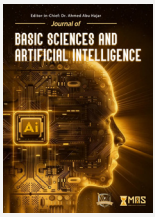
Journal of Basic Sciences and Artificial Intelligence (JBSAI)

DOI: 10.65098/jbsai.01.2026.18.23



RESEARCH ARTICLE

AraDigitalContent77: A Novel Arabic Intent Classification Dataset for Digital Content Platforms via Cross-Domain Schema Adaptation

Mohamad Hallak Fattal*

Department of Computer Engineering, Faculty of Electrical and Electronic Engineering, University of Aleppo, Aleppo 12212, Syria
**** Corresponding author: mohamadhallakf@gmail.com***

ARTICLE DETAILS

Article History:
Received 22 May 2026
Accepted 26 Aug 2026
Available online 31 Aug 2026

Online Article Code

ABSTRACT

Intent classification underpins conversational and recommendation systems, yet Arabic remains underserved by domain-specific benchmarks outside banking and finance. This paper introduces AraDigitalContent77, an Arabic intent classification dataset for digital content platforms, including streaming, podcasts, music, news, and audiobooks. The Banking77 schema was adapted through an Action-Object-Goal framework, producing 77 intents composed of 7 direct transfers, 34 analogical transfers, and 36 newly defined intents; 36 banking-specific intents were removed. Utterances were created through a five-stage hybrid pipeline combining human-written seeds, controlled large language model generation, automated quality filtering, selective human review, template-pattern transformation, and intentional linguistic-noise injection. The resulting corpus contains 13,090 utterances evenly distributed across 77 intents, with 170 utterances per intent, no exact duplicate sentences, and no missing values. A fine-tuned AraBERT classifier was evaluated using stratified 70/15/15 training, validation, and independent test splits. The final test set produced 96.59% accuracy and a macro F1-score of 96.61%. These results establish a strong within-dataset baseline while avoiding direct superiority claims across benchmarks that use different domains and evaluation protocols.

KEYWORDS

Arabic Natural Language Processing, Intent Classification, Schema Adaptation, Low-Resource Datasets, AraBERT, Digital Content Platforms, Synthetic Data Generation

1. INTRODUCTION

Intent classification is a foundational task in natural language understanding, enabling systems to map free-form user utterances to a finite set of actionable categories. It underlies customer support automation, chatbots, and information retrieval systems across domains such as banking, e-commerce, and telecommunications (Casanueva et al., 2020). While digital content platforms—including streaming, podcast, music, news, and audiobook services—represent a rapidly growing sector with heavy reliance on natural language interaction, no publicly available Arabic dataset exists for intent classification within this domain.

The scarcity of Arabic resources for intent classification is well documented. Banking77, introduced by Casanueva et al. (2020), remains the most cited benchmark in the field, comprising 13,083 English utterances across 77 fine-grained banking intents. Jarrar et al. (2023) later localized this benchmark into Arabic as ArBanking77, producing 31,404 utterances in Modern Standard Arabic and Palestinian dialect. However, both resources remain confined to the banking domain, leaving digital content platforms—a domain with materially different user intents (e.g., subscription management, content discovery, playback issues)—entirely unaddressed in Arabic.

This paper addresses this gap by introducing AraDigitalContent77, a novel Arabic intent classification dataset purpose-built for digital content platforms. Rather than collecting data from scratch, we adapt the well-established Banking77 schema through a systematic Action-Object-Goal transfer methodology, preserving the proven structural properties of the original benchmark (77 balanced, fine-grained intents) while replacing its domain content entirely. We further contribute a hybrid data generation pipeline that combines human-authored seed utterances, controlled large language model (LLM) generation, and multi-stage automated quality control to produce a large-scale, high-quality, and documented dataset without relying on costly crowdsourcing or machine translation.

The contributions of this paper are threefold: (1) a systematic schema adaptation methodology (Action-Object-Goal) for transferring intent classification taxonomies across domains; (2) a structured, documented data generation pipeline combining human seeds, LLM-based augmentation, and five-layer automated filtering; and (3) AraDigitalContent77 itself—a balanced, deduplicated, 13,090-utterance Arabic dataset released publicly on Hugging Face, together with a fine-tuned AraBERT baseline achieving a 96.61% test macro F1-score.

2. RELATED WORK

Efficient intent detection was popularized by Casanueva et al. (2020), who introduced Banking77, a single-domain dataset of 13,083 English utterances covering 77 banking intents, along with dual sentence encoder methods for few-shot and full-data settings. Banking77 has since become the most widely cited benchmark for intent detection research due to its fine granularity and balanced class distribution.

