamal@mail.ugm.ac.id (I. Amal)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-05

CEUR-WS.org/Vol-4315/paper19.pdf

attention mechanisms [16] and multimodal large language models [17, 18]. However, transferring these advances to lower-resource languages such as Indonesian remains challenging. Indonesian memes frequently contain local slang, code-mixed expressions, indirect references, and culturally grounded forms of humor that are underrepresented in available multilingual resources [19, 20, 21].

Several studies have begun to address multilingual and cross-lingual misogyny detection. The MIMIC dataset introduced Hindi–English code-mixed multimodal content [22], while the PINK system investigated cross-lingual and multimodal transformers for sexism detection in Spanish memes [23]. More recently, Wang et al. [24] evaluated Vision–Language Models across six languages and showed that translating culturally situated content before classification may remove pragmatic information and weaken detection performance. These findings emphasize the importance of preserving native linguistic and visual context. Nevertheless, limited attention has been given to settings in which no labeled training data are available for the target language and knowledge must instead be transferred from datasets in other languages.

Cross-lingual few-shot prompting provides a potential solution by placing labeled source-language examples in the context of a Multimodal Large Language Model (MLLM). However, the effectiveness of this transfer may depend strongly on how the demonstrations are selected. Static few-shot selection uses a fixed set of representative examples for every target instance and can provide stable, carefully curated guidance. In contrast, dynamic retrieval selects query-specific demonstrations based on semantic similarity and may adapt more closely to the content of each target meme. Previous studies have shown the potential of retrieval-based few-shot learning for text classification [25], social media classification [26], and cross-lingual transfer [27]. Retrieval augmentation has also been explored for multimodal hate detection in short videos [28]. However, it remains unclear whether query-adaptive retrieval is preferable to static demonstrations for cross-lingual misogynistic meme detection. It is also unclear whether demonstrations should be retrieved according to textual meaning alone or according to a joint representation of text and image.

To investigate these questions, this study presents a systematic exploration of cross-lingual few-shot transfer for Indonesian misogynistic meme detection using six source-language datasets and no labeled Indonesian examples for training. We compare three demonstration-selection strategies: **static few-shot selection**, **dynamic textual retrieval**, and **dynamic multimodal retrieval**. The source examples are drawn from MAMI [11], MDMD [29], EXIST [30], CMMD [31], and MAMITA [32]. In the static setting, a fixed and balanced set of source-language examples is used for every Indonesian query. In the dynamic settings, a native multimodal embedding model [33] retrieves query-specific examples using either textual representations or joint text–image representations. The selected demonstrations are then incorporated into an MLLM prompt to classify Indonesian memes without explicitly translating them.

Under a consistent MLLM-based classification framework, we evaluate the three few-shot strategies against zero-shot prompting across all six source-language settings. We further examine embedding similarity, retrieval concentration, and qualitative retrieval examples to understand how demonstration selection influences cross-lingual transfer. Our results show that cross-lingual demonstrations improve over zero-shot prompting in most configurations. Static selection provides the most consistent performance, whereas dynamic textual and multimodal retrieval offer complementary advantages in particular source-language settings. These findings suggest that no single demonstration-selection strategy is universally optimal; rather, its effectiveness depends on the relevance and diversity of the source examples, the retrieval modality, and the compatibility between source and target contexts.

