Polarization Detection at POLAR @ SemEval-2026: A MARBERT-based Approach for Arabic Polarization Detection

Milestone Report

Guanghui Ma
guanghui.ma@sabanciuniv.edu

Nuh Al Sharafi
nuh.sharafi@sabanciuniv.edu

Ali Khaled A. Ishtay Altamimi
ali.altamimi@sabanciuniv.edu

Hussein MH Nasser
h.nasser@sabanciuniv.edu

1 Introduction

Polarization refers to the divergence of political attitudes away from the center and towards more extreme views to the far-left or far-right. The internet and communication through it has allowed polarization to spread in ways previously unseen, the veil of anonymity encouraging more extreme behaviors and thoughts. This poses a threat to social cohesion as well as the proliferation of harmful thoughts.

This project's purpose is to detect online polarization as part of POLAR @ SemEval-2026 (Task 9): Detecting Multilingual, Multicultural, and Multievent Online Polarization. The task aims to develop models capable of identifying and classifying polarized content in text form. This project focuses on the Arabic language as part of Subtask 1: Polarization Detection. In this subtask, the goal is to determine whether a given text exhibits polarized opinion or not. Texts are classified into one of two categories: polarized or non-polarized.

Current Progress: Up to this stage, our work has included an in-depth literature review covering related papers in Arabic natural language processing and polarization detection. In parallel, we have carried out preliminary data exploration and analysis on the Arabic dataset provided by the POLAR @ SemEval-2026 challenge organizers, examining aspects such as label balance, text length distribution, and dialectal variation. We have also conducted preprocessing and exploratory data analysis on the datasets used for training as well as beginning to train and evaluate prospective models. This milestone report documents our progress to date and outlines our planned next steps toward the final submission.

2 Dataset Selection

For this project, we use the official Arabic dataset provided by the organizers of the POLAR @ SemEval-2026 (Task 9) challenge. This dataset has been specifically curated for the Polarization Detection subtask and contains social media texts labeled as either polarized or non-polarized. Each instance is annotated based on whether it expresses polarized attitudes, opinions, or language within its contextual setting, thereby removing the need for any additional manual labeling.

This dataset is particularly well suited to our objectives for several reasons. First, it was designed explicitly for polarization detection, ensuring that all included samples are directly relevant to our research goals. Second, the Arabic subset covers a broad range of dialects and regional varieties, offering rich linguistic and cultural diversity that reflects real-world online discourse. This diversity enhances our model's ability to generalize across different forms of Arabic expression. Finally, since all participating teams in the challenge utilize the same dataset, our results can be directly benchmarked against other approaches, allowing for fair comparison and meaningful evaluation of model performance across methodologies.

In addition, in order to have a wider range of data, we used the official English dataset provided by the organizers, and translated it into a variety of Arabic dialects using OpenL translation API to

serve as supplementary training data. This translated corpus expands our training pool, supporting the development of models with improved robustness and cross-linguistic understanding.

The Arabic dataset contains 3,380 entries with three columns: id, text, and polarization (with binary values 1 and 0). The English dataset, which we translated to Arabic, has the same structure with 2,676 entries, providing additional training data that complements the original Arabic corpus.

3 Exploratory Data Analysis and Preprocessing

To better understand the characteristics of the dataset and prepare it for model training, we conducted a multi-stage exploratory data analysis (EDA) followed by both basic and advanced preprocessing tailored to the challenges of Arabic user-generated text. This process allowed us to diagnose the dataset's structural properties, identify sources of noise, and iteratively refine the cleaning pipeline.

Our initial analysis focused on the distribution of the polarization labels. By examining counts and relative frequencies of polarized versus non-polarized instances, we observed an imbalance that has practical implications for model training, particularly when optimizing for metrics such as Macro F1. Visual summaries in the form of bar plots and pie charts were generated to more clearly inspect this skew (see Figure 1 in the Appendix). Alongside label distribution, we analyzed the "shape" of the textual data by computing character- and word-length distributions across the corpus (Figure 2). These distributions revealed that the dataset predominantly consists of short-format social media posts, with a pronounced long-tail of comparatively longer entries. This insight guided decisions about batching strategies and appropriate maximum sequence lengths for fine-tuning. Per-label comparisons of text length and word count are shown in Figure 2.

