People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
University of 8 May 1945-Guelma
Faculty of Mathematics, Computer Science, and Science of Matter
Department of Computer Science

Master Thesis

Specialty: Computer Science

Option: Science and technology of information and communication

Theme

Designing a Medical Chatbot Using Expert–Patient Conversations on Social Media and External Knowledge Integration

Presented by:

Abdelhalim KHALFOUN

Supervised by:

Dr. Djalila BOUGHAREB

Members Jury:

Dr. Hallaci Samir	Chairman
Dr. Boudria Asma	Examiner
Dr. Zahra Abbas	Incubator Representative

June 2025

Acknowledgments

First and foremost, I would like to express my heartfelt gratitude to Allah, whose mercy and guidance have granted me the strength and perseverance to complete this work. Without His blessings, none of this would have been possible.

I extend my sincere thanks to my supervisor, **Dr. Djalila Boughareb**, for her invaluable guidance, patience, and constant support throughout the course of this research. Her insightful feedback, encouragement, and belief in my potential have been a pillar of strength from the very beginning.

I am also grateful to the members of the thesis committee for their time, thoughtful suggestions, and constructive critiques, which have greatly contributed to refining this work.

My appreciation goes to my university for the academic environment, resources, and support provided during my studies. I also thank all the faculty members and staff who were part of this journey, offering help whenever needed.

To my beloved family, thank you for your unconditional love, sacrifices, and unwavering support during every step of this thesis. Your faith in me kept me going through the toughest moments.

Lastly, I wish to thank my friends who stood by me, lifted my spirits, and motivated me with their constant encouragement. Your presence made this journey lighter and more meaningful.

To everyone who played a part in this achievement thank you from the bottom of my heart.

Abstract

In recent years, conversational systems have gained considerable importance in the healthcare domain, particularly as the demand for real-time, accessible medical consultation continues to grow. Despite their utility, these systems often fall short in the Arab world, especially in regions that rely on local dialects such as Algerian Arabic. This shortfall is primarily due to the linguistic complexity of dialectal Arabic and the scarcity of curated, domain-specific datasets. This research addresses these challenges by designing and evaluating a multilingual medical chatbot capable of understanding and responding to user queries in Algerian dialect, Modern Standard Arabic (MSA), and French.

To construct the system, we assembled a dataset of approximately 74,918 question–context–answer triplets. This corpus was collected through web scraping from social media platforms including Facebook, YouTube, and TikTok, supplemented with 7,150 entries from existing Algerian and French medical datasets. Following anonymization, the dataset underwent a rigorous preprocessing pipeline, including orthographic normalization, spelling correction, dialectal standardization, and punctuation cleanup, ensuring linguistic consistency while preserving semantic intent.

The core of the system is a fine-tuned mT5-small transformer model, selected for its multilingual capabilities and encoder–decoder architecture. Training was performed with mixed-precision (FP16) optimization to improve computational efficiency without compromising model quality. Evaluation incorporated both classification metrics (Precision: 47%, Recall: 46%, F1-Score: 47%, Accuracy: 35%) and a generation metric (BLEU: 40%), revealing promising performance given the complexity of the task and linguistic diversity.

The model demonstrated its ability to generate semantically coherent and contextually relevant medical responses in multiple languages and dialects. A prototype web-based interface was developed for interaction with end users, highlighting the potential for practical deployment in low-resource healthcare environments.

Overall, this work advances the application of natural language processing in under-resourced languages, particularly in the medical domain. It lays a scalable foundation for future AI-driven digital health solutions that are inclusive of local linguistic realities and adaptable to multilingual settings.

Keywords: Medical chatbot, Algerian dialect, French, natural language processing, multilingual NLP, healthcare, low-resource languages.

ملخص

في السنوات الأخيرة أصبحت تطبيقات المحادثة ذات أهمية متزايدة في مجال الرعاية الصحية خصوصًا في ظل الحاجة المتزايدة لتوفير استشارات طبية فورية ومتاحة على مدار الساعة، ومع ذلك فإن هذه الأنظمة نادرا ما توفردعما فعليا في العالم العربي لا سيما في المناطق التي تستخدم اللهجات المحلية مثل الجزائر. وذلك بسبب التعقيد اللغوي ونُدرة البيانات الخاصة بهذه اللهجات. تهدف هذه الدراسة إلى سد هذه الفجوة من خلال تطوير روبوت محادثة طبي متعدد اللغات قادر على فهم الأسئلة الطبية المطروحة باللهجة الجزائرية ، اللغة العربية الفصحى والفرنسية وتوليد إجابات طبية مناسبة.

بدأ العمل بجمع ما يقارب 918,74 من الأسئلة والأجوبة. تتضمن سياقا طبيا تم استخراجها من المنتديات الطبية ومنصات التواصل الاجتماعي مثل يوتيوب وفيسبوك بالإضافة إلى مجموعات بيانات سابقة. خضعت البيانات لعمليات تنظيف شاملة شملت تصحيح الأخطاء الإملائية ، إزالة الحروف المكررة وتطبيق تقنيات معالجة اللغة الطبيعية لتطبيع النصوص. تم تدريب النموذج باستخدام الحساب المختلط الدقة لتقليل استهلاك الذاكرة وتسريع وقت التدريب دون التأثير سلبًا على جودة النتائج.

تم تقييم أداء النموذج باستخدام مقاييس التصنيف والتوليد معًا للتحقق من دقة الإجابات المُقدمة من قبل الروبوت. أظهرت النتائج أ ن الروبوت قادر على فهم الأسئلة وتوليد إجابات مناسبة حتى عندما تكون الأسئلة باللهجة الجزائرية أو بالعربية الفصحى. تمثلت النتائج مقاييس التصنيف في تحقيق دقة الترابط :47بالمئة,تذكر:46بالمئة,ف1: بنسبة 47بالمئة,دقة الاقتراب : 35بالمئة و ايضا مقاييس التوليد ببلو : 40بالمئة, تكشف هذه النتائج عن أداء واعد بالنظر إلى تعقيد المهمة والتنوع اللغوي لهذا البحث.

لا يقتصر هذا العمل على تحسين تقنيات المعالجة اللغوية الطبيعية في اللغة العربية والفرنسية فحسب، بل يوفر أيضًا أساسًا قابلًا للتوسع لتطبيقات الذكاء الاصطناعي المستقبلية التي تستهدف اللهجات المحلية كما يُبرز أهمية التصميم اللغوي الشامل والنظر في البنية التحتية عند نشر الذكاء الاصطناعي في البيئات الصحية متعددة اللغات والمحدودة الموارد .

الكلمات المفتاحية : روبوت الدردشة الطبية، اللهجة الجزائرية، الفرنسية، معالجة اللغة الطبيعية ، معالجة اللغة الطبيعية متعددة اللغات ،الرعاية الصحية، اللغات ذات الموارد المنخفضة .

Resume

Ces dernières années, les applications conversationnelles ont pris une importance croissante dans le secteur de la santé, notamment face à la nécessité de fournir des consultations médicales instantanées et accessibles en permanence. Toutefois, ces systèmes offrent rarement un soutien efficace dans le monde arabe, en particulier dans les régions utilisant des dialectes locaux comme l'algérien, en raison de la complexité linguistique et du manque de données dialectales. Ce travail vise à combler cette lacune en développant un chatbot médical multilingue capable de comprendre les questions médicales en dialecte algérien, en arabe standard moderne (MSA) et en français, et de générer des réponses médicales pertinentes.

Un corpus de 74 918 triplets question-contexte-réponse a été constitué à partir de données extraites de forums médicaux, de plateformes sociales comme YouTube et Facebook, ainsi que de jeux de données préexistants. Les données ont ensuite été nettoyées à l'aide de techniques de traitement automatique du langage (TAL), incluant la correction orthographique, la suppression des lettres répétées et la normalisation textuelle. L'entraînement du modèle a été réalisé en précision mixte (FP16), ce qui a permis de réduire la consommation mémoire et d'accélérer l'apprentissage sans compromettre la qualité des résultats.

Le modèle a été évalué à l'aide de métriques de classification (précision, rappel, F1-score, exactitude) et de génération (BLEU), afin de vérifier la qualité des réponses fournies. Les résultats montrent que le chatbot comprend efficacement les questions, même lorsqu'elles sont posées en dialecte algérien ou en arabe standard. Les performances obtenues sont les suivantes : 47% de précision, 46% de rappel, 47% de F1-score, 35% d'exactitude, et un score BLEU de 40%.

Ce projet contribue à l'avancement du TAL pour les langues arabe et française dans le domaine médical, tout en proposant une base solide pour des applications d'IA adaptées aux dialectes locaux. Il souligne l'importance d'une conception linguistique inclusive et de la prise en compte des infrastructures dans le déploiement de solutions de santé numérique dans des contextes multilingues et à faibles ressources.

Mots-clés : Chatbot médical, dialecte algérien, français, traitement du langage naturel, PNL multilingue, santé, langues à faibles ressources.

List of abbreviations

ML : Machine Learning
AI : Artificial Intelligence
DL : Deep Learning
NLP : Natural Language Processing
IE : Information Extraction
NER : Named Entity Recognition
MSA : Modern Standard Arabic
BERT : Bidirectional Encoder Representations from Transformers
GPT : Generative Pre-trained Transformer
T5 : Text-To-Text Transfer Transformer
mT5 : Multilingual Text-To-Text Transfer Transformer
BLEU : Bilingual Evaluation Understudy
LSTM : Long Short-Term Memory
BiLSTM : Bidirectional Long Short-Term Memory
NLTK : Natural Language Toolkit
TF-IDF : Term Frequency-Inverse Document Frequency
CRF : Conditional Random Fields
SVM : Support Vector Machine
RNN : Recurrent Neural Network
CNN : Convolutional Neural Network

Table des matières

Acknowledgments I

Abstract II

ملخص III

Resume IV

List of abbreviations V

TABLE OF Figures VII

List of Tables IX

General Introduction 1

Chapter I: Background and Literature Review 4

1.1 Introduction. 5

1.2 Definition of NLP 6

1.3 NLP Pipeline 7

1.4 NLP for Dialectal Arabic: Focus on Algerian Arabic 9

1.4.1 Modern Standard Arabic vs. Arabic Dialects 9

1.4.2 Linguistic Challenges of Algerian Dialect 11

1.4.3 Code-Switching and Informal Text 12

1.4.4 Existing Tools, Corpora, and Resources for Algerian Arabic 12

1.5 Named Entity Recognition (NER) 14

1.5.1 Definition of Named Entity Recognition 14

1.5.2 Traditional NER Approaches 15

1.5.3 Deep Learning Approaches for NER 17

1.6 Applications of NLP in Healthcare 19

1.6.1 Clinical, Biomedical and Public Health Use Cases 19

1.6.2 Health Discussions on Social Media Platforms 20

1.6.3 Challenges of Informal Health Texts 21

1.6.4 Language Challenges in Dialectal Arabic 23

1.7 Natural Language Processing Approaches 23

1.7.1 Machine Learning Based Approaches 23

1.7.2 Deep Learning Based Approaches 23

1.7.3 Transformers Based Approaches 24

1.8 Gaps in the Literature 28
1.8.1 Lack of Dialectal Arabic Medical Datasets 28
1.8.2 Limitations of Existing Models 29
1.9 Related Work 29
1.10 Conclusion 32
Chapter II. System Architecture and Methodology 34
2.1 Introduction 35
2.2 System Architecture 35
2.2.1 Data Acquisition 40
2.2.2. Data Annotation 41
2.2.3 Preprocessing and Normalization 41
2.2.4 Tokenization 42
2.2.5 Model Selection 42
2.2.6 Fine-Tuning and Optimization 43
2.2.7 Evaluation and Metrics 43
2.2.8 Answer Generation 45
2.2.9 Deployment 46
2.3 ChatBot System 46
2.4 Conclusion 49
Chapter III: Experimental Evaluation 50
3.1 Introduction 51
3.2 Experimental Setup 51
3.3 Examples of source codes 52
3.4 Case Studies and Sample Outputs 56
3.5 Results and Discussion 57
3.6 Interpretation of Results 60
3.7 Practical Implications 61
3.8 Safety, Diagnostic Accuracy and Usability Framework 62
3.9 Conclusion 63
General Conclusion 64
Perspectives 65
Bibliography 66

TABLE OF Figures

Figure 1 :NLP, Artificial Intelligence, Machine Learning and Deep Learning relationship 7
Figure 2 :NLP Pipeline 9
Figure 3 : Example of NER tagging 14
Figure 4 : Illustration of a BiLSTM-CRF model for NER 18
Figure 5 : BERT Architecture [30] 26
Figure 6 : GPT-2 Architecture [56] 27
Figure 7 below show T5 Tasks Examples 27
Figure 7 : T5 Domains Examples 28
Figure 8 : System Architecture 40
Figure 9 : TP,FP,FN,TN 44
Figure 10 : UML Diagram for Chatbot System 47
Figure 11 : how to import essential libraries 52
Figure 12 : Load and Prepare Dataset 52
Figure 13 : Dataset Tonkenization and Splitting 53
Figure 14 : Model and Tokenizer 53
Figure 15 : Training Parametres 54
Figure 16 : Generate Answer Function 55
Figure 17 : Result 1 55
Figure 18 : Result 2 55
Figure 19 : Model Interface 56
Figure 20 : Learning Curves for the mt5-small model 60

List of Tables

Table 1 : Comparison of Modern Standard Arabic and Arabic Dialects. 10
Table 2 : Resources and tools for Algerian Arabic NLP. 14
Table 3 : Comparison of traditional NER methods 17
Table 4 : NLP applications in healthcare 20
Table 5 : Examples of health-related social media platforms and activities 21
Table 6 : Challenges of processing informal health texts and NLP implications 22
Table 7 : Recent Works Comparison table 32
Table 8 : Comparative Analysis of a set of NLP Models 39
Table 9 : Examples of Raw and Preprocessed Algerian Arabic Input 42
Table 10 : Performance comparison of various transformer-based models 57
Table 11 : Medical Chatbots comparison 58

General Introduction

Social media platforms have become central channels for disseminating and accessing health information. Modern patients increasingly rely on online communities (forums, social networks, blogs) to ask medical questions, share personal experiences, and seek peer support [1]. This trend is reflected in the use of Twitter and Facebook for health discussions and in specialized patient communities (e.g. illness forums) where personal narratives circulate rapidly [11]. At the same time, advances in Natural Language Processing (NLP) have enabled a new generation of conversational agents in healthcare. AI-driven chatbots and virtual assistants are now employed to triage symptoms, provide medication guidance, and direct patients to relevant resources [12]. These systems can operate 24/7, extending medical advice into everyday settings and alleviating the burden on clinical staff [7][12]. In fact, recent surveys note the growing impact of NLP in public health and patient engagement, especially under conditions (e.g. pandemics) where remote access to clinicians is limited [38].

