iWAN-NLP at AHaSIS 2025: A Stacked Ensemble of Arabic Transformers for Sentiment Analysis on Arabic Dialects in the Hospitality Domain

Hend S. Al-Khalifa
iWAN Research Group, College of Computer and Information Sciences
King Saud University, Riyadh, Saudi Arabia
hendk@ksu.edu.sa

This paper details the iWAN-NLP system developed for participation in the AHaSIS 2025 shared task, "Sentiment Analysis on Arabic Dialects in the Hospitality Domain: A Multi-Dialect Benchmark." Our approach leverages a multi-model ensemble strategy, combining the strengths of MARBERTv2, SaudiBERT, and DarijaBERT. These pre-trained Arabic language models were fine-tuned for sentiment classification using a 5-fold stratified cross-validation methodology. The final predictions on the test set were derived by averaging the logits produced by each model across all folds and then averaging these combined logits across the three models. This system achieved a macro F1-score of 0.8055 (~0.81) on the official evaluation dataset and a cross-validated macro F1-score of 0.8513 (accuracy 0.8628) on the training set. Our findings highlight the effectiveness of ensembling regionally adapted models and robust cross-validation for Arabic sentiment analysis in the hospitality domain, ultimately securing first place in the AHaSIS 2025 shared task.

1 Introduction

The proliferation of user-generated content, particularly in the hospitality sector through reviews, provides a rich source of opinions and sentiments. For the Arabic language, this content is characterized by a complex interplay of Modern Standard Arabic (MSA) and numerous regional dialects. These dialects often exhibit substantial lexical, syntactic, and morphological divergence, posing considerable hurdles for Natural Language Processing (NLP) systems. The AHaSIS 2025 shared task, "Sentiment Analysis on Arabic Dialects in the Hospitality Domain: A Multi-Dialect Benchmark," aims to foster research in this area by providing a benchmark for evaluating systems on Arabic sentiment analysis within the context of hotel reviews, which can span MSA and various dialects.

Effective sentiment analysis of Arabic hotel reviews offers significant value for businesses in understanding customer satisfaction and for travelers in making informed decisions. However, the linguistic diversity, coupled with the nuances of sentiment expression (e.g., sarcasm, implicit feedback), makes this a challenging task. Transformer-based language models such as BERT (Devlin et al., 2019), have shown remarkable success in NLP by learning rich contextual representations, leading to the development of several Arabic-specific models.

In this paper, we present the iWAN-NLP system. Our system architecture is built upon an ensemble of multiple transformer models: MARBERTv2, chosen for its training on diverse Arabic social media; SaudiBERT, for its specialization in the Saudi dialect; and DarijaBERT, for its focus on North African Darija. Each base model was fine-tuned using 5-fold stratified cross-validation. Key training enhancements included a custom Hugging Face Trainer, label smoothing (0.1) via nn.CrossEntropyLoss, early stopping, FP16 mixed-precision training, and learning rate warmup. The final system predictions were generated by averaging model logits across folds and then across models. This approach yielded a macro F1-score of 81.0% on the

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025,
pages 7–13, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-002

official test set and a cross-validated macro F1-score of 85% on the training set. Notably, this performance earned our iWAN-NLP system the first-place ranking in the AHaSIS 2025 (Alharbi, Ezzini, et al., 2025) shared task on Sentiment Analysis on Arabic Dialects in the Hospitality Domain.

The remainder of this paper is organized as follows: Section 2 discusses related work. Section 3 describes the dataset. Section 4 details our methodology, including preprocessing, models, training, and ensemble strategy. Section 5 presents results. Section 6 offers an error analysis. Section 7 concludes and suggests future work.

2 Related Work

Sentiment analysis in Arabic has garnered significant attention in the NLP community. Early efforts often relied on lexicon-based approaches (Abdul-Mageed et al., 2012) or traditional machine learning algorithms like Support Vector Machines (SVMs) and Naive Bayes, typically using n-gram features or bag-of-words representations (El-Halees, 2011). While these methods provided initial breakthroughs, their performance was often limited by the morphological richness of Arabic and the challenges posed by dialectal variations.

The advent of deep learning, particularly Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), brought about improvements in capturing sequential and local features in text. However, the introduction of transformer-based models (Vaswani et al., 2017), especially BERT (Devlin et al., 2019), revolutionized the field. These models, pre-trained on massive text corpora, learn powerful contextual embeddings that have proven highly effective for a wide array of downstream tasks.