Jarrar et al. (2023) extended this benchmark to Arabic with ArBanking77, arabizing and localizing the original 13,083 English queries into 31,404 queries spanning Modern Standard Arabic and the Palestinian dialect. Their AraBERT-based model achieved F1-scores of 92.09% on Modern Standard Arabic (MSA) and 89.95% on the Palestinian dialect, establishing the first strong Arabic baseline for banking intent detection. While ArBanking77 demonstrates the value of Arabic-specific benchmarks, it remains restricted to the banking domain and relies partly on translation from English.

Data augmentation for intent detection using LLMs has received increasing attention. Lin et al. (2024) proposed a two-stage "Generate then Refine" approach for zero-shot intent detection, in which an open-source LLM generates candidate utterances that are subsequently refined by a smaller sequence-to-sequence model, improving data utility and diversity for unseen domains. Castillo-López et al. (2026) addressed a related failure mode—semantic ambiguity in LLM-generated augmentation data—with Disambiguated Data Augmentation for Intent Recognition, a method using sentence embeddings to detect and iteratively regenerate utterances that are semantically closer to a non-target intent than to their intended class. Pecher et al. (2026) conducted a broad multilingual study across 11 languages and four classification tasks, concluding that LLMs are generally more effective as synthetic data generators for training smaller downstream classifiers than as direct zero-shot classifiers themselves.

Shin et al. (2024) introduced One-Shot In-Context Intent Classification 2 (OSIC2), a one-shot in-context intent classification benchmark built from the Intent Classification collection, comprising 13 held-in datasets (748 intents) and 3 held-out datasets (109 intents) for evaluating cross-domain generalization. Their 7-billion-parameter model (OSIC2–7B) achieved an average accuracy of 81.25% on held-out domains, surpassing GPT-4-turbo, which reached 79.03% under the same one-shot evaluation protocol. This result underscores that domain adaptability, rather than raw model scale, is a key driver of intent classification performance in low-resource or unseen-domain settings—directly motivating our schema adaptation approach.

He et al. (2026) examined a related but distinct concern: the reliability of multilingual intent classification benchmarks constructed via machine translation. Using a real-world logistics customer-service dataset of approximately 30,000 queries organized into a two-level taxonomy (13 parent and 17 leaf intents) across English, Spanish, and Arabic, they showed that machine-translated evaluation can overestimate performance relative to native, human-written queries. This finding provides empirical support for our decision to construct AraDigitalContent77 as an originally authored and LLM-generated Arabic corpus rather than a translation of an English source.

Finally, Hossain et al. (2025) demonstrated that augmenting dialectal Arabic inputs with Modern Standard Arabic and English translations substantially improves intent detection accuracy, raising Tunisian dialect accuracy from 43.3% to 94% and Palestinian dialect accuracy from 87.7% to 93.5% under a multilingual input augmentation scheme. This result highlights the value of dialectal and multilingual augmentation as a promising direction for future extensions of the present work. The Arabic encoder used for the baseline in this study is AraBERT (Antoun et al., 2020), a transformer pretrained specifically for Arabic and widely used across Arabic Natural Language Processing tasks.

3. METHODOLOGY

3.1 Schema Adaptation: The Action-Object-Goal Framework

To transfer the Banking77 taxonomy to the digital content domain without inheriting irrelevant banking semantics, we developed a decision framework that decomposes every intent into three constituent elements: an Action (e.g., cancel, activate, inquire about), an Object (e.g., card, subscription, PIN), and a Goal (e.g., resolving a problem, obtaining information, changing a setting). Each of the 77 original Banking77 intents was evaluated against this framework and classified into one of four transfer categories:

- Direct Transfer (7 intents): the complete Action-Object-Goal triplet has a natural, unmodified equivalent in the digital content domain (e.g., freeze_account, apple_pay_or_google_pay).
- Analogical Transfer (34 intents): the Action and Goal are preserved while the Object is substituted with a domain-appropriate equivalent (e.g., change_pin → change_password; activate_my_card → activate_my_account).
- New Intent (36 intents): no equivalent exists in Banking77; these intents were authored from scratch to capture digital-content-specific needs such as video buffering issues or advertisement experience complaints.
- Deleted (36 intents): banking-specific intents (e.g., cash withdrawal, ATM support, currency exchange) with no meaningful analogue in the target domain were discarded.

This process yielded a final taxonomy of exactly 77 intents (7 direct + 34 analogical + 36 new). Of the 77 Banking77 intents, 41 were retained through direct or analogical transfer, and 36 were removed, resolving the complete source taxonomy without unaccounted categories.