Despite this potential, there is a mismatch between available NLP tools and the linguistic realities of many patient communities. In the Arab world, most user-generated health content appears in regional spoken dialects rather than Modern Standard Arabic (MSA) [13]. Algerian Arabic, for example, differs strongly from MSA in its vocabulary, morphology, and orthography [14][15]. In everyday social media and chat contexts, Algerian speakers often mix colloquial expressions with borrowed French or Berber terms, use Romanized transliterations (Arabizi), and omit classical grammatical markers [15][16]. Standard NLP models and pipelines are typically trained on well-formed MSA (or other high-resource languages) and therefore struggle to interpret such noisy, informal inputs [16]. Consequently, off-the-shelf systems tend to underperform on dialectal content, failing to correctly recognize intent or medical entities in these conversations.

These problems are exacerbated by resource scarcity. Arabic NLP has long been limited by a lack of high-quality annotated data, a situation that is even worse for dialectal and medical domains [17][24]. Publicly available medical datasets in Arabic are small and mostly in MSA (e.g. ARMed with only 350 questions, or CQA-MD with imbalanced QA pairs) [53][54], leaving most dialectal varieties unaddressed [24]. Algerian Arabic – spoken by over 40 million people – has virtually no

open medical corpora [17]. Privacy concerns further restrict the collection of patient–doctor dialogues in any language. In low-resource settings, models often overfit limited samples or fail to generalize to real-world queries. Overall, the combination of noisy, informal language and a dearth of labeled data makes extracting accurate medical information from social media a difficult task.

The central research problem of this thesis is therefore: **How can we design an NLP system that understands and generates medical advice in the face of informal, dialectal Arabic content, given scarce annotated resources?**

To address this, we propose a multilingual medical chatbot tailored to Algerian Arabic, MSA, and French.

Objectives

The primary objectives of this thesis are:

Medical Entity Extraction: Develop methods to accurately identify and extract core medical concepts (diseases, symptoms, treatments, medications, etc.) from informal user queries and conversational context.

Dialect and Language Handling: Adapt the system to recognize and interpret multiple Arabic dialects (with focus on Algerian Arabic) and French within a unified framework, including normalization of spelling variations and code-switching patterns.

Advanced Deep Learning Models: Leverage state-of-the-art transformer architectures (such as mT5, AraBERT, Bloom) that provide rich contextual embeddings and multilingual capabilities to better model informal and dialectal medical text.

Research Questions

The study will explore the following research questions:

Adaptation of Transformers: How can transformer-based models (e.g. mT5-small) be fine-tuned or adapted to handle the colloquial, informal style of dialectal Arabic health dialogues?

Multilingual Bridging: Can a multilingual model effectively bridge the linguistic gap between Algerian dialect, Modern Standard Arabic, and French, enabling cross-language understanding and answer generation?

Performance under Constraints: What levels of performance (precision, recall, F1-score, BLEU for generation, etc.) can be achieved on real-world dialectal medical QA tasks using limited training data and the proposed approach?

Thesis Structure

The remainder of this thesis is organized as follows.

Chapter I provides the technical and linguistic background underpinning this work. It begins with an overview of NLP (its tasks, methods, and pipelines) and then discusses the particular challenges of processing dialectal Arabic – especially Algerian Arabic including its diglossic relationship with MSA. The chapter introduces Named Entity Recognition (NER) as a key technique for structuring unstructured text and reviews applications of NLP in healthcare (from EHR mining to chatbots). It concludes with a survey of prior work on medical conversational agents and Arabic NLP, highlighting the gaps that motivate our approach.

Chapter II describes the architecture and methodology of the proposed system: data collection (including corpus creation from social media), preprocessing and annotation procedures, the choice of model (mT5-small) and training regimes, and the integration of NER.

Chapter III reports the experimental setup and results: the datasets and metrics used for evaluation, quantitative performance of the chatbot on entity extraction and answer generation, and qualitative analysis of its outputs.

Finally, the **General Conclusion** summarizes our contributions and discusses future research directions.

Chapter I: Background and Literature Review

Chapter I: Background and Literature Review

1.1 Introduction

This chapter lays out the theoretical foundation for our research at the intersection of NLP, healthcare, and dialectal language processing. We begin by formally defining Natural Language Processing and its core tasks, and then outline the major methodological paradigms (traditional machine learning, deep learning, and modern transformer-based models) that drive current language understanding systems. We describe a typical NLP processing pipeline from text cleaning to tokenization, parsing, and information extraction – with an emphasis on components relevant to medical text. Next, the chapter addresses the unique linguistic characteristics of Algerian Arabic: we compare it to Modern Standard Arabic and analyze its orthographic variability, grammar simplifications, and frequent code-switching with French. We discuss how these features complicate NLP and necessitate specialized normalization and modeling strategies.

Following the linguistic overview, we introduce Named Entity Recognition (NER) as a key tool for extracting structured information (such as disease names and drug terms) from unstructured text. We survey both classical and modern NER approaches, including rule-based systems, statistical sequence models (HMMs, CRFs), and deep learning architectures (BiLSTM-CRF, transformer-based taggers).

The chapter then examines real-world applications of NLP in healthcare. We review how NLP is used in clinical settings (e.g. mining EHRs for diagnoses and medications), biomedical literature mining, and public health surveillance through text analytics. We highlight patient-facing applications, especially chatbots and symptom-checkers that leverage language understanding to provide health advice. We also discuss the use of social media data for health monitoring (e.g. tracking outbreaks or public sentiment). Finally, this chapter concludes with a synthesis of related work on medical chatbots and Arabic NLP: recent efforts in building dialectal medical QA systems and conversational agents in Arabic, French, and English are compared, identifying their successes and limitations. This literature review makes clear the need for advanced, multilingual approaches that accommodate dialects and low-resource scenarios, which motivates our proposed system design.

1.2 Definition of NLP

Natural Language Processing (NLP) is the field of artificial intelligence that focuses on computational methods for **understanding and generating human language** [47]. It encompasses a wide range of tasks – from low-level text processing to high-level semantic interpretation – including tokenization, part-of-speech tagging, syntactic parsing, semantic role labeling, machine translation, summarization, question answering, and dialogue generation [47]. The goal of NLP is to bridge the gap between unstructured human language and structured machine representations.

In practice, an NLP system is often organized as a pipeline. Raw text input is first cleaned and preprocessed: this may involve lowercasing, removing punctuation or diacritics, and splitting the text into sentences and tokens. Each token may be linguistically annotated with attributes like its part-of-speech or lemma [47]. A model is then trained (using machine learning or deep learning) to perform the desired task on these features. For example, in a medical context an NLP system might extract symptom and disease mentions from a clinical note, or it might convert spoken patient queries into text via an automatic speech recognition frontend.

Modern NLP relies heavily on statistical and neural models. Early approaches treated tasks as classification or sequence-labeling problems, using hand-crafted features or word embeddings. Today, the state-of-the-art largely uses deep neural networks that learn distributed representations of words and contexts automatically [20]. A key innovation has been the transformer architecture, which uses self-attention to capture dependencies across an entire input text [25]. Transformer-based models like BERT and T5 (and their multilingual variants) have dramatically improved performance on many NLP tasks by enabling rich contextualization. In particular, encoder–decoder "text-to-text" models (such as T5 and mT5) recast every task (translation, summarization, QA, etc.) as text generation [8][25]. This unified approach simplifies system design and supports transfer learning across tasks and languages.

Figure 1 Shows the relationship between NLP, Artificial Intelligence, Machine Learning and Deep Learning.

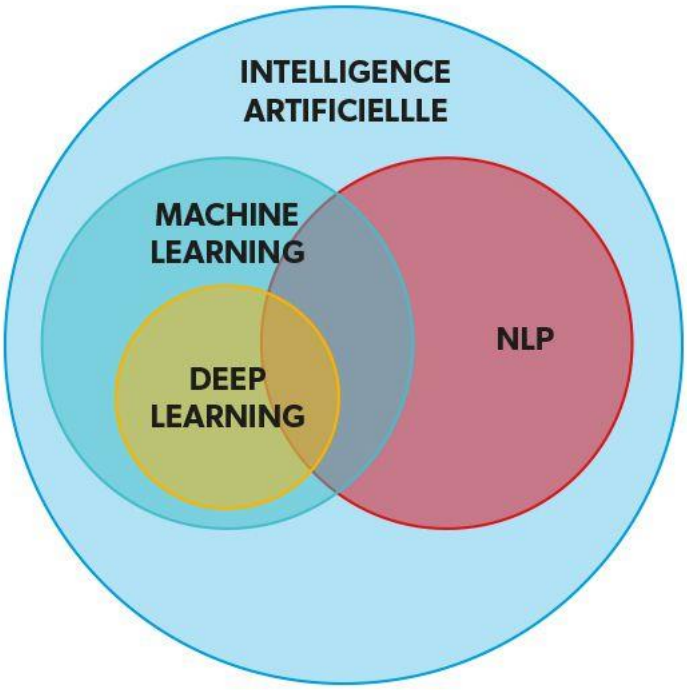

Figure 1 :NLP, Artificial Intelligence, Machine Learning and Deep Learning relationship

1.3 NLP Pipeline

A modern NLP pipeline typically includes the following components, each addresses specific challenges in preparing text for analysis:

Sentence Segmentation: The first step is to identify sentence boundaries. This can be difficult in practice due to abbreviations, ellipses, or nonstandard punctuation (especially in informal text). Rule-based cues (periods, question marks) are often combined with machine-learned classifiers to decide where one sentence ends and the next begins [47]. In mixed-language medical text (Arabic with French insertions), segmentation must handle different punctuation norms and code-switched constructs.

Word Tokenization: After segmenting into sentences, each sentence is split into tokens (words, symbols, or subwords). Tokenization is foundational: it determines the units on which downstream processing operates. For languages like English and French, whitespace and punctuation can be simple heuristics, but for Arabic tokenization is harder due to clitics (attached particles or affixes) and templatic morphology [47]. In our work, we use a subword tokenizer (such as SentencePiece)

that can handle multilingual input efficiently by encoding common subword units across languages [8]. This allows the model to manage rich vocabularies (including mixed Arabic and French content) with a fixed-size dictionary.

Stemming and Lemmatization: These procedures normalize words to their base or root form, reducing sparsity. Stemming applies heuristic rules (e.g. chopping off English suffixes) while lemmatization uses linguistic analysis to yield dictionary lemmas. For French, lemmatization tools like those based on POS tagging can convert inflected forms to their lemma (e.g. “mieux” → “bien”). For Arabic, specialized morphological analyzers (such as Farasa or MADAMIRA) are used to strip affixes and resolve ambiguity [47]. Proper normalization is particularly important in medical text to ensure that different forms of the same term are recognized.

Stop-word Removal: Function words (e.g. “and”, “the”, “من”, “le”) frequently occur but carry little semantic weight. Removing stop-words can reduce noise and dimensionality. However, in medical contexts even common words (e.g. negations like “لا”/“non”) can be crucial for meaning. Thus, our preprocessing retains semantically important tokens while filtering out truly non-informative ones [47].

