ADOR: Dataset for Arabic Dialects in Hotel Reviews: A Human Benchmark for Sentiment Analysis

Maram Alharbi[1,2], **Saad Ezzini**[3], **Tharindu Ranasinghe**[1]
Hansi Hettiarachchi[1] **and Ruslan Mitkov**[1]
[1]School of Computing and Communications, Lancaster University, UK
[2]Jazan University, Saudi Arabia
[3]King Fahd University of Petroleum and Minerals, Saudi Arabia
m.i.alharbi@lancaster.ac.uk

Abstract

Arabic machine translation remains a fundamentally challenging task, primarily due to the lack of comprehensive annotated resources. This study evaluates the performance of Meta's NLLB-200 model in translating Modern Standard Arabic (MSA) into three regional dialects: Saudi, Maghribi, and Egyptian Arabic using a manually curated dataset of hotel reviews. We applied a multi-criteria human annotation framework to assess translation correctness, dialect accuracy, and sentiment and aspect preservation. Our analysis reveals significant variation in translation quality across dialects. While sentiment and aspect preservation were generally high, dialect accuracy and overall translation fidelity were inconsistent. For Saudi Arabic, over 95% of translations required human correction, highlighting systemic issues. Maghribi outputs demonstrated better dialectal authenticity, while Egyptian translations achieved the highest reliability with the lowest correction rate and fewest multi-criteria failures. These results underscore the limitations of current multilingual models in handling informal Arabic varieties and highlight the importance of dialect-sensitive evaluation.

1 Introduction

Arabic is spoken by hundreds of millions across more than twenty countries, yet it remains significantly underrepresented in natural language processing (NLP) (Darwish et al., 2021; Premasiri et al., 2022). This is particularly acute for Arabic dialects, which diverge from Modern Standard Arabic (MSA) in terms of morphology, syntax, phonology, and vocabulary (Shoufan and Alameri, 2015). Dialects lack orthographic standardisation, exhibit wide regional variation, and are primarily used in informal and spoken contexts (El-Haj et al., 2024). As a result, NLP systems trained predominantly on MSA often perform poorly on dialectal data, limiting their effectiveness in real-world applications (Almansor and Al-Ani, 2017).

Machine translation (MT) of Arabic reflects these challenges. While MSA serves as the formal written standard, dialects are the primary medium of everyday communication across the Arab world. Their structural and lexical variation, combined with the absence of standardised norms, complicates the development of MT systems capable of handling the full spectrum of Arabic varieties (Zouidine and Khalil, 2025). Recent advancements in multilingual MT, such as Meta's NLLB-200 model, which incorporates FLORES-200 language codes, have extended support to low-resource languages, including Arabic dialects (Costa-jussà et al., 2022). Building on this, our study evaluates NLLB-200's performance in the reverse translation direction: from MSA into three major dialects; Saudi, Maghribi, and Egyptian. We introduce ADOR, (**A**rabic **D**ialects for H**o**tel **R**eviews) manually annotated dataset. ADOR assesses translation quality across four key dimensions: semantic correctness, dialect authenticity, sentiment preservation, and aspect category alignment. Using a structured human annotation protocol and error taxonomy, we offer both quantitative and qualitative insights into the capabilities and limitations of current MT systems.

The remainder of this paper is structured as follows: Section 2 reviews related work; Section 3 outlines the NLLB-200 architecture; Section 4 describes the dataset and preprocessing steps; Section 5 details the annotation framework; Section 6 presents the evaluation results; and Section 7 concludes with directions for future work.

2 Related Work

Meta's NLLB-200 model (Costa-jussà et al., 2022) supports over 200 languages and includes Arabic

Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, pages 187–191, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-100-4-017

dialects via FLORES-200 codes. However, its performance in dialectal settings remains limited. Atwany et al. (2024) evaluated NLLB-200 on dialect-to-MSA translation for divergent varieties including Gulf, Egyptian, Levantine, Iraqi, and Maghrebi, highlighting the inaccuracy of treating dialects as standard source languages due to their complexity.