During the EDA stage, we also examined linguistic and structural artifacts common in Arabic social media content. Notably, many texts contained multilingual elements, including Latin characters and both Arabic-Indic and Western digits, suggesting frequent code-switching. We quantified the presence of these elements and produced label-wise breakdowns to determine whether particular classes exhibited more multilingual behavior (see Figure 3). In addition, we evaluated the prevalence of social media markers such as URLs, user mentions, and hashtags. These elements appeared at non-trivial rates and were shown to vary modestly across classes (Figure 4). A similar analysis was conducted for emoji usage; while emojis occurred less frequently overall, a meaningful proportion of texts still included at least one emoji (Figure 5). We extracted representative sets and computed distribution statistics to determine whether emojis should be preserved, removed, or normalized during preprocessing.

With these insights established, we designed a two-tier preprocessing pipeline. In the first stage, the *ArabicBasicPreprocessor* standardizes orthographic variants-such as unifying different forms of alif and ya to reduce script-level inconsistency Alansary et al. (2025)-removes diacritics and tatweel to limit sparsity and improve token matching Al-Shammari (2007), and produces a normalized representation of each text. Empirically, this stage consistently reduced character-level noise in user-generated Arabic Al-Shammari (2007) and yielded a clean version of the dataset suitable for downstream modeling. Qualitative inspection indicated that core linguistic content was preserved and that label integrity remained unaffected Hasan et al. (2025).

Building on this foundation, we implemented a more fine-grained preprocessing workflow using the CAMeL Tools morphological analyzer. CAMeL Tools is an open-source Python toolkit from the CAMeL Lab that provides tokenization, morphological analysis and disambiguation, diacritization, dialect identification, and other utilities for Arabic NLP Obeid et al. (2020). We rely on its lexicon-backed analyzer to support clitic segmentation and expose morphological features across MSA and major dialects. This *advanced preprocessing* stage involved pronominal clitic segmentation, controlled retention of articles and particles, and enriched morphological analysis. Executing this pipeline in batch mode allowed us to generate expanded token sequences and derive new text variants, including fully segmented and morphologically annotated versions. The advanced processed datasets showed an expected increase in token counts due to explicit clitic separation (see Figure 6). To ensure the reliability of the morphological annotations, we performed targeted verification on a curated set of examples, examining

part-of-speech tags, lemmas, roots, glosses, and gender/number features. These checks confirmed stable analyzer performance for the majority of common dialectal and MSA constructions.

Together, the EDA findings and the layered preprocessing strategy provide a robust foundation for model training. They also highlight the importance of addressing the linguistic richness and inherent variability of Arabic social media text-factors that directly influence the effectiveness of downstream polarization detection.

4 Approach

Our approach for this project is guided by established conventions in Arabic Natural Language Processing and insights from polarization-related research. The main objective is to develop models capable of accurately classifying Arabic social media texts as either polarized or non-polarized while maintaining robust generalization across different dialects and contexts. This section outlines our planned methodology and the work completed to date.

4.1 Literature Foundations and Methodology

Polarization detection goes beyond standard sentiment or stance classification, as it focuses on identifying division, hostility, and opinion extremity rather than mere agreement or disagreement POLAR @ SemEval (2025). Previous research highlights that polarization is highly context-dependent and non-linear, necessitating deep, high-capacity models that can capture subtle linguistic, affective, and discourse cues Vasist et al. (2023). Early approaches relied on feature-engineering-based baselines, including lexical affect scoring and discourse structure modeling. While these methods offered interpretability, Transformer-based architectures have consistently outperformed them in capturing nuanced contextual relationships Wolf et al. (2020).

4.2 Selected Computational Framework

Based on comparative studies in Arabic stance detection and fake news classification, MARBERT has been identified as the most suitable pretrained language model for this task AlShenaifi et al. (2024); Al Hariri and Abu Farha (2024). Fine-tuning MARBERT is a practical choice given the limited computational resources available. It was trained on a large mixture of Modern Standard Arabic (MSA) and dialectal user-generated content, making it well suited to the noisy and informal language of online discussions. This alignment between training data and target domain reduces the need for extensive retraining while supporting the context-heavy analysis required for Arabic.