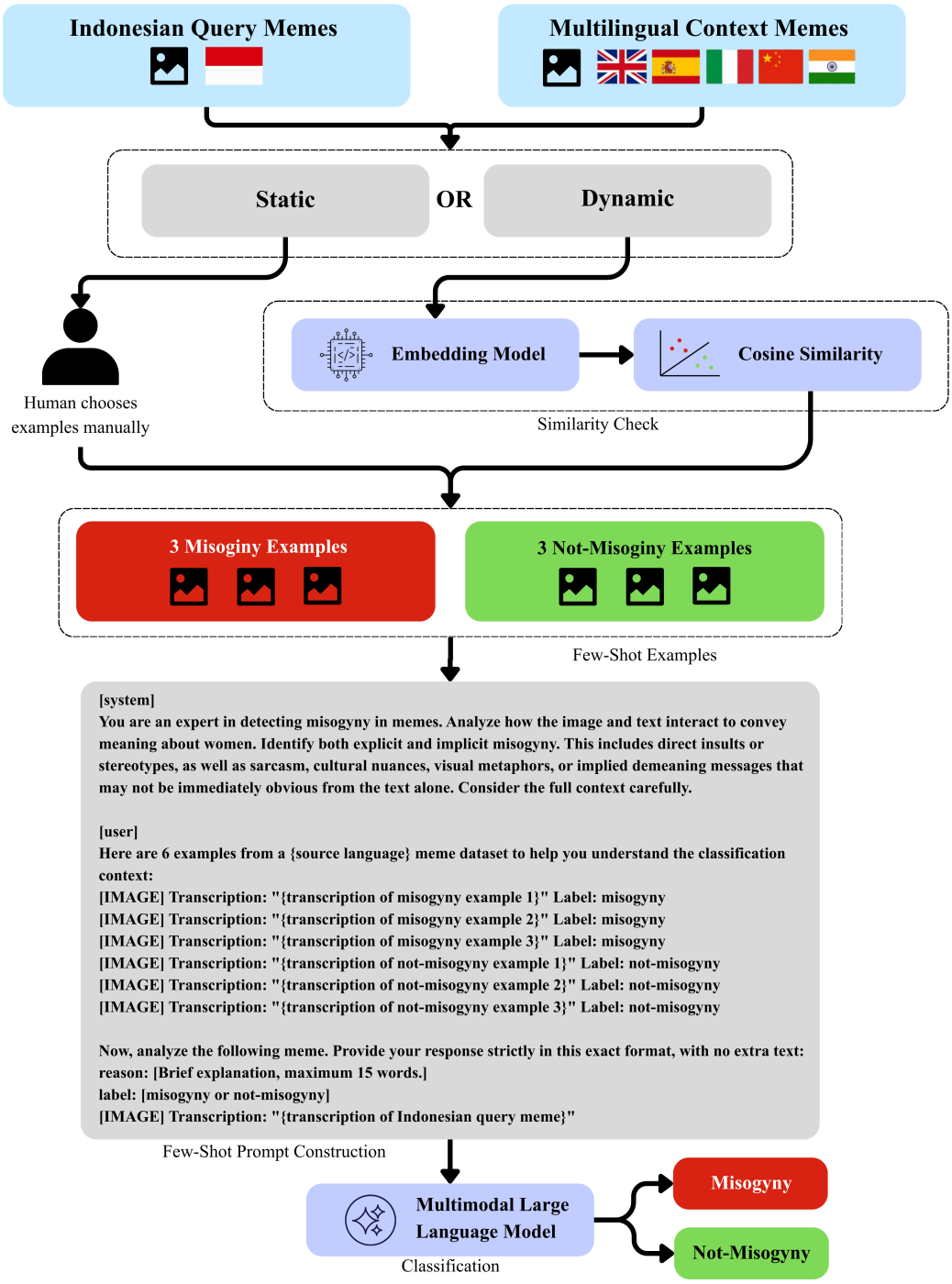


Figure 1: Overview of the cross-lingual few-shot framework comparing static and dynamic demonstration selection. In the static setting, examples are chosen manually by a human. In the dynamic setting, Indonesian query memes and multilingual context memes are embedded into a shared semantic space, and the top-3 most similar examples per class are retrieved via cosine similarity to construct a few-shot prompt, which is then classified by a Multimodal Large Language Model.

2. Methodology

2.1. Datasets

2.1.1. Indonesian Misogyny Dataset

For the Indonesian test set, we utilized the INDOMEME dataset [34], which comprises Indonesian memes collected from Facebook. From this dataset, we filtered memes that contain references to women, either visually or textually, resulting in 628 samples. These samples were then re-annotated by two independent annotators with expertise in Indonesian language and culture, labeling each meme as either *misogyny* or *not-misogyny*. The annotation focused on identifying content that contains negative stereotypes, generalizations, insults, or demeaning portrayals of women, whether expressed explicitly or implicitly through sarcasm, metaphor, or cultural references. The final dataset achieved a Cohen's Kappa score of 0.544, indicating moderate agreement, with disagreements resolved through consensus discussion. The dataset distribution consists of 207 misogyny samples and 421 not-misogyny samples. Given the limited size of this dataset, all 628 samples are used exclusively as the test set, with no Indonesian training data available. This motivates our use of multilingual meme datasets as the sole source of few-shot demonstrations without any target-language supervision.

2.1.2. Multilingual Source Datasets

To enable cross-lingual few-shot learning, we leveraged six existing misogynistic meme datasets from different languages. **MAMI** [11] is an English benchmark of 9,000 memes from SemEval-2022 Task 5. **MDMD** [29] covers two low-resource Indian languages, Tamil with 1,136 samples and Malayalam with 640 samples. **EXIST** [30] provides 1,712 Spanish memes annotated for sexist content using hard binary labels. **CMMD** [31] contains 1,190 Chinese memes addressing misogyny in East Asian cultural contexts. **MAMITA** [32] is an Italian dataset of 1,880 memes, and since it lacks text transcriptions, we generated them using Gemini 3.5 Flash. These datasets are summarized in Table 1. The pool deliberately spans both Latin-script languages (English, Spanish, Italian) and non-Latin scripts (Tamil, Malayalam, Chinese), which rigorously tests whether the embedding model aligns memes based on deep semantic and conceptual similarities rather than superficial character-level overlaps. The original dataset labels were mapped into a common binary *misogyny*/*not-misogyny* setting. However, differences in annotation definitions across datasets, particularly between sexism and misogyny, may introduce cross-dataset label variation.

