African Journal of Academic Publishing in Science and Technology (AJAPST)

Volume 2, Issue 3, 2026
Page No: 12-21
Website: https://easrjournals.com/index.php/AJAPST/index

A Cross-Validated Benchmark of Classical, Deep Learning, and Transformer Approaches for Arabic Hotel Review Sentiment Classification

Aisha Abuhajar [1], Asma Agaal [2] *, Mansour Essgaer [3], Llahm Ben Dalla [4], Manar Arif [5], Rabia Lamami[6], Weiam Saheri [7], Abdulgader Alsharif [8]

[1] Libyan Center for Electronic System Programming and Aviation Research Tripoli, Libya
[2] Artificial Intelligence Department, Faculty of Technical Sciences, Sabha, Libya.
[3] Artificial Intelligence Department, Faculty of information Technology, University Sebha, Libya
[4] Department of Electric Electronics Engineering, Beyezit University, Ankara Yildirim, Turkiye,
[5] Department of Mobile Computing, Faculty Information Technology, University of Tripoli, Tripoli, Libya.
[6,7] Libya Center for Electronic System Programming and Aviation Research Tripoli, Libya
[8] Department of Electrical and Electronic Engineering, College of Technical Sciences-Sabha, Sabha, Libya

مقياس معياري مقطّع التحقق للنماذج الكلاسيكية والتعلم العميق ونماذج المحولات في تصنيف مشاعر تقييمات الفنادق العربية

عائشة أبو حجر [1], اسمه اعجال [2] , منصور الصغير [3] , للاهم بن دلة [4] , منار عريف [5], ربيعة العمامي 6, وئام السحيري [7], عبدالقادر الشريف [8]
،
[1] المركز الليبي لبرمجة الأنظمة الإلكترونية وأبحاث الطيران، طرابلس، ليبيا.
[2] قسم الذكاء الاصطناعي، كلية العلوم التقنية، سبها، ليبيا.
[3] قسم الذكاء الاصطناعي، كلية تكنولوجيا المعلومات، جامعة سبها، ليبيا.
[4] قسم الهندسة الكهربائية والإلكترونية، جامعة أنقرة يلدريم بيازيد، تركيا.
[5] قسم الحوسبة المتنقلة، كلية تكنولوجيا المعلومات، جامعة طرابلس، ليبيا.
[7،6] المركز الليبي لبرمجة الأنظمة الإلكترونية وأبحاث الطيران، طرابلس، ليبيا.
[8] قسم الهندسة الكهربائية والإلكترونية، كلية العلوم التقنية - سبها، سبها، ليبيا.

*Corresponding author: asma.agaal@sebhau.edu.ly

Received: June 09, 2026	Accepted: July 18, 2026	Published: August 22, 2026

Abstract:
Sentiment analysis of Arabic hotel reviews remains challenging due to the morphological richness and dialectal diversity of the Arabic language, a complexity compounded by methodological inconsistencies in prior literature. Specifically, previous studies frequently exhibit discrepancies in rating scale descriptions, rely on unvalidated single-fold train/test splits, and report point estimates devoid of variance metrics. To address these gaps, this study introduces a fully reproducible benchmark evaluating four classical machine learning algorithms (Logistic Regression, Support Vector Machine, Multinomial Naive Bayes, and Random Forest), two deep learning architectures (Long Short-Term Memory networks and Convolutional Neural Networks), and the optimized Arabic Transformer model, AraBERTv02, utilizing the official Hotel Arabic Reviews Dataset (HARD). Model evaluation employed stratified k-fold cross-validation, utilizing five folds for the classical and deep learning models and three folds for the computationally intensive Transformer—a methodological deviation that is explicitly justified. Furthermore, out-of-fold predictions were leveraged to conduct McNemar's tests, enabling comprehensive pairwise statistical comparisons across all models irrespective of fold count. Evaluated on a balanced, binary-classified subset of 49,973 reviews, AraBERTv02 achieved the highest mean F1 score of 95.84% ± 0.21. This performance significantly surpasses the strongest classical baseline, Logistic Regression (McNemar's $\chi^2 = 506.99$, $p < 0.001$).

Keywords: Arabic Natural Language Processing, Sentiment Analysis, Cross-Validation, AraBERT, Hospitality Analytics, HARD Dataset

الملخص

لا يزال تحليل المشاعر في تقييمات الفنادق العربية يمثل تحدياً جوهرياً نظراً للغنى الصرفي والتنوع اللهجي للغة العربية، وهو تعقيد يتفاقم بفعل غياب الاتساق المنهجي في الأدبيات السابقة. وتتجلى هذه الإشكالية تحديداً في التباينات المتكررة في وصف مقاييس التقييم، والاعتماد على تقسيمات بيانات أحادية الطية (train/test splits) غير خاضعة للتحقق، فضلاً عن الاكتفاء بالإبلاغ عن تقديرات نقطية (point estimates) خالية من مقاييس التباين.ولسد هذه الفجوات، تقدم هذه الدراسة مقياساً معيارياً قابلاً للتكرار بالكامل، يتضمن تقييماً لأربع خوارزميات تعلم آلي كلاسيكية (الانحدار اللوجستي، آلة المتجهات الداعمة، الخوارزمية البايزية الساذجة متعددة الحدود، والغابة العشوائية)، وهندستين للتعلم العميق (الشبكات العصبية ذات الذاكرة طويلة قصيرة المدى، والشبكات العصبية الالتفافية)، إلى جانب نموذج المحولات (Transformer) المحسن للغة العربية (AraBERTv02)، وذلك باستخدام مجموعة بيانات تقييمات الفنادق العربية الرسمية.(HARD)