Several BERT-based models have been specifically developed for the Arabic language. AraBERT (Antoun et al., 2020) was one of the pioneering efforts, demonstrating strong performance on various Arabic NLP benchmarks. MARBERT (Abdul-Mageed et al., 2021), and its successor MARBERTv2, were trained on a significantly larger and more diverse dataset, including a substantial amount of Arabic social media text containing dialectal content, making them particularly well-suited for tasks like the AHaSIS challenge. Recognizing the limitations of general Arabic models when faced with specific dialects, researchers have also developed models tailored to regional variations. For instance, SaudiBERT (Qarah, 2024) was pre-trained exclusively on a large corpus of Saudi dialectal text, aiming to capture its unique linguistic characteristics. Similarly, DarijaBERT (Gaanoun et al., 2024) represents a significant step in NLP for the Moroccan Arabic dialect "Darija", providing the first set of BERT models specifically for this dialect.

Ensemble methods have a long history of improving machine learning model performance by combining the predictions of multiple individual learners. In NLP, ensembling has been successfully applied to various tasks, including sentiment analysis (Salah et al., 2019). While various ensemble techniques exist, including simple averaging/voting and more complex methods like stacking (Wolpert, 1992), the core idea is to leverage the diversity of multiple models. (Al Shamsi & Abdallah, 2023) explored ensemble techniques in the context of Arabic sentiment analysis, highlighting their potential for enhanced robustness.

To further optimize the training of large transformer models, several techniques have become standard practice. Label smoothing (Müller et al., 2019) is a regularization technique that prevents the model from becoming too confident in its predictions,

which can improve generalization. Early stopping is widely used to prevent overfitting by monitoring performance on a validation set and halting training when performance ceases to improve. FP16 mixed-precision training (Micikevicius et al., 2018) allows for faster training and reduced memory footprint by using 16-bit floating-point numbers for certain operations, often without sacrificing performance. Learning rate warmup (Goyal et al., 2018), where the learning rate is gradually increased at the beginning of training, helps stabilize the training process, especially for large models and batch sizes.

Our work builds upon these advancements by combining multiple state-of-the-art Arabic transformer models, including those specialized for dialects, within an averaging ensemble framework, and by employing a suite of modern training enhancements to maximize performance.

3 Data

The AHaSIS 2025 shared task organizers provided a dataset consisting of Arabic hotel reviews. Each review was annotated with one of three sentiment labels: *positive, negative, or neutral*. The dataset was pre-split into official training and test sets. For our experiments, we strictly adhered to these provided splits and did not incorporate any external datasets or perform external data augmentation.

Labels (positive, negative, neutral) were mapped to numeric IDs for model processing. Tokenization was performed using each base model's associated pre-trained tokenizer. A key aspect of our data handling was the use of 5-fold Stratified K-Fold Cross-Validation on the training set. This approach ensures that each fold maintains a similar class distribution to the overall training set, providing a more reliable estimate of model performance during development and robust out-of-sample predictions from each fold for evaluating ensemble strategies.

4 Methodology

Our system for Arabic sentiment classification employs a multi-model ensemble approach, with each base model fine-tuned using a rigorous cross-validation strategy and enhanced training techniques.

4.1 Preprocessing

The preprocessing steps applied to the review data were minimal:

1. **Label Mapping:** The textual sentiment labels ('positive', 'negative', 'neutral') were mapped to numerical IDs (0, 1, 2).

2. **Tokenization:** Each review text was tokenized using the specific pre-trained tokenizer associated with the respective base model (MARBERTv2, SaudiBERT, DarijaBERT). This ensures that the input format matches what each model expects from its pre-training phase.

No further extensive preprocessing steps such as stopword removal, emoji normalization, or detailed punctuation cleaning were performed, relying on the inherent capabilities of the transformer models to process relatively raw text.

4.2 Base Models

We utilized three pre-trained Arabic transformer models from the HuggingFace Transformers library (Wolf et al., 2020) as our base learners:

- **MARBERTv2 (UBC-NLP/MARBERTv2)[1]:** Chosen for its training on diverse Arabic social media

[1] https://huggingface.co/UBC-NLP/MARBERTv2

text, making it suitable for general dialectal Arabic.

- **SaudiBERT (faisalq/SaudiBERT)[2]:** A model specialized in the Saudi dialect.
- **DarijaBERT (SI2M-Lab/DarijaBERT)[3]:** Designed for North African Arabic dialects, particularly Moroccan Darija.

Each model was employed with its corresponding **AutoModelForSequenceClassification** head for the sentiment classification task.

4.3 Training Strategy

Each of the three base models was fine-tuned independently on the AHaSIS training dataset using a 5-fold stratified cross-validation strategy. This means the training data was divided into five folds, and for each fold, a model was trained on four folds and validated on the held-out fold. This process was repeated for each of the three base models.