Table 1
Multilingual meme datasets used as cross-lingual demonstration pools.

Dataset	Language	Script	Pool Size
MAMI [11]	English	Latin	9,000
MDMD (Tamil) [29]	Tamil	Non-Latin	1,136
MDMD (Malayalam) [29]	Malayalam	Non-Latin	640
EXIST [30]	Spanish	Latin	1,712
CMMD [31]	Chinese	Non-Latin	1,190
MAMITA [32]	Italian	Latin	1,880

2.2. Cross-Lingual Few-Shot Strategies

We compare three demonstration-selection strategies: static few-shot selection, dynamic textual retrieval, and dynamic multimodal retrieval, as illustrated in Figure 1.

2.2.1. Static Few-Shot Selection

For each source language, six examples were manually selected before evaluating the Indonesian test results, contained three misogynistic and three non-misogynistic examples. Selection aimed to cover diverse expressions in both classes while avoiding redundant or near-identical samples. The same six examples were subsequently used for all Indonesian queries in the corresponding source-language setting. Selected examples are shown in Appendix A.

2.2.2. Dynamic Cross-Lingual Retrieval

In the dynamic setting, query-specific examples are retrieved automatically using semantic similarity. The process consists of two components: an embedding model and a retrieval mechanism.

1. **Embedding Model**. We utilize Gemini Embedding 2 [33], a native multimodal embedding model that supports two distinct retrieval strategies. **Multimodal Embedding** jointly encodes both the meme image and its transcription into a single embedding vector, capturing the interaction between visual and textual modalities. **Textual Embedding** encodes only the transcription text, focusing purely on semantic meaning without visual influence. For each meme in both the Indonesian test set and multilingual training sets, embeddings are pre-computed under both strategies and normalized to unit length for cosine similarity computation.
2. **Retrieval Mechanism**. Given a query meme, we compute cosine similarity between its embedding and all embeddings in a source-language pool:

$$\text{similarity}(\vec{q}, \vec{e_i}) = \frac{\vec{q} \cdot \vec{e_i}}{\|\vec{q}\|\|\vec{e_i}\|} \tag{1}$$

   Retrieved samples are separated by label, and the top-3 most similar examples per class are selected, yielding 6 label-balanced examples (3 misogyny + 3 not-misogyny) for few-shot prompting.

2.2.3. Classification

Both static and dynamic strategies share the same classification setup. The few-shot prompt consists of a base instruction, followed by the 6 selected examples each presented with their meme image, transcription, and ground-truth label, and concludes with the Indonesian query meme to be classified, as illustrated in Figure 1. The base instruction is written in English, following the finding of Wang et al. [24] that English prompts consistently outperform native-language prompts for cross-lingual hate speech detection tasks. Full prompts are provided in Appendix B. We employ Gemini 3.5 Flash[1] for classification, chosen for its strong multilingual and multimodal capabilities. We additionally evaluate Gemma 4 E2B IT [35], a substantially smaller model to examine whether the effectiveness of cross-lingual few-shot demonstrations and retrieval strategies is consistent across models with different capacities. Temperature is set to 0 for more deterministic and reproducible predictions. We evaluate 19 configurations in total: one zero-shot baseline and 18 few-shot settings comprising six static, six dynamic textual, and six dynamic multimodal configurations, all measured using macro-F1 to account for class imbalance in the Indonesian test set.

3. Results and Discussion

Table 2 compares zero-shot prompting with static few-shot, dynamic textual retrieval, and dynamic multimodal retrieval across six source-language settings using two classifier models. Gemma 4 E2B IT, which has substantially fewer parameters than Gemini 3.5 Flash, exhibits limited benefits from cross-lingual demonstrations. All of its few-shot configurations perform below the zero-shot baseline of 0.5795, and no clear dominance is observed between static and dynamic few-shot strategies. The smaller

[1] https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/

Table 2
Macro-F1 scores for Indonesian misogyny detection using different source languages and few-shot strategies. Green indicates the best strategy per source language, bold indicates the best overall score for each classifier model.

Gemini 3.5 Flash			
Zero-shot Baseline		0.7227	
Source Language	**Static Few-Shot**	**Dynamic Multimodal**	**Dynamic Textual**
English	0.7536	0.7275	0.7328
Malayalam	0.7285	0.7192	0.7066
Tamil	0.7304	0.7275	0.7359
Spanish	**0.7549**	0.7306	0.7248
Italian	0.7472	0.7412	0.7424
Chinese	0.7314	0.7536	0.7349
Average	0.7410	0.7333	0.7296
Gemma 4 E2B IT			
Zero-shot Baseline		**0.5795**	
Source Language	**Static Few-Shot**	**Dynamic Multimodal**	**Dynamic Textual**
English	0.5010	0.4556	0.4830
Malayalam	0.4253	0.4404	0.5385
Tamil	0.5194	0.4187	0.5326
Spanish	0.4342	0.4700	0.2479
Italian	0.4841	0.4735	0.4544
Chinese	0.5142	0.5360	0.4859
Average	0.4797	0.4657	0.4571

model capacity may contribute to its limited ability to effectively utilize cross-lingual demonstrations, although further investigation is required to establish this relationship. In contrast, Gemini 3.5 Flash benefits substantially from cross-lingual demonstrations, with 16 of the 18 few-shot configurations outperforming the zero-shot baseline. Static few-shot selection is the most consistent strategy: it improves over zero-shot in all six settings, achieves the highest average macro-F1 of 0.7410, and performs best for four source languages. Its strongest result is obtained with Spanish examples, reaching 0.7549.

