AlHaram Analytics: A Multilingual Data Quality Taxonomy and Preprocessing Framework for Islamic Service Application Reviews

Naila Marir
Department of Computer Science
Effat University, Jeddah, Saudi Arabia
nmarir@effatuniversity.edu.sa

January 3, 2026

Abstract

Mobile applications supporting Islamic services—including pilgrimage management, religious guidance, and government services—generate millions of multilingual user reviews annually. However, existing Natural Language Processing (NLP) preprocessing pipelines fail to address the unique linguistic, cultural, and temporal characteristics of this domain. We present **AlHaram Analytics**, a comprehensive data quality taxonomy and preprocessing framework specifically designed for multilingual Islamic service application reviews. Our framework introduces two key innovations: (1) a four-level **Data Quality Taxonomy** encompassing lexical, linguistic, Islamic contextual, and demographic dimensions, and (2) a **Temporal-Cultural Context Injection** mechanism that leverages Islamic calendar periods as cross-linguistic semantic unifiers. We evaluate our framework on a dataset of 57,717 reviews from Saudi Arabian government and religious service applications, demonstrating significant improvements in data standardization, with 89.45% completeness rate and successful sentiment classification using CAMeL-BERT deep learning models. Our results show that Islamic temporal context (Hajj, Ramadan, Eid periods) provides a powerful shared semantic layer that enables meaningful cross-linguistic analysis across Arabic, English, and mixed-language reviews. The framework is released as an open-source toolkit with web interface, REST API, and Python library.

Keywords: Arabic NLP, multilingual text processing, sentiment analysis, Islamic calendar, mobile application reviews, data preprocessing, data quality taxonomy

1 Introduction

The Kingdom of Saudi Arabia hosts over 15 million pilgrims annually for Hajj and Umrah, supported by a sophisticated ecosystem of mobile applications spanning pilgrimage management (Nusuk, Eatmarna), transportation (Makkah Buses, Haramain Train), healthcare (Sehhaty), and government services (Tawakkalna). These applications generate vast quantities of user reviews in Arabic, English, and numerous other languages, representing an invaluable resource for understanding user experience and service quality in religiously significant contexts.

However, analyzing this data presents unique challenges that existing NLP preprocessing pipelines fail to address:

1. **Arabizi Prevalence**: Users frequently write Arabic using Latin characters with numeral substitutions (e.g., “7abibi” for “habibi”), requiring specialized transliteration
2. **Code-Mixing**: Reviews exhibit extensive Arabic-English mixing, Islamic terminology across languages, and dialectal variations
3. **Temporal-Religious Context**: User behavior and sentiment are profoundly influenced by Islamic calendar events (Hajj season, Ramadan, Eid celebrations)

4. **Cultural Semantics**: The same complaint carries different severity during peak pilgrimage versus regular periods

We argue that effective preprocessing of Islamic service reviews requires not merely linguistic normalization, but a comprehensive *data quality taxonomy* that explicitly models the cultural, temporal, and religious dimensions unique to this domain.

1.1 Research Contributions

This paper makes the following contributions:

1. **Data Quality Taxonomy**: We formalize a four-level taxonomy for multilingual Islamic service analytics: (1) Lexical Quality, (2) Linguistic Quality, (3) Islamic Contextual Quality, and (4) Demographic Quality
2. **Temporal-Cultural Context Injection**: We introduce a mechanism where Islamic calendar periods serve as cross-linguistic semantic anchors, enabling unified analysis across diverse languages
3. **Extraction-Based Text Cleaning**: Unlike traditional approaches that discard information, our text cleaner extracts URLs, emojis, hashtags, and mentions into separate columns while preserving them for downstream analysis
4. **Comprehensive Evaluation**: We evaluate our framework on 57,717 reviews, providing empirical evidence for preprocessing effectiveness and sentiment analysis performance
5. **Open-Source Toolkit**: We release AlHaram Analytics as a complete open-source framework with web interface, REST API, and Python library

2 Related Work

2.1 Arabic NLP and Preprocessing

Arabic text preprocessing has received significant attention in the NLP community. Farghaly and Shaalan [1] provided a comprehensive survey of Arabic NLP challenges, including morphological complexity, dialectal variation, and lack of diacritization in informal text. More recently, the development of AraBERT [2] and CAMeL-BERT [3] has advanced Arabic language understanding through pre-trained transformer models.