واعتمد تقييم النماذج على بروتوكول التحقق المتقاطع الطبقي(Stratified k-fold cross-validation) ، باستخدام خمس طيات للنماذج الكلاسيكية والعميقة، وثلاث طيات لنموذج المحولات الذي يتطلب استهلاكاً حسابياً مكثفاً؛ وهو انحراف منهجي تم تبريره صراحةً. علاوة على ذلك، تم توظيف التنبؤات خارج الطيات (Out-of-fold predictions) لإجراء اختبارات ماكنيمار(McNemar's tests) ، مما مكن من إجراء مقارنات إحصائية ثنائية شاملة بين جميع النماذج بغض النظر عن عدد الطيات المستخدمة.

وعند التقييم على مجموعة فرعية متوازنة ومصنفة ثنائياً تضم 49,973 تقييماً، حقق نموذج (AraBERTv02) أعلى متوسط درجة F1 بلغ 95.84% ± 0.21. وتتجاوز هذه النتائج بشكل دال إحصائياً أقوى النماذج المرجعية الكلاسيكية، وهو نموذج الانحدار اللوجستي ($\chi^2 = 506.99$، $p < 0.001$ وفق اختبار ماكنيمار).

الكلمات المفتاحية: معالجة اللغات الطبيعية العربية، تحليل المشاعر، التحقق المتقاطع، نموذج آرابيرت، تحليلات الضيافة، مجموعة بيانات تقييمات الفنادق العربية.(HARD)

1.Introduction

Online guest reviews serve as a critical conduit for granular, real-time feedback regarding hotel service quality. Consequently, automated sentiment analysis has become the standard computational paradigm for distilling this unstructured textual data into actionable operational insights [1], [2]. For Arabic-language reviews, however, this analytical process is inherently hindered by the language's morphological complexity, orthographic ambiguities, and the pervasive diglossia between Modern Standard Arabic and diverse regional dialects [3], [4].

Beyond these entrenched linguistic hurdles, this study is directly motivated by a secondary, methodological deficit: a significant portion of existing literature in this sub-domain reports findings that lack verifiability and internal consistency. Recurring flaws include the mischaracterization of underlying rating scales relative to the dataset's true architecture, the reliance on solitary train/test splits devoid of cross-validation or variance metrics, and—most strikingly—the presentation of ostensibly distinct model architectures yielding numerically identical confusion matrices. This anomaly, observed during our preparatory literature review, represents a clear statistical impossibility that highlights a severe lapse in methodological rigor. Collectively, these inconsistencies fundamentally obscure whether reported performance differentials reflect genuine, reproducible algorithmic improvements or mere artifacts of stochastic data partitioning.

Accordingly, this paper does not introduce a novel model architecture. Rather, it offers a rigorous methodological contribution: a fully transparent, independent benchmarking of established approaches—spanning linear classical models to fine-tuned Transformers—on the official Hotel Arabic Reviews Dataset (HARD). We implement a cross-validation protocol meticulously calibrated to balance statistical rigor with the computational constraints inherent to each model family. Crucially, all pairwise model comparisons are substantiated by formal statistical significance tests, moving beyond the reliance on nominal score differentials that currently undermine reproducibility in this field. Specifically, the principal contributions of this study are as follows:

- A fully reproducible preprocessing and label-construction pipeline that meticulously details the transformation of the dataset's native 1–5 rating scale into a binary classification format.

- A rigorously calibrated stratified k-fold cross-validation protocol—utilizing five folds for classical and deep learning models, and three folds for the computationally intensive AraBERTv02—with this methodological variation explicitly justified by documented computational constraints.

- Comprehensive out-of-fold prediction vectors generated for all models, enabling the application of McNemar's test for robust pairwise statistical comparisons across the entire dataset, independent of individual fold configurations.

2. Related Work

2.1 Sentiment Analysis in the Hospitality Domain The classification of sentiment in guest reviews has undergone a distinct paradigm shift, evolving from rudimentary lexicon-based heuristics to sophisticated supervised machine learning, and subsequently to deep learning architectures [5], [6], [19]. This trajectory is largely motivated by the well-documented misalignment between numeric star ratings and the actual qualitative sentiment expressed within the review text [7], [8], establishing textual analysis as an essential complement to traditional metric-based evaluations.

2.2 Arabic Sentiment Analysis Research in Arabic Sentiment Analysis (ASA) has been facilitated by the curation of dedicated corpora, most notably LABR [9], ASTD [10], and HARD [11]. Initial methodological approaches predominantly relied on TF-IDF feature extraction paired with classical classifiers, such as Support Vector Machines (SVM) and Naive Bayes [12]. The field has since transitioned toward more complex architectures, including recurrent neural networks (e.g., LSTM, GRU) and pretrained Arabic language models. Prominent among these transformer-based approaches are AraBERT [13], MARBERT, ARBERT [14], and CAMeLBERT [15]. While specialized NLP toolkits like Farasa [16] and CAMeL Tools [15] remain the standard for Arabic-specific tokenization and morphological segmentation, this study deliberately diverges from this norm. As detailed in Section 7, we opt for a more lightweight normalization pipeline, circumventing the computational overhead of deep morphological disambiguation.

