

Nom et prénom	Etablissement	Grade	Qualité
Pr. ELFAR Mohamed	Faculté des Sciences Dhar El Mahraz, Fès	PES	Président
Pr. JELLOULI Ismail	Ecole Supérieure de Technologie, Kénitra	PES	Rapporteur
Pr. AGOUTANE Badraddine	Faculté des Sciences, Meknès	PES	Rapporteur
Pr. SABRI My Abdelouahed	Faculté des Sciences Dhar El Mahraz, Fès	PES	Rapporteur
Pr. RIFFI Jamal	Faculté des Sciences Dhar El Mahraz, Fès	MCH	Examineur
Pr. BENNANI Mohamed Taj	Faculté des Sciences Dhar El Mahraz, Fès	MCH	Examineur
Pr. EL GAROUANI Said	Faculté des Sciences Dhar El Mahraz, Fès	PES	Directeur de thèse

Dans ce contexte, cette thèse se concentre d’abord sur l’établissement d’un cadre fondamental pour le traitement de l’arabe dialectal, avec un intérêt particulier pour le dialecte marocain. Cette démarche repose sur la construction de jeux de données de haute qualité visant à pallier le manque de ressources dans ce domaine, suivie de l’étude de modèles mieux adaptés aux caractéristiques linguistiques de l’arabe dialectal. Sur cette base, plusieurs modèles spécialisés sont proposés et développés.

En combinant la création de jeux de données robustes, l'adaptation des modèles et des stratégies d'évaluation réalistes intégrant les notions de robustesse et d'équité, cette thèse propose une chaîne de traitement complète permettant d'améliorer la fiabilité et l'applicabilité des systèmes de traitement automatique du langage naturel pour l'arabe. Elle contribue ainsi au développement de systèmes intelligents plus robustes, évolutifs et sensibles au contexte, capables de prendre en compte la complexité et la diversité de l'arabe dialectal dans des applications réelles.

Mots-clés :

Traitement automatique du langage naturel (NLP), chatbot arabe, chatbot médical, reconnaissance automatique de la parole, reconnaissance des émotions, détection des discours haineux, langue marocaine (darija), apprentissage par attaques adversariales.



DEVELOPMENT OF AN INTELLIGENT ARABIC LANGUAGE SPEECH RECOGNITION CHAT BOT USING AI

Abstract :

This thesis aims to enhance the field of Arabic speech recognition by developing a robust chatbot and virtual agent. To achieve this objective, it investigates the field of Arabic natural language processing, providing a comprehensive overview of the various approaches employed, the current state of the art, and the key challenges that continue to limit its progress.

In particular, this work places strong emphasis on dialectal Arabic, which poses unique linguistic and computational difficulties due to its variability, lack of standardization, and limited available resources. The thesis examines a wide range of proposed solutions to address these challenges, with a focus on recent advancements in machine learning and deep learning techniques.

By integrating insights from speech recognition and NLP, this research aims to contribute to the development of more effective, scalable, and context-aware conversational systems for Arabic, ultimately supporting the creation of intelligent virtual agents capable of understanding and interacting in dialectal Arabic.

Despite significant efforts, the gap between high-resource languages and Arabic, particularly dialectal Arabic, remains substantial across almost all subdomains of natural language processing. One of the primary challenges is the scarcity of annotated and high-quality linguistic resources, which severely limits the development, training, and reliable evaluation of robust models. This limitation is further amplified by the lack of standardized datasets and benchmarks, making it difficult to compare approaches or reproduce results consistently across studies.

Moreover, the generalizability of existing approaches is often constrained, as the majority of prior research has been conducted on Modern Standard Arabic. While it is widely used in formal contexts, it differs significantly from dialectal Arabic, which is the dominant form in daily communication. These differences span multiple linguistic levels, including vocabulary, syntactic structures, pronunciation variations, and even pragmatic and cultural expressions. As a result, models trained on Modern Standard Arabic data frequently struggle when applied to real-world dialectal inputs, leading to performance degradation in practical applications.

In addition, dialectal Arabic itself is highly diverse, with substantial variation across regions and countries, further complicating the development of unified models. This diversity introduces additional challenges related to normalization, transliteration, and ambiguity resolution, especially in informal text and speech data. Consequently, existing systems often fail to capture the richness and variability of dialectal usage, limiting their effectiveness in real-world scenarios.

Therefore, this thesis first focuses on establishing a foundational framework for dialectal Arabic processing, with a major emphasis on the Moroccan dialect. This is achieved through the construction of high-quality datasets that address the scarcity of resources in this domain, followed by the investigation of models that are better aligned with the linguistic characteristics of dialectal Arabic. Building on this foundation, the thesis proposes and develops dedicated models tailored to this language setting. The resulting framework is applied and evaluated across multiple subdomains, including speech recognition, emotion recognition, offensive language detection, and healthcare applications. Furthermore, we



addressed a critical issue related to real-world deployment, which is often overlooked in prior work. We found that model performance can drop significantly in practical settings compared to controlled test environments, in some cases by up to 80%. This reveals a strong bias between benchmark evaluation and real-world usage. To tackle this issue, we proposed an evaluation framework that integrates adversarial attacks to simulate real-world scenarios, where the model is exposed to different types of perturbations.

This pipeline resulted in the development of several models that achieved strong and consistent performance across multiple subdomains, thereby addressing key challenges in Arabic natural language processing, particularly for dialectal Arabic. Overall, the proposed framework successfully contributes to reducing the gap between high-resource languages and Arabic by tackling core limitations such as data scarcity, lack of standardization, and limited generalizability of existing approaches.

By integrating robust dataset creation, model adaptation, and realistic evaluation strategies that integrate robustness and fairness of models, this thesis provides a comprehensive pipeline that enhances the reliability and applicability of Arabic Natural language processing systems. Ultimately, it contributes toward building more robust, scalable, and context-aware intelligent systems capable of handling the complexity and variability of dialectal Arabic in real-world applications.

Keywords: NLP, Arabic Chatbot, Healthcare chatbot, Speech recognition, Emotion recognition, Hate speech recognition, Moroccan Language, Adversarial attack training.