However, existing Arabic preprocessing tools focus primarily on Modern Standard Arabic (MSA) and fail to address domain-specific challenges such as Arabizi conversion in user-generated content or Islamic terminology normalization.

2.2 Multilingual Sentiment Analysis

Cross-lingual sentiment analysis has evolved from translation-based approaches [4] to multilingual transformer models like mBERT and XLM-RoBERTa [5]. For Arabic sentiment analysis, studies have explored both lexicon-based [6] and deep learning approaches [7].

Our work differs by introducing *contextual modulation* where sentiment interpretation is adjusted based on Islamic temporal context—a "crash" during Hajj carries different semantic weight than during regular periods.

2.3 Domain-Specific Data Quality

Data quality frameworks in NLP typically focus on generic dimensions: accuracy, completeness, consistency, and timeliness [8]. Domain-specific taxonomies have been developed for healthcare [9] and financial NLP [10], but no comprehensive framework exists for Islamic service analytics.

2.4 Islamic Calendar in Computing

The Hijri calendar has been studied primarily for date conversion algorithms [11]. To our knowledge, no prior work has leveraged Islamic calendar periods as semantic features for NLP tasks or used them as cross-linguistic unification mechanisms.

3 Data Quality Taxonomy

We propose a four-level taxonomy specifically designed for multilingual Islamic service application reviews. Each level addresses distinct quality dimensions that must be satisfied for effective downstream analysis.

3.1 Level 1: Lexical Quality

Lexical quality ensures consistent character-level representation across diverse writing systems and conventions.

3.1.1 Script Normalization

Arabizi Resolution: We implement a comprehensive mapping for Arabic text written in Latin characters with numeral substitutions:

Table 1: Arabizi Character Mapping

Numeral	Arabic	Sound	Example
2	/	Hamza	a2mad → ahmad
3		Ain	3ali → ali
5		Kha	5aled → khaled
7		Ha	a7mad → ahmad

Arabic Character Normalization: We unify Alef variants (, , , →), normalize Alef Maksura (→), and remove diacritics (tashkeel) while flagging their presence.

3.1.2 Identity Normalization

Usernames require cleaning that preserves cultural patterns while standardizing format:

- Remove trailing digits (“Hassan855” → “Hassan”)
- Strip emojis and special characters
- Apply Arabizi conversion
- Mark names <3 characters as “Anonymous”

3.1.3 Information Extraction

Unlike traditional cleaners that discard elements, we *extract* information to dedicated columns:

$$\text{clean}(t) \rightarrow \{t_{clean}, E_{urls}, E_{emojis}, E_{hashtags}, E_{mentions}\} \tag{1}$$

This preserves emojis for sentiment analysis, URLs for reference tracking, and hashtags for topic analysis.

3.2 Level 2: Linguistic Quality

Linguistic quality addresses language identification and code-mixing patterns.

3.2.1 Language Detection Algorithm

We employ a multi-strategy approach combining the langid library with Unicode range analysis:

Algorithm 1 Language Detection

```
**Input:** Text $t$
$arabic\_ratio \leftarrow \frac{|\{c \in t : c \in [U+0600, U+06FF]\}|}{|t|}$
$latin\_ratio \leftarrow \frac{|\{c \in t : c \in [A-Za-z]\}|}{|t|}$
**if** $arabic\_ratio > 0.5$ **then**
    **return** “Arabic”
**else if** $latin\_ratio > 0.5$ **then**
    $lang \leftarrow$ langid.classify($t$)
    **return** $lang$
**else if** $arabic\_ratio > 0.1$ AND $latin\_ratio > 0.1$ **then**
    **return** “Mixed”
**else**
    **return** “Unknown”
**end if**
```

3.2.2 Quantitative Text Features

We extract 14 quantitative features for downstream analysis:

Table 2: Text Feature Categories

Category	Features
Length	char_count, word_count, sentence_count
Script	arabic_ratio, latin_ratio, digit_ratio
Lexical	unique_words, lexical_diversity (TTR)
Punctuation	exclamation_count, question_count

3.3 Level 3: Islamic Contextual Quality

This level introduces domain-specific enrichment that captures the unique temporal and service context of Islamic service applications.