Part-of-Speech (POS) Tagging: POS tagging assigns each token a grammatical category (noun, verb, adjective, etc.), which is useful for subsequent steps like lemmatization or parsing. State-of-the-art POS taggers use neural models (e.g. BiLSTM-CRF or transformers) trained on annotated corpora [47]. Accurate POS tagging helps identify key medical terms (often nouns for conditions, verbs for actions) and can improve downstream NER and syntactic analysis.

Dependency Parsing: Dependency parsing builds a syntactic structure by linking words with directed arcs that indicate grammatical relations (subject, object, modifier, etc.) [47]. For example, in “the patient reports pain,” a dependency parser would link “reports” (the head verb) to “pain” (its object). In clinical questions, dependency relations help disambiguate the roles of words in multi-clause sentences or nested queries.

Named Entity Recognition (NER): NER is the process of locating and classifying spans of text into predefined categories (e.g. Disease, Symptom, Drug, BodyPart) [20]. In medical NLP, NER transforms raw text into structured semantic units. For instance, in “fever and cough” the system

would tag “fever” and “cough” as symptom entities. We employ pre-trained NER models fine-tuned on medical data, capable of recognizing entities even in noisy, dialectal input [20]. Accurate NER is critical for extracting patient-specific information from unstructured dialogue.

Figure 2 below illustrate NLP Pipeline Steps.

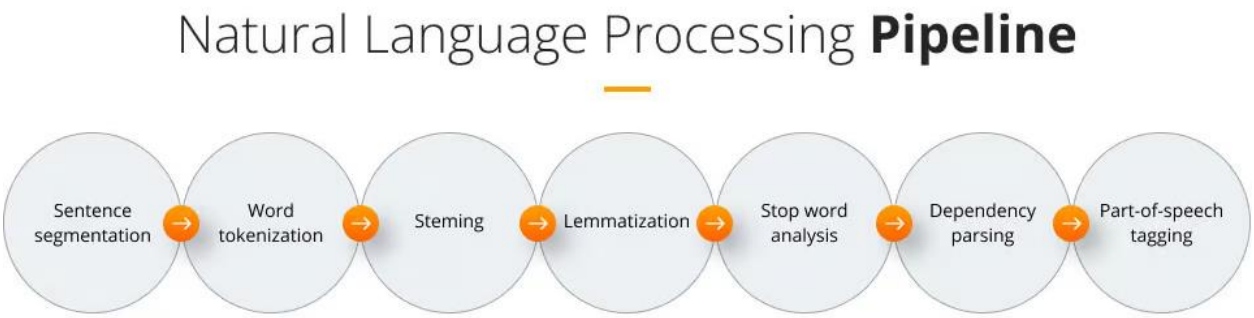

Figure 2 :NLP Pipeline

1.4 NLP for Dialectal Arabic: Focus on Algerian Arabic

1.4.1 Modern Standard Arabic vs. Arabic Dialects

Arabic exhibits a **diglossic** relationship between Modern Standard Arabic (MSA) and the various regional dialects [13]. MSA is a standardized, formal register taught in schools, used in news media, literature, and official communication [14]. It has a fixed grammar, rich inflection, and a largely uniform written form with prescribed spelling rules. In contrast, dialectal Arabic comprises the informal spoken varieties used in everyday conversation; these include Egyptian, Levantine, Gulf, Maghrebi (North African) dialects, among others. Dialects are learned natively and exhibit high variability: each has its own lexicon, simplified morphology, and pronunciation norms [13][14]. Importantly, dialects have no single standard orthography. The same word may be written in multiple ways even within the Arabic script, and dialect speech is often transcribed using Latin characters (Arabizi) as well.

Table 1 (below) compares key features of MSA versus an exemplar dialect (Algerian Arabic). It highlights, for instance, that dialects often omit case endings and dual forms present in MSA, and incorporate many loanwords (e.g. French or Berber terms) [15]. MSA enforces uniform spelling, whereas dialect writing is inconsistent – for example, the colloquial word for "gold" might appear as *dhahab* or *dhab* in Arabic script. This mismatch is sociolinguistic as well as linguistic: MSA is associated with formal education and pan-Arab communication, while dialects carry local identity and are preferred in casual contexts. The gulf between MSA and dialects can greatly hinder mutual intelligibility and NLP processing, since most NLP resources (taggers, parsers, lexicons) are built for MSA [14].

Feature	**Modern Standard Arabic (MSA)**	**Dialectal Arabic (e.g. Algerian)**
Status	Official, standardized language used in formal writing, media, education	Informal spoken variety; no official status or standardized orthography
Usage	Employed in news broadcasts, literature, and formal lectures; learned in schools	Used in everyday conversation, popular culture, and local media
Vocabulary	Largely fixed lexicon inherited from Classical Arabic; neologisms from pan-Arab sources	Rich local vocabulary with many loanwords (e.g. French, Berber influences)
Phonology	Maintains classical sounds and pronunciation rules taught academically	Varied phonology by region (e.g. Algerian Arabic realizes ق/q/ as [q],[g],[k] in different areas)
Grammar	Complex inflection with case endings and dual/plural forms	Simplified grammar; many dialects drop case endings and use different pluralization strategies
Orthography	Uniform spelling and orthographic conventions (Arabic script with diacritics)	No standard spelling; written forms vary (e.g. **dhab** vs **dhahab** for 'gold')

Table 1: Comparison of Modern Standard Arabic and Arabic Dialects.

1.4.2 Linguistic Challenges of Algerian Dialect

Algerian Arabic Dialec poses several linguistic obstacles for NLP.

First, there is **orthographic ambiguity**. Algerian dialect has no codified writing system, so words can be spelled in multiple ways. Writers may use informal spellings or standard ones interchangeably. for example, the dialectal word for **'gold'** might appear as **dhahab** or **dhab**, and vowel length can be marked inconsistently [15]. Moreover, social media text often contains Romanized Arabic (Arabizi), where Arabic words are written in Latin letters (e.g. *bslama* for "بسلامة") or mixed script. In one Algerian tweet corpus most messages were in Arabic script, but also a big amount were written with Latin characters [17]. This inconsistency makes preprocessing difficult. Normalization of inconsistency in the use of certain letters, dialect words is a key challenge in social-media Arabic text [16].

Second, there is **lack of standardization** in vocabulary and grammar.Algerian dialect has been influenced over centuries by other languages Takes a lot From Berber, Turkish, Italian, Spanish and French due to their history,but there is no authoritative dictionary for these terms. which complicates lexical coverage [15]. Additionally, morphosyntactic rules vary: for instance, dual noun forms have largely disappeared in Algerian dialect [15]. These dialect-specific patterns such as verb conjugations and word order are under-documented, making taggers and parsers prone to error.

Third, **phonetic and phonological variability** is high in Algerian Arabic. A single phoneme in MSA can have multiple dialectal realizations. For example, the classical Arabic letter ق **(/q/)** is pronounced as [q], [g], or [k] in different Algerian regions [15]. Such variation means speech and text can deviate unpredictably from standard pronunciations. In written dialectal Arabic even in Arabic script, short vowels are typically omitted (as in MSA), further increasing ambiguity. Together, these issues inconsistent spelling, borrowed lexicon, and variable phonology make Algerian Arabic a challenging target for standard NLP tools.

1.4.3 Code-Switching and Informal Text

A further complication is **code-switching** and informality in online communication. Algerian speakers frequently insert French words or phrases into their Arabic dialogue without translation. During French colonial rule, many French terms entered local parlance (e.g. *contrôle*, *danger*, *merci*) and remain in everyday use [15]. In digital chat, a user might write a sentence mixing Arabic and French seamlessly. For NLP, this means models must handle mid-sentence language shifts and foreign tokens.

Online user-generated text adds additional noise. Social media messages can include **Romanized Arabic**, emojis, abbreviations, elongated letters (e.g. "salaaaam"), and non-standard punctuation [16]. Spelling errors and creative orthography (e.g. "frrrouge" to exaggerate "red" in French) are common. Language identification itself becomes difficult: a dialectal Arabic sentence might be tagged as French or "other" if written in Latin script. In one study, every token of an Algerian tweet had to be classified into categories (Algerian, MSA, French, Berber, English, other). The prevalence of code-switching and Arabizi means that many off-the-shelf tools (designed for clean Arabic text) simply break. We therefore incorporate robust normalization in our pipeline: collapsing character repetitions, mapping slang to formal equivalents, and using dialect-aware lexicons. Even so, processing informal text requires special attention. Overall, the mixture of dialect, French, and orthographic creativity in Algerian social media poses a major challenge for NLP systems [15][16].

1.4.4 Existing Tools, Corpora, and Resources for Algerian Arabic

Until recently, Algerian Arabic was under-resourced. However, researchers have begun to compile dialectal datasets. In the **speech domain**, for example, Harrat et al. collected several corpora: **AMCASC** (735 telephone dialogues across three Algerian regions, ~72 hours) and **ALG-DARIDJAH** (a parallel multi-dialect read-speech corpus, ~4.5 hours) [42]. A larger project, **KALAM'DZ**, crawled online media to assemble 104 hours of transcribed Algerian speech from 4,881 speakers across major dialect areas [42]. These speech corpora support automatic speech

recognition (ASR) and dialect ID tasks, which can be leveraged to preprocess spoken queries. (Table 2 summarizes these and other resources.)

In the text domain, **Twifil** and **Narabizi** corpora, containing thousands of Algerian tweets annotated for sentiment and topic [58]. These cover both Arabic-script and Romanized messages, reflecting real usage. Furthermore, an Algerian NER dataset (DziriNER) has been created to label entities in dialect text [23]. On the modeling side, the DziriBERT language model represents a major breakthrough: Abdaoui et al. pretrained a BERT variant on 1 million Algerian Arabic tweets and achieved state-of-the-art performance on dialectal sentiment and topic classification [17]. This shows the value of dialect-specific pretraining. Open-source toolkits are also evolving: for instance, CAMeL Tools now includes an Algerian dialect identification component [34], and there are ongoing efforts to extend morphological analyzers (e.g. FARASA) to Maghrebi varieties.

In summary, while resources for Algerian Arabic have been limited, recent work has produced valuable corpora and models [42][17]. Speech and text datasets now exist for dialect identification and normalization, and models like DziriBERT offer strong baselines.

Table 2 highlights key available datasets both speech and text and tools.

Resource	**Type**	**Description**	**Ref**
AMCASC	Speech corpus	3 regional Algerian dialects; 735 speakers, ~72 h of telephone recordings	[42]
ALG-DARIDJAH	Speech corpus	Parallel corpus of Algerian sub-dialects; 109 speakers, ~4.5 h (direct recordings)	[42]
KALAM'DZ	Speech corpus	Web-mined Algerian dialect (8 regions); 4,881 speakers, 104.4 h of segmented speech	[42]
Twifil	Text corpus	9,437 Algerian tweets (Arabic & Roman script) annotated for sentiment & emotion	[58]
Narabizi	Text corpus	1,279 Algerian Arabic sentences (Roman script) annotated for sentiment & topic	[58]
DziriBERT	Language model	BERT-based model pretrained on 1M Algerian dialect tweets; achieves state-of-art	[17]

		on dialect tasks	
CAMeL Tools (Dialect ID)	Toolkit	Dialect identification module (25 Arabic city dialects including Algerian)	[59]
DZDialect	Corpus	A new dataset comprising 117569 annotated Comments in Algerian Dialect	[57]

Table 2: Resources and tools for Algerian Arabic NLP.

1.5 Named Entity Recognition (NER)

1.5.1 Definition of Named Entity Recognition

Named Entity Recognition (NER) is the NLP task of detecting and classifying **named entities** in text into predefined categories [19]. Typical categories include **Person** (PER), **Location** (LOC or GPE for geopolitical entities), **Organization** (ORG), and other classes such as **dates/times**, **quantities**, or **miscellaneous**. For example, in the Figure Below "Mary" is tagged as a PERSON, "The Ritz" as an ORG, and "London" as a GPE. NER transforms unstructured text into structured information, which benefits downstream tasks such as information retrieval, question answering, and text summarization [19] [20]. In practical NLP pipelines, NER is often used after tokenization and POS-tagging to enrich text with semantic labels for knowledge graph construction or summary generation. Figure 3 Shows an example processed using NER.

Figure 3: Example of NER tagging

Common entity categories are illustrated below named entity types are typically abbreviated in corpora and models:

Person (PER): Names of people (e.g. Albert Einstein, صلاح).

Location (LOC/GPE): Geographical places (e.g. Paris, Guelma).

Organization (ORG): Companies, institutions, agencies (e.g. UNESCO, سوناطراك).

Miscellaneous (MISC) or **Other**: Entities not fitting the above (dates, events, products, etc., e.g. “5th May 2025” as DATE, “DZD 100” as MONEY) .

NER answers the “who, where, what, when” of text. By identifying names of people, places, organizations, etc., it also provides structured annotations that can, for example, improve the relevance of search results or guide summarization by extracting key entities for summary focus.