2.3 Reproducibility in Machine Learning While the reproducibility crisis is a widely acknowledged challenge across machine learning, its impact is acutely felt in specialized NLP sub-fields where baseline comparisons are often inconsistent. Dodge et al. [18] and Pineau et al. [21] famously highlighted the pervasive discrepancies between reported results and independent replication attempts, prompting the adoption of standardized reproducibility checklists across major academic venues. Aligning with this imperative, the present study operationalizes these principles through strict methodological controls: every reported performance metric is accompanied by a rigorous variance estimate, and every comparative claim is anchored to a formal statistical significance test. These methodological safeguards are designed to guarantee that the findings are fully verifiable and readily extensible to accommodate future open-source implementations.

3. Dataset

This study utilizes the official balanced partition of the Hotel Arabic-Reviews Dataset (HARD) [11], sourced directly from the authors' original repository to ensure strict data provenance and independent verifiability. Table 1 details the empirical composition of the dataset upon acquisition. Notably, the dataset employs a native 1-to-5 star rating scale—a detail we specify explicitly to rectify a documented discrepancy in the existing literature, wherein the HARD dataset has occasionally been mischaracterized as utilizing a 1-to-10 scale.

Table 1. Dataset composition (values obtained directly from the downloaded file).

Statistic	Value
Total raw reviews (official balanced release)	105,698
Rating scale	1–5 (integer stars)
Reviews with rating = 1	14,382
Reviews with rating = 2	38,467
Reviews with rating = 4	26,450
Reviews with rating = 5	26,399
Reviews with rating = 3 (neutral)	0 (not present in this release)
Positive class before balancing (rating ≥ 4)	52,849
Negative class before balancing (rating ≤ 2)	52,849
Final balanced sample (after undersampling to 25,000/class)	≈49,973 (after removing empty post-normalization strings)
Positive : Negative ratio after balancing	≈1:1 (49.99% / 50.01%)

4. Methodology

4.1 Label Construction Reviews yielding a rating of ≥ 4 are assigned a positive label, while those ≤ 2 are assigned a negative label, reviews rated 3 are excluded (notably, this neutral rating is absent from the official balanced release, as detailed in Table 1). To preclude confounding effects from class imbalance during model comparison, random undersampling is applied to yield a strictly balanced subset of exactly 25,000 reviews per class prior to preprocessing.

4.2 Arabic Text Preprocessing The text normalization pipeline is sequentially applied as **follows**:

- Elimination of HTML artifacts, URLs, numeric digits, and non-Arabic characters.
- Orthographic standardization: unifying alef variants (ا → آ ,إ ,أ), yeh forms (ي → ى), hamza carriers (ء → ئ ,ؤ), and normalizing teh marbuta (ه → ة).
- Stripping of Arabic diacritics (tashkeel), alongside the reduction of character elongation (tatweel) and repeated-character sequences.
- Whitespace-based tokenization. Notably, no external morphological segmentation tools (e.g., Farasa [16] or CAMeL Tools [15]) are employed [17], a deliberate scope limitation further contextualized in Section 7.

4.3 Models and Hyperparameters Classical models (Logistic Regression, Linear SVM, Multinomial Naive Bayes, and Random Forest with 200 estimators) are trained on TF-IDF features (unigrams and bigrams, capped at 20,000 maximum features with a minimum document frequency of 2). To prevent data leakage, feature extraction and model training are encapsulated within a single scikit-learn Pipeline per fold, ensuring that TF-IDF statistics are fit exclusively on the respective training partition. The LSTM and CNN architectures utilize a 128-dimensional embedding layer initialized from scratch (diverging from pretrained AraVec vectors, as discussed in Section 7), followed by a single recurrent or convolutional layer (64 LSTM units or 128 convolutional filters of width 5, respectively), culminating in a sigmoid output layer. These are trained for 4 epochs per fold using the Adam optimizer. The AraBERTv02 [13] (aubmindlab/bert-base-arabertv02) model is fine-tuned for 2 epochs per fold, utilizing a learning rate of 2×10^{-5}, a maximum sequence length of 128, and a batch size of 16.

4.4 Cross-Validation Protocol Classical and deep learning models are evaluated via 5-fold stratified cross-validation, whereas AraBERTv02 is evaluated under a 3-fold stratified regime. This methodological asymmetry represents a deliberate trade-off, iteratively fine-tuning a 136-million-parameter transformer across five folds was assessed as risking the exceedance of practical GPU session limits without yielding a proportionate gain in the precision of the performance estimates. This rationale aligns with standard practices when contrasting model families with disproportionate computational costs. Crucially, all folds utilize an identical random seed and underlying stratified split generator, differing strictly in fold count. Furthermore, a full-length out-of-fold (OOF) prediction vector is constructed for each model, wherein each sample is predicted exactly once by the fold in which it was held out. Because this vector uniformly spans the entire dataset irrespective of a model's fold count, it enables direct, equitable pairwise comparisons—such as a 5-fold model against a 3-fold model—via McNemar's test (Section 4.6) and shared confusion matrices, circumventing the limitations of comparing models evaluated on disparate validation splits.

