Gender Identification of Arabic Names: A Comparative Analysis of Morphological, Semantic and Deep Learning Approaches

Minseok Kim
New York University Abu Dhabi
mk7545@nyu.edu

Abstract

This project presents the development and evaluation of a machine learning framework for the binary classification of Arabic first names. Utilizing a dataset of over 13,000 labeled names, we initially conduct a comparative analysis of three distinct methodologies: a logistic regression model based on morphological features, a geometric approach using word embeddings, and a fine-tuned pre-trained Transformer. While our results indicated that the morphological approach achieved the highest baseline accuracy at 76.69% – suggesting that explicit linguistic patterns are the primary predictors – it faced coverage limitations regarding ambiguous names. To address this, we developed a Hybrid Approach that cascades from high-precision morphological rules to a deep learning fallback. This composite model significantly outperforms individual baselines, achieving an accuracy of **91.68%** and effectively combining the structural precision of linguistics with the robustness of neural networks.

1 Introduction

Gender identification of Arabic names constitutes a fundamental prerequisite for several essential tasks in Natural Language Processing (NLP). Accurate gender classification contributes significantly to syntactic and morphological correctness in processing Arabic.

The potential utility of a reliable Arabic name gender classifier extends across several Arabic NLP domains:

Machine Translation (MT) A primary application lies in improving the robustness of Machine Translation systems that handle Arabic. While the gender of a name may be inferred in context, a pre-classified name allows the MT system to maintain grammatical coherence when translating from Arabic to other languages. Crucially, the classifier ensures correct inflections for verbs, adjectives, and pronouns that must agree with the named entity's gender.

Conversational AI and Dialogue Systems In a dialogue system, identifying the gender of the user's name is essential for natural and grammatically correct interaction in Arabic. Unlike English, Arabic requires distinct morphological agreements for the second person. A name-to-gender classifier allows chatbots and virtual assistants to appropriately conjugate verbs and select gender-specific pronouns, thereby guaranteeing linguistic fluency.

Pronoun Assignment Gender information is crucial for linking pronouns to the correct names in Arabic text. By classifying an ambiguous name (like *Nour*) as male or female, the system can accurately match it to the correct subsequent pronoun, such as *huwa* (he) or *hiya* (she). This allows for a deeper understanding of the document's structure and meaning.

Linguistic Challenge Despite its utility, the classification of gender in Arabic names presents a unique challenge. Unlike simpler linguistic systems (such as Spanish) where explicit suffixes largely determine gender, Arabic bases on a complex morphological system. While some clues exist, such as the *Teh Marbuta*, many names are gendered based on implicit patterns or traditional usage rather than a clear ending. This difficulty is worsened by the lack of diacritics and the existence of unisex names.

This study aims to develop a reliable classification framework capable of distinguishing masculine and feminine Arabic names with high precision. We conduct a comparative analysis of three methodologies: a linear approach using Logistic Regression with engineered morphological features, a geometric approach using the semantic vector space of word embeddings, and a transfer learning approach utilizing pre-trained Arabic

Transformers. By systematically evaluating these methods, this work establishes a robust baseline for Arabic NLP applications.

2 Related Work

Gender classification based on proper names has been studied in English (Cassidy et al., 1999) and Korean contexts (Yoon et al., 2008). However, research focused specifically on gender identification of Arabic names remains limited. In Arabic NLP, much of the research related to gender has focused on gender rewriting (adapting phrases to match a desired gender) (Alhafni et al., 2022) or author gender identification (inferring gender from the style and content of a tweet or post) (Alsmearat et al., 2017). In contrast, the fundamental task of classifying the gender based solely on the orthographic and morphological structure of the Arabic name itself has received comparatively little research attention.

