Enhancing Arabic Dialectal Sentiment Analysis through Advanced Data Augmentation Techniques

Md. Rafiul Biswas
Hamad Bin Khalifa University
Doha, Qatar
mbiswas@hbku.edu.qa

Wajdi Zaghouani
Northwestern University in Qatar
Education City, Doha, Qatar
wajdi.zaghouani@northwestern.edu

Abstract

This work addresses the challenge of Arabic sentiment analysis in the hospitality domain in all dialects by using data augmentation techniques. We created a pipeline with three simple techniques: context-based paraphrasing, pattern-based sentence generation, and domain-specific word replacement. Our method preserves the original dialect features, meanings, and key classification details while adding diversity to the training data. It also includes automatic fallback between methods to handle challenges effectively. We used the Fanar API for dialectal data augmentation in the hospitality domain. The AraBERT-Large-v02 model was fine-tuned on original and augmented data, showing improved performance. This study helps solve the problem of limited dialect data in Arabic NLP and offers an effective framework that is useful for other Arabic text analysis tasks.

1 Introduction

In the Arabic-speaking world, most of the content on social networks and online reviews is written in regional dialects rather than Modern Standard Arabic (MSA). Arabic dialectal sentiment analysis aims to identify emotions (positive, negative, neutral) in social media texts written in regional Arabic dialects. These dialects, such as Egyptian, Saudi, and Moroccan Arabic, exhibit significant variations in vocabulary, grammar, and syntax, making sentiment analysis a complex task due to the lack of standard spelling and limited NLP tools (Salloum, 2021; Baly et al., 2017).

In the hospitality industry, which includes hotels, restaurants and tourism services, sentiment analysis is crucial to analyze customer feedback to improve service quality and customer satisfaction (Musanovic et al., 2021; Kim et al., 2022). Analyzing sentiments from reviews written in different Arabic dialects can provide valuable insights for companies operating in Arabic-speaking countries (Al-Thubaity et al., 2018). Unlike formal Modern Standard Arabic (MSA), these dialects are used in daily chats and reviews, but they do not follow strict rules, making it hard for computers to understand them (El-Naggar et al., 2017). In addition, there are not many tools or datasets built specifically for these dialects. Dataset for context-specific dialect like hospitality is essential to build machine learning tool. The release of a dialectal sentiment dataset and the organization of a shared task in the Hospitality Domain by the Ahasis organizing team is a pivotal contribution to the Arabic NLP community (Alharbi et al., 2025a,b). It will encourage the research community to develop NLP tool in specific domain.

2 Dataset Description

Ahasis shared task (Alharbi et al., 2025a) includes Arabic sentiment analysis texts from two dialects: Moroccan Darija and Saudi Arabic. The training set contains 860 balanced sentences, with 430 samples from each dialect. Both subsets share identical sentiment distributions: 39.07% negative, 25.12% neutral, and 35.81% positive. This balance helps prevent the model from favoring any dialect or sentiment class. The test set maintains this dialectal balance with 108 samples each, ensuring robust evaluation. See Table 1 for more details.

This dataset offers unique advantages for Arabic NLP sentiment analysis by covering dialectal diversity beyond standard Arabic, ensuring consistent sentiment annotations, and representing all sentiment classes adequately. Its dual-dialect structure supports both dialect-specific, cross-dialect experiments, industry applications like customer feedback analysis, review summarization, and reputation management.

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025, pages 24–28, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-004

However, the dataset has some limitations. The relatively small size (860 training samples) may limit the model's ability to generalize to more complex patterns. It is insufficient size for training large-scale models, particularly deep neural networks, without overfitting risks. The absence of contextual topic metadata also tied to specific topics may be missed during modeling.

Table 1: Train Dataset Distribution

Dialect	Negative	Neutral	Positive	Total
Darija	168	108	154	430
Saudi	168	108	154	430

The aim of this shared task is to correctly predict the sentiment based on dialect. The test set provides dialects for each text to correctly predict sentiment. Research suggests that hybrid methods, which combine word lists with machine learning, often perform better than traditional approaches. Recent advances, such as transformer models, show promising results, but require more dialect-specific resources.

3 System Description

Transformer based model requires a good amount of data for training the model. However, the task lacks of original data. Data augmentation plays a crucial role in enhancing the performance and generalizability of text classification models, especially in low-resource scenarios (Shah et al., 2024). Our system addresses these challenges through a multi-faceted approach to text augmentation. We explained the system in the following sections.

3.1 Dataset Preprocessing

Data preprocessing is essential to preserve the semantic integrity of dialectal Arabic for optimizing the with state-of-the-art language models (e.g., AraBERT). The preprocessing pipeline involved several steps to clean and standardize the textual data. HTML markup was removed to eliminate irrelevant formatting tags. URLs, email addresses, and social media mentions were replaced with special tokens, ensuring a consistent structure across diverse inputs. To normalize the text, Arabic diacritics (tashkeel) and elongation characters (tatweel) were stripped. Whitespace was optionally inserted around punctuation and special characters to facilitate better tokenization. Emojis were retained to preserve sentiment-related cues, and extra spaces were removed to maintain clean and consistent formatting throughout the dataset.