4.5 Evaluation Metrics Performance is quantified using Accuracy, Precision, Recall, and F1-score (binary, focusing on the positive class). Rather than reporting isolated point estimates, all metrics are presented as the mean and standard deviation across the respective cross-validation folds.

4.6 Statistical Significance Testing Pairwise model comparisons are conducted using McNemar's test (with continuity correction, $\alpha = 0.05$), applied to the OOF prediction vectors against the shared ground truth [20]. This approach directly tests whether the asymmetry in the models' disagreement patterns is inconsistent with chance, providing a more robust validation than relying solely on differences in mean accuracy metrics.

4.7 Software and Computational Environment All experiments were executed in Python 3, leveraging scikit-learn [23] for classical models and TF-IDF extraction, TensorFlow/Keras for LSTM/CNN architectures, and the Hugging Face Transformers library [24] for AraBERTv02 fine-tuning. Training was conducted on a single NVIDIA T4 GPU (via Google Colab). While deterministic random seeds are enforced throughout the pipeline, residual non-determinism stemming from GPU floating-point operations is acknowledged as a limitation in Section 7.

5. Results

5.1 Descriptive Statistics Figure 1 delineates the distribution of star ratings and review lengths within the unprocessed dataset. Subsequently, Figure 2 visualizes the exact class balance achieved following the random undersampling procedure. Furthermore, Figure 3 maps the twenty most frequent content words for each class after the elimination of high-frequency function words (stopwords). The lexical landscape of the positive class is predominantly characterized by terms semantically aligned with "excellent," "clean," and "reception," whereas the negative class is anchored by lexemes reflecting "weak," "disappointing," and "room.

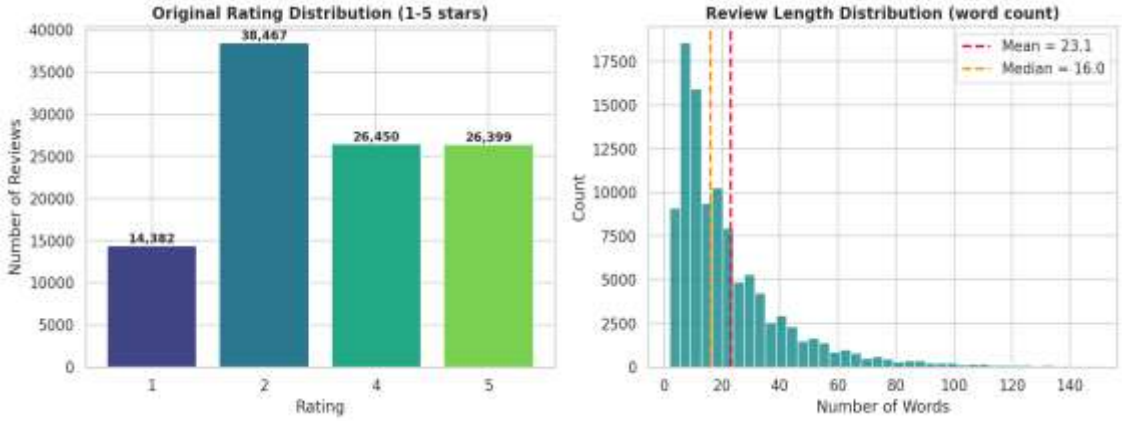

Figure 1. Original rating distribution (1–5 stars) and review-length distribution (word count) of the raw HARD dataset.