Mousi et al. (2025), while not focused on translation, benchmarked performance across dialects and highlighted substantial disparities, reinforcing the need for dialect-aware evaluation protocols. A complementary perspective is offered by Yakhni and Chehab (2025), who studied Lebanese dialect-to-English translation. Their findings showed that NLLB ability to preserve cultural nuance in informal, idiomatic content is limited.

Finally, Boughorbel et al. (2024) addressed English-to-Arabic translation by translating the TinyStories dataset using NLLB-3B. They found that relying solely on MT introduced linguistic and cultural noise, and showed that further pre-training on a small corpus of native-generated Arabic stories improved output quality.

Collectively, these studies demonstrate that existing MT systems struggle with dialectal Arabic across multiple translation directions. They also emphasise the need for native speaker evaluation, cultural grounding, and dialect-specific benchmarks. Our work builds on these insights by evaluating NLLB-200 translations from MSA into three arabic dialects using a structured human-annotated framework.

3 Meta's No Language Left Behind (NLLB-200)

The No Language Left Behind (NLLB-200) model (Costa-jussà et al., 2022), developed by Meta, is a multilingual encoder–decoder transformer architecture designed to improve machine translation (MT) for low-resource languages. It supports direct translation between over 200 languages using FLORES-200 language codes, including several Arabic dialects.

NLLB-200 uses a unified encoder to process source text and a decoder that generates output conditioned on the target language or dialect. In this study, we use the model to translate from MSA into three Arabic dialects: Saudi, Maghribi and Egyptian. These dialects are explicitly supported within the model's language inventory, enabling direct generation without intermediate normalisation to MSA.

4 Data

This study uses the Arabic hotel reviews dataset from SemEval 2016 Task 5 on Aspect-Based Sentiment Analysis (ABSA) (Pontiki et al., 2016). The original dataset comprises over 10,000 sentences written in Modern Standard Arabic (MSA), each annotated with one or more aspect terms, sentiment polarities (positive, negative, neutral), and aspect categories. To ensure consistency and relevance for downstream translation and manual annotation, the dataset underwent several preprocessing steps.

Sentence Deduplication Duplicate entries were removed by grouping reviews with identical sentence text. For each unique sentence, all associated sentiment, aspect target, and category annotations were aggregated to retain the full range of opinions tied to that sentence.

Text Cleaning The text was normalised using the Ruqya library by removing special characters, hashtags, and diacritics (tashkīl). This step ensured uniform formatting and reduced noise, making the data more suitable for input into translation models.

Length Filtering Sentences with fewer than six words were excluded, as they typically lacked the contextual richness necessary for meaningful translation and sentiment analysis.

Polarity Consolidation Since a single sentence could be annotated with multiple aspect-level polarities, a rule-based approach was applied to assign one consolidated sentiment label:

- Sentences containing both positive and negative polarities were labelled as neutral.
- If a neutral polarity co-occurred with either positive or negative, the non-neutral polarity was retained.
- If only one polarity was present, it was used as the sentence-level label.

All consolidated sentiment labels were then manually reviewed to ensure correctness and internal consistency.

Following these preprocessing steps, the resulting dataset consisted of 538 sentences, balanced across sentiment classes: 200 positive, 200 negative, and 138 neutral.

5 Annotation Framework

Following the dialectal translation process, each sentence was manually evaluated by a native speaker of the corresponding dialect. The purpose of the annotation task was to assess the quality of the machine-generated translations across multiple linguistic and semantic dimensions.

Each annotator received a structured annotation template in CSV format containing the original MSA sentence, the machine-translated dialectal sentence, the sentiment label, and the associated aspect categories. One native speaker was assigned per dialect to ensure linguistic authenticity. Annotations were conducted independently for each dialect.

Annotators were instructed to assess each translation according to six criteria:

1. *Translation Correctness*: Does the translation accurately convey the meaning of the original sentence?

2. *Dialect Accuracy*: Is the sentence rendered in the appropriate dialect?

3. *Sentiment Preservation*: Is the original sentiment polarity maintained in the translation?

4. *Target Preservation*: Is the aspect or subject of the sentence correctly preserved?

5. *Corrected Sentence*: If needed, a revised version of the translated sentence.

6. *Target Correctness*: A corrected version of the aspect/target if it was omitted, distorted, or unclear.