3.3.1 Temporal-Religious Tagging

We convert Gregorian dates to Hijri calendar and tag reviews with relevant Islamic periods:

3.3.2 Service Domain Classification

Applications are classified into service categories based on predefined mappings:

Table 3: Islamic Period Definitions

Period	Hijri Date	Significance
Hajj Season	Dhul Hijjah 1-15	Peak pilgrimage
Eid al-Adha	Dhul Hijjah 10-13	Festival of Sacrifice
Ramadan	Month 9 (full)	Fasting month
Eid al-Fitr	Shawwal 1-3	End of Ramadan
Regular	Other dates	Normal operation

- **Pilgrimage**: Nusuk, Eatmarna, Tawafa
- **Transportation**: Makkah Buses, HHR Train
- **Healthcare**: Sehhaty, Asaafni
- **Government**: Tawakkalna, Absher
- **Religious**: Qibla Finder, Haramain Quran

3.4 Level 4: Demographic Quality (Optional)

Gender prediction from usernames using ensemble transformer models:

$$g_{final} = \begin{cases} g_1 & \text{if } g_1 = g_2 \\ g_1 & \text{if } c_1 \geq 0.80 \\ g_2 & \text{if } c_2 \geq 0.80 \\ \text{unknown} & \text{otherwise} \end{cases} \quad (2)$$

where g_i and c_i are predictions and confidences from model i.

4 Temporal-Cultural Context Injection

4.1 Core Insight

We propose that Islamic temporal context serves as a **cross-linguistic semantic unifier**. The same review text carries different semantic weight depending on the Islamic period during which it was written.

4.2 Cross-Linguistic Alignment Hypothesis

We hypothesize that:

$$\text{Sim}(R_{L_1}^{Hajj}, R_{L_2}^{Hajj}) > \text{Sim}(R_{L_1}^{Hajj}, R_{L_1}^{Regular}) \quad (3)$$

Period	Severity	Context
Regular	Medium	User can retry later
Hajj Season	**Critical**	Religious obligation at stake
Ramadan	High	Fasting, time-sensitive
Eid	High	Family gathering disrupted

Figure 1: Semantic Modulation by Islamic Period for the review: “The app crashed and I couldn’t find my bus”

That is, reviews from different languages during the same Islamic period are more semantically similar than reviews from the same language during different periods.

4.3 Cultural Signal Encoding

The period tag encodes multiple implicit signals:

- **+spiritual_urgency**: Hajj, Ramadan periods
- **+crowd_stress**: Hajj, Eid, Friday prayers
- **+time_sensitivity**: Ramadan iftar times, Hajj rituals
- **+foreign_user_likelihood**: Hajj, Umrah seasons
- **+family_context**: Eid celebrations

5 Sentiment Analysis

5.1 Deep Learning Approach

We employ CAMeL-BERT (CAMeL-Lab/bert-base-arabic-camelbert-mix-sentiment), a transformer model fine-tuned for Arabic sentiment classification that handles both Modern Standard Arabic and dialectal variations.