A primary reference in this domain is the work of (Almelabi and Baashirah, 2022), who evaluated the performance of six machine learning algorithms, such as Support Vector Machines, Random Forest, Naive Bayes, etc., alongside a deep learning model (LSTM) for gender prediction. Their study utilized a dataset of approximately 21,000 Arabic names and treated names as simple bag-of-word tokens. They reported high accuracy rates, particularly with Random Forest (98%).

While they established a strong benchmark for name-to-gender classification based on neural networks algorithms, our study diverges and advances in three distinct ways:

Feature Depth vs. Surface Statistics Almelabi and Baashirah's approach treats the name as a string, ignoring the internal derivational structure of Arabic. In contrast, our Logistic Regression approach will take morphological features into account by utilizing **CAMeL Tools** to decompose the structure. We extract the roots and morphological variations, allowing our model to learn that specific templates are inherently gendered, rather than simply memorizing character strings.

Transformer Architectures vs. RNN The deep learning component of the previous study utilized Long Short-Term Memory (LSTM) networks, which utilize RNN. While they are effective for sequence modeling, they are often deprecated for modern attention-based models. Our study moves beyond RNNs to utilize Transformer-based Pre-trained Model **AraBERT**. This model employs learning from billions of Arabic tokens, allowing for a more robust semantic understanding of names that may be ambiguous.

Semantic Approach Finally, previous work has focused exclusively on supervised classification of name strings. We introduce a semantic approach by constructing **Gender Centroids** in the vector space. This allows us to measure gender associations via Cosine Similarity, offering an interpretable view of how names cluster semantically without the need for a neural network.

3 Data

To evaluate our proposed models, we utilize the open-source Muslim Names Dataset sourced from the Hugging Face repository.[1]

3.1 Data Statistics

The raw dataset consists of $14,585$ rows. Following preprocessing, the cleaned dataset used for training and evaluation consists of $13,622$ names. The distribution is approximately balanced: Male: 6,864 (50.4%) and Female: 6,758 (49.6%).

Table 1 provides a sample of the raw data structure. Each row consists of Arabic name, English name, Gender and Meaning. The string length of the Arabic names ranges from 1 to 12 characters; length of English names from 2 to 19 characters; the associated meanings from 2 to 320 characters.

Arabic	**English**	**Gender**	**Meaning**
عبد	Aabad	Male	A great worshiper...
ابن	Aaban	Male	Clear, Eloquent...

Table 1: Sample entries from the Muslim Names Dataset showing the mapping between orthography and gender.

3.2 Preprocessing and Splitting

Prior to feature extraction, we apply a cleaning pipeline to ensure data quality:

- **Noise Filtering:** We remove entries with value null to prevent the error.
- **Deduplication:** Duplicate name entries mapping to the same gender are removed to prevent inaccuracy.

[1] https://huggingface.co/datasets/takiuddinahmed/muslim-names-dataset

Following preprocessing, the cleaned dataset consisted of 13622 rows – male: 6864; female: 6758. Then, the data is partitioned into three sets to ensure robust evaluation: Training (80%), Validation (10%), and Testing (10%).

3.3 Evaluation Metrics

To rigorously assess the performance of our classifiers, we evaluate based on the following metrics: Accuracy, Precision, Recall, and F1-Score. These are calculated based on the elements of the confusion matrix: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Accuracy Accuracy measures the ratio of correctly predicted observations to the total observations. While useful, it can be misleading if the dataset classes are imbalanced.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision and Recall Precision (or Positive Predictive Value) quantifies the accuracy of positive predictions, while Recall (Sensitivity) measures the ability of the classifier to find all positive samples.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-Score To balance the trade-off between Precision and Recall, we utilize the F1-Score, which is the harmonic mean of the two.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

We report the **Macro-Averaged F1-Score**, which calculates the F1 score for each class (Male and Female) independently and then takes the unweighted mean. This ensures that the model is penalized if it performs poorly on one gender, preventing bias.

Since the dataset used is fairly balanced, accuracy serves as the appropriate overall evaluation metric.