The first four criteria were assessed using binary labels (Yes/No) to reduce subjectivity and enforce consistency. The final two were free-text fields used only when corrections were necessary.

All annotators received a detailed guideline document outlining each evaluation criterion. For clarity, definitions were provided alongside examples of both accurate and flawed translations, helping annotators distinguish between acceptable variation and critical errors. The guidelines also included instructions for identifying mismatches, particularly in sentiment and aspect categories, emphasising the importance of accurately preserving key content from the original MSA sentence. Annotators were familiarised with common translation error patterns, such as literal translations of idioms, the use of formal MSA constructions in dialectal output, and inappropriate lexical choices. In such cases, they were expected to revise translations to align with dialect norms while maintaining the intended meaning. Finally, procedures were outlined for handling ambiguous or incomplete translations, encouraging annotators to flag unclear outputs and consult the MSA source sentence before making corrections. Regular check-ins ensured consistency across dialects and allowed for immediate resolution of ambiguities. Upon completion, all annotation files were manually reviewed.

Although each dialect was annotated by a single native speaker, the structured process, guided instructions, and direct oversight helped ensure a high level of annotation reliability.

6 Evaluation Results and Discussion

The evaluation is based on a structured annotation framework designed to assess semantic correctness, dialectal fidelity, and sentiment preservation. It provides both quantitative metrics and qualitative insights into the limitations of NLLB-200 in translating MSA into regional Arabic dialects.

6.1 Overview of Annotation Outcomes

Table 1 summarises annotator judgments across four binary evaluation criteria. While all three dialects show high sentiment and aspect category preservation scores, there is a clear disparity in translation correctness and dialect accuracy. Egyptian translations were rated highest in both criteria, 84.7% correctness and 76.6% dialect accuracy, indicating stronger adaptation by NLLB-200 for Egyptian Arabic. Maghribi follows with 77.6% correctness and 44.4% accuracy, while Saudi trails with the lowest scores. These results suggest that NLLB-200 is more effective at generating fluent and regionally appropriate outputs for Maghribi than for Saudi Arabic.

6.2 Sentiment Agreement

Given that each sentence in the dataset was pre-annotated with sentiment labels prior to translation, the Sentiment Preservation score can be interpreted as a measure of annotator agreement. Specifically, a "Yes" label indicates that the human reviewer agreed that the sentiment expressed in the translated sentence matches that of the original MSA input. Agreement rates were high for all the three

Criterion	Saudi		Maghribi		Egyptian	
	Yes (%)	No (%)	Yes (%)	No (%)	Yes (%)	No (%)
Translation Correctness	68.4	31.6	77.6	22.4	84.7	15.3
Dialect Accuracy	5.2	94.8	44.4	55.6	76.6	23.4
Sentiment Preservation	98.9	1.1	90.9	9.1	94.6	5.4
Target Preservation	91.4	8.6	90.5	9.5	92.3	7.7

Table 1: Binary annotation outcomes across key criteria.Values represent the proportion of "Yes" and "No" judgments for each dialect.

Dialect	Avg Criteria Score	Correction Rate (%)	$\geq$ 2 Failures (%)	$\geq$ 3 Failures (%)
Saudi	0.684	95.17	26.21	5.02
Maghribi	0.776	55.58	18.22	11.34
Egyptian	0.867	31.04	13.01	2.97

Table 2: Summary of translation evaluation statistics across dialects. Criteria score reflects average binary ratings for translation correctness, dialect accuracy, sentiment preservation, and target preservation.

dialects; 98.88% for Saudi, 94.60 for Egyptian and 90.71% for Maghribi. That affirm the reliability of the original sentiment labels and the annotators' consistency, lending additional credibility to the overall annotation process.

6.3 Error Distribution and Correction Analysis

Table 2 presents aggregated error metrics, including average criteria scores, correction rates, and the frequency of compound evaluation failures. Among the three dialects, Egyptian outputs required the fewest corrections (31.04%) and exhibited the lowest rate of multi-criteria failures, indicating greater reliability and better alignment with dialectal norms. In contrast, Saudi Arabic translations had a significantly higher correction rate (95.17%) compared to Maghribi (55.58%), reinforcing earlier observations that Saudi outputs demanded more extensive post-editing. This finding is consistent with the low dialect accuracy score and highlights systemic challenges in the model's ability to generate fluent and authentic Saudi vernacular.