3.2 Hybrid Data Generation Pipeline

Given the absence of any existing Arabic corpus for this domain, utterances were generated through a five-stage hybrid pipeline designed to balance scale, quality, and naturalness:

- Stage 1 Manual Seed Generation: 2,310 utterances (30 per intent) were hand-written to establish stylistic ground truth and guide subsequent automated generation.
- Stage 2 LLM-Based Expansion: the 30 seeds per intent were used as in-context examples to prompt LLMs for expansion. Multiple candidate generators (including GPT-5, Claude, Qwen, and DeepSeek) were considered during dataset construction. The final corpus used a multi-model generation approach, primarily GPT-5 and Claude, as a design choice intended to diversify generation sources rather than as a claim of experimentally demonstrated superiority over single-model generation.
- Stage 3—Multi-Layer Automated Filtering: generated utterances passed through five sequential filters—(1) a language filter removing dialectal or uncontrolled text, (2) a repetition filter reducing over-represented lexical patterns, (3) a length filter removing utterances shorter than 5 or longer than 18 words, (4) a semantic drift filter removing utterances more similar to a non-target intent than to their labeled intent, and (5) an intent-name filter removing utterances that trivially embed the literal intent label.
- Stage 4—Selective Human Review: a targeted qualitative review was used to inspect borderline class-boundary cases for label consistency, semantic drift, and linguistic plausibility. This step is reported as quality control rather than as a formal blind multi-annotator annotation study; no inter-annotator agreement statistic is reported.
- Stage 5—Lexical Diversity Enhancement: approximately 577 template-heavy sentences were paraphrased to reduce repetitive surface patterns, and approximately 72 naturally occurring linguistic errors (e.g., hamza-drop spelling variants) were deliberately injected to reflect authentic user-generated text.

The linguistic-noise cases were introduced intentionally after creation and filtering of the clean corpus. They include selected spelling variations, word repetition, and imperfect constructions intended to approximate noisy user-generated queries rather than annotation errors.

3.3 Model Training and Evaluation Protocol

The final corpus was divided before tokenization using a two-stage stratified split with a fixed random seed of 42. The training set contained 9,163 utterances (70%), the validation set contained 1,963 utterances (15%), and the independent test set contained 1,964 utterances (15%).

Original row identifiers were retained during splitting, and pairwise disjointness checks confirmed that no record appeared in more than one split. All 77 intents were represented in every split.

The aubmindlab/bert-base-arabertv2 checkpoint was fine-tuned for sequence classification. Text was truncated or padded to 128 tokens. Training used a batch size of 8, a learning rate of 5×10^{-5}, weight decay of 0.01, and 10 epochs in a Graphics Processing Unit-enabled Google Colab environment. Model evaluation and checkpoint saving occurred after every epoch. The checkpoint with the highest validation macro F1-score was selected, after which the independent test set was evaluated once. Accuracy, macro precision, macro recall, macro F1, and weighted F1 were calculated with zero-division handling. The implementation used Transformers 4.48.3, Datasets 3.2.0, Accelerate 1.3.0, and scikit-learn 1.6.1.

4. DATASET STATISTICS AND QUALITY CONTROL

The final release of AraDigitalContent77 comprises 13,090 utterances distributed evenly across 77 intents (170 utterances per intent), with zero duplicate sentences and zero missing values. Table 1 summarizes key corpus statistics.

To characterize lexical diversity beyond a single corpus-level value, per-intent TTR was recalculated across all 77 intents using Unicode word tokens (punctuation excluded; no stemming or lemmatization). The mean per-intent TTR was 0.1368 (median = 0.1346, sample SD = 0.0184), with values ranging from 0.0977 for profile_not_loading to 0.1976 for platform_subscription_fee. These statistics show that lexical diversity is not uniform across intents and provide a more fine-grained description of the corpus than the mean value alone.

The released split files contain 9,163 training, 1,963 validation, and 1,964 test utterances. Because 170 examples per intent cannot be divided into equal integer 15% subsets, individual validation and test supports are 25 or 26 examples per intent while preserving stratification and full class coverage.

Table 1 Summary Statistics for AraDigitalContent77

Metric	Value
Total utterances	13,090
Number of intents	77
Utterances per intent	170 (perfectly balanced)
Duplicate sentences	0 (0%)
Missing values	0
Average sentence length	9.3 words
Average per-intent Type-Token Ratio (TTR)	0.137
Unique vocabulary size	3,834 tokens
Unique opening word trigrams	4,047

5. EXPERIMENTAL RESULTS

Table 2 reports training loss and validation performance across 10 epochs. Epoch 8 produced the highest validation macro F1-score (97.25%) and was selected before the independent test set was evaluated.