Dynamic retrieval provides complementary, language-dependent benefits for Gemini 3.5 Flash. Textual retrieval achieves the best performance for Tamil with a macro-F1 of 0.7359, whereas multimodal retrieval performs best for Chinese with 0.7536. On average, multimodal retrieval also outperforms textual retrieval across all source languages. However, neither retrieval modality is consistently superior across languages. Malayalam is the only setting in which both dynamic strategies fall below the zero-shot baseline, suggesting that query-specific similarity does not always produce informative cross-lingual demonstrations. These results indicate that static selection offers greater consistency, while dynamic retrieval can be valuable when the retrieved examples are well aligned with the target query.

Figure 2 illustrates how the two dynamic strategies select different examples for the same Indonesian query. The query presents an appearance-based comparison framed as a before-and-after transformation. For the misogyny class, both strategies retrieve examples related to appearance judgment, although multimodal retrieval emphasizes visual resemblance while textual retrieval emphasizes semantic similarity. In contrast, the retrieved not-misogyny examples are less closely related to the query. This example shows that class-balanced retrieval guarantees label balance but does not necessarily provide equally relevant demonstrations from both classes.

Figure 3 shows that multimodal embeddings produce higher mean cosine similarity scores ($\approx$0.70) than textual embeddings ($\approx$0.58) across all source languages. However, higher similarity does not

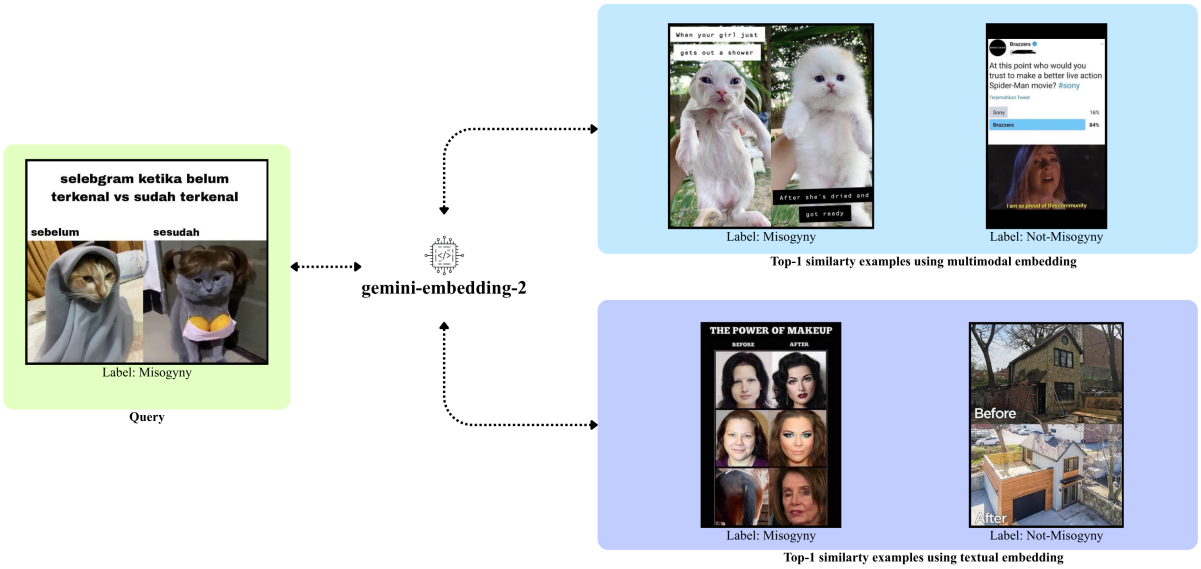


Figure 2: Top-1 examples retrieved for each class using multimodal and textual embeddings for the same Indonesian query meme.