The model produces three outputs:

- **sentiment**: Categorical label (positive, neutral, negative)
- **sentiment_score**: Continuous score in [-1, +1]
- **sentiment_confidence**: Model confidence in [0, 1]

5.2 Lexicon-Based Fallback

For environments without GPU support, we provide a lexicon-based analyzer using curated Arabic and English sentiment word lists plus emoji sentiment mapping:

$$s(t) = \frac{\sum_{w \in t} \text{polarity}(w) + \sum_{e \in t} \text{emoji\_sentiment}(e)}{|t| + |e|} \tag{4}$$

6 System Architecture

AlHaram Analytics is implemented as a modular Python framework with three interfaces:

6.1 Component Architecture

- **Preprocessing Module**: TextCleaner, UsernamePreprocessor, LanguageDetector
- **Feature Engineering**: TextFeatureExtractor, PeriodTagger, ServiceClassifier, DeviceMapper
- **Sentiment Module**: SentimentAnalyzer (deep learning), SimpleSentimentAnalyzer (lexicon)
- **Analytics Module**: DatasetAnalyzer, DatasetVisualizer
- **Gender Prediction**: HuggingFace ensemble classifier (optional)

6.2 Interfaces

1. **Web Application**: Flask-based UI with drag-drop upload, step-by-step preview, and interactive visualization
2. **REST API**: Programmatic access with endpoints for upload, transform, download, and analytics
3. **Python Library**: Direct integration via `from alharam_analytics import *`

6.3 Scikit-learn Compatibility

All transformers implement the scikit-learn interface:

```
class TextCleaner:
    def fit(self, X, y=None):
        return self
    def transform(self, X):
        # Processing logic
        return X_transformed
```

7 Experimental Evaluation

7.1 Dataset

We collected 57,717 user reviews from Saudi Arabian mobile applications:

Table 4: Dataset Statistics

Metric	Value
Total Reviews	57,717
Original Columns	22
Post-Processing Columns	63
Memory Usage	90.41 MB

7.2 Data Quality Results

Table 5: Data Quality Metrics

Metric	Value
Completeness Rate	89.45%
Duplicate Rate	0.00%
Valid Samples	57,717 (100%)

7.3 Language Distribution

Table 6: Language Distribution

Language	Count	Percentage
English	47,296	81.94%
Arabic	8,922	15.46%
Unknown	1,018	1.76%
Mixed	481	0.83%

7.4 Sentiment Analysis Results

Using CAMeL-BERT deep learning model:

Table 7: Sentiment Distribution

Sentiment	Count	Percentage
Positive	44,105	76.42%
Negative	11,289	19.56%
Neutral	2,323	4.02%
Avg. Score	0.4045	

Table 8: Islamic Period Distribution

Period	Count	Percentage
Regular	31,860	55.20%
School Summer	16,672	28.89%
Ramadan	5,828	10.10%
Hajj Season	3,096	5.36%
Eid al-Fitr	261	0.45%

7.5 Period Distribution

7.6 Sentiment Evaluation

We evaluated sentiment predictions against star ratings as proxy ground truth (1-2⋆ = negative, 3⋆ = neutral, 4-5⋆ = positive):

Table 9: Sentiment Classification Performance

Class	Precision	Recall	F1
Negative	0.158	0.133	0.144
Neutral	0.068	0.060	0.064
Positive	0.731	0.773	0.751
Macro Avg	0.319	0.322	0.320
Weighted Avg	0.568	0.592	0.579
Accuracy	59.20%		

Note: The 59.20% accuracy reflects *text-rating alignment*, not model accuracy. Star ratings reflect overall user satisfaction, while sentiment analysis captures text content—these may legitimately diverge (e.g., "App crashes but I love it" = positive text, could be any rating).

Table 10: Text Feature Statistics

Metric	Value
Avg. Word Count	7.4
Median Word Count	2.0
Avg. Character Count	39.4
Avg. Arabic Ratio	16.0%
Avg. Lexical Diversity	0.96

Table 11: Device Platform Distribution

Platform	Count	Percentage
Android	54,772	94.90%
iOS	2,945	5.10%

Table 12: Processing Time (57,717 reviews)

Component	Time
Text Cleaning	∼2 min
Username Processing	∼30 sec
Language Detection	∼3 min
Feature Extraction	∼1 min
Period Tagging	∼2 min
Sentiment (CAMeL-BERT)	∼30 min
Total Pipeline	∼40 min