4 Linguistic Facts

To overcome the limitations of shallow analysis, we employ a feature engineering strategy that utilizes CAMeL Tools to extract useful linguistic features. By applying the CAMeL Tools, we decompose each name in the dataset to get the morphological analyses, in order to extract the relevant features.

Morphological Features Our approach is founded on the hypothesis that gender is encoded in the internal features of the name. We focus on the following core morphological components derived from the analyzer's output:

- **Root:** The underlying root is extracted. This feature captures semantic families that often cluster by gender (e.g., roots related to strength or virtue are frequently masculine).

- **Pattern:** The template of the name is extracted. The pattern is a strong indicator that certain templates are historically restricted to masculine or feminine names.

- **Gender:** The internal gender tag assigned by the morphological analyzer itself is not included as a feature to prevent potential bias.

Orthographic & Surface Features In addition to morphology, we compute explicit surface features that are historically predictive of gender. These features serve to complement the deep morphological features and capture patterns in names that may not conform strictly to classical Arabic morphology:

- **Suffix:** We check for the presence of specific feminine suffix indicators: Teh Marbuta (<ة >), Alif Maqsura (<ى>), and Hamza (<ء>) at the end of the name string.

- **Character N-grams:** We extract N-grams (unigrams and bigrams) of the final 1-2 characters of the name. This captures local orthography, such as common feminine name endings that may not be a formal Teh Marbuta but still signal gender (e.g., "-a" for feminine name).

5 Approach

We propose a comparative study leveraging three distinct techniques to address the challenge of Arabic name gender classification. These approaches range from linear models to high-dimensional deep learning architectures.

5.1 Logistic Regression

Our first approach is grounded in linguistic theory. We hypothesize that the gender of an Arabic name is not arbitrary but is encoded in its morphology – specifically its Root and Pattern. By explicitly modeling these components, we aim to construct an interpretable classifier that can identify which morphological templates typically signal masculinity or femininity.

5.2 Word Embedding

The second approach shifts from explicit rules to implicit semantic representation. We utilize the vector space of pre-trained Arabic Language Models. The core hypothesis here is that even without fine-tuning, pre-trained models cluster names in the vector space based on gendered contexts observed during their initial pre-training on massive corpora. We treat classification as a geometric problem, measuring the distance between a name and gender prototypes.

5.3 Finetuning Pre-trained Transformer

The third approach represents the deep learning technique. We adapt a pre-trained Transformer specifically for this task. Unlike the geometric approach which uses the model as a static feature extractor, this method updates the model's internal weights.

5.4 Hybrid Pipeline

To address the limitations of individual models, we developed a cascading hybrid system. This approach prioritizes high-precision linguistic rules while utilizing the deep learning model as a fallback for ambiguous cases. The pipeline proceeds as follows:

1. **Morphological Check:** The name is analyzed using CAMeL Tools. If the Part-of-Speech (POS) tag is identified as a `Noun`, `Proper Noun`, or `Adjective`, the Logistic Regression prediction is accepted.

2. **Deep Learning Fallback:** If the morphological analyzer fails to return a valid analysis or tags the word as a non-nominal class (e.g., Preposition), the system falls back to the Fine-Tuned AraBERT model for prediction.

6 Implementation

This section details the specific tools, algorithms, and hyperparameters employed to execute the approaches defined above.

6.1 Logistic Regression Pipeline

For the morphological approach, we utilize **CAMeL Tools**, a suite of open-source Python tools for Arabic NLP. We train a binary Logistic Regression classifier to predict gender based on the following engineered features:

Feature Extraction We use the CAMeL Tools to decompose each name. We utilized 36 distinct feature tags, resulting in a high-dimensional space of 77,655 vectorized features. From the analysis candidates, we extract:

- **Roots and Patterns:** These categorical features are vectorized.
- **Suffix Indicators:** We implement binary flags for the presence of Teh Marbuta (ة), Alif Maqsura (ى), and Hamza (ء).
- **N-Grams:** Unigrams and bigrams of the final 1-2 characters are generated.