A custom Hugging Face Trainer class was utilized for the fine-tuning process. The key training parameters and enhancements were:

- **Optimizer:** AdamW optimizer.
- **Loss Function:** nn.CrossEntropyLoss with label smoothing applied at a factor of 0.1. Label smoothing helps to regularize the model and prevent overconfidence.
- **Learning Rate:** 2e-5, with a learning rate warmup schedule.
- **Batch Size:** 8 (for training, constrained by GPU memory).
- **Epochs:** Up to 6 epochs per fold, with early stopping based on validation F1-score to prevent overfitting and save the best model checkpoint.
- **FP16 Mixed-Precision Training:** Enabled to accelerate training and reduce memory consumption.
- **Hardware:** Experiments were conducted on Google Colab (single GPU).

This training regimen was repeated for each of the 5 folds for all three models (MARBERTv2, SaudiBERT, and DarijaBERT).

4.4 Ensemble Strategy

Our ensemble strategy focused on combining the predictions (logits) from the fine-tuned base models.

Validation Phase (within Cross-Validation): During the 5-fold cross-validation on the training set, for each validation fold, we obtained logits from each of the three models (MARBERTv2, SaudiBERT, DarijaBERT) trained on the other four folds. To evaluate an intermediate ensemble performance during development, a soft-voting ensemble was considered: the logits from the three models for the validation set samples were averaged, and the class with the highest average logit was chosen as the prediction. This allowed for monitoring the ensemble's potential during the cross-validation process.

Final Test Set Prediction: To generate the final predictions for the official AHaSIS test set, the following procedure was used:

1. **Per-Model Averaging Across Folds:** For each base model (MARBERTv2, SaudiBERT, DarijaBERT), the predictions (logits) on the test set were generated by each of the 5 fine-tuned instances of that model (one from each fold of the cross-validation). These 5 sets of logits for the test set were then averaged to get a single, more stable set of logits for each of the three base models.
2. **Cross-Model Averaging:** The averaged logits for the test set from MARBERTv2, SaudiBERT, and DarijaBERT were then further averaged together.
3. **Final Label Determination:** The final sentiment label for each test instance was determined by applying an argmax function to these final averaged logits.

[2] https://huggingface.co/faisalq/SaudiBERT

[3] https://huggingface.co/SI2M-Lab/DarijaBERT

The class corresponding to the highest logit value was chosen as the predicted sentiment.

This multi-stage averaging ensemble aims to combine the strengths of the diverse models and smooth out variations from individual training runs, leading to a more robust final prediction.

5 Results

Our IWAN-NLP system was evaluated on both the training data (via cross-validation) and the official AHaSIS 2025 test set. The primary evaluation metric was the macro F1-score.

Training Set Performance (5-Fold Cross-Validation): Across the 5-fold stratified cross-validation on the training set, our ensemble approach (evaluating out-of-sample predictions from each fold) yielded the following average performance:

- Cross-Validated Macro F1-score: 0.8513
- Cross-Validated Accuracy: 0.8628

These results on the training data indicated strong performance and good generalization before evaluating on the unseen test set.

Official Test Set Performance: Our final system, employing the described averaging ensemble strategy (averaging logits across folds per model, then averaging these across models for the test set), achieved the highest macro F1-score of 81%, which was the top-ranking performance among all participating systems in the shared task. This result confirms the effectiveness of our ensemble approach in leveraging the diverse strengths of the chosen Arabic language models and the robustness introduced by cross-validation and prediction averaging.

6 Error Analysis

Upon qualitative examination of the predictions made by our ensemble system, a notable pattern of confusion emerged, primarily between the neutral and negative sentiment classes. Several factors could contribute to this observation:

1. **Ambiguity of Neutrality**: Neutral sentiment itself can be inherently ambiguous. Reviews classified as neutral might contain subtle negative undertones or vice-versa, making it challenging for the models to draw a clear distinction, especially with the minimalist preprocessing approach.
2. **Sarcasm and Irony**: Social media text (and even formal reviews) can contain sarcasm and irony, where the literal meaning of words contradicts the intended sentiment. Our system did not explicitly model sarcasm detection, which is a complex NLP task in itself. Sarcastic reviews expressing negative sentiment in a seemingly positive or neutral way (or vice-versa) could easily be misclassified.
3. **Dialectal Nuances**: While we employed dialect-specific models, the interplay of different dialects within a single review or the presence of code-switching (mixing MSA with dialects, or dialects with English) could still pose challenges. Certain dialectal expressions might carry sentiment connotations that are not universally captured even by specialized models, leading to confusion, particularly with neutral or subtly negative statements.
4. **Implicit Sentiment**: Some reviews might express sentiment implicitly rather than through explicit sentiment-bearing words. For example, a factual statement about a service failure could imply negative sentiment without using overtly negative language. Detecting such implicit sentiment requires a deeper level of contextual understanding.