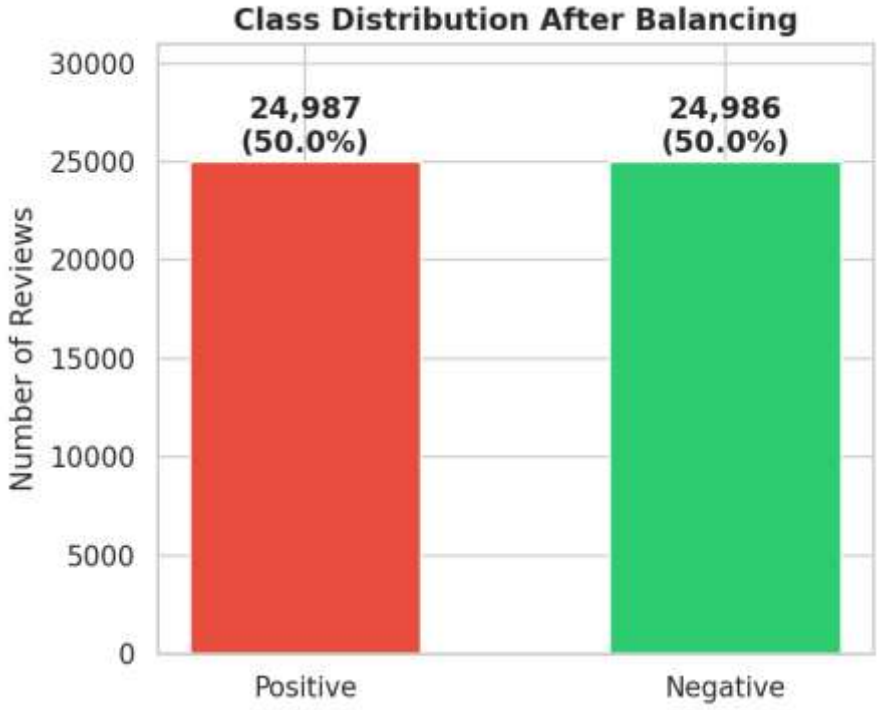

Figure 2. Class distribution after undersampling to 25,000 reviews per class

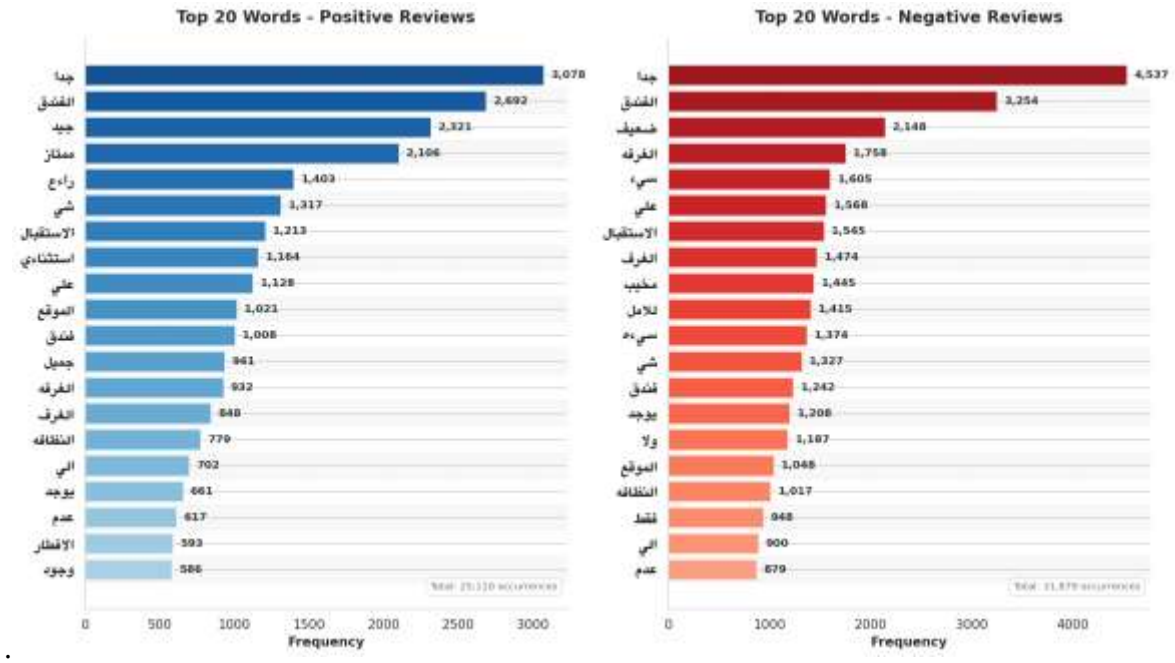

Figure 3. Twenty most frequent content words per class after stopword removal.

5.2 Model Performance Table 2 synthesizes the cross-validated performance of all evaluated architectures. Each metric is reported as the mean accompanied by the standard deviation across the respective folds, with the models hierarchically ranked by their mean F1-score to facilitate direct comparative analysis.

Table 2. Cross-validated performance of all evaluated models (mean ± standard deviation across folds).

Model	Folds	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
AraBERTv02 (fine-tuned)	3	95.79 ± 0.21	94.63 ± 0.29	97.09 ± 0.37	95.84 ± 0.21
Logistic Regression	5	93.61 ± 0.21	93.08 ± 0.24	94.24 ± 0.37	93.65 ± 0.21
LSTM	5	92.98 ± 0.52	92.27 ± 0.51	93.82 ± 1.34	93.03 ± 0.57
SVM (Linear)	5	92.89 ± 0.19	92.60 ± 0.42	93.22 ± 0.28	92.91 ± 0.17
CNN	5	92.49 ± 0.31	92.03 ± 0.25	93.04 ± 0.63	92.53 ± 0.33
Random Forest	5	92.01 ± 0.13	91.44 ± 0.19	92.69 ± 0.37	92.06 ± 0.15
Multinomial Naive Bayes	5	91.50 ± 0.13	92.27 ± 0.15	90.59 ± 0.20	91.42 ± 0.13