6.2 Static Embedding Construction

For the geometric approach, we utilize a specific model: `AraBERT-v02-base`.

Centroid Computation We extract the vector representation from the model. We construct Centroids for each class using the training split. The Male Centroid ($\vec{V}_{male}$) is computed as:

$$\vec{V}_{male} = \frac{1}{N_m} \sum_{i=1}^{N_m} \vec{v}_i \tag{5}$$

where $\vec{v}_i$ is the embedding of the i-th male name in the training set. The Female Centroid is computed similarly.

Inference For a new name ($name_{new}$), we calculate the Cosine Similarity between its vector $\vec{v}_{new}$ and the centroids, then use argmax. The classification rule is:

$$Pred = argmax_{c \in \{m,f\}} \left(\frac{\vec{v}_{new} \cdot \vec{V}_c}{||\vec{v}_{new}|| \, ||\vec{V}_c||} \right) \tag{6}$$

6.3 Fine-Tuning Transformer

We fine-tuned the **AraBERT-v02-base** model using the Hugging Face Transformers library. The dataset was split 80/10/10. The model was trained for 5 epochs. We observed that the model reached optimal performance at Epoch 2, after which it began to overfit.

7 Results

We evaluated all four approaches on the held-out test set. Table 2 summarizes the accuracy of each method. The Hybrid approach demonstrated superior performance, validating that combining structural rules with deep learning approach yields the best results.

Approach	Accuracy
Morphological (Logistic Reg.)	0.7669
Deep Learning (Fine-Tuning)	0.7400
Semantic (Embeddings)	0.6500
Hybrid (Cascade)	**0.9168**

Table 2: Comparison of Accuracy across all methodologies.

7.1 Morphological Approach Performance

The morphological approach achieved the highest accuracy among the baseline models. The detailed breakdown of precision and recall is shown in Table 3.

Class	Precision	Recall	F1-Score
Female	0.81	0.73	0.77
Male	0.72	0.81	0.77

Table 3: Performance metrics for the Morphological Approach.

We observed a distinct performance trade-off: the model exhibits higher precision for Female names (fewer false positives) but higher recall for Male names (fewer false negatives). Feature importance analysis revealed that the **Last 2 Characters** of the name are the primary dictators of gender.

Interestingly, while specific morphological patterns (such as 12ي3ةٌ) served as strong predictors, the explicit binary flags for *Teh Marbuta* and *Alif Maqsura* were found to be less crucial than expected. This suggests that the model captures these gender signals more effectively through the broader N-gram context (the last 2 characters) rather than through isolated suffix flags.

7.2 Semantic Approach Performance

The Semantic approach (using static embeddings) performed the lowest with an accuracy of 0.65. As shown in Table 4, a deeper look at the class-wise performance reveals a significant disparity.

Class	Precision	Recall	F1-Score
Female	**0.68**	0.59	0.63
Male	0.64	**0.72**	0.68

Table 4: Performance metrics for the Semantic Approach. Note the bias towards Male names in overall F1-score.

This disparity suggests a "Semantic Gender Bias," where male names are more centrally located or predictable within the pre-trained vector space of AraBERT. While the model achieves higher precision for female names (meaning it is careful when predicting "Female"), its lower recall indicates it misses many actual female names, defaulting them to the male class.

7.3 Deep Learning Performance

The Fine-Tuning approach utilizing AraBERT demonstrated a substantial improvement over the baseline configuration. As detailed in Table 5, the fine-tuned model achieved an accuracy of **0.7473** and an F1-Score of **0.7454**, significantly outperforming the baseline which operated near chance levels (Accuracy 0.5102).