Although Maghribi translations were more accurate on average, they exhibited a slightly higher rate of cases with three or more simultaneous evaluation failures (11.34%), which indicate that, despite closer alignment with dialect norms, certain Maghribi translations required broader structural revisions to address subtler fluency or coherence issues.

6.4 Interpretation

These results collectively illustrate the need for human-centered, dialect-sensitive evaluation frameworks in Arabic MT. While NLLB-200 demonstrates promising performance on some dimensions, it struggles with dialectal fluency and semantic fidelity.

The data also highlights the importance of manual correction and targeted annotation, as many outputs superficially appear fluent but fail under semantic or dialectal review. These insights underscore the limitations of automatic metrics in low-resource dialect contexts and support the value of qualitative human validation as part of the evaluation process.

7 Conclusion and Future Work

This study presented a structured evaluation of NLLB-200's ability to translate MSA into three major Arabic dialects: Saudi, Maghribi, and Egyptian. Using **ADOR**, a manually annotated benchmark grounded in linguistic and semantic criteria, we identified systematic translation errors and highlighted performance variability across dialects. The Findings revealed that while NLLB-200 achieves high rates of sentiment and aspect preservation, its performance on dialect accuracy and translation correctness remains inconsistent. Saudi Arabic translations exhibited a high dependency on human correction, pointing to the model's difficulty in handling dialects that are lexically and syntactically distant from MSA. In contrast, Maghribi translations demonstrated better dialect fidelity and required fewer revisions. Notably, Egyptian outputs achieved the highest overall reliability, with the lowest correction rate and the fewest multi-criteria failures, suggesting stronger alignment between NLLB-200 output and Egyptian dialect norms.

Building on this work, future efforts will expand the dialectal coverage of the dataset to include ad-

ditional varieties such as Levantine and Yemeni dialects. To enhance annotation reliability, multiple annotators will be recruited per dialect, enabling inter-annotator agreement analysis and reducing subjectivity in evaluation.

References

Ebtesam H. Almansor and Ahmed Al-Ani. 2017. Translating dialectal arabic as low resource language using word embedding. In *International Conference Recent Advances in Natural Language Processing, RANLP*, volume 2017-September, pages 52–57. Incoma Ltd.

Hanin Atwany, Nour Rabih, Ibrahim Mohammed, Abdul Waheed, and Bhiksha Raj. 2024. OSACT 2024 task 2: Arabic dialect to MSA translation. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 98–103, Torino, Italia. ELRA and ICCL.

Sabri Boughorbel, MD Rizwan Parvez, and Majd Hawasly. 2024. Improving language models trained on translated data with continual pre-training and dictionary learning analysis.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. https://ai.facebook.com/research/no-language-left-behind/. Meta AI.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab world. *Commun. ACM*, 64(4):72–81.

Mo El-Haj, Sultan Almujaiwel, Damith Premasiri, Tharindu Ranasinghe, and Ruslan Mitkov. 2024. DARES: Dataset for Arabic readability estimation of school materials. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 103–113, Torino, Italia. ELRA and ICCL.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Damith Premasiri, Tharindu Ranasinghe, Wajdi Zaghouani, and Ruslan Mitkov. 2022. DTW at qur'an QA 2022: Utilising transfer learning with transformers for question answering in a low-resource domain. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 88–95, Marseille, France. European Language Resources Association.

Abdulhadi Shoufan and Sumaya Alameri. 2015. Natural language processing for dialectical Arabic: A survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 36–48, Beijing, China. Association for Computational Linguistics.

Silvana Yakhni and Ali Chehab. 2025. Can LLMs translate cultural nuance in dialects? a case study on Lebanese Arabic. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 114–135, Abu Dhabi, UAE. Association for Computational Linguistics.

Mohamed Zouidine and Mohammed Khalil. 2025. Large language models for arabic sentiment analysis and machine translation. *Engineering, Technology and Applied Science Research*, 15:20737–20742.