7.7 Text Statistics

7.8 Device Distribution

7.9 Processing Performance

8 Discussion

8.1 Key Findings

1. **English Dominance**: Despite targeting Saudi Arabian applications, 81.94% of reviews are in English, reflecting the international user base during Hajj/Umrah seasons

2. **Positive Sentiment Majority**: 76.42% positive sentiment aligns with generally favorable app store ratings for government-supported applications

3. **Temporal Clustering**: 44.80% of reviews occur during religious or holiday periods (Hajj, Ramadan, Eid, School Summer), confirming the importance of temporal context

4. **Short Reviews**: Median word count of 2.0 indicates predominantly brief feedback, requiring specialized NLP approaches for short text

5. **Android Dominance**: 94.90% Android users reflects regional device preferences

8.2 Cross-Linguistic Analysis Potential

The period-tagged dataset enables novel research questions:

- Do Indonesian and Arabic speakers report similar pain points during Hajj?

- Does sentiment negativity correlate with pilgrimage stage across languages?

- Which service categories receive most criticism during peak periods?

8.3 Limitations

1. **Ground Truth**: Sentiment evaluation used star ratings as proxy; human annotation would provide more accurate assessment

2. **Language Coverage**: Framework optimized for Arabic/English; other languages (Urdu, Indonesian, Turkish) require additional validation

3. **Temporal Scope**: Dataset may not capture all Islamic events or regional variations in observance

9 Conclusion

We presented AlHaram Analytics, a comprehensive data quality taxonomy and preprocessing framework for multilingual Islamic service application reviews. Our key contributions—the four-level data quality taxonomy and temporal-cultural context injection—address previously unmet needs in Arabic NLP and domain-specific text processing.

The framework successfully processes 57,717 reviews with 89.45% completeness, extracts 14 quantitative text features, tags reviews with Islamic calendar periods, and performs deep learning sentiment analysis using CAMeL-BERT. Our extraction-based approach to text cleaning preserves information that traditional methods discard, enabling richer downstream analysis.

We release AlHaram Analytics as open-source software, providing the research community with tools for Islamic service analytics and establishing a foundation for culturally-situated, data-centric NLP research.

9.1 Future Work

1. **Topic Modeling**: Automatic extraction of discussion topics using BERTopic

2. **Aspect-Based Sentiment**: Fine-grained sentiment on specific aspects (UI, performance, reliability)

3. **Cross-Linguistic Validation**: Empirical testing of the cross-linguistic alignment hypothesis

4. **Real-Time Pipeline**: Streaming processing for continuous review monitoring

Acknowledgments

The author thanks Effat University for supporting this research and the Saudi government agencies that develop the applications studied in this work.

References

[1] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Transactions on Asian Language Information Processing*, vol. 8, no. 4, pp. 1–22, 2009.

[2] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," in *Proc. 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, 2020, pp. 9–15.

[3] G. Inoue, B. Alhafni, N. Baimber, H. Bouamor, and N. Habash, "The interplay of variant, size, and task type in Arabic pre-trained language models," in *Proc. Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 92–104.

[4] X. Zhou, X. Wan, and J. Xiao, "Cross-lingual sentiment classification with bilingual document representation learning," in *Proc. 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1403–1412.

[5] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451.

[6] M. Abdul-Mageed, M. Diab, and S. Kübler, "SAMAR: Subjectivity and sentiment analysis for Arabic social media," *Computer Speech & Language*, vol. 28, no. 1, pp. 20–37, 2014.

[7] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, "Deep learning for Arabic sentiment analysis: A systematic literature review," *IEEE Access*, vol. 7, pp. 7092–7117, 2019.

[8] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.

[9] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 144–151, 2013.

[10] T. Loughran and B. McDonald, "Textual analysis in accounting and finance: A survey," *Journal of Accounting Research*, vol. 54, no. 4, pp. 1187–1230, 2016.

[11] E. M. Reingold and N. Dershowitz, *Calendrical Calculations: The Ultimate Edition*. Cambridge University Press, 2018.