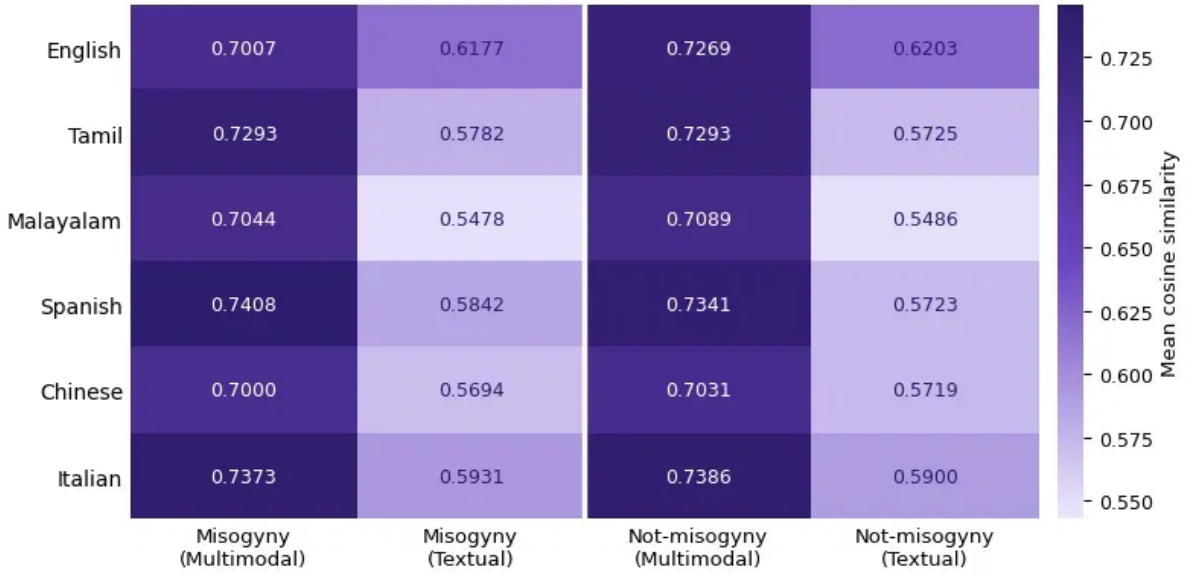


Figure 3: Mean cosine similarity between Indonesian queries and retrieved examples across source languages and embedding strategies.

consistently correspond to better classification performance. For example, multimodal retrieval produces the best result for Chinese, whereas textual retrieval performs better for Tamil and several other languages. Thus, the magnitude of embedding similarity alone is insufficient to determine whether retrieved demonstrations will be useful for classification.

To further examine retrieval diversity, we calculate the repetition ratio. For each class, the retrieval process produces $N_{\text{total}} = 628 \times 3 = 1{,}884$ retrieval slots:

$$\text{RepetitionRatio} = \frac{N_{\text{total}} - |\mathcal{U}|}{N_{\text{total}}}, \tag{2}$$

where $|\mathcal{U}|$ denotes the number of unique retrieved examples. A lower value indicates greater diversity

in the retrieved demonstrations across Indonesian queries.

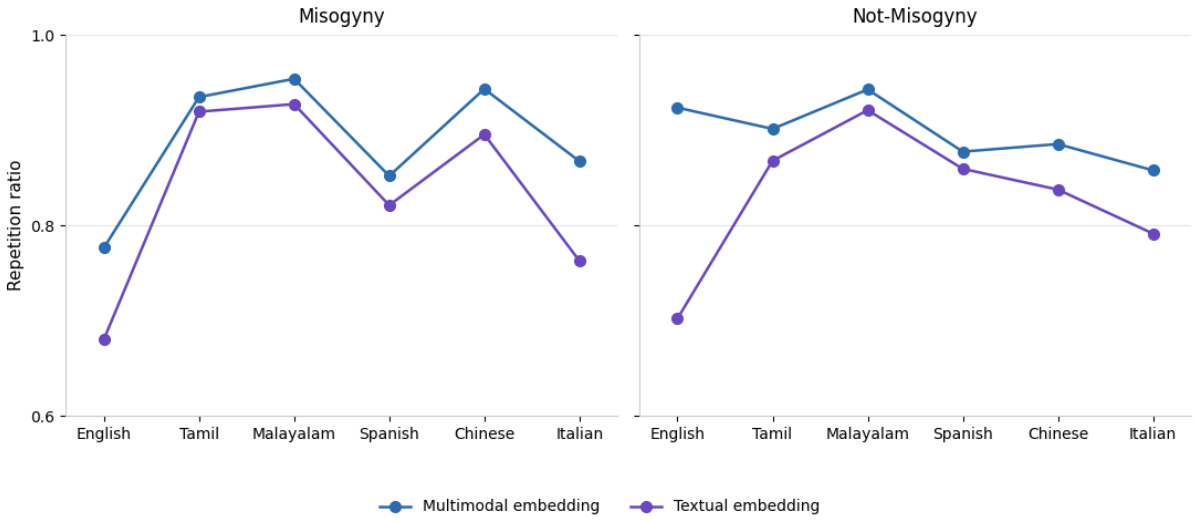


Figure 4: Repetition ratio across source languages and embedding strategies for misogyny (left) and not-misogyny (right).