The first model, referred to as Smol-MARBERT, is being fine-tuned exclusively on the Arabic portion of the SemEval dataset. In contrast, a second model, BigData-BERT, is under development to leverage translated datasets from English into Arabic. Since the polarization rubric for Arabic is not publicly available, this approach expands the available training data while approximating polarization labels. Several challenges accompany this strategy:

- **Relevance filtering** is necessary since many English datasets include political content (e.g., Biden/Trump) that is irrelevant in Arabic contexts.
- **Translation quality** must be verified, as the accuracy of machine translation varies; validation by speakers of each source language is required to ensure semantic fidelity.
- **Mixed-language comments** are common in Arabic social media, where users frequently code-switch with English or French, requiring specialized preprocessing to handle multilingual text effectively.

Finally, we will replicate the same training process using MARBERT V2. By comparing the performance and accuracy of both versions, we will determine which model is better suited for this task. The more efficient version will then be optimized and adopted for the final system.

4.3 Handling Arabic Linguistic Challenges

Arabic presents unique linguistic challenges due to diglossia, morphological complexity, and orthographic variation Farghaly and Shaalan (2009). To mitigate these issues, we employ character normalization to standardize inconsistent letter forms (e.g., various forms of alif) and subword tokenization by retaining SentencePiece from MARBERT's pretrained tokenizer, which is effective for morphologically rich and agglutinative languages Qarah and Alsanoosy (2024).

4.4 Dialect Coverage

The model training and evaluation will explicitly account for major Arabic dialects present on social media based on distribution estimates. Details are provided in Appendix Table 1.

5 Next Steps

With the EDA and Preprocessing stages largely complete and initial model training underway, our next steps focus on refining and optimizing our polarization detection system. Key areas of emphasis include hyperparameter tuning, advanced model architectures, and comprehensive evaluation across dialects.

The distribution of the workload and their status are as follows:

Data Exploration and Analysis Perform exploratory data analysis (EDA) to examine label distribution, text characteristics, and class imbalance issues in the dataset, providing insights that will inform model training and evaluation. *Status: Partially Complete. Responsible: Ali Khaled A. Ishtay Altamimi*

Preprocessing and Data Preparation Implement and validate the Arabic-specific preprocessing pipeline, including character normalization, tokenization, and handling of dialectal text variations to ensure data consistency and model readiness. *Status: Partially Complete. Responsible: Guanghui Ma*

Model Training Fine-tune the BDB model on the official Arabic and English dataset using both weighted and contrastive loss configurations to handle class imbalance and improve classification performance. *Status: In Progress. Responsible: Nuh Al Sharafi*

Testing and Optimization Evaluate the trained models on the development and test sets, perform hyperparameter tuning, and apply optimization strategies to enhance the Macro F1 score and overall model robustness. *Status: In Progress. Responsible: Hussein MH Nasser*

6 Conclusion

This milestone report has outlined our current progress in developing a MARBERT-based polarization detection system for Arabic social media text. We have established a solid foundation through literature review, dataset preparation, and initial model development. Our preliminary work demonstrates the feasibility of our approach, though significant optimization and evaluation work remains. The next phase of our project will focus on refining our models, conducting comprehensive experiments, and extending our work to Subtask 2. A final report will document our complete methodology, experimental results, and findings upon completion of the shared task.

Acknowledgments

We thank the organizers of POLAR @ SemEval-2026 for providing the datasets and evaluation framework for this shared task.

References

N. AlShenaifi, N. Alangari, and H. Al-Negheimish. 2024. Rasid at StanceEval: Fine-tuning MARBERT for Arabic Stance Detection. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 828–831, Bangkok, Thailand.

Y. Al Hariri and I. Abu Farha. 2024. SMASH at StanceEval 2024: Prompt Engineering LLMs for Arabic Stance Detection. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 800–806, Bangkok, Thailand.

J. Gatto, O. Sharif, and S. M. Preum. 2023. Chain-of-Thought Embeddings for Stance Detection on Social Media. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 4154–4161.

Z. Zhang, J. Zhang, H. Xu, J. Guo, and X. Cheng. 2025. MPRF: Interpretable Stance Detection through Multi-Path Reasoning Framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 454–470, Suzhou, China.

