# CAMeL Tools: Short Technical Report

Project: CAMeL Tools (Arabic Natural Language Processing Toolkit)
**Sajjad Kadhim & Zainab Jassim**

## 1. Problem Statement

Arabic Natural Language Processing (NLP) faces unique challenges due to the language's complex morphology, orthographic ambiguity (lack of diacritics), and significant dialectal variation (Diglossia). Existing tools often lack support for dialects or fail to bridge the gap between traditional linguistic rules and modern Artificial Intelligence, creating a need for a comprehensive, modular, and accurate toolkit.

## 2. Datasets

The project leverages several high-quality linguistic resources and corpora for training and validation:

- **Morphology & Disambiguation:** Trained on the **Penn Arabic Treebank (PATB)** and utilized the **ALMOR** database for analysis.
- **Dialect Identification:** Utilized the **MADAR corpus** and other dialectal datasets referenced in *Salameh et al. (COLING 2018)*, covering up to 26 distinct Arabic dialects.
- **Pre-trained Models:** Leveraged **AraBERT** and **mBERT** (Multilingual BERT) datasets for fine-tuning deep learning components.

## 3. Method

The system employs a hybrid architecture combining Deep Learning and traditional Machine Learning:

- **Deep Learning (BERT):** Used for context-sensitive tasks (Disambiguation, Named Entity Recognition, Sentiment Analysis). The system fine-tunes BertForTokenClassification and BertForSequenceClassification models to align WordPiece tokens with morphological features.
- **Machine Learning (Naive Bayes):** Used for Dialect Identification. A Multinomial Naive Bayes classifier is trained on feature unions of word n-grams, character n-grams (TF-IDF), and Language Model probability scores (KenLM).
- **Rule-Based Utilities:** Finite-State Transducers (FST) and Python-based logic are used for transliteration, dediacritization, and text cleaning.

## 4. Results & Accuracy

The implementation demonstrates State-of-the-Art (SOTA) performance across key metrics:

- **Disambiguation:** The BERT-based unfactored model significantly outperforms legacy Maximum Likelihood Estimation (MLE) approaches in resolving morphological ambiguity.
- **Dialect ID:** The "Model-6" implementation (Salameh et al.) achieves high precision in distinguishing between fine-grained dialects (e.g., distinguishing Aleppine from Damascene) compared to standard baselines.
- **Efficiency:** The integration of caching (LFUCache) and GPU support ensures the system remains performant for production-level workloads.

## 5. Conclusion

CAMeL Tools successfully addresses the complexities of Arabic NLP by providing a centralized, open-source Python suite. By integrating Transformer-based architectures for semantic tasks and robust statistical models for classification, it offers a scalable solution that supports both Modern Standard Arabic and various dialects, facilitating advanced research and application development in the field.