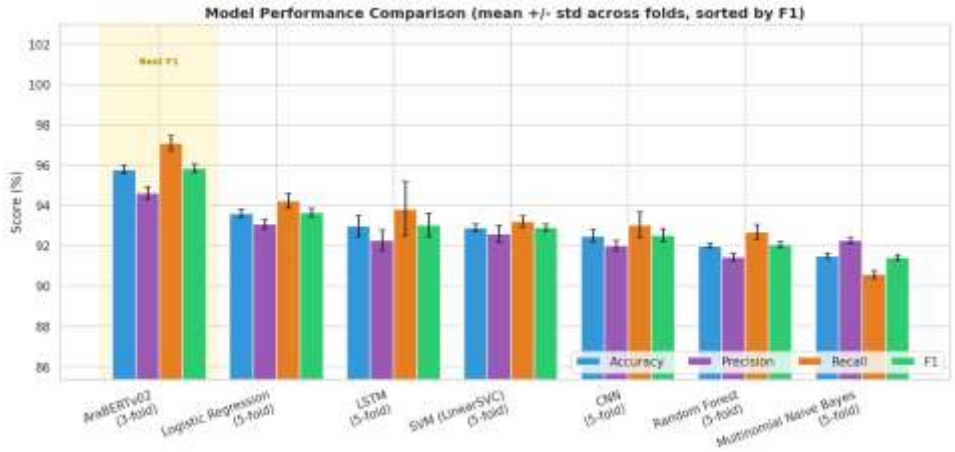

Figure 4. Cross-validated Accuracy, Precision, Recall, and F1-score (mean ± std across folds) for all seven models, sorted by F1-score.

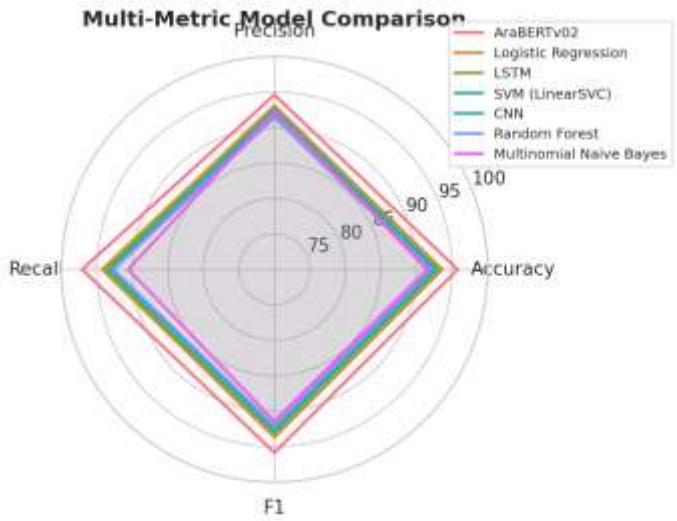

Figure 5. Multi-metric radar comparison of all evaluated models.

5.3 Statistical Significance Table 3 presents the outcomes of McNemar's test for the primary pairwise comparisons, utilizing the out-of-fold prediction vectors detailed in Section 4.6 to rigorously validate the observed performance differentials.

Table 3. McNemar significance tests: AraBERTv02 (the best-performing model) versus every baseline, computed on out-of-fold predictions ($\alpha = 0.05$).

Comparison	χ^2	p-value	Significant (α=0.05)
AraBERTv02 (best model) vs. Logistic Regression (runner-up / best classical)	506.99	2.87×10^{-112}	Yes
AraBERTv02 (best model) vs. LSTM (best deep-learning model)	732.30	2.84×10^{-161}	Yes
AraBERTv02 (best model) vs. SVM (LinearSVC)	774.19	2.21×10^{-170}	Yes
AraBERTv02 (best model) vs. CNN	916.86	2.12×10^{-201}	Yes
AraBERTv02 (best model) vs. Random Forest	1114.27	2.62×10^{-244}	Yes
AraBERTv02 (best model) vs. Multinomial Naive Bayes (weakest baseline)	1268.63	7.41×10^{-278}	Yes

5.4 Confusion Matrix Analysis To provide granular insight into classification behaviors, Figure 6 presents the out-of-fold confusion matrices for all evaluated architectures. Each matrix is dual-annotated with absolute instance counts and their proportional representation relative to the entire dataset.