POLAR @ SemEval. 2025. Task Description: Detecting Multilingual, Multicultural, and Multievent Online Polarization.

H. Najadat, M. Tawalbeh, and R. Awawdeh. 2022. Fake news detection for Arabic headlines-articles news data using deep learning. *International Journal of Electrical and Computer Engineering (IJECE)*, 12(4):3951–3959.

P. N. Vasist, D. Chatterjee, and S. Krishnan. 2023. The Polarizing Impact of Political Disinformation and Hate Speech: A Cross-country Configural Narrative. *Information Systems Frontiers*, pages 1–26.

T. Wolf, L. Debut, V. Sanh, and J. Chaumond. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP Demos)*, pages 38–45.

F. Qarah and T. Alsanoosy. 2024. A Comprehensive Analysis of Various Tokenizers for Arabic Large Language Models. *Applied Sciences*, 14(13):5696.

A. Farghaly and K. Shaalan. 2009. Arabic Natural Language Processing: Challenges and Solutions. In *Proceedings of the 2009 Workshop on Arabic Natural Language Processing*.

O. Obeid, K. Eryani, N. Zalmout, S. Khalifa, D. Taji, B. Alhafni, B. AlKhamissi, and N. Habash. 2020. CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, Marseille, France.

S. Alansary, et al. 2025. Arabic Natural Language Processing: Challenges and Solutions.

E. Al-Shammari. 2007. Arabic Text Preprocessing for Natural Language Processing Applications. Princess Sumaya University for Technology.

M. Hasan, et al. 2025. Arabic NLP: A Survey of Pre-Processing and Representation Techniques. *Jurnal UMSU*.

Appendix: Data Figures and Outputs

Dialect	Code	Est. %
Standard Arabic	ar	12.5
Egyptian Arabic	arz	20.0
North Levantine Arabic	apc	20.0
Sudanese Arabic	apd	2.5
Gulf Arabic	afb	10.0
Iraqi Arabic	acm	7.5
Hejazi Arabic	acw	10.0
Najdi Arabic	ars	2.5
Yemeni Arabic	ayn	5.0
Maghrebi Arabic	ary	10.0

Table 1: Arabic dialect distribution estimates in social media data.

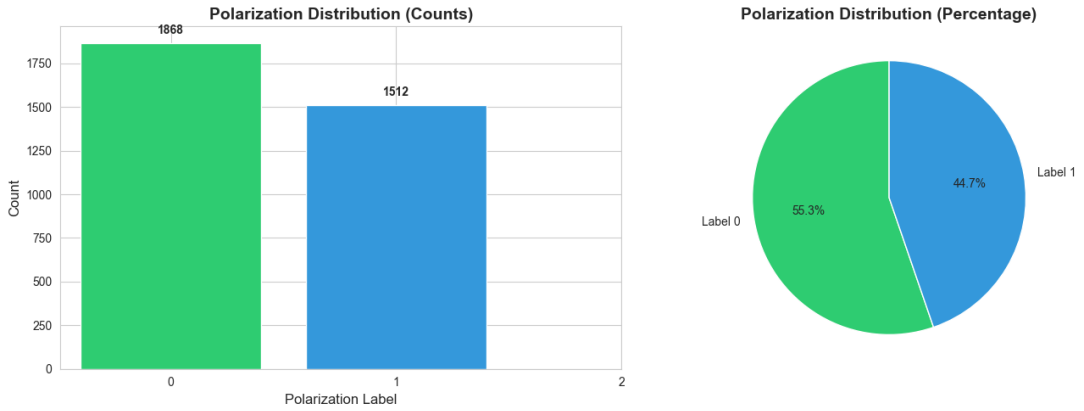

Figure 1: Bar and Pie chart showing the distribution and percentage of polarized vs non-polarized labels in the training dataset.