1.5.2 Traditional NER Approaches

Early NER systems relied on handcrafted rules or statistical sequence models [50]. **Rule-based** methods use regular expressions, grammars, and gazetteers (lists of known names) to spot entities. These can achieve high precision for patterns they cover, but require extensive expert effort and often have very low recall on unseen data [50]. For instance, a rule-based Arabic NER system might recognize personal names using patterns of letters but fail whenever a new name appears or if the input is noisy. **Dictionary lookup** is a simple variant: any exact match to an entries database is tagged. This is fast but inflexible: it misses any entity not in the list.

By contrast, **statistical machine learning** treats NER as a sequence-labeling problem. Models such as Hidden Markov Models (HMMs), Maximum Entropy classifiers, or Conditional Random Fields (CRFs) learn from annotated corpora [50]. A CRF-based NER tagger can model dependencies between neighboring token labels (e.g. “B-Disease” followed by “I-Disease”). Such models do not need hand-coded rules and can capture contextual cues. In Arabic, classical NER systems often used features like word prefixes/suffixes, part-of-speech tags, and lexicons. For example, a CRF tagger trained on an Arabic newswire corpus might learn that capitalized words are likely locations or organizations. Studies show CRFs generally outperform simpler Maximum Entropy models on NER when enough data is available [20].

However, statistical models have limitations. They require substantial annotated data to learn good feature weights, and feature engineering is labor-intensive. In low-resource or highly inflected languages like Arabic, out-of-vocabulary words and complex morphology can degrade performance. Early Arabic NER used news corpora (ACE, ANERCorp) to get sufficient training examples [20], but dialectal or medical text often contains vocabulary absent from those corpora.

In summary, traditional NER approaches laid important groundwork but struggle with recall and domain adaptation [50][20].

Table 3 (below) compares these traditional methods in terms of their strengths and weaknesses

Approach	**Methods (Examples)**	**Strengths**	**Limitations**	**Ref**
Rule-based / Gazetteers	Hand-crafted rules, regexes, gazetteer lookups (e.g., LaSIE-II, NetOwl, FASTUS, RENAR)	No training data; high precision on known entities	Labor-intensive; low recall; brittle; not easily portable	[50]
Dictionary-based	Pure lexicon/gazetteer matching	High precision for known items; fast	Cannot generalize; coverage limited by lexicon scope	[50]
Statistical Supervised ML	HMM, MaxEnt, SVM, CRF (e.g., Bikel et al. 1999; Shen et al. 2003)	Captures context; CRF excels at sequence modeling; reliable with enough annotated data	Needs labeled corpora; extensive feature engineering; struggles with OOV words	[51]
Semi-supervised	Co-training, self-training, bootstrapping from seeds (e.g., Collins & Singer 1999; Watanabe 2003)	Leverages unlabeled data; reduces labeling cost	Error propagation risk; sensitive to seed quality	[52]
Unsupervised / Clustering	Pattern extraction, clustering, bootstrapping (e.g., Etzioni 2005 KNOWITALL; Nadeau et al. 2006)	No supervision needed; customizable to new domains	Low precision; inconsistent; unpredictable; requires tuning	[50], [51]
Hybrid	Rule + ML, cascaded systems (e.g., rules	Combines rule precision with ML flexibility;	Architecturally complex; still dependent on	[20], [50]

	feeding CRF/SVM; grammar + HMM combinations)	strong overall performance	annotated data and expert-crafted rules	

Table 3: Comparison of traditional NER methods

1.5.3 Deep Learning Approaches for NER

Word embedding are fed into a bidirectional LSTM, whose outputs are passed to a CRF layer that predicts entity tags jointly for the sequence [21]. In recent years, deep neural models have become dominant in NER. Architectures typically use word embeddings and character-level encodings as input, feeding these through sequence models like recurrent neural networks. The BiLSTM-CRF framework a bidirectional LSTM followed by a CRF layer is now a standard architecture: the bidirectional LSTM captures context from both directions, and the CRF layer enforces valid tag sequences [21]. Character-level encoders are often used to model sub-word morphology, which is critical for morphologically rich languages like Arabic. Overall, deep models learn features automatically and can be trained end-to-end, reducing the need for manual feature engineering and often yielding more robust representations than linear models [20].

Convolutional Neural Networks (CNNs) have also been applied – typically to extract n-gram features or embed character sequences – but most NER systems use CNNs in combination with RNNs [21]. The move to deep learning for Arabic NER has substantially improved performance. For example, fine-tuning pre-trained models on Arabic NER corpora achieves higher accuracy: the CAMeL Lab fine-tuned an **AraBERT** model on the Arabic ANERcorp dataset and reached 83% F1, outperforming a CRF-based system by about 4 percentage points [20]. Figure 4 is an Illustration of a BiLSTM-CRF model for NER.

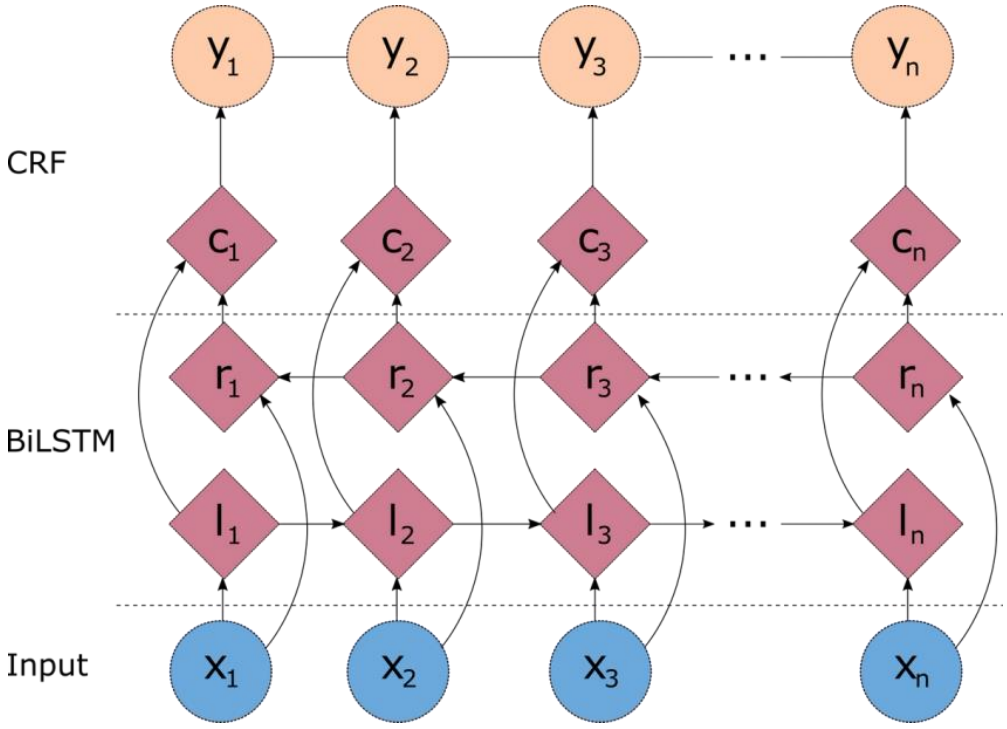

Figure 4: Illustration of a BiLSTM-CRF model for NER

Following BERT's success, researchers have developed Arabic BERT models and dialectal BERTs. These transformers are first trained on large Arabic text corpora and then fine-tuned for NER. Such models inherently capture rich contextual information and mitigate some Arabic-specific issues (e.g. missing short-vowel diacritics). As one survey notes, models like AraBERT and AraELECTRA can relieve the ambiguity caused by absent diacritics during NER. In practice, AraBERT and similar models generally outperform earlier approaches on Modern Standard Arabic NER [20]. Other BERT variants include MARBERT and ARBERT, which provide better handling of dialectal or informal text.

For dialectal Arabic and low-resource variants, specialized models have been created. For instance, **DziriBERT** is a transformer pre-trained on Algerian Arabic text. Recent work on Algerian NER compared several models and found that these dialect-aware models can be fine-tuned for the task [22]. However, dialectal NER remains challenging. The Arabic language has over twenty regional dialects with significant spelling, vocabulary, and syntactic variations [23]. There is no standard orthography for many dialects and annotated data is extremely scarce. As a result, even powerful deep models struggle with unseen dialectal forms or code-switched Arabic [23]. In summary, deep learning – especially transformer-based models has greatly advanced Arabic NER performance, but dialect-specific challenges require ongoing adaptation and data collection.

1.6 Applications of NLP in Healthcare

1.6.1 Clinical, Biomedical and Public Health Use Cases

NLP has become a critical technology in healthcare by enabling the extraction and interpretation of unstructured text data [7]. Clinical settings generate vast free-text data (doctor's notes, discharge summaries, lab reports) that contain valuable patient information. Advanced NLP systems can automatically identify and structure key entities from these texts: diagnoses, medications, procedures, patient history, risk factors, etc. For example, an NLP pipeline might extract "Type II diabetes" from a narrative and map it to a standardized disease code. This structured data can then feed into clinical decision support tools or patient stratification models, enhancing efficiency and reducing manual data entry [7]. Speech-to-text systems also play a role, converting dictated clinical narratives into digital records in real time, which improves documentation completeness and physician workflow. Additionally, NLP is used to match patients to clinical trials by analyzing their EHRs against trial eligibility criteria[7], dramatically speeding up recruitment and ensuring more accurate matching. These applications demonstrate that NLP can directly improve care quality and clinician productivity.

Beyond individual care, NLP is invaluable in biomedical research. It powers literature mining tools that sift through thousands of research papers, identify emerging trends, and summarize findings [10]. For instance, NLP-based text classification and information retrieval help researchers track the latest developments in a disease area or assemble evidence for systematic reviews [10]. This accelerates knowledge discovery, such as finding novel drug targets or monitoring new treatments across journals. In epidemiology and public health surveillance, NLP analyzes reports, news, and social media to detect disease outbreaks early [11]. Algorithms trained to spot symptom mentions or unusual patterns in text streams have been used to give early warning of influenza, dengue, and even COVID-19 spread, often before official alerts. By mining large text corpora, public health authorities can respond more proactively to emerging health threats.

On the **patient-facing** side, NLP underlies many virtual health assistants and chatbots [12] [38]. These conversational agents interpret natural language queries about symptoms or medications and generate contextually appropriate responses. Common functions include symptom triage (guiding a

patient on whether to seek care), medication reminders, appointment scheduling, and providing 24/7 access to basic health advice [12]. NLP-based symptom-checkers take user descriptions (e.g. “I have a headache and fever”) and suggest possible causes or next steps based on medical knowledge. Moreover, sentiment and topic analysis of patient feedback or social media can reveal patient attitudes toward treatments and alert providers to emerging dissatisfaction or psychosocial issues [7]. Overall, NLP enables personalized, scalable healthcare interactions outside of the clinic.

Table 4 (below) summarizes representative NLP applications across healthcare domains.

Application	**Example / Description**	**References**
EHR Information Extraction	Automated extraction of diagnoses, medications, procedures from clinical notes	[7][10]
Pharmacovigilance	Detecting adverse drug reactions from social media posts	[11]
Clinical Decision Support	Matching patients to trials via EHR search; evidence summarization	[7][10]
Patient-Facing Chatbots	AI chatbots for symptom checking, reminders, scheduling	[10]
Public Health Surveillance	Monitoring outbreaks via social media and news feed analysis	[7][10]

Table 4: NLP applications in healthcare

1.6.2 Health Discussions on Social Media Platforms

Social media and online forums are now important venues for patients and caregivers to share health experiences and advice [11]. Users may post about personal illness journeys, chronic conditions, or preventive health behaviors. Platforms like Twitter, Facebook, and Instagram (general networks) see rapid spreading of health news and public campaigns, while dedicated communities (DailyStrength, PatientsLikeMe, health subreddits) allow peer-to-peer support and detailed discussions [11]. For example, cancer survivor forums and mental health groups provide emotional support and practical tips; on platforms like Reddit or Health StackExchange, lay users can pose questions to medical professionals or informed volunteers. Hashtags and trending topics can even transform into citizen-driven public health alerts (e.g. #flu hashtag spikes).

NLP analysis of these discussions has yielded valuable insights for public health. Mining social media text has been used to gauge public sentiment on vaccinations, monitor side effects of new medications, and track symptoms during outbreaks [11]. For instance, during the COVID-19

pandemic researchers applied NLP to millions of tweets to model disease prevalence and vaccine attitudes [11]. Patient-generated content also constitutes a form of "social support data"; analyses have shown that engagement in online patient communities can correlate with better self-reported outcomes. However, social media data also pose challenges: they contain misinformation, and user demographics may not be representative of all populations. Nonetheless, the sheer volume of health-related posts makes them an increasingly important data source. Crucially for this thesis, social media health discussions are often **multilingual and informal**, underscoring the need for robust multilingual NLP models (like mT5) to process content from global platforms [9][38].

Table 5 (below) lists examples of social media platforms and their health-related uses.