As shown in Figure 4, all observed repetition ratios exceed 0.60, indicating that all settings repeatedly select a relatively small subset of source examples. Malayalam exhibits particularly high repetition and is also the only source language for which both dynamic strategies underperform zero-shot prompting. This suggests that limited retrieval diversity may reduce the usefulness of query-specific examples. Nevertheless, repetition is partly constrained by the size and class distribution of each source dataset, so this relationship should be interpreted descriptively rather than causally. Overall, the analysis shows that effective cross-lingual transfer depends not only on retrieving highly similar examples, but also on obtaining relevant and sufficiently diverse demonstrations.

4. Conclusion

This study explored cross-lingual few-shot transfer for Indonesian misogynistic meme detection by comparing static selection, dynamic textual retrieval, and dynamic multimodal retrieval across six source languages. Most few-shot configurations outperformed zero-shot prompting, showing that multilingual demonstrations can support low-resource meme classification without explicit translation or labeled Indonesian training data. Static selection provided the most consistent performance, while textual and multimodal retrieval achieved the best results in particular language settings. These findings indicate that no single demonstration-selection strategy is universally optimal. Effective transfer depends on the relevance, diversity, and modality of the selected examples. As the study uses a single embedding model and MLLM, future work will evaluate alternative models and investigate more culturally informed strategies for selecting cross-lingual demonstrations.

Acknowledgments

This work was funded by the Lembaga Riset dan Inovasi, Universitas Muhammadiyah Surakarta. The financial support was provided under the National Research Collaboration (Riset Kerjasama Nasional) scheme, Contract Number 236.24/A.3-III/LRI/VII/2025.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to perform grammar and spelling checks, improve writing style, and assist with formatting. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Safe and Responsible Innovation Statement

This paper contains examples of misogynistic memes, including offensive language and imagery directed at women, which are presented solely for research purposes. This work may support safer content moderation, but cross-lingual differences in culture, annotation, and data distribution can introduce bias and misclassification. Model outputs should therefore support, not replace, human judgment. Responsible use requires privacy protection, transparent error monitoring, appeal mechanisms, and care for annotators exposed to harmful content. The approach should not be used for surveillance or punitive moderation without appropriate safeguards.

References

[1] A. Bondielli, G. Bonetta, E. Leonardelli, I. Siragusa, Preface to the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026) co-located with 24th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2026), 2026.

[2] S. He, W. Zhang, J. Luo, P. Zhang, K. Zhao, D. D. Zeng, Modeling the co-diffusion of competing memes in online social networks, Decision Support Systems 187 (2024) 114324. doi:10.1016/j.dss.2024.114324.

[3] K. M. Duchscherer, J. F. Dovidio, When memes are mean: Appraisals of and objections to stereotypic memes., Translational Issues in Psychological Science 2 (2016) 335–345. doi:10.1037/tps0000080.

[4] M. Paciello, F. D'Errico, G. Saleri, E. Lamponi, Online sexist meme and its effects on moral and emotional processes in social media, Computers in Human Behavior 116 (2021) 106655. doi:10.1016/j.chb.2020.106655.

[5] G. Brookes, Language and online misogyny, in: H. Nesi, P. Milin (Eds.), International Encyclopedia of Language and Linguistics (Third Edition), third edition ed., Elsevier, London, 2026, pp. 299–303. URL: https://www.sciencedirect.com/science/article/pii/B9780323955041006773. doi:https://doi.org/10.1016/B978-0-323-95504-1.00677-3.

[6] G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, E. Fersini, Recognizing misogynous memes: Biased models and tricky archetypes, Information Processing & Management 60 (2023) 103474. doi:10.1016/j.ipm.2023.103474.

[7] D. S. M. Pandiani, E. T. K. Sang, D. Ceolin, 'toxic' memes: A survey of computational perspectives on the detection and explanation of meme toxicities, Online Social Networks and Media 47 (2025) 100317. doi:10.1016/j.osnem.2025.100317.

[8] R. Gomez, J. Gibert, L. Gomez, D. Karatzas, Exploring hate speech detection in multimodal publications, in: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1459–1467. doi:10.1109/WACV45572.2020.9093414.

[9] A. Muti, F. Ruggeri, K. A. Khatib, A. Barrón-Cedeño, T. Caselli, Language is scary when over-analyzed: Unpacking implied misogynistic reasoning with argumentation theory-driven prompts, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 21091–21107. doi:10.18653/v1/2024.emnlp-main.1174.

[10] B. Sheppard, A. Richter, A. Cohen, E. Smith, T. Kneese, C. Pelletier, I. Baldini, Y. Dong, Biasly: An expert-annotated dataset for subtle misogyny detection and mitigation, in: Findings of the

Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 427–452. doi:10.18653/v1/2024.findings-acl.24.

[11] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, Semeval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, 2022, pp. 533–549. doi:10.18653/v1/2022.semeval-1.74.

[12] A. R. Rao, A. Rao, ASRTrans at SemEval-2022 task 5: Transformer-based models for meme classification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 597–604. doi:10.18653/v1/2022.semeval-1.82.