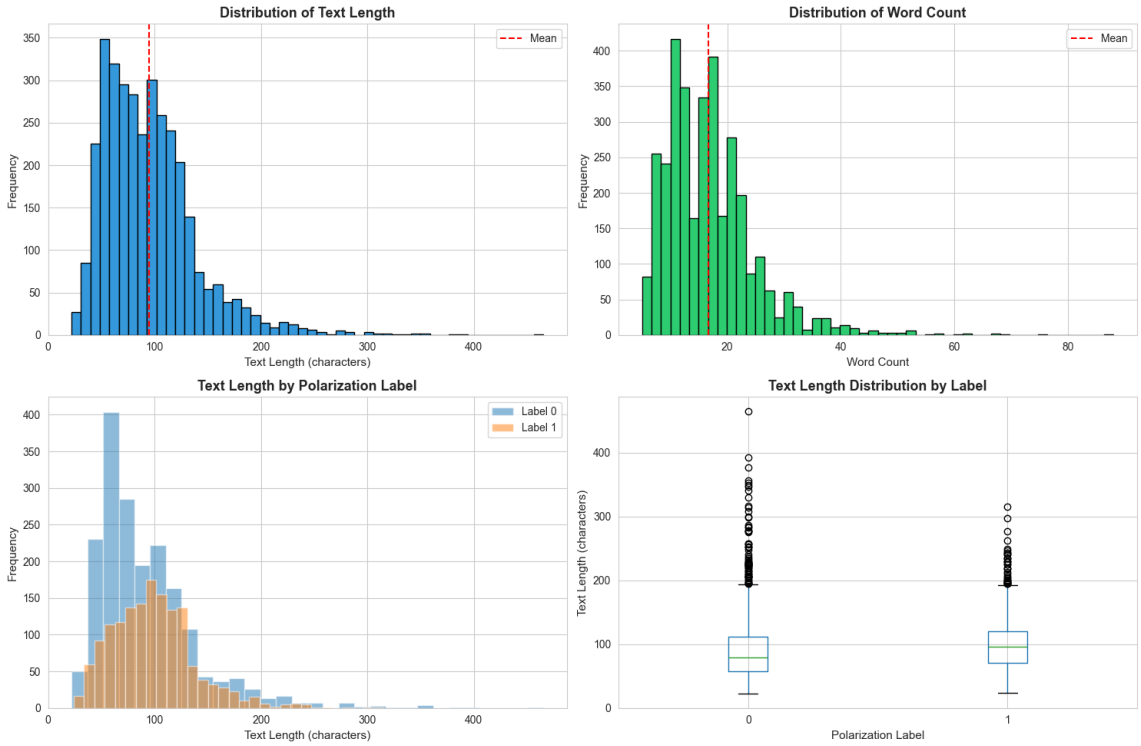

Figure 2: Distribution of text length, distribution word count per text, and text length by label across the entire dataset showing the variations in text length for polarized and non-polarized classes.

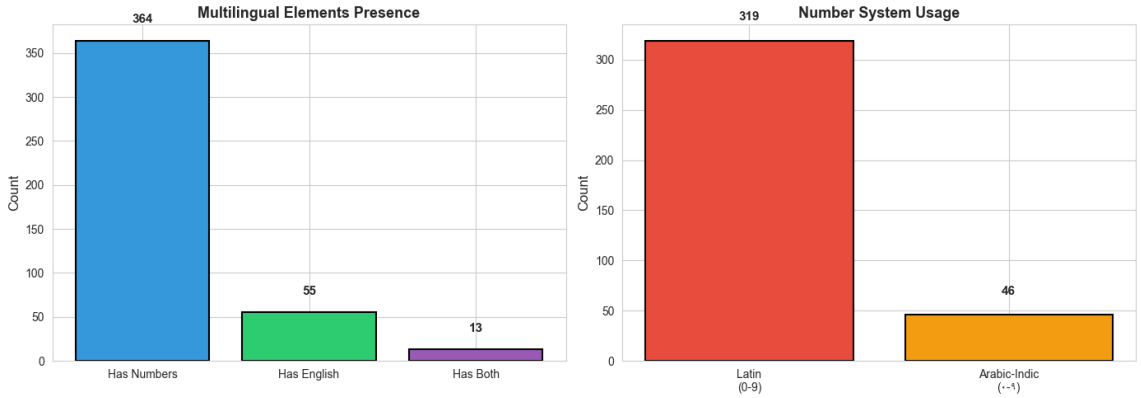

Figure 3: Presence and frequency of numbers, latin characters in texts, presence of Latin and Arabic-Indic digits across the dataset by label, showing code-switching behavior.