Platform / Community	**Health-related Use**
General Social Networks (Twitter, Facebook, Instagram)	Sharing personal health experiences and news updates; public health campaign monitoring
Online Health Communities (DailyStrength, MedHelp, PatientsLikeMe)	Peer-to-peer advice and support for chronic illness, medication, lifestyle
Patient Advocacy Forums (e.g., cancer forums, mental health forums)	Emotional support, coping strategies, patient education (via user posts)
Professional Q&A Forums (e.g., Health StackExchange, r/AskDocs)	Anonymous Q&A with health professionals or informed volunteers

Table 5: Examples of health-related social media platforms and activities

1.6.3 Challenges of Informal Health Texts

Health-related text on social media and online communities is often *informal* and noisier than clinical text. Common challenges include nonstandard language, misspellings, slang, abbreviations, and emoticons. For example, users may write "diareah" or use "smh" to express symptoms or emotions. Many social media posts lack correct grammar or punctuation, making it hard to parse sentence structure [11]. This necessitates additional preprocessing to map lay expressions to formal terms. In health forums specifically, laypeople often use imprecise descriptions: they might say "stomach bug" instead of "gastroenteritis" or "high blood" instead of "hypertension" [11]. These

lexical mismatches require NLP systems to have robust dictionaries or contextual models to interpret user language.

Multilingual challenges compound these issues. Informal posts may code-switch between languages or dialects, especially in global platforms. Furthermore, domain-specific training data are scarce for many languages. Large corpora of medical text in languages other than English are often inaccessible, limiting model training [9]. In short, the informal nature of social media text (slang, errors..) and the multilingual diversity of user posts present significant obstacles. NLP systems for health data must handle noise, adapt to varied vocabularies, and often rely on transfer learning or language models pre-trained on large multilingual datasets to mitigate these challenges [9] [11]. Table 6 demonstrates the Challenges of processing informal health texts and NLP implications:

Challenge	**Description**	**Implication**	**Ref**
Informal Language	Slang, abbreviations, emojis, and poor grammar are common (e.g. “been 2 hosp” for hospital).	Requires normalization and noise-robust parsing; standard models often fail.	[11]
Lay Terminology	Users use non-technical descriptions (“upset stomach” vs. “gastritis”).	Need mapping from colloquial to clinical terms (semantic lexicons or context models).	[11]
Data Sparsity (Multilingual)	Limited labeled medical text in many languages; code-switching between languages.	Necessitates cross-lingual transfer (e.g. multilingual models like mT5) and creative annotation strategies.	[9]
Misspellings and Typos	Frequent spelling errors (e.g. “headake” for headache).	Requires spell-correction or phonetic matching to avoid vocabulary mismatch.	[9]

Table 6: Challenges of processing informal health texts and NLP implications

Overall, these challenges mean that off-the-shelf NLP tools (trained on formal text) often underperform on user-generated health data [9].

1.6.4 Language Challenges in Dialectal Arabic

Arabic dialects differ significantly from Modern Standard Arabic in vocabulary, grammar, and spelling. This poses unique challenges for our chatbot. Unlike MSA (which has a standardized orthography), dialectal Arabic has no single official spelling. Users may write dialect words in Arabic script or in Latin-based transliteration, and spelling can vary widely for the same word [33]. For example, the letter ("ق", "qaf") may be pronounced as /q/, /g/, or as a glottal stop in different dialects, leading speakers to spell words like ("قال" ,"he said") with different conventions: e.g., *gāl* vs *'āl* [33].

Linguistically, dialects exhibit **lexical** and **syntactic** differences from MSA. Dialectal Arabic often uses different words for common concepts and has simpler inflections. For instance, the MSA phrase *Almuwadafan akala* (أَلْمُوَظَّفَانِ أَكَلَا, "the two employees ate") uses dual-number marking and case endings. In Egyptian dialect this becomes *al-muwadifīn al-athnin akalu* ("the two employees ate") – dual marking is lost and replaced by a plural form [33]. Dialects also omit many of MSA's grammatical markers and rely on critics heavily. In summary, lexical, syntactic, semantic and pragmatic variations abound between MSA and Dialectal Arabic [33]. These variations make tokenization and normalization difficult and require tailored preprocessing.

1.7 Natural Language Processing Approaches

1.7.1 Machine Learning Based Approaches

Traditional NLP largely relied on **machine learning** algorithms with hand-crafted features. Text would be represented via features such as TF-IDF, bag-of-words, n-grams, or counts of linguistic attributes (e.g. POS tags, morphological prefixes) [47]. Classical classifiers (Support Vector Machines, Naive Bayes, Decision Trees) could then be trained on these features for tasks like text classification or basic sequence labeling. These methods are efficient on small to medium datasets and offer interpretability (since feature weights indicate importance). In the context of NER and sequence labeling, **Conditional Random Fields (CRFs)** became a standard ML solution [50]. CRFs are discriminative probabilistic models that learn to predict a sequence of labels given a

sequence of observed features, capturing dependencies between labels (e.g. ensuring that an "Inside" tag follows a "Begin" tag). CRFs were widely used in early Arabic NLP systems for tasks like token tagging, especially when no deep learning models were available [20].

In low-resource settings (such as dialectal Arabic), ML methods still play a role. They require far less computational power than deep networks and can work with smaller amounts of data. Early experiments in Arabic NER on social media used CRFs, Hidden Markov Models, and even rule-based hybrids as baselines [20][22]. These methods provided reference points for later improvements. However, feature-driven ML approaches have well-known drawbacks: manual feature engineering is laborious, and models struggle to capture long-range context or subtle semantics. Ambiguity and polysemy in language are difficult to encode with fixed features. As a result, purely ML-based systems have largely been supplanted by deep learning in recent research, although they may still be used as components (e.g. a CRF layer on top of neural embeddings) or in pipeline integration where transparency is needed.

1.7.2 Deep Learning Based Approaches

Deep learning Approaches revolutionized NLP by replacing manual features with neural networks that learn representations automatically [21]. Two common neural architectures are Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). RNNs, and especially Long Short-Term Memory networks (LSTMs), process text sequentially, maintaining a "memory" of past tokens. Bidirectional LSTMs (BiLSTMs) process input in both forward and backward directions, effectively capturing context around each word [21]. CNNs, on the other hand, apply sliding filters to capture local n-gram features.

In practice, many NLP tasks (including NER, sentiment analysis, and text classification) have employed BiLSTM or CNN models. For example, a BiLSTM–CRF architecture (BiLSTM followed by a CRF output layer) became a state-of-the-art solution for NER [21]. In Arabic medical NLP, DL models have also shown promise. CNNs have been used to detect clinical entities or categorize note sections [10]. Generally, DL models outperform classical ML when sufficient labeled data is available, because they can capture complex patterns without explicit feature design.

However, traditional DL models have limitations. RNNs are inherently sequential and can be slow to train on long texts; they may struggle with very long dependencies despite LSTM gating mechanisms. Both RNNs and CNNs lack explicit global attention, making it harder to focus on relevant parts of long documents. The advent of **pretrained word embeddings** (Word2Vec, GloVe, fastText) helped DL models by providing semantic vectors learned from large corpora, which improved generalization and convergence. But these static embeddings could not model polysemy (words with multiple meanings) or adapt easily to new contexts.

In Arabic NLP, LSTM-based models have been widely used because they handle morphological richness well. Early dialectal NER models, for instance, used BiLSTM layers to avoid manual lexicon creation [22]. Yet, deep models still require large annotated datasets to reach their potential – a challenge in low-resource domains. This shortfall eventually led to the adoption of transformer-based models, which we discuss next.

1.7.3 Transformers Based Approaches

Transformer-based architectures have fundamentally transformed the landscape of natural language processing by introducing self-attention mechanisms capable of modeling global dependencies within input sequences. Originally proposed in 2017 [25], the Transformer model replaced recurrence with fully parallelized self-attention, dramatically improving both training efficiency and representational power.

The original Transformer architecture consists of encoder and decoder blocks stacked with multi-head attention layers and feedforward networks. This structure enabled the development of three broad classes of models:

1)Encoder-only models: e.g. BERT [30]. and its variants. BERT uses a transformer encoder stack and is pretrained on a masked language modeling objective, learning bidirectional context. It excels at classification and understanding tasks (textual entailment, NER, etc.) because it captures deep contextual representations. Multilingual BERT (mBERT) covers over 100 languages in one model but is not generative. Later Arabic models (AraBERT, Arabic Electra) refined this approach, improving performance on Arabic by training on large Arabic corpora [30].

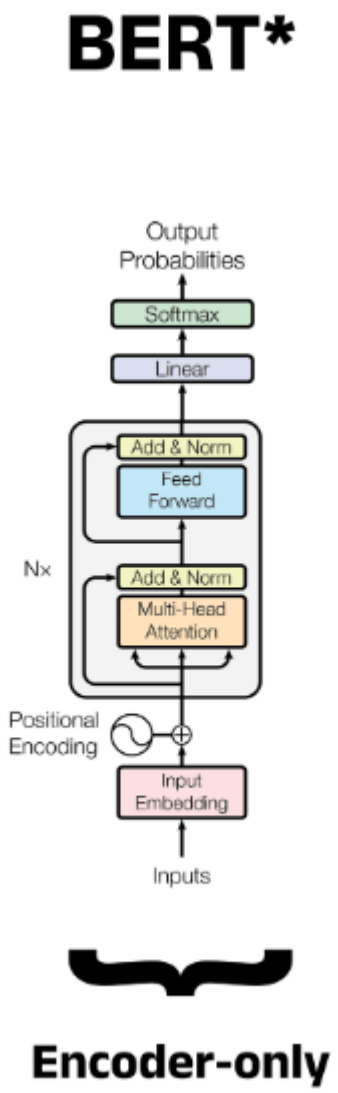

Figure 5: BERT Architecture [30]

2)Decoder-only models: e.g. GPT-2 [56]. These use the transformer decoder stack and are pretrained with an autoregressive objective (predict next word). They are optimized for fluent text generation (completions, open-ended dialogue), but their unidirectional training can limit understanding of complex context. GPT-2 became famous for generating coherent paragraphs of English text, though it is not bilingual by default.

Figure 6 demonstrates the GPT-2 Architecture.

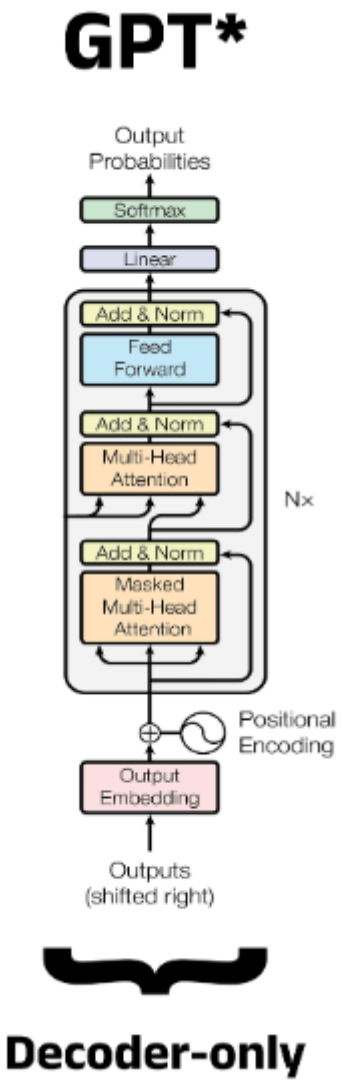

Figure 6: GPT-2 Architecture [56]

3)Encoder–decoder (Seq2Seq) models: e.g. T5 and its multilingual variant mT5 [8][25]. These models combine an encoder stack and a decoder stack, making them naturally suited for tasks that require both understanding an input and generating an output (such as machine translation, summarization, and question answering). T5 reframed many NLP tasks as "text-to-text": for example, translating, summarizing, and answering questions are all cast as generating a target text conditioned on a prompt. mT5 extends T5 by pretraining on mC4, a massive multilingual dataset covering 101 languages [8]. As such, mT5 is capable of cross-lingual transfer: it can apply knowledge learned in one language to another, which is valuable for a multilingual chatbot.

