When Bigger Isn't Better: Evaluating LLMs for Arabic Sentiment Analysis

Mohamed Ibrahim, Abdullah Makki, Youssef Barakat, Nour Samy, Sarah AlHumoud

Wittify.ai, Riyadh, Saudi Arabia

{mohamed.ibrahim, abdullah.makki, youssef.barakat, nour.samy, sarah}@wittify.ai

1. Abstract

This study evaluates the performance of a fine-tuned Arabic sentiment transformer (CAMeL-MSA) against eight large language models (LLMs). Using zero-shot prompting across six Arabic sentiment datasets, we compare a specialized, task-specific approach against generalized model capabilities. Results show that the fine-tuned baseline substantially outperformed all LLMs on five of the six datasets in both accuracy and Macro F1-score. While LLMs offer versatility, this comparison highlights the continued practical superiority of task-specific fine-tuning over zero-shot prompting.

Keywords: Arabic sentiment analysis, pretrained transformers, large language models, CAMeL-MSA, CAMeL-BERT, multilingual NLP, dialectal Arabic, evaluation protocol, Macro F1-score

2. Introduction

Arabic sentiment analysis remains a challenging task due to the language's rich morphology, dialectal diversity, and frequent code-switching with other languages. While large language models (LLMs) have demonstrated strong performance on general NLP tasks, their effectiveness in specialized tasks such as Arabic sentiment classification remains less clearly understood, particularly under realistic evaluation settings.

In practical applications, developers often face a choice between using general-purpose LLMs with prompting or deploying smaller, task-specific models that have been fine-tuned for sentiment analysis. However, this decision is rarely evaluated systematically across multiple datasets under consistent evaluation protocols.

This work addresses the following research question: Do pretrained transformer models fine-tuned for Arabic sentiment analysis outperform general-purpose LLMs when those LLMs are used in a zero-shot, prompt-based setting, especially on complex and diverse datasets?

To answer this question, we evaluate one fine-tuned transformer model (CAMeL-MSA) and eight LLMs across six Arabic sentiment datasets. These datasets cover multiple domains, including standard benchmarks (LABR, ASTD, ArSAS), ASR-transcribed customer feedback, synthetic conversational data, and multilingual multi-dialect text.

Our contributions are as follows:

(1) A systematic multi-dataset evaluation under a unified protocol: eight LLMs × six corpora (54 runs), with CAMeL-MSA evaluated on the same six corpora for comparison;

(2) Analysis of the impact of dataset characteristics such as dialect and label consistency;

(3) Transparent documentation of evaluation settings, including prompting and decoding strategies for LLMs. To facilitate reproducibility while maintaining controlled access, the full prompt templates and configuration files are available upon request[1].

3. Related Work

Arabic sentiment analysis has been studied extensively using datasets that vary in domain, dialect, and annotation quality. Common benchmarks include LABR (Aly and Atiya, 2013), ASTD (Nabil et al., 2015), and ArSAS (Al-Twairesh et al., 2017). These datasets differ significantly in their linguistic characteristics and labeling methodologies. In particular, datasets such as LABR require mapping from rating scales to sentiment classes, which introduces potential inconsistencies across studies.

Pretrained transformer models have demonstrated strong performance on Arabic NLP tasks. Models such as AraBERT (Antoun et al., 2020) and CAMeL-BERT (Inoue et al., 2021) have been widely used for sentiment analysis due to their ability to capture contextual representations. Other architectures, such as AraXLNet (Alduailej and Alothaim, 2022), have also shown competitive performance in this domain. These models benefit from supervised fine-tuning on task-specific datasets.

More recently, large language models (LLMs) have been applied to sentiment analysis using prompting-based approaches. However, prior work suggests that zero-shot or few-shot prompting may not always match the performance of fine-tuned models, particularly in low-resource or domain-specific settings. For example, Alahmadi (2025) compares human

[1] PI: sarah@wittify.ai

annotations, fine-tuned transformers, and zero-shot LLMs for Arabic sentiment analysis, showing strong performance for specialized models. Similarly, Polock (2025) demonstrates that LLMs often require adaptation techniques such as few-shot prompting or fine-tuning to achieve competitive results across low-resource languages. Our study focuses strictly on zero-shot LLM use, which is a tighter and more challenging baseline for LLMs than settings that allow few-shot or tuning. Hasan et al. (2024) further highlight that performance differences depend heavily on language and evaluation protocol.