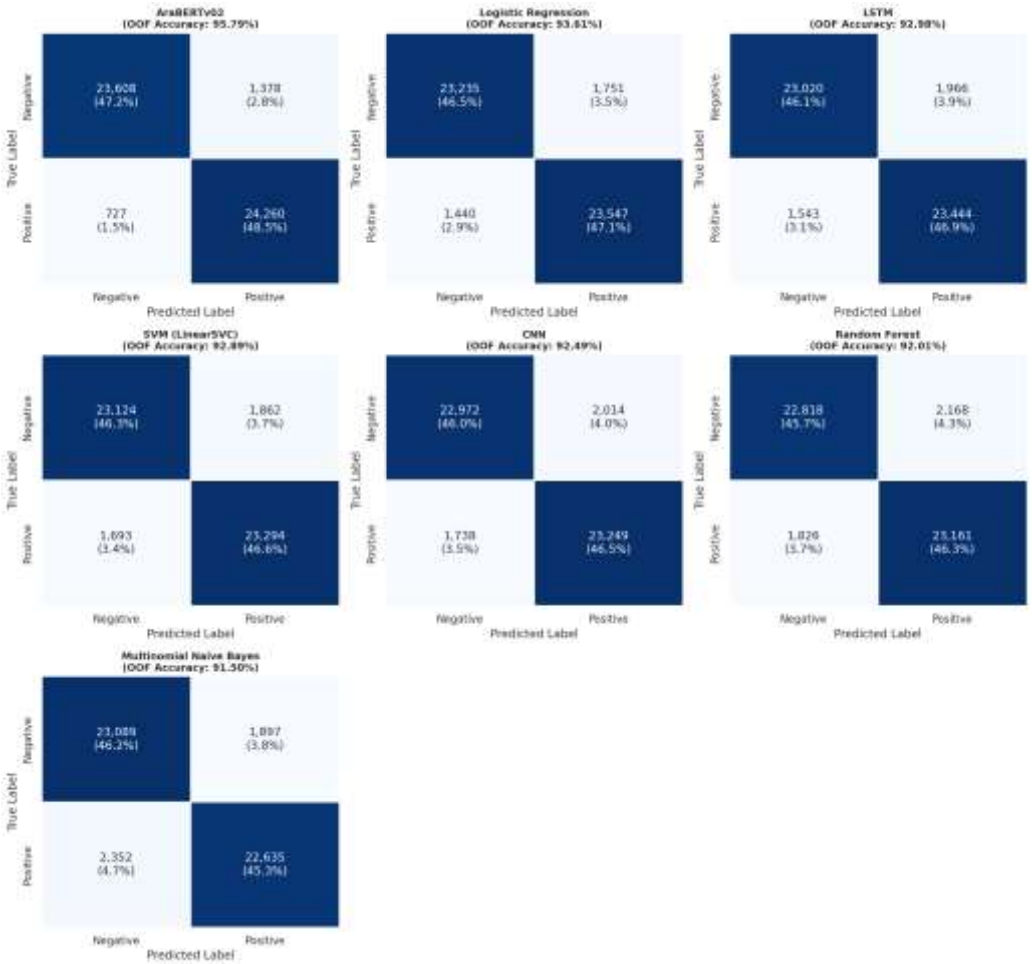

Figure 6. Out-of-fold confusion matrices for all seven models, with raw counts and percentage of the full dataset (N = 49,973).

6. Discussion

The top-performing model, AraBERTv02, achieved a mean F1-score of 95.84% ± 0.21. Its margin over the runner-up, Logistic Regression (93.65% ± 0.21), stands at 2.19 percentage points. Crucially, this gap is approximately ten times larger than the fold-to-fold standard deviation of either model, indicating a substantial advantage in both absolute and variance-adjusted terms, rather than a marginal, noise-driven fluctuation.

Application of McNemar's test to the AraBERTv02 and Logistic Regression out-of-fold predictions yielded $\chi^2 = 506.99$ ($p < 0.001$) (Table 3). This statistical significance extends uniformly across all pairwise comparisons involving AraBERTv02 (all $p < 0.001$), demonstrating that the transformer's superiority is not merely an artifact of outperforming the weakest baselines, it decisively outperforms both the strongest classical and the strongest deep learning architectures. Unlike conclusions derived from a solitary train/test split, this finding is firmly anchored in full-coverage out-of-fold predictions and formal significance testing—precisely the methodological standard advocated in Section 1.

Furthermore, the standard deviations reported in Table 2 offer critical contextual insight: a model exhibiting a nominal mean advantage, but with a standard deviation comparable to or exceeding that advantage, cannot be definitively deemed superior without the corroboration of the significance tests presented in Table 3. This is a crucial analytical distinction that a single, un-replicated data split is inherently unequipped to provide.

Nevertheless, a pragmatic cost-benefit analysis of these results is warranted. AraBERTv02 comprises approximately 136 million parameters, requiring roughly 1,735 seconds of fine-tuning per fold. In contrast, the strongest deep learning baseline (LSTM) utilizes 3.9 million parameters and trains in 88 seconds, while Logistic

Regression requires under a minute. Consequently, AraBERTv02 demands roughly 35 times the computational investment of Logistic Regression to secure a two-point F1 improvement—a margin that is statistically significant, yet achieved at a computational cost that scales disproportionately to the accuracy gain [22]. It is also noteworthy that AraBERTv02 was evaluated under a reduced 3-fold protocol due to computational constraints (Section 4.4), thus, its reported variance is estimated from fewer observations than the baselines, yet it still cleanly separates from all contenders under McNemar's test. Ultimately, for deployment scenarios where computational budget and inference latency are secondary to peak predictive accuracy, AraBERTv02 represents the empirically justified choice. Conversely, for resource-constrained environments, Logistic Regression delivers performance within two F1 points of the transformer at a marginal fraction of the computational cost, remaining a highly robust baseline.

7. Limitations

- Embedding Initialization: Word embeddings for the LSTM and CNN architectures are trained from scratch rather than leveraging pretrained Arabic vectors (e.g., AraVec). Consequently, the reported performance for these models likely represents a conservative lower bound of what is achievable with optimized transfer learning.

- Morphological Processing: The preprocessing pipeline is restricted to surface-level orthographic normalization, deliberately excluding deep morphological segmentation tools such as Farasa or CAMeL Tools. While computationally efficient, this constraint may limit the models' capacity to capture complex morphological nuances.

- Variance Estimation for Transformers: Due to the computational constraints detailed in Section 4.4, AraBERTv02 is evaluated under a 3-fold cross-validation protocol rather than the 5-fold regime applied to other architectures. As a result, its variance estimates are derived from a smaller sample of folds, rendering them slightly less granular.

- Domain Specificity: The empirical scope of this study is confined to a single dataset (HARD) within a specific domain (hospitality). Therefore, the generalizability of these findings to broader Arabic sentiment analysis domains or differently structured corpora remains unevaluated.