Figure 7 below show T5 Tasks Examples

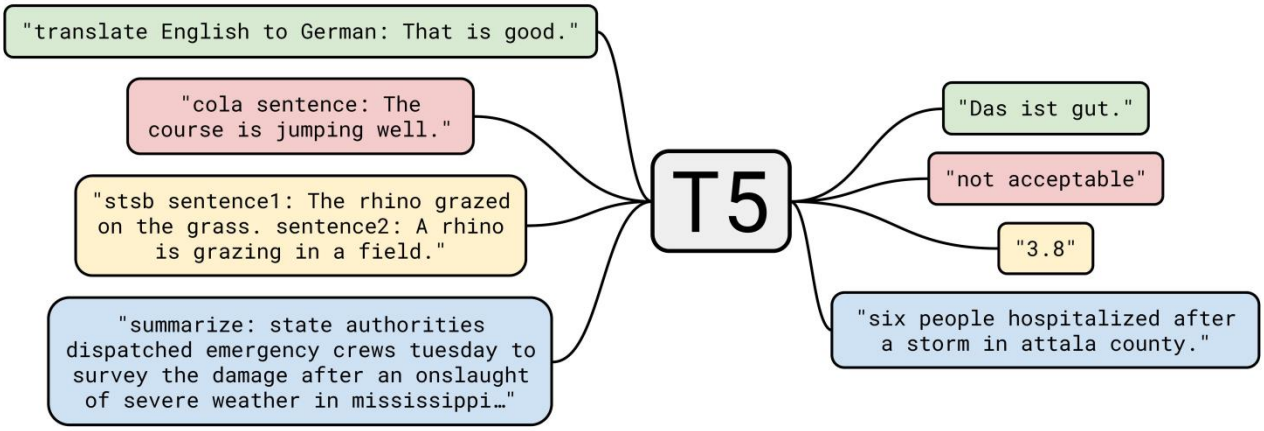

Figure 8: T5 Domains Examples

1.8 Gaps in the Literature

1.8.1 Lack of Dialectal Arabic Medical Datasets

Existing datasets for Arabic medical NLP are exceedingly scarce and generally confined to MSA or limited domains. For example, the only publicly known Arabic medical question-answering corpora are ARMed (350 consumer questions) and CQA-MD (57,764 QA pairs) [24]. CQA-MD itself is highly imbalanced: only 1,943 of its 57,764 answers are directly relevant to queries (≈4%). Also note that even these resources leave many dialects unaddressed, reflecting "the lack of resources, the morphological complexity of Arabic, and the multiple forms and dialects" used in online medical communication. In contrast, the new AraMed dataset (270k QA pairs from the Altibbi forum) provides a much larger MSA corpus, but it still does not capture regional dialectal expressions or local health idioms [24].

In practice, no annotated corpus for medical text in Algerian Arabic is publicly available. The Algerian dialect "is spoken by 44 million people but lacks publicly available datasets" [17]. This dearth holds across all medical tasks: there are no known dialectal Arabic corpora for clinical narrative, symptom-checking dialogs, or patient–doctor communications. Moreover, privacy and sensitivity of health records mean even fewer opportunities to collect data [24]. In sum, researchers in Arabic medical NLP must rely on either small MSA datasets or synthetic data, highlighting that none cover dialectal Arabic:

ARMed : 350 Arabic medical questions (mixed dialects) [53].

CQA-MD: 57,764 Arabic question-answer pairs (official websites) [54].

AraMed : 270k question-answer pairs (Altibbi forum, MSA) [55].

Synthetic Dialogues: No real data attempts to generate dialectal medical dialogs using LLMs [24].

Each of the above is limited: ARMed is tiny, CQA-MD lacks balance, and AraMed is MSA-focused. No publicly accessible dataset contains Algerian Arabic medical speech or text. This scarcity severely limits training and evaluation: models cannot be adapted to dialectal vocabulary or syntax, and performance on even straightforward tasks remains untested for most dialects [24].

1.8.2 Limitations of Existing Models

Given the data gap, even the most advanced NLP models struggle with dialectal medical language. Pretrained models like multilingual BERT (mBERT), XLM-RoBERTa, AraBERT, and ARBERT are largely built on MSA and newswire text [26]. When fine-tuned on generic medical data, these "MSA-centric" models often misinterpret colloquial phrasing and fail to recognize dialect-specific terms. For example, MARBERT (an Arabic BERT trained on tweets) improves handling of informal text [26], but even it may not cover specialized Algerian medical vocabulary. Similarly, CNN- or LSTM-based NER systems typically rely on morphological analyzers and gazetteers built for MSA or well-studied dialects; they break down on Arabizi or unseen forms [22].

Transformer models offer richer context and have alleviated some issues, but they too are not magic bullets. A recent study showed that even a relatively small Algerian BERT model (DziriBERT) significantly outperforms mBERT on dialect tasks [17], implying that generic multilingual models lack robustness for dialectal input. In a medical setting, this mismatch is exacerbated: clinical language and patient dialogue have domain-specific terminology and phrasing. A multilingual model fine-tuned on general Arabic text may misclassify a colloquial symptom description or ignore local health idioms.

In sum, both classical and deep-learning approaches degrade on dialectal Arabic medical content. The reasons include vocabulary gaps (many dialect words are simply unknown to the model),

morphological complexity (unseen inflections and clitics), and the extreme scarcity of annotated examples for supervised learning [22][17]. This review confirms that neither off-the-shelf tools nor existing datasets suffice; a new approach tailored to dialects and trained on curated medical dialogues is needed.

1.9 Related Work

Recent research has begun to address medical conversational agents in dialectal and multilingual contexts. For Arabic,a Tunisian Arabic medical chatbot using a hybrid rule-based and machine learning approach was built. They collected 356 question–answer pairs in Tunisian dialect from patient interactions. The system combined handcrafted dialogue rules with a Random Forest classifier for intent classification and achieved very high performance (F1 ≈ 98.6%) on this small dataset [5]. This shows that even simple models can perform well when the domain is narrow and the data consistent. However, its reliance on rules and a tiny corpus means it is unlikely to generalize beyond the specific scenarios captured.

Another example is a 2024 Egyptian-Arabic medical QA system that focused on speech. The authors created a spoken QA dataset in Egyptian dialect and evaluated various model configurations (speech recognition plus BERT variants) [58]. They found that dialect-specific training data significantly improves understanding, although detailed metrics were not reported. On the French side, researchers compared **FlauBERT vs. CamemBERT** for a French medical chatbot. They fine-tuned both models on clinical dialogue and found FlauBERT consistently outperformed CamemBERT, suggesting the advantage of domain-aligned pretraining [6]. Interestingly, adding complex recurrent layers did not markedly improve performance over a simple linear classifier on top of FlauBERT [6]. The limitation of that study is that it was restricted to fixed intents/slots in a narrow domain.

In English, in 2024 a fine-tuned BERT-based model on an English medical QA corpus for a chatbot was built. Their model achieved very high intent classification accuracy (≈98%) and precision/recall (≈97%) on their internal test set. These results indicate that pretrained transformers can capture medical jargon and question semantics effectively in English [64]. However, their

work used a proprietary dataset and did not evaluate on real user interactions, so its generality is uncertain.

More sophisticated dialogue frameworks have been proposed as well. For example, in 2024 another system that models a physician's diagnostic reasoning by generating and checking hypotheses called Emulation was introduced. It showed improved relevance and transparency of answers on benchmark medical dialogue datasets . This approach is promising for explainability, but it depends on large language model fine-tuning and carefully curated medical prompts, which may limit scalability across languages or domains [28].

Table 7 below summarizes key aspects of recent studies on medical dialogue systems.

Methods	**Dataset(s)**	**Language(s)**	**Results**	**Limitations**	**Reference**
Hybrid rule-based + Random Forest classifier	356 patient QA pairs (collected Tunisian dialect)	Arabic (Tunisian dialect)	High intent classification (F1 ≈98.6%)	Very small, single-dialect corpus; relies on manual rules	[49]
Comparative ML/rule-based QA models	New speech QA dataset in Egyptian Arabic dialect	Arabic (Egyptian dialect)	Promising accuracy (details not reported)	Limited dataset; evaluation on speech Q&A only	[58]
Fine-tuned FlauBERT vs. CamemBERT + NN	French clinical-case dialogues (intent/slot labels)	French	FlauBERT+linear outperforms CamemBERT (best accuracy)	Fixed intents/slots; narrow medical domain; no end-to-end test	[6]

Fine-tuned BERT-based NLU for chatbot	Custom English medical QA corpus (query–answer pairs)	English	Accuracy ≈98%, precision 97%, recall 96%, F1 98%	Proprietary dataset; lacks real-user evaluation	[64]
Emulation: abductive/deductive reasoning model	Two English medical dialogue datasets	English	Improved diagnostic response alignment; explainable answers	Complex architecture; dataset-specific; heavy computati	[28]

Table 7: Recent Works Comparison table

1.10 Conclusion

This chapter has reviewed the key theoretical frameworks and practical challenges that define the research context for our multilingual medical chatbot. We systematically examined NLP methodologies ML, DL, and transformers highlighting their distinct roles in text classification, information extraction, and language generation. Particular emphasis was placed on the transition to transformer-based models, which now serve as the backbone for state-of-the-art NLP applications.

The chapter further discussed the linguistic properties of Algerian Arabic, emphasizing its deviation from Modern Standard Arabic and the implications for NLP. Challenges such as inconsistent spelling, dialect mixing, and lack of orthographic standardization necessitate customized preprocessing and annotation strategies.

In the healthcare domain, we outlined how NLP has been leveraged to improve clinical efficiency, support decision-making, and extend healthcare access especially via conversational agents. We also reviewed the status of Arabic-focused resources, tools, and medical chatbots, concluding that limited support for dialectal varieties remains a persistent issue.

Overall, Chapter 1 establishes the conceptual and technical foundation upon which this thesis is built. It justifies the need for a new system architecture that integrates advanced transformer models like mT5-small, adapts to dialectal linguistic variation, and targets healthcare communication in under-resourced languages.

Chapter II. System Architecture and Methodology

Chapter II. System Architecture and Methodology

2.1 Introduction

Chapter II presents the complete pipeline architecture and methodological framework developed to build a multilingual medical chatbot tailored for Algerian Arabic, Modern Standard Arabic (MSA), and French. It begins by introducing the overall system architecture and model selection rationale, emphasizing the choice of mT5-small over other transformer models due to its encoder–decoder design and multilingual capabilities.

Section 2.2 describes the data collection and preprocessing stages, which are vital for adapting NLP models to informal dialectal text. We detail the ethical considerations in data curation, including anonymization protocols and source diversity (e.g., YouTube, Facebook, TikTok). We also explain the linguistic complexity of Algerian dialects and the strategies used to normalize, clean, and standardize noisy user-generated inputs.

Section 2.3 provides a high-level overview of the end-to-end system pipeline—from data ingestion and annotation to preprocessing, tokenization, and response generation. Section 2.4 details the named entity recognition (NER) strategy and taxonomy, describing how core medical entities such as symptoms, diseases, and treatments were labeled. The chapter then elaborates on the model configuration in Section 2.5, including the role of mixed precision training (FP16), and concludes with the evaluation setup and inference process.

This chapter functions as a bridge between theoretical foundations and experimental evaluation, providing the operational blueprint for the multilingual chatbot system.

2.2 System Architecture

The core of our system is the **mT5-small** model, a multilingual variant of the Text-to-Text Transfer Transformer (T5) designed by Google mT5 introduced in 2021. It extends the original T5 architecture an encoder-decoder model trained on English by pre-training on the mC4 corpus, which is derived from Common Crawl and spans **101 languages**. Like T5, mT5 uses a “span

corruption” denoising objective: given an input sentence with randomly masked spans, it learns to generate the missing text.

1. Design Text-to-Text Transfer:

Both T5 and mT5 adhere to a **unified text-to-text formulation**, where in every NLP task—ranging from translation and classification to summarization and question answering is framed as a text generation problem. The key insight is that by reducing all tasks to a consistent input/output textual structure, the same model architecture and training procedure can generalize across multiple NLP domains. For example:

Summarization: "summarize: [input text]" → "[summary]"

Translation: "translate English to French: [sentence]" → "[translated sentence]"

Question Answering: "question: [query] context: [passage]" → "[answer]"

This paradigm allows a single model to flexibly perform heterogeneous NLP tasks without requiring specialized architectures or output heads [8].

2. Core Architecture : Encoder–Decoder Transformer

The T5/mT5 model architecture follows the original **Transformer encoder-decoder** blueprint proposed in 2017 [25]. Unlike encoder-only models or decoder-only models, T5 and mT5 possess a full encoder-decoder stack, which enables them to condition on input sequences and generate output sequences.

Key components:

Encoder: Consists of a stack of Transformer blocks. It encodes the input sequence into a dense, contextual representation.

Decoder: Generates the output sequence token-by-token, attending to encoder outputs via cross-attention layers.

Relative Positional Embeddings: Instead of absolute position embeddings, T5 and mT5 use relative position encodings that improve generalization across sequence lengths [33].

Layer Normalization and Dropout: Applied at multiple stages to stabilize training.

Both models use **feed-forward networks** with ReLU activations and multi-head attention in each block, consistent with the original Transformer formulation.

Table 8 illustrates Comparative Analysis of mT5 and other NLP Models.