The selected checkpoint achieved 96.59% accuracy and 96.61% macro F1 on the 1,964-utterance independent test set. The equality of macro and weighted F1 after rounding is consistent with the nearly balanced per-intent test support (Table 3).

5.1 Per-Intent and Error Analysis

The classifier made 67 errors among 1,964 test utterances. The lowest per-intent F1-scores were observed for subscription_fee_charged (86.21%) and platform_subscription_fee (87.50%), reflecting semantic overlap among fee and charge categories. The most frequent confusion was watch_history_not_updated being predicted as watch_history_missing (four cases). Three unexpected_charge_on_invoice utterances were predicted as subscription_fee_charged. Several other confusions occurred twice, including download_limit_reached versus concurrent_streams_limit_reached and content_sharing_issue versus change_privacy_settings. These patterns indicate that remaining errors are concentrated mainly at closely related class boundaries rather than broadly distributed across the taxonomy (Table 4).

5.2 Contextual Comparison with Prior Benchmarks

Casanueva et al. (2020) reported approximately 93% on the English Banking77 benchmark, while Jarrar et al. (2023) reported an F1-score of 92.09% for the MSA portion of ArBanking77. AraDigitalContent77 produced a 96.61% macro F1-score on its independent test split. These values provide context only: the datasets differ in language composition, domain, construction process, data split, and evaluation protocol. Consequently, the present result is treated as a baseline for AraDigitalContent77 and not as evidence that the model or dataset is intrinsically superior to Banking77 or ArBanking77.

Table 2 Training and Validation Performance by Epoch

Epoch	Training loss	Validation loss	Validation accuracy	Validation macro F1
1	2.3372	0.8975	75.90%	73.97%
2	0.5686	0.3716	89.91%	89.44%
3	0.2140	0.2501	93.68%	93.30%
4	0.1023	0.2662	94.55%	94.50%
5	0.0633	0.2306	96.13%	96.12%
6	0.0402	0.2075	96.79%	96.79%
7	0.0187	0.2179	96.94%	96.94%
8 (best)	0.0114	0.2067	97.25%	97.25%
9	0.0043	0.2265	97.15%	97.14%
10	0.0027	0.2305	96.99%	96.99%

Table 3 Final Independent Test-Set Performance

Metric	Test result
Accuracy	96.59%
Macro precision	96.80%
Macro recall	96.59%
Macro F1-score	96.61%
Weighted F1-score	96.61%

Table 4 Most Frequent Inter-Intent Confusions

Confused intent pair	Number of test errors
watch_history_not_updated → watch_history_missing	4
unexpected_charge_on_invoice → subscription_fee_charged	3
download_limit_reached / concurrent_streams_limit_reached	2
content_sharing_issue / change_privacy_settings	2

6. DISCUSSION

The independent test result indicates that the adapted taxonomy supports reliable intent separation within the digital-content domain.

The reduction from 97.25% validation macro F1 to 96.61% test macro F1 is modest and expected when performance is measured on data excluded from both training and model selection. This evaluation design provides a more defensible estimate than the previous validation-only result.

The error analysis also clarifies where the taxonomy remains challenging. Confusions cluster around semantically adjacent categories, especially fee types, watch-history states, access limits, and privacy-related sharing actions. These findings support future refinement of intent definitions and boundary-focused data collection. The result should nevertheless be interpreted within this dataset rather than compared directly across unrelated benchmark protocols. Per-intent TTR varies across the taxonomy (0.0977–0.1976), indicating lexical heterogeneity; separately, the observed misclassifications provide direct evidence of difficulty at specific, semantically adjacent class boundaries. TTR is therefore used here as a lexical-diversity measure rather than as a proxy for semantic separability.

The choice to construct AraDigitalContent77 through original Arabic authorship and LLM generation, rather than machine translation from

an English source, is further supported by He et al.'s (2026) finding that translated evaluation sets can overestimate real-world model robustness by a measurable margin. While our dataset was not evaluated against a separately collected native test set, this consideration motivated our reliance on human-written seeds and Arabic-native LLM generation throughout the pipeline, rather than translating Banking77's original English utterances.

7. LIMITATIONS

The dataset is restricted to Modern Standard Arabic; no dialectal variants are currently included, despite evidence that dialectal augmentation can improve robustness (Hossain et al., 2025).

Most utterances were generated or expanded synthetically. Although the pipeline included human-written seeds, automated filtering, selective human review, and intentional linguistic noise, the test set is not an independently collected native-user corpus. In addition, the semantic-drift filter was intended to improve label consistency but may also remove naturally ambiguous borderline utterances, potentially producing cleaner class boundaries than unconstrained real-world user traffic.