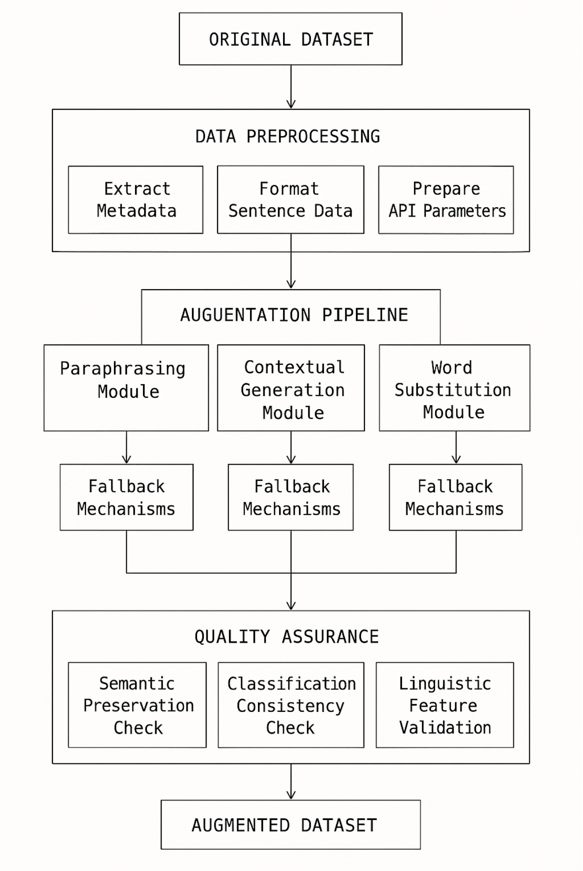

Figure 1: Data Augmentation

3.2 Data Augmentation

We have proposed an innovative approach for data augmentation using Large Language Model (LLM) to mitigate low resource challenges. Figure 1 describes the data augmentation scenario. The system is designed to preserve critical linguistic features including dialectal characteristics, domain context, and classification-relevant patterns. We performed data augmentation in three ways (Text Paraphrasing using LLM, Contextual Text Generation, and Word Substitution). Finally, we combined all the augmented data.

Text Paraphrasing using LLM: We integrated FANAR API (Team et al., 2025) to paraphrase the raw data. The paraphrasing module generates alternative versions of existing Arabic sentences while preserving their essential characteristics: preserving dialectal features, maintaining of domain context (tourism, hotel and reservations), and retenting tonal characteristics. We have created instructed prompt to instruct the FANAR model. We provided

10 example with system prompt, user prompt, text, dialect, sentiment. Finally, we asked the FANAR API to perform text paraphrasing.

Contextual Text Generation: We avoided sensitive content by focusing on language structure (e.g., hate speech, vulgarity, offensive language). We analyzed example sentences to find patterns, and then created new sentences from these patterns. If standard generation faces issues, it rebuilds sentences directly from identified patterns. This module creates new sentences matching patterns from the original dataset. It increases dataset variety while keeping original language features, generating text based on topic (hospitality), matching dialect (Saudi, Darija), and keeping important features for classification overall pattern.

Word Substitution: The module slightly modifies sentences without changing important meanings that affect classification. We replaced key words with suitable synonyms, and kept original sentence meaning and structure. We made minimal changes to avoid classification errors. We used synonyms related to the specific topic (Hospitality) ensuring sentence meaning stays consistent.

Fallback Mechanism: In all three cases mentioned above, we trained and evaluated model performance on the validation dataset. Our key innovation lies in the implementation of comprehensive fallback mechanisms (Liu et al., 2023), which ensure robustness when primary augmentation strategies face challenges. A fallback mechanism automatically activates when the preferred option is unavailable or fails. To maintain data augmentation quality, we assessed semantic preservation, classification consistency (i.e., label distribution across positive, negative, and neutral), and linguistic integrity through human evaluation. We validated the augmentation outputs manually and applied fallback procedures when necessary to improve outcomes. This process enabled us to achieve high-quality data augmentation for model training. Table 2 summarizes the total number of augmented samples.

Table 2: Data Augmentation

Dialect	Negative	Positive	Neutral
Darija	647	613	421
Saudi	617	587	394

4 Experimental Setup

Our sentiment analysis leveraged with AraBERT-Large-v02 ("aubmindlab/bert-large-arabertv02") (Antoun et al., 2020) and CAMeLBERT mix(Inoue et al., 2021), optimized for dialect-specific Arabic preprocessing. We stratified data sampling into (80% training, 20% validation) to retain sentiment class proportions. We employed employed AdamW (learning rate: 2e-5, epsilon: 1e-8) with linear scheduling (no warmup). The experiment continued for 10 epochs, batch size 16, and maximum token length 128, alongside gradient clipping (threshold: 1.0) to stabilize training.