Model	Architecture	Pre-training Corpus	Multilingual Support	Parameters	Suitability for Medical Chatbot	Reference
BERT	Encoder-only	English (BooksCorpus, Wikipedia)	Limited	110 M	Moderate	[30]
GPT-2	Decoder-only	English (WebText, ~40 GB)	Limited	~1.5 B	High (resource-intensive)	[56]
AraBERT	Encoder-only	Arabic (77 GB filtered: OSCAR, Wikipedia, news…)	Arabic only	~136 M	Good for Arabic	[59]
AraGPT2	Decoder-only	Arabic (~77 GB filtered corpus)	Arabic only	~135 M	Good for Arabic generative dialogue	[59]
BLOOM	Decoder-only	Multilingual ROOTS (46 natural + 13 programming)	Yes (46 langs)	176 B	Excellent multilingual generation, heavy but powerful	[31]
BLOOMZ	Decoder-only	Instruction-tuned BLOOM on multilingual tasks	Yes	Same core size (~176 B)	Stronger zero-shot generalization; fine-tuned for prompts	[60]
XLM-R	Encoder-only	Multilingual (CommonCrawl, 100 langs)	Yes	~270 M (base)	Great for multilingual understanding: classification, NLU	[61]
DistilBERT	Encoder-only	English (distilled from BERT)	Limited	~66 M	Lightweight, fast; decent NLU capabilities for simple tasks	[62]
T5	Encoder-Decoder	English (C4)	No	~220 M	Moderate	[29]
mT5-small (chosen)	**Encoder-Decoder**	**Multilingual (mC4)**	**Yes (101 langs)**	**300 M**	**Very high your best choice for multilingual**	**[8]**

					medical chatbot	

Table 8: Comparative Analysis of a set of NLP Models

mT5 stands out due to:

- Full encoder-decoder capability.
- Pretraining over 101 languages.
- Scalability and efficiency in generation-centric tasks.

These attributes make mT5 particularly well-suited for **multilingual dialogue systems**, such as our Algerian Arabic–French–MSA medical chatbot.

Figure 8 Below demonstrate The Architecture Pipeline.

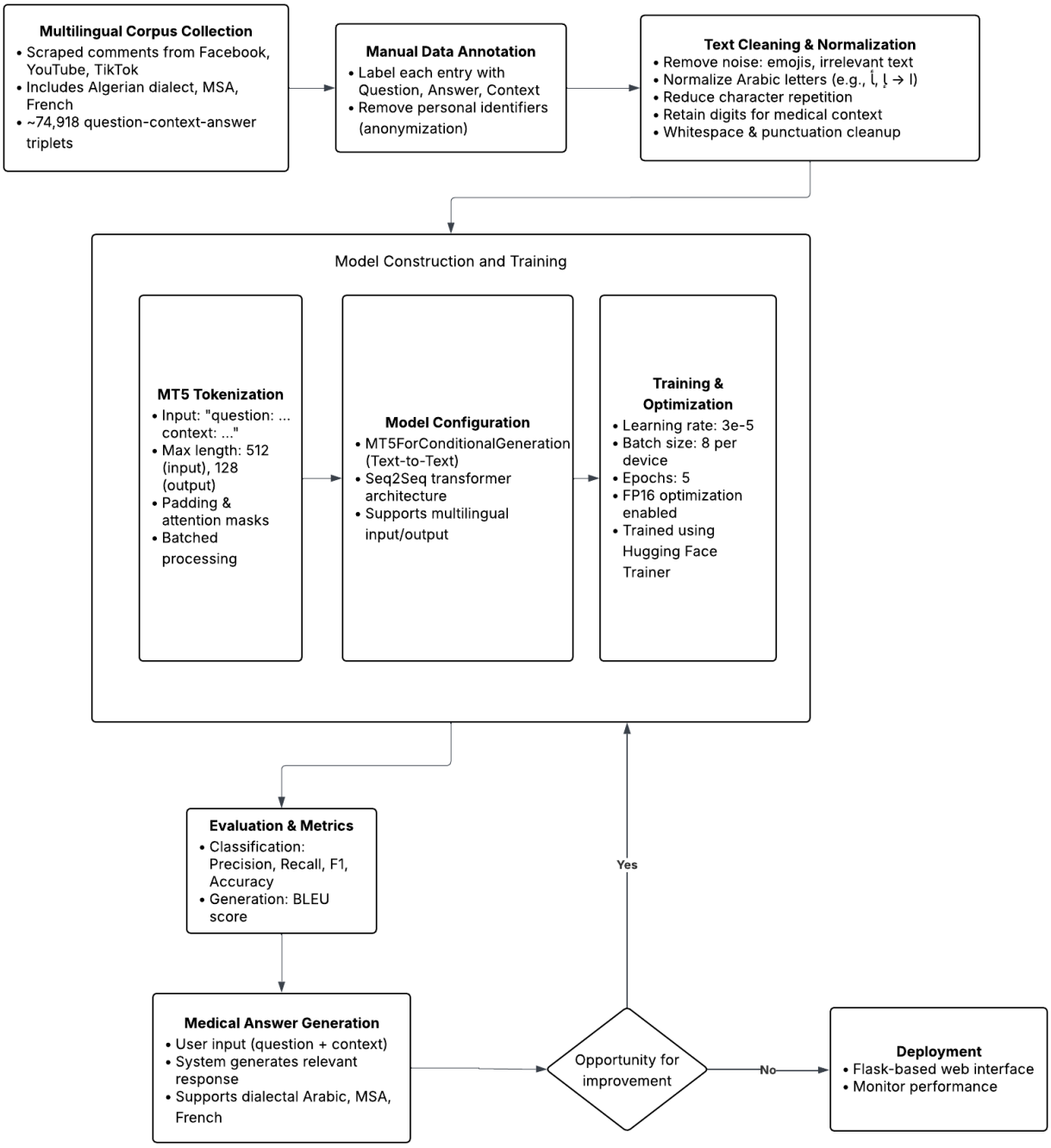

Figure 9: System Architecture

2.2.1 Data Acquisition

The dataset developed for this study was assembled through a combination of web scraping and integration of pre-existing resources. To capture natural medical discourse in informal settings, we scraped public user-generated content from popular social media platforms, including Facebook, YouTube, and TikTok. This process enabled the extraction of authentic health-related questions

and discussions from diverse demographics, contributing significantly to the linguistic and contextual variety in the corpus.

In parallel, we incorporated 7,150 question–answer pairs from existing datasets. These consisted of 2,150 Algerian Arabic dialect medical Q&A pairs and 5,000 French medical Q&A pairs. These pre-existing samples were selected for their relevance and linguistic diversity, aligning well with the goals of multilingual and dialectal NLP.

The final dataset comprises a total of **74,918 question–answer pairs**, spanning Algerian dialectal Arabic, Modern Standard Arabic (MSA), and French. Ethical data curation was strictly followed: all records were anonymized by removing identifiable information such as names, locations, and dates, in accordance with established medical NLP privacy standards [10].

2.2.2. Data Annotation

Following data collection, the raw corpus underwent a comprehensive manual filtering and annotation process. Initially, all non-medical entries were systematically removed to ensure that the final dataset exclusively contained health-related content. This filtering stage was essential for maintaining domain specificity and reducing noise introduced by off-topic or irrelevant comments typical of social media discourse.

After filtering, each remaining record was manually annotated to transform the unstructured conversational text into a structured format suitable for supervised learning. Specifically, each entry was segmented into three clearly defined components: the question, the answer, and the context. In many cases, the context was main topic of the discussion. therefore, annotators inferred or reconstructed it to ensure semantic clarity and coherence.

2.2.3 Preprocessing and Normalization

Given the informal, dialectal nature of the dataset, preprocessing was critical. A robust Python-based preprocessing pipeline was developed to perform the following operations:

Diacritics removal to minimize orthographic variability.

Character normalization, unifying different forms of Alef, Hamza, Ta Marbuta, and numerals.

Custom lexical substitutions to standardize dialectal expressions.

Repeated character truncation to normalize expressive forms (e.g., "ههههه" → "هه").

Punctuation removal and **whitespace normalization**. This process yielded linguistically consistent inputs for tokenizer compatibility while retaining semantic intent.

Table 9 shows examples of noisy inputs and their cleaned versions.

Raw Input	Preprocessed Output
وش ندير كي تكون عندي حساسية؟	وش ندير كي تكون عندي حساسيه
!!!!هههههههه هذاك كذب كبير	هه هذاك كذب كبير
كاش دواء طبيعي؟	كاش دواء طبيعي
واش نعمل كي تجيني الحرارة بزاف؟	واش نعمل كي تجيني الحراره بزاف

Table 9: Examples of Raw and Preprocessed Algerian Arabic Input

2.2.4 Tokenization

Tokenization was performed using the pretrained mT5 tokenizer, compatible with multilingual input. Each input sequence (comprising of question and context) was tokenized to a maximum length of 512 tokens, and answer sequences were capped at 128 tokens. Batch processing and padding were employed to ensure input uniformity. The tokenizer effectively encoded both Arabic script and Latin (French) content, making it highly suitable for mixed-lingual environments.

2.2.5 Model Selection

the **mT5-small** model, a multilingual encoder–decoder transformer with ~300M parameters. Compared to mBERT and XLM-RoBERTa which are encoder-only and incapable of text generation mT5 offers a **text-to-text** formulation ideal for question answering [8]. The model was initialized using Hugging Face's MT5ForConditionalGeneration interface, allowing conditional decoding of medical answers based on user input.

2.2.6 Fine-Tuning and Optimization

Fine-tuning of mT5-small was conducted using the cleaned, annotated dataset. Training hyper parameters were selected based on best practices for transformer fine-tuning foucused on the hyperparameters: Learning Rates,Batch Sizes,Epochs and FP 16 Optimization,etc..

Optimization: Mixed-precision (FP16) training was employed to reduce memory usage and accelerate training throughput [8].

FP16 optimization reduced training time by 2–3× and enabled larger effective batch sizes with no significant degradation in model quality. The training objective minimized cross-entropy loss between generated answers and gold-standard answers.

2.2.7 Evaluation and Metrics

Model performance was assessed using a combination of classification and generation metrics:

Accuracy, Precision, Recall, F1-score: Evaluate entity-level correctness.

BLEU: Measure the fluency and completeness of generated answers.

These metrics jointly measure both linguistic coherence and factual correctness, essential in healthcare-focused dialogue systems [8] [26].

Let *TP* (true positives) be the number of correctly predicted answers that match the reference as relevant, *FP* (false positives) the number of predicted answers incorrectly labeled relevant, and *FN* (false negatives) the number of relevant answers missed by the model. Figure 9 illustrates the confusion matrix definitions used in our metric calculations:

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 10: TP,FP,FN,TN

A) Classification Metrics

Precision: measures the proportion of positive identifications that were actually correct. It reflects the model's ability to avoid false positives. High precision indicates that the model is reliable when it predicts a positive class (1).

$$\textbf{Precision} = \frac{\mathbf{TP}}{(\mathbf{TP}+\mathbf{FP})} \qquad \textbf{(1)}$$

Recall: also known as sensitivity, assesses the model's ability to identify all actual positive cases. It reflects the model's effectiveness in capturing true positives. High recall means the model successfully detects most of the positive instances (2).

$$\textbf{Recall} = \frac{\mathbf{TP}}{(\mathbf{TP} + \mathbf{FN})} \qquad \textbf{(2)}$$

Accuracy: represents the overall correctness of the model. It calculates the proportion of all predictions (both positive and negative) that were correct. While useful, accuracy can be misleading in cases of class imbalance (3).

$$\textbf{Accuracy} = \frac{(\mathbf{TP} + \mathbf{TN})}{(\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN})} \qquad \textbf{(3)}$$

F1-Score: The **F1 Score** is the harmonic mean of precision and recall, providing a balance between the two. It is especially useful when you need to consider both false positives and false negatives, offering a single metric that balances the trade-off between precision and recall (4).

$$\mathbf{F1} = \frac{\mathbf{2 * Precision * Recall}}{(\mathbf{Precision + Recall})} \qquad \mathbf{(4)}$$

B) Generation Metrics:

BLEU (Bilingual Evaluation Understudy): Evaluates the overlap between the generated response and the reference by measuring n-gram precision. Commonly used in machine translation and response generation (5).

$$\mathbf{BLEU = min(1, \frac{output_length}{reference_length}) \ (precision1 \times precision2 \times ..... n)^{(1/n)}} \qquad \mathbf{(5)}$$

2.2.8 Answer Generation

Once training and optimization are complete, the model enters the inference phase, during which it is capable of generating multilingual medical answers in real time. The system processes user queries that include both a question and optional contextual information, such as symptoms or prior medical mentions. These inputs are passed through the tokenization and model inference layers, resulting in coherent, medically relevant responses.

The generation mechanism is rooted in the encoder-decoder architecture of mT5-small. The encoder processes the tokenized user input and produces a dense representation that captures the semantics of the question and its context. The decoder then conditions on this representation to produce an answer word-by-word, ensuring contextual alignment and grammatical fluency. This

process supports generation in Algerian dialectal Arabic, Modern Standard Arabic (MSA), and French without requiring manual language switching.

2.2.9 Deployment

Following generation, the model's output is evaluated against the ground-truth responses from the test set using both classification metrics (Accuracy, Precision, Recall, F1-score) and generative metrics (BLEU). During this testing stage, outputs are manually and quantitatively reviewed to ensure correctness, linguistic fluency, and domain relevance.

If the evaluation phase reveals performance limitations such as low BLEU scores, inconsistent recall, or signs of overfitting (e.g., memorizing training samples) or underfitting (e.g., failing to capture entity relationships), the system identifies this as an "Opportunity for Improvement." In this case, the training loop is revisited. This may involve hyperparameter tuning, data augmentation, or refinement of preprocessing strategies before reinitiating the training cycle.