The stratified random split prevents exact record overlap, but the dataset does not contain generation-lineage identifiers that would support group-based separation of paraphrases derived from the same seed. Consequently, semantic-family overlap across partitions cannot be ruled out, and the reported 96.61% macro F1 should be interpreted as a record-disjoint within-dataset baseline rather than evidence of seed-family-independent generalization.

The reported baseline is based on one AraBERT configuration and one fixed random seed. Multi-seed evaluation, comparisons with additional Arabic encoders, full component ablation of the generation pipeline, threshold-sensitivity analysis, embedding-based inter-intent similarity analysis, and a prospectively designed blinded human naturalness evaluation remain future work.

8. CONCLUSION AND FUTURE WORK

This paper introduced AraDigitalContent77, an Arabic intent classification dataset for digital content platforms, together with an Action-Object-Goal schema adaptation methodology and a five-stage hybrid data-generation pipeline. The 13,090-utterance corpus contains 77 balanced intents. Under a stratified 70/15/15 evaluation protocol, the selected AraBERT checkpoint achieved 96.59% test accuracy and 96.61% test macro F1. These results establish a strong within-dataset baseline while the per-intent analysis identifies remaining ambiguity among closely related intent boundaries.

Future work will extend the dataset to Arabic dialects, collect a smaller native human-authored test set, preserve generation lineage for group-based splitting, evaluate multiple random seeds and additional encoders, perform controlled ablation and threshold-sensitivity studies on retained intermediate corpora, add embedding-based inter-intent similarity analysis and a prospectively designed blinded human naturalness evaluation, and refine the intent pairs identified by the error analysis.

DATA AVAILABILITY STATEMENT

The dataset and reproducible splits are publicly released under a CC BY 4.0 license.

AUTHOR CONTRIBUTIONS

The sole author performed the conceptualization, methodology design, data curation, software development, validation, formal analysis, investigation, visualization, and manuscript preparation, and approved the final version.

FUNDING STATEMENT

This research received no external funding.

CONFLICT OF INTEREST STATEMENT

The author declares no conflict of interest.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable. The study did not involve human participants, private user conversations, or personally identifiable information. The utterances were created manually and through controlled artificial-intelligence-assisted generation.

DECLARATION OF GENERATIVE AI USE

LLMs were used as controlled generators of candidate utterances within the documented data-construction methodology. Generated outputs were subjected to automated filtering and selective human review. Artificial intelligence systems are not listed as authors and bear no responsibility for the work.

ACKNOWLEDGMENTS

The author thanks Dr. Sawsan Asjia for her academic supervision and guidance during this research.

ORIGINALITY DECLARATION

This manuscript is original, has not been published previously, and is not under consideration by another journal or conference.

REFERENCES

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

Casanueva, I., Temčinas, T., Gerz, D., Henderson, M., & Vulić, I. (2020). Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for Conversational AI* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.nlp4convai-1.5

Castillo-López, G., Lombard, A., Semmar, N., & de Chalendar, G. (2026). How DDAIR you? Disambiguated data augmentation for intent recognition. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 274–286). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.eacl-short.20

He, H., Zhuang, J., Chu, H., Yu, S., J&T AI Group, Wang, H., & Han, K. (2026). *From synthetic to native: Benchmarking multilingual intent classification in logistics customer service* (arXiv:2603.23172). arXiv. https://doi.org/10.48550/arXiv.2603.23172

Hossain, S., Shammary, F., Shammary, B., & Afli, H. (2025). Enhancing dialectal Arabic intent detection through cross-dialect multilingual input augmentation. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)* (pp. 44–49). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.wacl-1.5

Jarrar, M., Birim, A., Khalilia, M., Erden, M., & Ghanem, S. (2023). ArBanking77: Intent detection neural model and a new dataset in modern and dialectical Arabic. In *Proceedings of ArabicNLP 2023* (pp. 276–287). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.arabicnlp-1.22

Lin, I.-F., Hasibi, F., & Verberne, S. (2024). Generate then refine: Data augmentation for zero-shot intent detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 13138–13146). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.768

Pecher, B., Cegin, J., Belanec, R., Srba, I., Simko, J., & Bielikova, M. (2026). *Better as generators than classifiers: Leveraging LLMs and synthetic data for low-resource multilingual classification* (arXiv:2601.16278). arXiv. https://doi.org/10.48550/arXiv.2601.16278

Shin, J., Ahn, Y., Won, S., & Choi, S. J. (2024). Learning to adapt large language models to one-shot in-context intent classification on unseen domains. In *Proceedings of the 2024 CustomNLP4U Workshop* (pp. 182–197). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.customnlp4u-1.15