In this work, we extend prior research by providing a multi-dataset comparison between a fine-tuned Arabic sentiment model and eight contemporary LLMs under a consistent zero-shot evaluation setup. Unlike studies that focus on a single dataset or model family, we evaluate across six datasets, including synthetic and ASR-based data, to better reflect real-world deployment scenarios.

Recent work also compares modern LLMs on Arabic tasks such as sentiment and translation (Zouidine and Khalil, 2025), underscoring rapid change in model availability and the need for up-to-date, task-specific benchmarks.

4. Methodology

The methodology in this study includes the models evaluated and the metrics used for evaluation, as will be explained in the following.

4.1 Models Evaluated

Nine models are evaluated in this study. First, CAMeL-BERT (Inoue et al. 2021): a pretrained transformer model (110M parameters) developed by CAMeL Lab at New York University Abu Dhabi, UAE. It is fine-tuned specifically for Arabic sentiment analysis. This model serves as the baseline for domain-specific pretrained models. Second, Qwen2.5-3B:(Yang et al., 2024) a general-purpose multilingual LLM with 3B parameters, developed by Alibaba Cloud, which was evaluated using zero-shot prompting for sentiment classification. Third, Qwen2.5-7B-Instruct (Yang et al., 2024): an instruction-tuned variant of Qwen2.5 with 7B parameters, optimized for following instructions and task-specific prompts. Fourth, Qwen3-8B (Yang et al., 2025): an 8 billion parameter multilingual LLM from Alibaba Cloud. It was evaluated in a zero-shot setting for Arabic sentiment analysis. Fifth, AraGPT2-medium (Antoun et al., 2021): an Arabic pretrained generative language model based on the GPT-2 architecture, trained on large-scale Arabic corpora and evaluated using prompting-based sentiment classification.Sixth, BLOOMz-1b7 (Muennighoff et al., 2022): multilingual LLM with 1.7 billion parameters developed by the BigScience project. It was evaluated to assess the performance of the LLM for sentiment analysis. Seventh, BLOOMz-3b (Muennighoff et al., 2022): a 3 billion parameter instruction-tuned multilingual model from BigScience, assessed for sentiment classification in a zero-shot setting. Eighth, BLOOMz-7b1 (Muennighoff et al., 2022): a 7.1 billion parameter multilingual instruction-tuned model from BigScience, evaluated to examine the effect of increased parameter scale on Arabic sentiment analysis performance.Ninth, Gemma-3-4b-it (Gemma Team, 2025): a 4 billion parameter instruction-tuned model developed by Google, evaluated using zero-shot prompting for sentiment classification.

4.2 Evaluation Protocol

We evaluate all models using a consistent evaluation protocol: the labels are three-class sentiment classifications (positive, negative, neutral). The metrics used are Accuracy and Macro F1-score. For the LLM evaluation, a zero-shot method using enhanced prompts was implemented to guide the models in sentiment classification. For transformers, batch processing and checkpointing systems are implemented to ensure evaluation robustness and to allow resumption.

5. Datasets

5.1 LABR (Large-Scale Arabic Book Reviews)

The LABR dataset (Aly and Atiya 2013) contains 7,056 book reviews (after balancing). With 1-5 star ratings. The reviews are in Modern Standard Arabic (MSA) and various dialects. The labels mapping is as follows: 1-2 stars is negative, 3 stars is neutral, and 4-5 stars is positive.

The LABR dataset exhibits label inconsistencies that affect evaluation reproducibility. The original data set uses a rating system of 1 to 5 stars. Some annotators use 1 star to indicate the best rating (positive), and others use 5 stars to indicate the best rating (positive). In this evaluation, we follow the standard convention: 4 or 5 stars indicate positive sentiment, 3 stars indicate neutral sentiment, and 1 or 2 stars indicate negative sentiment. However, it should be noted by researchers that different preprocessing steps may use different label mappings, leading to inconsistent results. We suggest that future work should aim to normalize the label mapping or use a dataset where sentiment is already labeled.