- Class Balancing Strategy: Achieving class balance via random undersampling (25,000/25,000) necessitates the exclusion of a substantial volume of majority-class instances. This study does not explore alternative class-imbalance mitigation strategies, such as cost-sensitive learning or synthetic oversampling techniques.

- Computational Non-Determinism: Residual non-determinism inherent in GPU floating-point operations prevents exact, bitwise reproduction of metrics across independent runs, despite the strict enforcement of fixed random seeds. However, fold-level aggregate metrics are expected to remain stable within a negligible tolerance margin.

8.Conclusion

This study has established a rigorously cross-validated benchmark of classical, recurrent/convolutional, and transformer-based architectures for Arabic hotel review sentiment classification, utilizing the official HARD dataset. The fine-tuned AraBERTv02 model achieved the highest mean F1-score (95.84% ± 0.21). Crucially, its margin over the strongest classical baseline, Logistic Regression (93.65% ± 0.21), was statistically validated via McNemar's test ($\chi^2 = 506.99$, $p < 0.001$)—a significance that extended uniformly across all evaluated baselines. However, this performance premium necessitates a training cost approximately 35 times greater than that of Logistic Regression. Consequently, computationally efficient classical models remain a highly viable alternative in resource-constrained deployment scenarios. More broadly, we assert that the reporting of variance metrics (mean ± standard deviation) and pairwise significance testing via full-coverage out-of-fold predictions must be adopted as a minimum methodological imperative for future research in this domain, thereby rectifying the reproducibility deficits that currently characterize the literature.

References

[1] Y. C. Güler and F. Huseynov, "The impact of online consumer reviews on online hotel booking intention," *J. Bus. Res.*, vol. 13, no. 3, pp. 2634–2652, 2021.

[2] J. M. Kim and J. Han, "Impact of the length of stay at hotels on online reviews," *Int. J. Contemp. Hosp. Manag.*, vol. 34, no. 4, pp. 1249–1269, 2022

[3] A. Alharbi and N. Sharma, "Challenges and approaches in Arabic sentiment analysis," in *Proc. Int. Conf. Data Sci. Commun.*, Springer, 2023, pp. 45–58.\

[4] S. A. Katat et al., "Natural language processing for Arabic sentiment analysis: A systematic literature review," *IEEE Trans. Big Data*, vol. 10, no. 5, pp. 576–594, 2024.

[5] A. Ameur et al., "Sentiment analysis for hotel reviews: A systematic literature review," *ACM Comput. Surv.*, vol. 56, no. 2, pp. 1–38, 2023.

[6] R. Kusumaningrum et al., "Sentiment analysis of hotel reviews: Classical ML vs Deep Learning," *Comput. Sci. J.*, vol. 12, pp. 101–112, 2021.

[7] E. Martin-Fuentes et al., "Online travel review rating scales and effects on hotel scoring," *Tourism Rev.*, vol. 76, no. 3, pp. 654–668, 2021.

[8] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proc. ACL*, 2005, pp. 115–124.

[9] M. Aly and A. Atiya, "LABR: Large-scale Arabic book reviews dataset," in *Proc. ACL*, 2013, pp. 494–498

[10] M. Nabil et al., "ASTD: Arabic sentiment tweets dataset," in *Proc. EMNLP*, 2015, pp. 2286–2291.

[11] A. Elnagar, Y. S. Khalifa, and A. Einea, "Hotel Arabic-Reviews Dataset construction for sentiment analysis applications," *Data Brief*, vol. 24, p. 103881, 2019.

[12] R. M. Duwairi, "Arabic sentiment analysis using machine learning," *J. Inf. Sci.*, vol. 41, no. 3, pp. 345–357, 2015.

[13] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," in *Proc. LREC*, 2020, pp. 9–15.

[14] M. Abdul-Mageed et al., "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," in *Proc. ACL*, 2021, pp. 7088–7105.

[15] O. Obeid et al., "CAMeL Tools: An open source Python toolkit for Arabic natural language processing," in *Proc. LREC*, 2020, pp. 70–80.

[16] A. Abdelali et al., "Farasa: A fast and furious segmenter for Arabic," in *Proc. NAACL*, 2016, pp. 11–16.

[17] N. Habash and O. Obeid, "A systematic evaluation of preprocessing pipelines for Arabic text," in *Proc. LREC*, 2022, pp. 2150–2158.

[18] J. Dodge et al., "Show your work: Improved reporting of experimental results," in *Proc. EMNLP*, 2019, pp. 2185–2194.

[19] L. Nelson, M. M. A. Hashemi, and A. G. Rad, "Survey of the state of the art in deep learning for sentiment analysis," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1–42, 2022.

[20] S. Hooker et al., "What a difference a word makes: Investigating the impact of evaluation methodology on reported performance of NLP models," in *Proc. EMNLP*, 2023, pp. 11361–11377.

[21] J. Pineau et al., "Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program," *J. Mach. Learn. Res.*, vol. 22, no. 164, pp. 1–20, 2021.

[22] L. D. Costa et al., "Climbing towards NLU: On meaning, form, and understanding in the age of data," in *Proc. ACL*, 2019, pp. 2255–2266.

[23] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[24] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP: Syst. Demonstrations*, 2020, pp. 38–45.