Institut National de Statistique
et d'Economie Appliquée

Centre des Etudes Doctorales
Sciences, Ingénierie
et Développement Durable

Avis de soutenance de thèse de Doctorat

Monsieur Kamel GAANOUN

Doctorant au laboratoire de recherche

« Systèmes d'Information, Systèmes Intelligents et Modélisation Mathématique »

(SI2M)

Spécialité : ***Systèmes d'Information et Systèmes Intelligents***

Soutiendra publiquement sa thèse de Doctorat

Le jeudi 25 décembre 2025 à 09h

à la salle de conférence de l'INSEA

Intitulé de la thèse

« Leveraging Transformer Models and Data-Centric Approaches for Arabic Dialect Processing : Focus on Moroccan Arabic »

Devant le jury composé de :

Président :
Pr. Ahmed ZELLOU, PES, ENSIAS-Rabat

Directeur de thèse :
Pr. Imad BENELALLAM, PES, INSEA-Rabat

Membres du jury :
Pr. Mohamed Nabil SAIDI, PES, INSEA-Rabat
Pr. Anwar Nabil ADIL, PES, EMI-Rabat
Pr. Moulay Hafid YAZIDI, PH, FSJES – Rabat
Pr. Ibtissam ABNANE, PH, ENSIAS-Rabat
Pr. Redouane EZZAHIR, PES, Université Ibn Zohr

INSEA
Institut National de Statistique et d'Économie Appliquée
CEDOC-SIDD

Les Résumés de la thèse (F1)

Date : 20/01/2025

Réservé à l'administration

N°de thèse : .....

Nom : GAANOUN

Prénom : KAMEL

Résumé

Les récents progrès en apprentissage profond, en particulier l'émergence des architectures *Transformer*, ont révolutionné le traitement automatique des langues (TAL), entraînant des améliorations significatives dans diverses applications linguistiques. Cependant, ces avancées ont principalement bénéficié aux langues disposant de ressources abondantes, tandis que l'arabe dialectal (AD), et plus particulièrement l'arabe Marocain (Darija), demeure largement sous-représenté. Cette disparité est exacerbée par les défis spécifiques au traitement du Darija, notamment son importante divergence linguistique par rapport à l'arabe standard moderne (ASM), la prévalence du *code-switching* et, surtout, la rareté critique de données annotées de haute qualité. Un défi particulièrement notable réside dans l'utilisation répandue de l'Arabizi (arabe translittéré en caractères latins) dans la communication numérique, qui ajoute une complexité supplémentaire en raison de son absence de standardisation et de son ambiguïté au niveau des caractères.

Pour réduire cet écart technologique, cette thèse propose une approche hybride combinant la puissance des architectures *Transformer* avec des méthodologies centrées sur les données. La solution apportée vise à répondre au défi fondamental de la rareté des données par des stratégies innovantes de collecte et d'augmentation des données, tout en tirant parti des capacités computationnelles des architectures modernes d'apprentissage profond. Notamment, cette recherche inaugure le développement de modèles *Transformer* spécialisés pour les variations en écriture arabe et Arabizi du Darija, constituant la première tentative de fournir un support complet en TAL pour ces deux systèmes d'écriture largement utilisés dans la communication numérique au Maroc.

Les contributions principales de cette recherche incluent le développement de DarijaBERT, les premiers modèles basés sur des *Transformers* spécifiquement conçus pour le Darija dans les deux scripts, ainsi que la création de jeux de données complets couvrant plusieurs domaines et applications linguistiques. Elle démontre l'efficacité de diverses approches centrées sur les données, y compris les techniques d'augmentation *Mix* et la génération de données synthétiques, pour améliorer les performances des modèles malgré des ressources limitées. Les résultats expérimentaux montrent des améliorations significatives dans plusieurs tâches spécifiques au TAL pour le Darija, notamment la synthèse de textes. De plus, ce travail introduit des contributions pionnières dans des applications

spécifiques au Darija, telles que la détection du sarcasme, l'identification des dialectes et la classification thématique, soutenues par le premier jeu de données marocain spécialement conçu pour la catégorisation des sujets. Le succès de ces applications valide l'hypothèse selon laquelle la combinaison des architectures neuronales modernes avec une curation et une augmentation rigoureuse des données peut efficacement répondre aux défis du traitement des dialectes à faibles ressources.

Ce travail fait progresser l'état de l'art dans le traitement de la langue Darija et propose un cadre pour le développement de systèmes de TAL pour d'autres langues et dialectes. Les méthodologies, modèles bi-scriptes et jeux de données complets développés dans le cadre de cette recherche établissent une base solide pour les travaux futurs sur le traitement des dialectes arabes et démontrent le potentiel des approches centrées sur les données dans des contextes à faibles ressources.

INSEA
CEDOC-SIDD

Institut National de Statistique et d'Économie Appliquée

Abstract

Recent advances in Deep Learning, particularly the advent of Transformer architectures, have revolutionized Natural Language Processing (NLP), leading to unprecedented improvements in various language tasks. However, these advancements have primarily benefited high-resource languages, while Dialectal Arabic (DA), especially Moroccan Arabic (Darija), remains underserved. This disparity is further amplified by the unique challenges of processing Darija, including its significant linguistic divergence from Modern Standard Arabic (MSA), the prevalence of code-switching, and most critically, the severe scarcity of high-quality annotated data. A particularly significant challenge is the widespread use of Arabizi (Latinized Arabic script) in digital communication, which introduces additional complexity due to its non-standardized nature and character-level ambiguity.

