Sifal KLIOUI

TIZI-OUZOU, ALGERIA

sifal.klioui@gmail.com Linkedin Github HuggingFace

Education

Mouloud MAMMERI University **Sep. 2019 – jul 2022**

Bachelor in Computer Systems

Mouloud MAMMERI University **Sep. 2022 – jul 2024**

Masters in Intelligent Computer Systems

Relevant Coursework

- Deep Learning
- Internet of Things
- Probability and Statistics
- Linear Algebra
- Computer Architecture
- Operating systems
- Algorithms and Data-structures
- Meta-heuristics and Evolutionary Algorithms

Professional Experience

Omdena **May 2023 - Sep 2023**

Remote Natural Language Processing Engineer

- Played a pivotal role in creating a comprehensive hate speech detection dataset by employing various techniques, including web scraping, assisted annotation tools, and data cleaning.
- Utilized advanced models, such as BERT and T5, to develop state-of-the-art natural language processing solutions.
- Conducted an extensive hyperparameter search within a constrained search space, informed by prior literature, to optimize model performance.
- Performed rigorous experimentation to monitor the learning process, evaluating model performance through metrics and loss analysis.

Projects

From-Scratch Deep Learning Architectures | *Pytorch* **2022**

- Designed and implemented various Deep learning architectures from scratch, showcasing a deep understanding of the underlying principles.
- Developed Generative Adversarial Networks (GANs) to generate realistic data and images, including Wasserstein GANs (WGANs) for improved stability and convergence.
- Built custom Transformer models, demonstrating expertise in sequence-to-sequence tasks and natural language processing.
- Created a lightweight language model, Nano GPT, optimized for resource-constrained environments, showcasing versatility in model design.
- Applied these self-built architectures to solve real-world problems and research tasks, gaining hands-on experience in model development and fine-tuning.

Smart Environment Activity Recognition | *Pandas, PyTorch, Scikit-Learn,matplotlib* **Jul 2023**

- Implemented data preprocessing and modeling based on the research by Singla, Cook, and Schmitter-Edgecombe.
- Engineered features, cyclically encoding time attributes and employing one-hot encoding for categorical columns.
- Ensured robust reproducibility by meticulously controlling random seed values, applied consistently throughout dataset splitting and model training.
- Employed a range of model configurations to identify optimal architectures for activity recognition in smart environments.

Model Pruning with Dziribert | *Pytorch, Transformers* **Aug 2023**

- Demonstrated model pruning using the Dziribert language model, optimizing it for Arabic token processing.
- Identified and marked unwanted tokens, including non-alphabetic characters and specific tokens for removal.
- Created a modified tokenizer by removing unwanted tokens from the vocabulary.
- Pruned the model's embeddings to match the modified vocabulary, reducing memory usage.
- Saved the final pruned model, ready for efficient use in resource-constrained applications.

Hate Speech Detection with AraT5v2 | *Dataset, PyTorch Lightning, Transformers, Metrics, Wandb, Optuna* **Aug 2023**

- Developed a hate speech detection model based on the T5 architecture using PyTorch and Transformers.
- Fine-tuned the model for accurate hate speech classification.
- Implemented a custom Dataset class for hate speech detection, utilizing the T5 tokenizer.

- Utilized PyTorch Lightning to define the model, loss function, and training loop.
- Conducted hyperparameter search using Optuna to optimize model performance.
- Logged training progress using WandB (Weights and Biases).
- Evaluated the model's performance on a test dataset, computing metrics like accuracy, precision, recall, and F1-score.
- Included clear instructions in the notebook for running the code and training custom hate speech detection models.

Dzarashield: Hate Speech Detection in Arabic Text | *Datasets, Transformers, PyTorch, Wandb, Optuna* **Sep 2023**

- Demonstrated fine-tuning of a BERT-based model for hate speech detection in Arabic text.
- Utilized multiple Arabic hate speech datasets and leveraged the Hugging Face Transformers library for model training and evaluation.
- Performed data preprocessing, including data cleaning, merging, and tokenization.
- Defined a custom classification model based on the pre-trained BERT model.
- Explored various hyperparameter configurations and used Optuna for hyperparameter optimization, resulting in improved model performance.
- Tracked and visualized training progress with Wandb, allowing for real-time monitoring of loss, accuracy, and other metrics.
- Demonstrated model inference with sample Arabic sentences for hate speech detection.
- Provided a comprehensive notebook serving as a template for building Arabic hate speech detection models.

Neural Machine Translation: English to Kabyle | *PyTorch, Transformers, tokenizers, flair, wandb* **In Progress**

- Collecting and preprocessing parallel English-Kabyle corpora for training the translation model.
- Conducting extensive experiments with different subword tokenization strategies and vocabulary sizes to optimize model input representations.
- Incorporating part-of-speech embeddings to improve translation quality and linguistic accuracy.
- Experimenting with BERT embeddings in the encoder to enhance the model's understanding of the context and semantics.
- Optimizing the beam search strategy for better translation results
- Tracking the improvement in the BLEU score, a standard metric for machine translation quality.

Publications

[1] **Hate speech detection in Algerian dialect using deep learning**: Dihia Lanasri, Juan Olano, **Sifal Klioui**, Sin Liang Lee, Lamia Sekkai

Accepted at NeurIPS 2023 Workshop: North Africans in Machine Learning Workshop - NAML 2023

Community Involvement

Kabyle Language Database

[1] Created and actively maintain a comprehensive database of the Kabyle language, a Berber language spoken in North Africa. The dataset serves as a valuable resource for linguistic research, language preservation, and cultural diversity promotion.

Certifications and Courses

DL Specialization, AI for Medical Prognosis & Diagnosis, GANs, ML Specialization | *Coursera* **Dec 2022**

Deep Learning with PyTorch | *DataCamp* **Jan 2023**

SQL (Advanced) Certificate | *HackerRank* **Jan 2023**

Language Proficiency

- **English** | *Full professional proficiency*
- **French** | *Native or bilingual proficiency*
- **Arabic** | *Full professional proficiency*
- **Kabyle** | *Native or bilingual proficiency*