[13] J. Zhang, Y. Wang, SRCB at SemEval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 585–596. doi:10.18653/v1/2022.semeval-1.81.

[14] J. García-Díaz, C. Caparros-Laiz, R. Valencia-García, UMUTeam at SemEval-2022 task 5: Combining image and textual embeddings for multi-modal automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 742–747. doi:10.18653/v1/2022.semeval-1.103.

[15] N. Jindal, P. K. Kumaresan, R. Ponnusamy, S. Thavareesan, S. Rajiakodi, B. R. Chakravarthi, MISTRA: Misogyny detection through text–image fusion and representation analysis, Natural Language Processing Journal 7 (2024) 100073. doi:10.1016/j.nlp.2024.100073.

[16] M. Z. U. Rehman, S. Zahoor, A. Manzoor, M. Maqbool, N. Kumar, A context-aware attention and graph neural network-based multimodal framework for misogyny detection, Information Processing & Management 62 (2025) 103895. doi:10.1016/j.ipm.2024.103895.

[17] B. Grasso, V. La Gatta, V. Moscato, G. Sperlì, KERMIT: Knowledge-empowered model in harmful meme detection, Information Fusion 106 (2024) 102269. doi:10.1016/j.inffus.2024.102269.

[18] P. Kapil, A. Ekbal, A transformer based multi task learning approach to multimodal hate speech detection, Natural Language Processing Journal 11 (2025) 100133. doi:10.1016/j.nlp.2025.100133.

[19] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in Twitter: a multilingual and cross-domain study, Information Processing & Management 57 (2020) 102360. doi:10.1016/j.ipm.2020.102360.

[20] E. Pamungkas, D. Putri, A. Fatmawati, Hate speech detection in Bahasa Indonesia: Challenges and opportunities, International Journal of Advanced Computer Science and Applications 14 (2023) 1175–1181. doi:10.14569/IJACSA.2023.01406125.

[21] M. O. Ibrohim, I. Budi, Hate speech and abusive language detection in Indonesian social media: Progress and challenges, Heliyon 9 (2023). doi:10.1016/j.heliyon.2023.e18647.

[22] A. Singh, D. Sharma, V. K. Singh, MIMIC: Misogyny identification in multimodal Internet content in Hindi-English code-mixed language, ACM Transactions on Asian and Low-Resource Language Information Processing (2024). doi:10.1145/3656169.

[23] G. Rizzi, D. Gimeno-Gómez, E. Fersini, C. D. Martínez-Hinarejos, Pink at exist2024: A cross-lingual and multi-modal transformer approach for sexism detection in memes, in: Conference and Labs of the Evaluation Forum, 2024.

[24] M. Wang, K. Ren, P. Jalan, A. Ashraf, T. V. Vu, R. Seetharaman, S. Nawaz, U. Naseem, From native memes to global moderation: Cross-cultural evaluation of vision–language models for hateful meme detection, in: Proceedings of the ACM Web Conference 2026, WWW '26, Association for Computing Machinery, New York, NY, USA, 2026, p. 9788–9799. doi:10.1145/3774904.3793007.

[25] G. Yu, L. Liu, H. Jiang, S. Shi, X. Ao, Retrieval-augmented few-shot text classification, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 6721–6735. doi:10.18653/v1/2023.findings-emnlp.447.

[26] Y. Wang, H. Li, F. Teng, L. Chen, GORAG: Graph-based online retrieval augmented generation for dynamic few-shot social media text classification, in: Proceedings of the ACM Web Conference

2026 (WWW '26), Association for Computing Machinery, New York, NY, USA, 2026, p. 9148–9159. doi:10.1145/3774904.3793011.

[27] M. S. Bari, B. Haider, S. Mansour, Nearest neighbor few-shot learning for cross-lingual classification, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 1745–1753. doi:10.18653/v1/2021.emnlp-main.131.

[28] J. Lang, R. Hong, J. Xu, Y. Li, X. Xu, F. Zhou, Biting off more than you can detect: Retrieval-augmented multimodal experts for short video hate detection, in: Proceedings of the ACM Web Conference 2025, WWW '25, Association for Computing Machinery, Sydney NSW, Australia, 2025. doi:10.1145/3696410.3714560.

[29] R. Ponnusamy, S. Rajiakodi, B. Sivagnanam, A. Kizhakkeparambil, D. Sharma, P. Buitelaar, B. Raja Chakravarthi, Fuse-md: A culturally-aware multimodal model for detecting misogyny memes, Natural Language Processing Journal 14 (2026) 100197. doi:https://doi.org/10.1016/j.nlp.2026.100197.

[30] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9–12, 2025, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2025, p. 266–289. doi:10.1007/978-3-032-04354-2_16.

[31] B. R. Chakravarthi, R. Ponnusamy, P. Du, X. Zhuang, S. Rajiakodi, P. Buitelaar, B. Premjith, B. Sivagnanam, A. Kizhakkeparambil, S. Lavanya, An overview of the misogyny meme detection shared task for chinese social media, in: Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, Unior Press, Naples, Italy, 2025, pp. 200–208. URL: https://aclanthology.org/2025.ltedi-1.32/.