5.2 ASTD (Arabic Sentiment Tweets Dataset)

The size of the ASTD dataset is 2,415 tweets (after balancing) (Nabil et al. 2015). The dataset of Arabic tweets with sentiment labels includes Egyptian and Gulf dialects. For the label mapping, it is a direct mapping: positive, negative, and neutral, with objective/mixed mapped to neutral.

5.3 ArSAS (Arabic Saudi Sentiment)

The size of the ArSAS dataset (Al-Twairesh et al. 2017) is 13,200 tweets (after balancing) in the Saudi Arabian dialect, annotated with sentiment labels. The label mapping is direct: positive, negative, and neutral.

5.4 Transcribed Dataset

This dataset includes 97 audio samples of customer feedback in the Saudi dialect and was collected in a natural setting. There are three types of audio samples in this dataset: positive samples (35), negative samples (35), and neutral samples (27). The audio samples were supplied by a voice actor from Saudi Arabia and were designed to resemble real customer scenarios. Unlike other datasets, the audio samples were not manually transcribed but were instead transcribed by our ASR system. This was done to simulate real-world transcription errors and test the sentiment analysis system in real-world scenarios.

Model	LABR (7k)	ArSAS (13k)	Transcribed (97)	Multilingual (15k)	Saudi (300)	ASTD (2.4k)
CAMeL-MSA	47.00	86.70	93.33	97.22	100.00	18.39
Qwen2.5-3B	36.32	34.11	31.00	33.37	33.33	33.42
Qwen2.5-7B	32.57	41.50	28.33	30.71	35.67	33.87
Qwen3-8B	34.17	33.80	33.00	35.16	30.00	32.46
AraGPT2-medium	32.26	36.28	39.00	46.55	19.33	34.53
BLOOMz-1b7	32.78	36.23	25.00	32.90	34.00	34.53
BLOOMz-3b	34.06	44.14	32.33	33.33	29.00	33.33
BLOOMz-7b1	32.54	34.47	34.67	17.98	33.33	33.37
Gemma-3-4b-it	33.50	32.46	34.33	33.33	33.33	34.00

Table 1. Accuracy by model and dataset (%)

Model	LABR (7k)	ArSAS (13k)	Transcribed (97)	Multilingual (15k)	Saudi (300)	ASTD (2.4k)
CAMeL-MSA	0.446	0.866	0.933	0.972	1.000	0.187
Qwen2.5-3B	0.278	0.215	0.174	0.248	0.167	0.172
Qwen2.5-7B	0.207	0.332	0.226	0.178	0.213	0.271
Qwen3-8B	0.199	0.259	0.187	0.291	0.159	0.260
AraGPT2-medium	0.250	0.244	0.291	0.372	0.133	0.276
BLOOMz-1b7	0.209	0.288	0.198	0.264	0.239	0.255
BLOOMz-3b	0.222	0.353	0.262	0.167	0.150	0.226
BLOOMz-7b1	0.178	0.197	0.213	0.146	0.167	0.170
Gemma-3-4b-it	0.253	0.171	0.187	0.167	0.167	0.227

Table 2. Macro F1-score by model and dataset

5.5 Saudi Dataset

The dataset consists of 300 samples of synthetic conversational Arabic text designed for sentiment analysis tasks. Every sample was generated using the GPT-4.1 Nano LLM to simulate authentic dialogue patterns across various Arabic dialects. The label mapping for this collection is direct and categorizes sentiment into three distinct classes: positive, negative, and neutral.

5.6 Multilingual Multi-Dialect Dataset

This data is generated synthetically using GPT-4.1 Nano LLM. It comprises multiple Arabic dialects, including Egyptian, Saudi, Moroccan, Levantine, and MSA, in addition to mixed Arabic–English. The size of the dataset is 15,000 sentences from daily-life interaction with sentiment labels and direct mapping: positive, negative, and neutral.