To bridge this technological gap, this dissertation proposes a hybrid approach combining the power of Transformer architectures with data-centric methodologies. My solution addresses the fundamental challenge of data scarcity through innovative data curation and augmentation strategies, while simultaneously leveraging the computational capabilities of modern deep learning architectures. Notably, I pioneer the development of specialized Transformer models for both Arabic script and Arabizi variations of Darija, marking the first attempt to provide comprehensive NLP support for both writing systems commonly used in Moroccan digital communication.

The primary contributions of this research include the development of DarijaBERT, the first Transformer-based models specifically designed for Moroccan Arabic in both scripts, alongside the creation of comprehensive datasets spanning multiple domains and tasks. I demonstrate the effectiveness of various data-centric approaches, including Mix augmentation techniques and synthetic data generation, to enhance model performance despite limited resources. My experimental results show significant improvements across several previously unexplored tasks for Darija, including text summarization. Moreover, I introduce pioneering work in Darija-specific tasks such as sarcasm detection, dialect identification, and topic classification, supported by the first Moroccan Arabic dataset specifically curated for topic categorization. The success of these applications validates my hypothesis that combining modern neural architectures with careful data curation and augmentation can effectively address the challenges of processing low-resource dialects.

This work not only advances the state-of-the-art in Darija language processing but also provides a framework for developing NLP systems for other under-resourced languages and dialects. The methodologies, dual-script models, and comprehensive datasets developed through this research establish a foundation for future work in Arabic dialect processing and demonstrate the potential of data-centric approaches in low-resource scenarios.

***INSEA** B.P. 6217, Madinat Al Irfane, Rabat-Instituts, Rabat, Maroc*
***Tél.** : +212 (0)5 37 77 48 59/60 **Fax** : + 212 (0) 537 77 94 57*

INSEA
CEDOC-SIDD

Institut National de Statistique et d'Économie Appliquée

ملخص

أحدثت التطورات الأخيرة في التعلم العميق، ولا سيما ظهور معماريات المحولات(Transformer) ، ثورة في مجال معالجة اللغات الطبيعية(NLP) ، مما أدى إلى تحسينات ملحوظة في العديد من التطبيقات اللغوية. ومع ذلك، استفادت من هذه التطورات بشكل رئيسي اللغات ذات الموارد الوفيرة، بينما ظلَّت اللغة العربية العامية (الدارجة)، وخاصة الدارجة المغربية(Darija) ، غير مدعومة بشكل كافٍ. وتتفاقم هذه الفجوة بسبب التحديات الخاصة بمعالجة الدارجة، والتي تشمل الاختلاف اللغوي الكبير عن اللغة العربية الفصحى الحديثة، وانتشار استخدام المزج اللغوي(code-switching) ، والأهم من ذلك النقص الحاد في البيانات عالية الجودة والموسومة. ومن أبرز التحديات استخدام كتابة "العربيزية (Arabizi) "على نطاق واسع في التواصل الرقمي، مما يزيد من تعقيد المعالجة بسبب غياب معايير ثابتة وغموض النصوص على مستوى الحروف.

لجسر هذه الفجوة التقنية، تقترح هذه الأطروحة منهجًا هجينًا يجمع بين قوة معماريات المحولات وأساليب موجهة نحو البيانات. تهدف هذه الحلول إلى معالجة التحدي الأساسي المتمثل في ندرة البيانات من خلال استراتيجيات مبتكرة لجمع البيانات وزيادة حجمها، مع الاستفادة من القدرات الحسابية للنماذج الحديثة في التعلم العميق. تُعد هذه الدراسة الأولى من نوعها التي تطور نماذج محولات متخصصة لمعالجة الدارجة المغربية بنوعيها الكتابي (الأبجدية العربية والعربيزية)، مما يمثل خطوة رائدة نحو تقديم دعم شامل لمعالجة اللغة الطبيعية لأنظمة الكتابة المستخدمة في التواصل الرقمي في المغرب.

تشمل المساهمات الرئيسية لهذا البحث تطوير نموذج"DarijaBERT" ، وهو أول نموذج قائم على المحولات صُمم خصيصًا للدارجة المغربية بالخطين العربي والعربيزية، بالإضافة إلى إنشاء مجموعات بيانات شاملة تغطي العديد من المجالات والتطبيقات اللغوية. يُظهر هذا البحث فعالية مجموعة متنوعة من الأساليب الموجهة نحو البيانات، بما في ذلك تقنيات الزيادة المختلطة (Mix augmentation)وتوليد البيانات الاصطناعية، لتحسين أداء النماذج في ظل محدودية الموارد. أظهرت النتائج التجريبية تحسينات ملحوظة في العديد من المهام اللغوية للدارجة، بما في ذلك التلخيص الآلي للنصوص. علاوة على ذلك، يُقدم هذا العمل إسهامات رائدة في مهام خاصة بالدارجة مثل اكتشاف السخرية، تحديد اللهجات، وتصنيف الموضوعات، مدعومًا بأول مجموعة بيانات مغربية مخصصة لتصنيف المواضيع. تؤكد هذه النتائج الفرضية القائلة بأن الجمع بين النماذج العصبية الحديثة وأساليب العناية بالبيانات وزيادتها يمكن أن يعالج بفعالية تحديات معالجة اللهجات منخفضة الموارد.

لا يقتصر هذا العمل على تعزيز المستوى الحالي في معالجة لغة الدارجة المغربية، بل يقدم أيضًا إطار عمل لتطوير أنظمة معالجة اللغة الطبيعية للغات واللهجات الأخرى التي تعاني من نقص الموارد. تُشكل المنهجيات، النماذج ثنائية النصوص، ومجموعات البيانات الشاملة التي طُورت من خلال هذا البحث أساسًا متينًا للأعمال المستقبلية في معالجة اللهجات العربية، وتوضح الإمكانات الكبيرة لأساليب موجهة نحو البيانات في سياقات محدودة الموارد.