[32] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, Mamita: Benchmarking misogyny in italian memes, in: Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), CEUR Workshop Proceedings, 2025, pp. 445–455. URL: https://aclanthology.org/2025.clicit-1.44/.

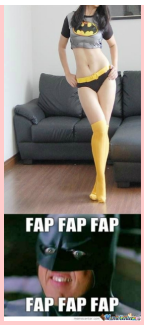
[33] M. Shanbhogue, Z. Li, S. Zhang, G. H. Ábrego, S.-C. Huang, A. Jain, D. Salz, S. Goenka, C. Hegde, J. Ma, F. Chen, J. Wu, T. Dabral, B. Samari, K. Poulet, D. Cer, K. Chen, P. Suganathan, H. Hui, J. Andonov, P. Schlattner, J. Han, I. Naim, W. Lowe, V. Pchelin, A. Yang, Y.-T. Chen, Z. Ding, G. Zhang, G. Heigold, Y. Chen, A. Reveillon, B. Mccloskey, W. Zhou, D. Kim, R. Meng, E. Wang, J. Zheng, H. Fede, Z. Yang, K. Mosley, B. Potetz, S. Dua, H. S. Vera, S. Gao, H. Zhang, A. Hess, H. Ying, A. Montes, K. Gill, M. Choi, S. Russo, A. Hauth, J. Lee, M. Boratko, M. Barnes, V. Rao, C. Musat, C. Allauzen, E. Variani, S. Kumar, T. Bagby, J. Jiao, Y. Gu, T. Li, A. Agrawal, R. Santana, D. Nath, S. Karukas, S. Han, L. Loher, A. Twu, N. Vyas, S. Bhai, F. P. Gomez, W. Zhang, C. Liu, J. Yang, S. Qiu, S. Zhang, S. Kulkarni, S. Rothe, S. Nakamoto, R. Hoffmann, Z. Gleicher, Y. Sung, Q. Yin, T. Duerig, M. Seyedhosseini, Gemini embedding 2: A native multimodal embedding model from gemini, 2026. doi:10.48550/arXiv.2605.27295.

[34] E. W. Pamungkas, C. S. Wahyuni, I. Amal, D. Purworini, B. S. Rintyarna, Decoding hate in memes: multimodal and multitask approaches for low-resource indonesian social media, PeerJ Computer Science 12 (2026) e3736. doi:10.7717/peerj-cs.3736.

[35] G. Team, Gemma 4 technical report, 2026. URL: https://arxiv.org/abs/2607.02770. arXiv:2607.02770.

A. Static Few-Shot Examples

Table 3
Static few-shot examples used per source language.

Language	Misogyny			Not-Misogyny		
English						
Tamil						
Malayalam						
Spanish						
Chinese						
Italian						

B. Classification Prompts

Table 4
Zero-shot and few-shot prompts used for misogyny classification.

Zero-Shot Prompt

[system]

You are an expert in detecting misogyny in memes. Analyze how the image and text interact to convey meaning about women. Identify both explicit and implicit misogyny. This includes direct insults or stereotypes, as well as sarcasm, cultural nuances, visual metaphors, or implied demeaning messages that may not be immediately obvious from the text alone. Consider the full context carefully.

[user]

Provide your response in this exact format, with no extra text:

```
reason: [Brief explanation, maximum 15 words.]
label: [misogyny or not-misogyny]
```

[IMAGE] Transcription: "*{transcription of Indonesian meme}*"

Few-Shot Prompt

[system]

You are an expert in detecting misogyny in memes. Analyze how the image and text interact to convey meaning about women. Identify both explicit and implicit misogyny. This includes direct insults or stereotypes, as well as sarcasm, cultural nuances, visual metaphors, or implied demeaning messages that may not be immediately obvious from the text alone. Consider the full context carefully.

[user]

Here are 6 examples from a *{source language}* meme dataset to help you understand the classification context:

[IMAGE] Transcription: "*{transcription of misogyny example 1}*" Label: `misogyny`
[IMAGE] Transcription: "*{transcription of misogyny example 2}*" Label: `misogyny`
[IMAGE] Transcription: "*{transcription of misogyny example 3}*" Label: `misogyny`
[IMAGE] Transcription: "*{transcription of not-misogyny example 1}*" Label: `not-misogyny`
[IMAGE] Transcription: "*{transcription of not-misogyny example 2}*" Label: `not-misogyny`
[IMAGE] Transcription: "*{transcription of not-misogyny example 3}*" Label: `not-misogyny`

Now, analyze the following meme. Provide your response strictly in this exact format, with no extra text:

```
reason: [Brief explanation, maximum 15 words.]
label: [misogyny or not-misogyny]
```

[IMAGE] Transcription: "*{transcription of Indonesian meme}*"