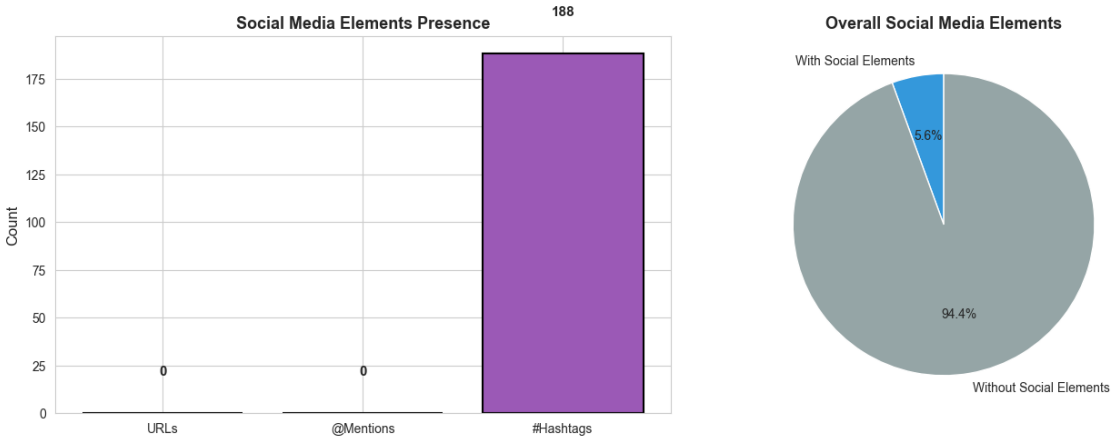

Figure 4: Frequency and presence of social elements (URLs, @mentions, and #hashtags) across labels, indicating user interaction patterns.

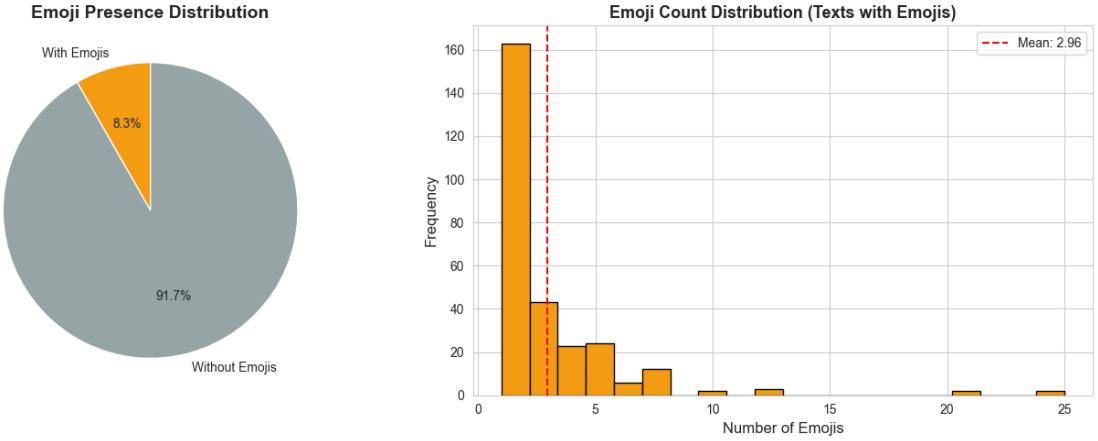

Figure 5: Presence of emojis in texts by label, and distribution of emoji counts per text showing emotional expression trends.

```
================================================================================
TEST 5: Comparison - No Preprocessing vs Basic vs Advanced
================================================================================

Complex sentence: وبكتابهم المدرسي فسيدرسونها للطلاب كالمعلمين
(Translation: And with their school book, they will teach it to students like the teachers)

1. No preprocessing:
   وبكتابهم المدرسي فسيدرسونها للطلاب كالمعلمين

2. Basic preprocessing (normalization, diacritics removal):
   وبكتابهم المدرسي فسيدرسونها للطلاب كالمعلمين

3. Advanced - selective clitic segmentation:
   و+ ب+ كتابهم المدرسي ف+ س+ يدرسونها ل+ لطلاب ك+ المعلمين
   Tokens: ['و+', 'ب+', 'كتابهم', 'المدرسي', 'ف+', 'س+', 'يدرسونها', 'ل+', 'لطلاب', 'ك+', 'المعلمين']

4. Advanced - with lemmatization:
   و+ ب+ كُتَّاب مَدْرَسِيّ ف+ س+ دَرَس ل+ طالب ك+ مَعْلَم
   Tokens: ['و+', 'ب+', 'كُتَّاب', 'مَدْرَسِيّ', 'ف+', 'س+', 'دَرَس', 'ل+', 'طالب', 'ك+', 'مَعْلَم']
```