Metric	Baseline	Fine-Tuned
Accuracy	0.5102	**0.7473**
F1-Score	0.3379	**0.7454**
Precision	0.2551	0.7518
Recall	0.5000	0.7458

Table 5: Comparison of the Transformer model performance before (Baseline) and after Fine-Tuning.

While the deep learning model proved robust, it did not reach the precision of the feature-engineered morphological model (0.76). Training logs indicated that the model reached optimal performance quickly (at Epoch 2) before beginning to overfit, a behavior likely attributed to the dataset size (13,622 samples) being relatively small for a complex Transformer architecture.

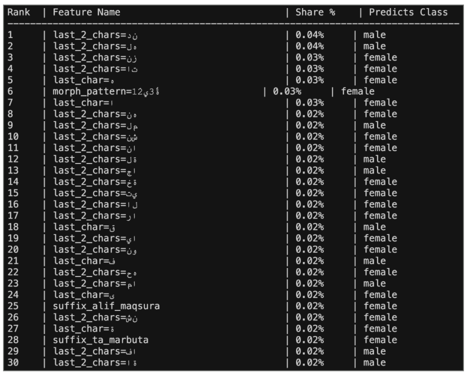
7.4 Hybrid Approach Performance

The Hybrid Cascade model achieved the highest overall performance with an accuracy of **0.9168**.

By strictly applying the Morphological classifier only to names identified as Nouns, Proper Nouns, or Adjectives, and delegating the remaining ambiguous cases to the Fine-Tuned Transformer, the system effectively minimized the weaknesses of both individual models. This result represents a substantial 15% absolute improvement over the single best baseline (Morphological: 0.7669), confirming that the cascading strategy successfully integrates the high precision of linguistic rules with the broad coverage of deep learning.

8 Analysis

Feature Importance An analysis of the logistic regression weights reveals that the **Last 2 Characters** are the single most decisive factor. As shown in Figure 1, specific suffix n-grams occupy the majority of the top feature slots. The `morph_pattern` feature appears at rank 6, confirming that while derivational templates are significant, the explicit ending of the name remains the primary signal for gender in this dataset.



```
Rank  | Feature Name                            | Share %    | Predicts Class
--------------------------------------------------------------------------
1     | last_2_chars=دن                          | 0.04%     | male
2     | last_2_chars=له                          | 0.04%     | male
3     | last_2_chars=نز                          | 0.03%     | female
4     | last_2_chars=ات                          | 0.03%     | female
5     | last_char=ه                              | 0.03%     | female
6     | morph_pattern=12ي3ة                  | 0.03%     | female
7     | last_char=ا                              | 0.03%     | female
8     | last_2_chars=نه                          | 0.02%     | female
9     | last_2_chars=لم                          | 0.02%     | male
10    | last_2_chars=نش                          | 0.02%     | female
11    | last_2_chars=نا                          | 0.02%     | female
12    | last_2_chars=لة                          | 0.02%     | male
13    | last_2_chars=جا                          | 0.02%     | male
14    | last_2_chars=خة                          | 0.02%     | female
15    | last_2_chars=تي                          | 0.02%     | female
16    | last_2_chars=ال                          | 0.02%     | female
17    | last_2_chars=را                          | 0.02%     | female
18    | last_char=ق                              | 0.02%     | male
19    | last_2_chars=يا                          | 0.02%     | female
20    | last_2_chars=نو                          | 0.02%     | female
21    | last_char=ف                              | 0.02%     | male
22    | last_2_chars=حه                          | 0.02%     | female
23    | last_2_chars=ما                          | 0.02%     | male
24    | last_char=ى                              | 0.02%     | female
25    | suffix_alif_maqsura                      | 0.02%     | female
26    | last_2_chars=شن                          | 0.02%     | female
27    | last_char=ة                              | 0.02%     | female
28    | suffix_ta_marbuta                        | 0.02%     | female
29    | last_2_chars=فا                          | 0.02%     | male
30    | last_2_chars=اة                          | 0.02%     | male
```

Figure 1: Top 30 most important features for Gender Prediction. The 'last_2_chars' feature dominates the ranking.