All datasets are mainly balanced to ensure fair evaluation across sentiment classes. We use stratified sampling to create balanced subsets when the original datasets are imbalanced.

6. Results

We evaluated one pretrained model (CAMeL-MSA) and eight LLMs on six datasets, giving 54 model–dataset combinations. (see Section 4.1).

The pretrained model (CAMeL-MSA) substantially outperformed all LLMs on five of six datasets. On ArSAS (N=13,200), CAMeL-MSA reached 86.7% accuracy and 0.87 Macro F1-score, compared to the best LLM (BLOOMz-3b) at 44.1% and 0.35. On LABR (N=7,056), CAMeL-MSA achieved 47% accuracy and 0.45 Macro F1-score versus the best LLM (Qwen2.5-3B) at 36.3% and 0.28. On two other datasets (Multilingual, N=15,000 and

Transcripted, N=97), CAMeL-MSA reached 97.2% and 93.3% and 100% on one Saudi dataset, while LLMs remained in the 28–46% range. On ASTD (N=2,415), both the pretrained model and the LLMs performed poorly (CAMeL-MSA 18.4%, best LLM AraGPT2 34.5%), consistent with known domain and label issues on that benchmark.

7. Discussion

Our evaluation reveals several key findings about the performance of pretrained transformers versus LLM-based models for Arabic sentiment analysis.

7.1 Pretrained Models vs LLM-Based

Results show that the pretrained transformer (CAMeL-MSA), despite having far fewer parameters (110M vs 1.7B–8B), consistently outperforms the evaluated LLMs when the latter are used in a zero-shot, prompt-based way. As shown in Tables 1 and 2, on five of six datasets, CAMeL-MSA achieves substantially higher accuracy and Macro F1-score than the best LLM (e.g., 86.7% vs 44.1% on ArSAS, 47% vs 36.3% on LABR). This indicates that domain-specific fine-tuning provides substantial advantages over general-purpose LLMs for Arabic sentiment analysis.

It is important to note that the evaluated LLMs are used in a zero-shot setting without task-specific fine-tuning. Therefore, the comparison reflects differences in adaptation strategies rather than model size alone. Increasing model size does not consistently improve performance under zero-shot prompting, as observed across several datasets. This suggests that task-specific adaptation plays a more critical role than parameter scale in this setting. Under our protocol, zero-shot prompting alone is often insufficient to match the fine-tuned baseline on several corpora, which has implications for reliable Arabic sentiment classification in resource-sensitive deployments.

On LABR, larger Qwen variants do not monotonically improve Macro F1-score (Table 2); BLOOMZ ordering also varies by dataset.

The ASR-transcribed subset (N = 97) and the synthetic Saudi set (N = 300) are small or synthetically generated; very high scores on some columns warrant caution (limited diversity, possible generator bias). Treat as complementary stress tests, not sole evidence of production readiness.

7.2 Impact of Dataset Characteristics

The performance also shows considerable variation across datasets. Better performance is observed for simpler and more consistent datasets (e.g., Saudi Sentences) compared to complex and dialectal ones (e.g., LABR and ASTD). So, it is essential that the model be tested with relevant and diversified data, which reflects the complexity of real-life data. The LABR dataset consists of two types of challenges. First, the nature of book reviews. The sentiment in book reviews is often expressed indirectly and is nuanced in the form of critique. As such, the sentiment is not necessarily expressed through direct positive and negative words. As such, the review often contains praise and criticism, metaphor, comparison, and the use of literary words. This sentiment is also implicit and context-dependent, which makes it more difficult for the model to learn than in the case of simpler review datasets. The second challenge is label inconsistencies, as discussed in Section 5.1. This challenge reflects a larger problem in the evaluation of Arabic NLP systems: the absence of standardized preprocessing and label mapping conventions. We recommend that future work explicitly document label mapping strategies and, where possible, use datasets with explicit sentiment labels rather than star ratings.

8. Limitations

This study compares a fine-tuned transformer model with LLMs evaluated in a zero-shot setting. While this reflects common practical usage, it does not isolate the effect of model size from adaptation strategy.

We do not include few-shot prompting or fine-tuning of LLMs, which may improve their performance. Future work should explore these settings for a more comprehensive comparison.