Figure 6: Comparison of original text, basic preprocessed output, and advanced preprocessed output (with clitic segmentation), demonstrating the effect of each preprocessing stage.

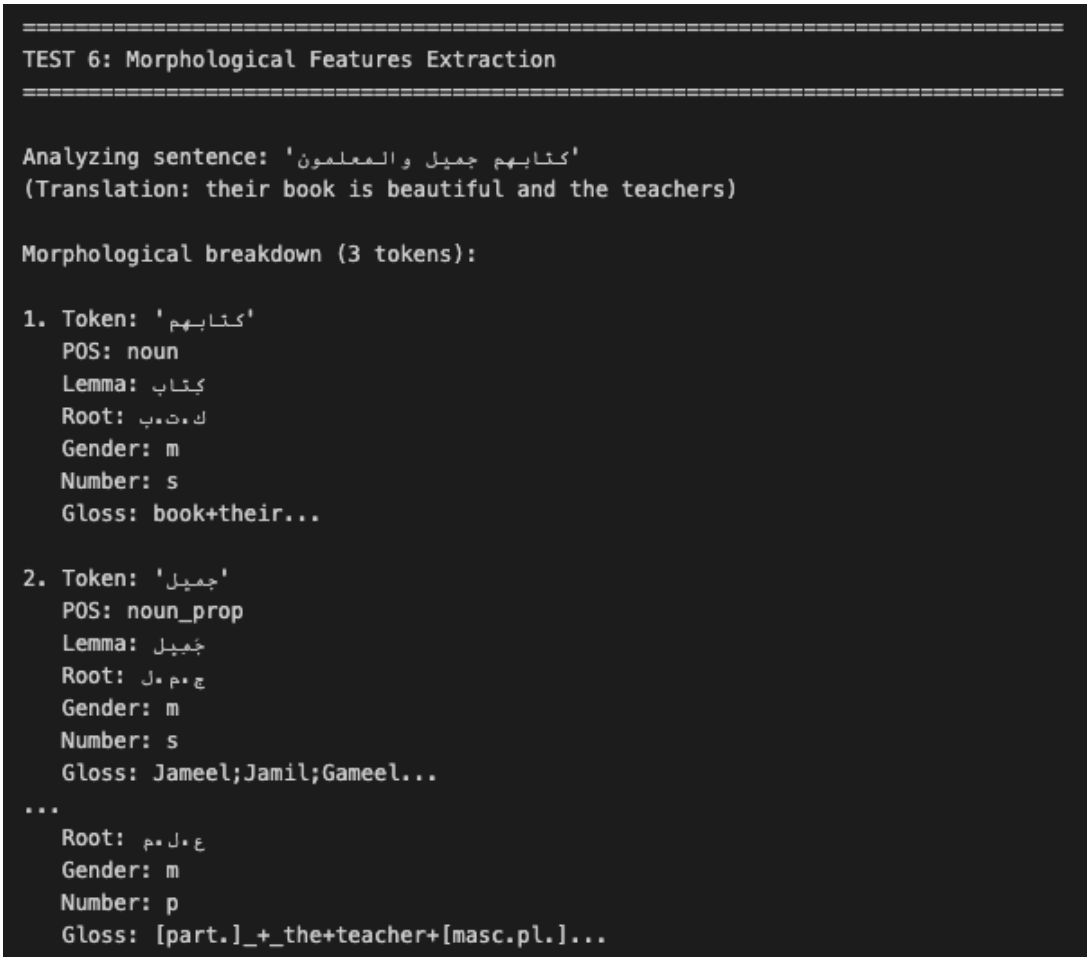

Figure 7: Sample morphological features extracted using CAMeL Tools, including part-of-speech tags, lemmas, roots, and gender