5. **Dataset Artifacts**: The nature of the annotation process or inherent biases within the dataset could also contribute to certain types of systematic errors. For instance, if borderline cases were predominantly labeled as neutral, the model might struggle to differentiate them from slightly negative instances. Furthermore, it was observed that some texts labeled as Darija might not be pure Darija or could be mixed with other dialects/MSA, potentially impacting the performance of the Darija-specific model on these instances and contributing to confusion.

Future work could aim to address these issues. Integrating a dedicated sarcasm detection module could be beneficial. More advanced preprocessing techniques, or alternatively, models even more robust to noisy text, might help. Furthermore, exploring multi-task learning frameworks, where the system is jointly trained to detect sentiment and other related linguistic phenomena (like sarcasm or emotion), could lead to improved performance and a better understanding of nuanced expressions.

7 Conclusion

In this paper, we described the iWAN-NLP system developed for the AHaSIS 2025 shared (Alharbi, Chafik, et al., 2025). Our system utilized an averaging ensemble of three pre-trained transformer models: MARBERTv2, SaudiBERT, and DarijaBERT. Each base model was fine-tuned using 5-fold stratified cross-validation with a custom Hugging Face Trainer, incorporating label smoothing (0.1), early stopping, FP16 mixed-precision training, and learning rate warmup. The final ensemble, derived by averaging logits across folds and models, achieved a macro F1-score of 81.0% on the official test set, and a cross-validated macro F1 of 0.8513 on the training set. This performance distinguished our system, leading to its recognition as the first-place submission in the AHaSIS 2025 shared task.

Our results demonstrate the advantages of leveraging a diverse set of language models, including those adapted to specific regional dialects, within a robust ensemble framework. The meticulous cross-validation and prediction averaging strategy proved effective in combining the strengths of these individual models and enhancing overall performance. The inclusion of modern training best practices further contributed to the stability and generalization capabilities of our system.

Future research directions could involve exploring more sophisticated ensemble techniques, investigating advanced data augmentation tailored for Arabic dialects, and incorporating explicit mechanisms for handling linguistic phenomena such as sarcasm and implicit sentiment. Our participation in AHaSIS 2025 highlights the potential of combining specialized NLP models and principled ensembling to tackle the complexities of sentiment analysis in the diverse Arabic linguistic landscape.

References

Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2021). ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 7088–7105). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.551

Abdul-Mageed, M., Kuebler, S., & Diab, M. (2012). SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social Media. In A. Balahur, A. Montoyo, P. M. Barco, & E. Boldrini (Eds.), *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis* (pp. 19–28). Association for Computational Linguistics. https://aclanthology.org/W12-3705/

Al Shamsi, A. A., & Abdallah, S. (2023). Ensemble Stacking Model for Sentiment Analysis of Emirati and Arabic Dialects. *Journal of King Saud University - Computer and Information Sciences*, *35*(8), 101691. https://doi.org/10.1016/j.jksuci.2023.101691

Alharbi, M., Chafik, S., Ezzini, S., Mitkov, R., Ranasinghe, T., & Hettiarachchi, H. (2025). AHaSIS: Shared Task on Sentiment Analysis for Arabic Dialects. *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Alharbi, M., Ezzini, S., Hettiarachchi, H., Ranasinghe, T., & Mitkov, R. (2025). Evaluating Large Language Models on Arabic Dialect Sentiment Analysis. *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based Model for Arabic Language Understanding. *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 9–15. https://www.aclweb.org/anthology/2020.osact-1.2

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

El-Halees, A. (2011, December 11). *Arabic Opinion Mining Using Combined Classification Approach*. https://www.semanticscholar.org/paper/Arabic-Opinion-Mining-Using-Combined-Classification-El-Halees/4bc8728992fcf5b26ae80286b9524ff115e7d329

Gaanoun, K., Naira, A. M., Allak, A., & Benelallam, I. (2024). DarijaBERT: A step forward in NLP for the written Moroccan dialect. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-023-00498-2

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., & He, K. (2018). *Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour* (arXiv:1706.02677). arXiv. https://doi.org/10.48550/arXiv.1706.02677

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., & Wu, H. (2018). *Mixed Precision Training* (arXiv:1710.03740). arXiv. https://doi.org/10.48550/arXiv.1710.03740

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help? *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper_files/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html

Qarah, F. (2024). *SaudiBERT: A Large Language Model Pretrained on Saudi Dialect Corpora* (arXiv:2405.06239). arXiv. https://doi.org/10.48550/arXiv.2405.06239

Salah, Z., Al-Ghuwairi, A.-R. F., Baarah, A., Aloqaily, A., Qadoumi, B., Alhayek, M., & Alhijawi, B. (2019). A systematic review on opinion mining and sentiment analysis in social media. *International Journal of Business Information Systems*, *31*(4), 530–554. https://doi.org/10.1504/IJBIS.2019.101585

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1