Additionally, some datasets used in this study are small or synthetically generated, which may limit generalizability. Synthetic datasets may introduce biases that do not fully reflect real-world language use.

Finally, LLM performance may be sensitive to prompt design and decoding parameters. Although we standardize these settings, variations may lead to different results.

9. Conclusion

This paper evaluated one pretrained transformer (CAMeL-MSA) and eight large language models on six Arabic sentiment datasets. The results show that the pretrained model substantially outperforms zero-shot LLMs on five of six datasets, with large gaps in accuracy and Macro F1-score (e.g. 47–97% vs 31–44% on the same data), despite having far fewer parameters. We identify key factors affecting performance: dataset complexity, dialectal variation, and label consistency. Another contribution is the identification of inconsistencies in the LABR set's labeling and their impact on the reproducibility of the evaluation process. Future research directions include: (1) standardizing the evaluation process and the labeling scheme, (2) extending the evaluation to larger models and datasets, (3) examining the impact of

domain-specific pre-training on model performance, and (4) improving the evaluation metrics for multilingual and multi-dialect content.

10. References

Alahmadi, D. (2025). Human Versus AI: A Comparative Study of Zero-Shot LLMs and Transformer Models Against Human Annotations for Arabic Sentiment Analysis. *International Journal of Advanced Computer Science and Applications*, 16(8). *(Match pages/DOI to the published PDF.)*

Alduailej, A., & Alothaim, A. (2022). AraXLNet: Pre-trained language model for sentiment analysis of Arabic. *Journal of Big Data, 9*(1), 72. https://doi.org/10.1186/s40537-022-00625-z

Al-Twairesh, N., Al-Khalifa, H., & Al-Salman, A. (2017). ArSAS: An Arabic Saudi sentiment analysis dataset. In *Proceedings of the 24th International Conference on Neural Information Processing (ICONIP).*

Aly, M., & Atiya, A. (2013). LABR: A large-scale Arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 494–498). Association for Computational Linguistics. https://aclanthology.org/P13-2088

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection* (pp. 9–15). European Language Resources Association. https://aclanthology.org/2020.osact-1.2

Antoun, W., Baly, F., & Hajj, H. (2021). AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop* (pp. 196–207). Association for Computational Linguistics. https://aclanthology.org/2021.wanlp-1.21

Gemma Team. (2025). *Gemma 3 technical report*. arXiv Preprint arXiv:2503.19786. https://arxiv.org/abs/2503.19786

Hasan, M. A., Das, S., Anjum, A., Alam, F., Anjum, A., Sarker, A., & Noori, S. R. H. (2024). Zero- and few-shot prompting with LLMs: A comparative study with fine-tuned models for Bangla sentiment analysis. In *Proceedings of LREC-COLING 2024* (pp. 17808–17818).

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., & Habash, N. (2021). The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop* (pp. 92–104). Association for Computational Linguistics. https://aclanthology.org/2021.wanlp-1.10

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Le Scao, T., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., & Raffel, C. (2022). *Crosslingual generalization through multitask finetuning*. arXiv Preprint arXiv:2211.01786. https://arxiv.org/abs/2211.01786

Nabil, M., Aly, M., & Atiya, A. (2015). ASTD: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2515–2519). Association for Computational Linguistics. https://aclanthology.org/D15-1299

Polock, S. I. S. (2025). *Sentiment analysis for low-resource languages using large language models with zero-shot, N-shot prompting, and LoRA fine-tuning* [Master's thesis, SIS Polock]. *https://oulurepo.oulu.fi/handle/10024/56755*

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., ... Qiu, Z. (2024). *Qwen2.5 technical report*. arXiv Preprint arXiv:2412.15115. https://arxiv.org/abs/2412.15115

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., ... Qiu, Z. (2025). *Qwen3 technical report*. arXiv Preprint arXiv:2505.09388. https://arxiv.org/abs/2505.09388

Zouidine, M., & Khalil, M. (2025). Large language models for Arabic sentiment analysis and machine translation. *Engineering, Technology & Applied Science Research, 15*(2), 20737–20742.