We utilized Google Colab Pro, NVIDIA A100 GPUs, enabling efficient model optimization with accelerated processing capabilities suitable for transformer-based architectures like AraBERT-Large-v02. This infrastructure significantly reduced training duration and facilitated rapid experimentation and hyperparameter tuning.

5 Results

Figure 2 illustrates the training and validation loss alongside validation performance metrics over 10 epochs. Training loss consistently decreased and validation loss reached stability around epoch 6, reflecting a balanced training without notable overfitting. Our fused model demonstrated consistent performance across both validation and test datasets for Arabic dialectal sentiment classification. The results (see Table 3) show a slight decrease in performance metrics when moving from validation to test data. The slight performance decrease on the test set within expected ranges and indicates good generalization capabilities.

Table 3: Validation vs. Test Performance Comparison

Metric	Validation	Test
Accuracy	0.780	0.750
Precision	0.780	0.750
Recall	0.780	0.750
F1-Score	0.780	0.750

The competition winner achieved (F1-score 0.81) where we achieved (F1-score 0.75) showing 0.6 back to winner. This indicates that while our model is competitive, further improvements in data augmentation, model tuning, or handling of dialectal variations may help bridge this gap.

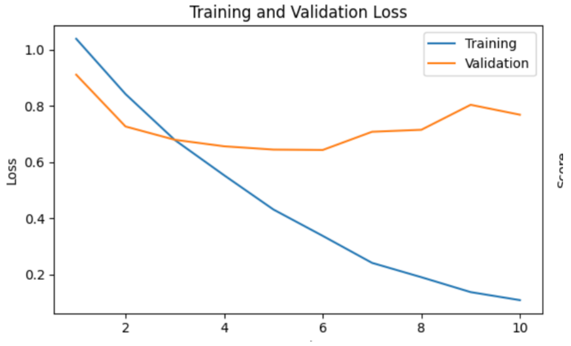

Figure 2: Training loss Vs Validation Loss

6 Conclusion

In this study, we explore the analysis of Arabic dialectal sentiment using AraBERT-Large-v02, supported by domain-specific preprocessing and controlled data augmentation. Our results demonstrate that careful handling of dialectal features and balanced data splitting are crucial to achieving reliable sentiment classification. The model's ability to maintain consistent performance across Darija and Saudi dialects suggests that it could be deployed in real-world applications requiring sentiment monitoring across diverse Arabic dialectal contexts. Future work should expand to include additional Arabic dialects and explore multitask learning approaches, advanced augmentation, or domain adaptation techniques to further improve classification performance.

Acknowledgments

This study was supported by the grant NPRP14C-0916-210015, awarded by the Qatar Research, Development and Innovation Council (QRDI).

References

Abdulmohsen Al-Thubaity, Qubayl Alqahtani, and Abdulaziz Aljandal. 2018. Sentiment lexicon for sentiment analysis of saudi dialect tweets. *Procedia computer science*, 142:301–307.

Maram Alharbi, Salmane Chafik, Saad Ezzini, Ruslan Mitkov, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2025a. Ahasis: Shared task on sentiment analysis for arabic dialects. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Maram Alharbi, Saad Ezzini, Hansi Hettiarachchi, Tharindu Ranasinghe, and Ruslan Mitkov. 2025b. Evaluating large language models on arabic dialect sentiment analysis. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Ramy Baly, Georges El-Khoury, Rawan Moukalled, Rita Aoun, Hazem Hajj, Khaled Bashir Shaban, and Wassim El-Hajj. 2017. Comparative evaluation of sentiment analysis methods across arabic dialects. *Procedia Computer Science*, 117:266–273.

Nadine El-Naggar, Yasser El-Sonbaty, and Mohamad Abou El-Nasr. 2017. Sentiment analysis of modern standard arabic and egyptian dialectal arabic tweets. In *2017 computing conference*, pages 880–887. IEEE.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 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*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Taekyung Kim, Hwirim Jo, Yerin Yhee, and Chulmo Koo. 2022. Robots, artificial intelligence, and service automation (raisa) in hospitality: sentiment analysis of youtube streaming data. *Electronic Markets*, 32(1):259–275.

Ye Liu, Semih Yavuz, Rui Meng, Meghana Moorthy, Shafiq Joty, Caiming Xiong, and Yingbo Zhou. 2023. Modeling uncertainty and using post-fusion as fallback improves retrieval augmented generation with llms. *arXiv preprint arXiv:2308.12574*.

Jelena Musanovic, Raffaella Folgieri, Maja Gregoric, et al. 2021. Sentiment analysis and multimodal approach applied to social media content in hospitality industry. *TOURISM IN SOUTH EAST EUROPE...*, 6:533–544.

SA Salloum. 2021. Sentiment analysis in dialectal arabic: a systematic review. *Advanced machine learning technologies and applications: proceedings of AMLTA*.

Uzair Shah, Md Rafiul Biswas, Marco Agus, Mowafa Househ, and Wajdi Zaghouani. 2024. Mememind at araieval shared task: generative augmentation and feature fusion for multimodal propaganda detection in arabic memes through advanced language and vision models. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 467–472.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. Fanar: An arabic-centric multimodal generative ai platform.