The Structural Advantage The superiority of the Morphological approach (0.77) over the Deep Learning baseline (0.74) highlights the regular structure of Arabic names. Patterns and suffixes serve as high-precision indicators that neural networks may struggle when trained on small datasets.

Analyzer Limitations and the "Blind Spot" Despite its precision, the morphological approach suffered from a critical "blind spot." Error analysis revealed that 94% of the names in our dataset were not identified as proper nouns by the CAMeL Tools analyzer. A frequent failure mode occurred when names coincided with particles; for instance, the name "Ali" was analyzed as the preposition "on" (عليّ), leading to an incorrect or null gender tag.

Success of the Hybrid Strategy The dramatic accuracy jump to **91.68%** with the Hybrid Pipeline confirms that these two models have complementary strengths. The success stems from using Part-of-Speech tags as a confidence filter:

- **Precision Maintenance:** By restricting the Morphological model to only clear nominal tags (Noun, Proper Noun, Adjective), we retained the high-precision rule-based predictions for structured names.
- **Coverage Correction:** For the ambiguous tokens where the analyzer failed (e.g., classifying "Ali" as a preposition), the system effectively "fell back" to the Deep Learning model.

The Deep Learning model, relying on semantic embeddings rather than rigid dictionary lookups, correctly handled these "residual" names, effectively covering the morphological blind spots.

9 Conclusions and Future Work

This study presented a comprehensive comparison of three methods for Arabic name gender identification. We found that a **Morphological Approach** utilizing Logistic Regression and engineered features achieved the highest performance for this dataset (Accuracy 0.77), surpassing both Semantic (0.65) and Fine-Tuned Transformer (0.74) baselines.

Our analysis confirms that Arabic gender is strongly encoded in the suffixes and morphological templates. However, the semantic bias found in embedding models suggests that male names are more centrally located in the vector space of standard Arabic corpora. Future work will focus on examining the source of the gender confidence bias in the semantic prediction approach.

Limitations

Our study faces two primary limitations. First, the **Dataset Size**: while 13,000 names are sufficient for linear models, it is relatively small for fine-tuning large Transformer models, leading to rapid overfitting. Second, the **Binary Constraint**: this work treats gender as a binary classification (Male/Female), which does not account for non-binary identities or unisex names that may vary by context without explicit markers.

Ethics Statement

We acknowledge the ethical implications of building gender classification systems. Assigning gender based on names can be reductive and may misgender individuals, particularly in the case of unisex names or individuals whose gender identity does not align with traditional naming conventions. This tool is intended for linguistic analysis and grammatical agreement tasks (e.g., machine translation correctness) and should not be used for profiling or inferring sensitive personal attributes without user consent. Furthermore, we recognize that the pre-trained models used (AraBERT) may contain inherent biases from their training corpora, which could be reflected in the semantic classification results.

References

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. User-centric gender rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States. Association for Computational Linguistics.

Mohammed Almelabi and Rania Baashirah. 2022. Gender prediction based on Arabic names with machine learning techniques. In *International Conference on Business and Technology*, pages 1248–1255. Springer.

Kholoud Alsmearat, Mahmoud Al-Ayyoub, Riyad Al-Shalabi, and Ghassan Kanaan. 2017. Author gender identification from Arabic text. *Journal of Information Security and Applications*, 35:85–95.

Kimberly Wright Cassidy, Michael H Kelly, and Lee'at J Sharoni. 1999. Inferring gender from name phonology. *Journal of Experimental Psychology: General*, 128(3):362.

Hee-Geun Yoon, Seong-Bae Park, Yong-Jin Han, and Sang-Jo Lee. 2008. Determining gender of Korean names with context. In *2008 International Conference on Advanced Language Processing and Web Information Technology*, pages 121–126. IEEE.