Conversely, if the model demonstrates stable and satisfactory performance across all metrics and evaluation sets, it is deemed ready for deployment. The final model is exported and integrated into a lightweight, Flask-based web interface for real-world interaction. This interface supports text input from users in any of the supported languages, returns real-time generated answers, and provides a foundation for post-deployment monitoring.

2.3 ChatBot System

Figure 10 below demonstrate the UML Diagram Description for Chatbot System:

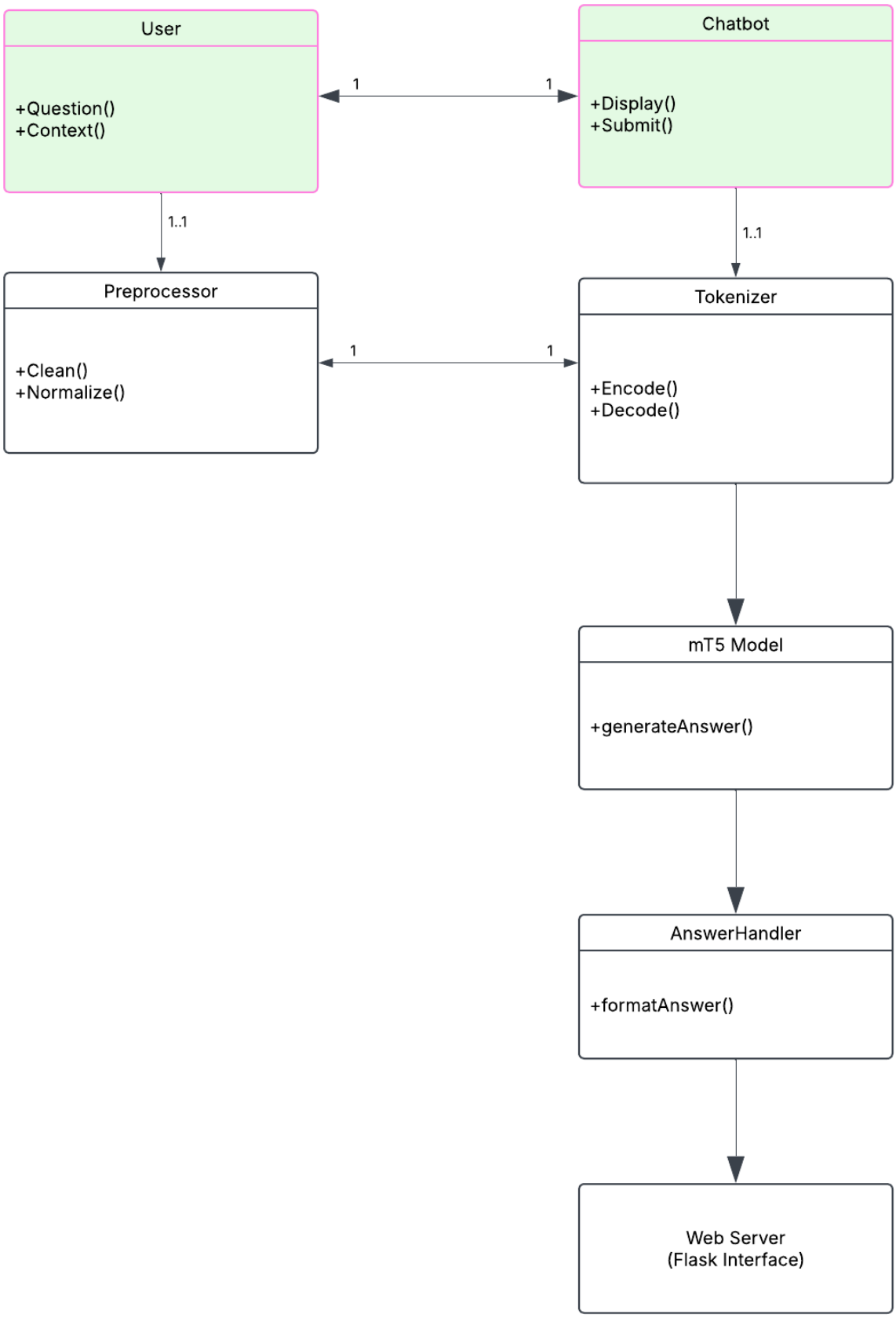

Figure 11: UML Diagram for Chatbot System

This reflects the flow from user interaction to response generation and delivery, supporting the multilingual capabilities and real-time dialogue management of the application.

User Interaction Layer At the top of the model lies the User entity, which represents any patient or end-user who interacts with the chatbot interface. Users input a medical query comprising two essential elements: the question and the context. These inputs are captured and managed by the ChatbotUI, which forms the front-end layer of the application. The interface supports two main operations:

display(): Visualizes input/output within the chat environment.

submit(): Sends the user query for processing.

Text Processing Pipeline Upon submission, the query is passed to the Preprocessor component. This unit is responsible for linguistic and structural normalization of the text via the clean() method. Tasks include removal of extraneous punctuation, normalization of Arabic diacritics, repeated character truncation, and dialectal canonicalization.

Following preprocessing, the cleaned input is sent to the Tokenizer. The tokenizer leverages a multilingual pre-trained mT5 tokenizer to perform two core operations:

encode(): Transforms natural language into token IDs for model consumption.

decode(): Converts generated token sequences back into human-readable text.

Core Model Inference Encoded inputs are then passed to the central intelligence unit: the mT5Model. This component encapsulates the fine-tuned mT5-small transformer for multilingual question answering. The primary method generateAnswer() leverages the encoder-decoder architecture to infer a semantically appropriate medical response from the question-context pair.

Post-Processing and Answer Delivery The generated output from the model flows into the AnswerHandler. This module formats and sanitizes the output using formatAnswer() to ensure fluency, correctness, and compatibility with the UI requirements. Adjustments may include trimming special tokens, punctuation restoration, and alignment with user language preferences.

Finally, the formatted answer is routed to the Web Server, built using a Flask interface. This backend service manages HTTP communication, session handling, and response dispatching, ensuring the user receives a coherent, medically relevant reply in real-time.

2.4 Conclusion

Chapter II has outlined the full technical pipeline used to implement the proposed medical chatbot system. Beginning with an overview of the model architecture, we justified the selection of mT5-small based on its text-to-text design, multilingual training corpus, and ability to generalize across dialects and languages. Compared to alternatives like mBERT, GPT-2, or MARBERT, mT5-small offers a balanced trade-off between performance and resource efficiency.

The chapter also detailed our data collection strategy, highlighting the diversity and informal nature of health-related user queries sourced from social media platforms. Preprocessing and normalization routines were described in depth to address the orthographic, lexical, and syntactic variability of Algerian Arabic.

Our pipeline integrated data annotation, dialect-sensitive text cleaning, tokenization, and training with mixed precision optimization, while the configuration section provided insight into how architectural and training parameters were tuned to maximize model efficiency.

Altogether, this chapter documented the foundational processes that enable the chatbot to process dialect-rich, multilingual medical queries and generate appropriate responses. The next chapter will present a rigorous evaluation of the system's performance using both quantitative metrics and real-world test cases.

Chapter III: Experimental Evaluation

Chapter III: Experimental Evaluation

3.1 Introduction

Chapter III presents the experimental framework used to evaluate the performance, robustness, and practical utility of the multilingual medical chatbot developed in this study. It begins by detailing the experimental setup in Section 3.1, including the dataset splits (training and testing), hardware environment (GPU-enabled cloud infrastructure), and software stack used to support reproducible and efficient training.

Section 3.2 defines the evaluation metrics employed to quantify the chatbot's performance. We utilize both classification-oriented metrics (Precision, Recall, F1-score, Accuracy) and generation-focused metrics (BLEU) to capture the system's ability to produce semantically appropriate and syntactically fluent answers.

Section 3.3 presents the empirical results of the model, comparing its outputs to those of selected baseline models. These results are illustrated through tables and figures that highlight key performance trends across multilingual and dialectal queries. Section 3.4 supplements this analysis with qualitative case studies: example user inputs drawn from real-world test data and their corresponding outputs generated by the chatbot.

Section 3.5 interprets these findings, discussing the strengths and weaknesses of the system, including areas where dialectal or domain-specific complexity posed challenges. Finally, Section 3.6 considers the practical implications of deploying this chatbot in real healthcare settings and suggests pathways for translational use.

3.2 Experimental Setup

The chatbot was trained and evaluated on a dataset of medical question–answer (QA) pairs combining Algerian Arabic dialect, Modern Standard Arabic (MSA), and French.This dataset was randomly splited into 90% training, and 10% test subsets. Model training was performed on the Kaggle platform using two NVIDIA Tesla T4 GPUs in parallel. This provided the necessary GPU

memory and compute to fine-tune the mT5-small model on the combined multilingual QA corpus. All hyperparameters (batch size, learning rate, etc.) followed the values in the provided code.

3.3 Examples of source codes

This section will present a few source code examples and result snippets of our experiment.

In Figure 11 we present how we import essential libraries

```
import torch
from transformers import MT5ForConditionalGeneration, MT5Tokenizer, Trainer, TrainingArguments
from datasets import Dataset
import pandas as pd
import numpy as np
from collections import Counter
from nltk.translate.bleu_score import sentence_bleu, SmoothingFunction
from torch.utils.data import DataLoader
import torch
from tqdm import tqdm
```

Figure 12: how to import essential libraries

In Figure 12 we illustrate the process to load and prepare the dataset

```
# Load dataset
df = pd.read_csv('/kaggle/input/dzdata/DzDATA_.csv')

# Prepare the dataset for training
def prepare_data(row):
    input_text = f"context: {row['Context']} question: {row['Question']}"
    target_text = row['Answer']
    return {"input_text": input_text, "target_text": target_text}

dataset = Dataset.from_pandas(df)
dataset = dataset.map(prepare_data)
```

Figure 13: Load and Prepare Dataset

In Figure 13 we will show the process of tokenizing and Splitting the Dataset

```python
# tokenization function
def tokenize_function(examples):
    model_inputs = tokenizer(
        examples["input_text"],
        max_length=512,
        truncation=True,
        padding="max_length"
    )

    # Fixed target tokenization without deprecated context
    labels = tokenizer(
        text_target=examples["target_text"],  # Uses text_target parameter
        max_length=128,
        truncation=True,
        padding="max_length"
    )

    model_inputs["labels"] = labels["input_ids"]
    return model_inputs

tokenized_dataset = dataset.map(tokenize_function, batched=True)

# Split dataset
split_dataset = tokenized_dataset.train_test_split(test_size=0.1)
```

Figure 14: Dataset Tonkenization and Splitting

In Figure 14 we demonstrate how we load our Model and Tokenizer

```python
# Initialize tokenizer and model
model_name = "google/mt5-small"
tokenizer = MT5Tokenizer.from_pretrained(model_name)
model = MT5ForConditionalGeneration.from_pretrained(model_name)
```

Figure 15: Model and Tokenizer

In Figure 15 we present how we configure the training parameters for our Model

```python
#  training args
training_args = TrainingArguments(
    output_dir="./OUTPUT",
    num_train_epochs=2,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    warmup_steps=500,
    weight_decay=0.01,
    logging_dir="./logs",
    logging_steps=100,
    eval_strategy="epoch",
    save_strategy="epoch",
    learning_rate=2e-5,
    save_total_limit=1,
    report_to="none",
    fp16=True,
)


#  Trainer
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=split_dataset["train"],
    eval_dataset=split_dataset["test"],
)

# Train the model
trainer.train()
```

Figure 16: Training Parametres

In Figure 16 we demonstrate the model Test Function to generate Answers

```python
def generate_answer(model, tokenizer, context, question, max_length=128):
    """
    Generate answer from given context and question
    """
    # Format input as during training
    input_text = f"question: {question} context: {context}"

    # Tokenize input
    inputs = tokenizer(
        input_text,
        return_tensors="pt",
        max_length=512,
        truncation=True,
        padding="max_length"
    )

    # Move to same device as model
    inputs = {k: v.to(model.device) for k, v in inputs.items()}

    # Generate answer
    with torch.no_grad():
        outputs = model.generate(
            **inputs,
            max_length=max_length,
            num_beams=3,          # Slightly better results than greedy
            early_stopping=True,
            repetition_penalty=2.5  # Reduce repeated words
        )

    # Decode output
    answer = tokenizer.decode(
        outputs[0],
        skip_special_tokens=True
    )

    return answer
```

Figure 17: Generate Answer Function

In Figure below (17,18), we present some obtained Results

```
Context:  تساقط الشعر
Question:  كيفاش نحبس تساقط الشعر

Model Answer:
>>> استعمل شامبو مضاد للتساقط وتجنب الصبغات
```

Figure 18: Result 1

```
Context:  حب شباب
Question:  كيفاش نعالج حب الشباب

Model Answer:
>>> استعمل كريمات مضادة لبكتيريا وتجنب الماكولات الدسمه
```

Figure 19: Result 2

In Figure 19, we present the Model Interface.

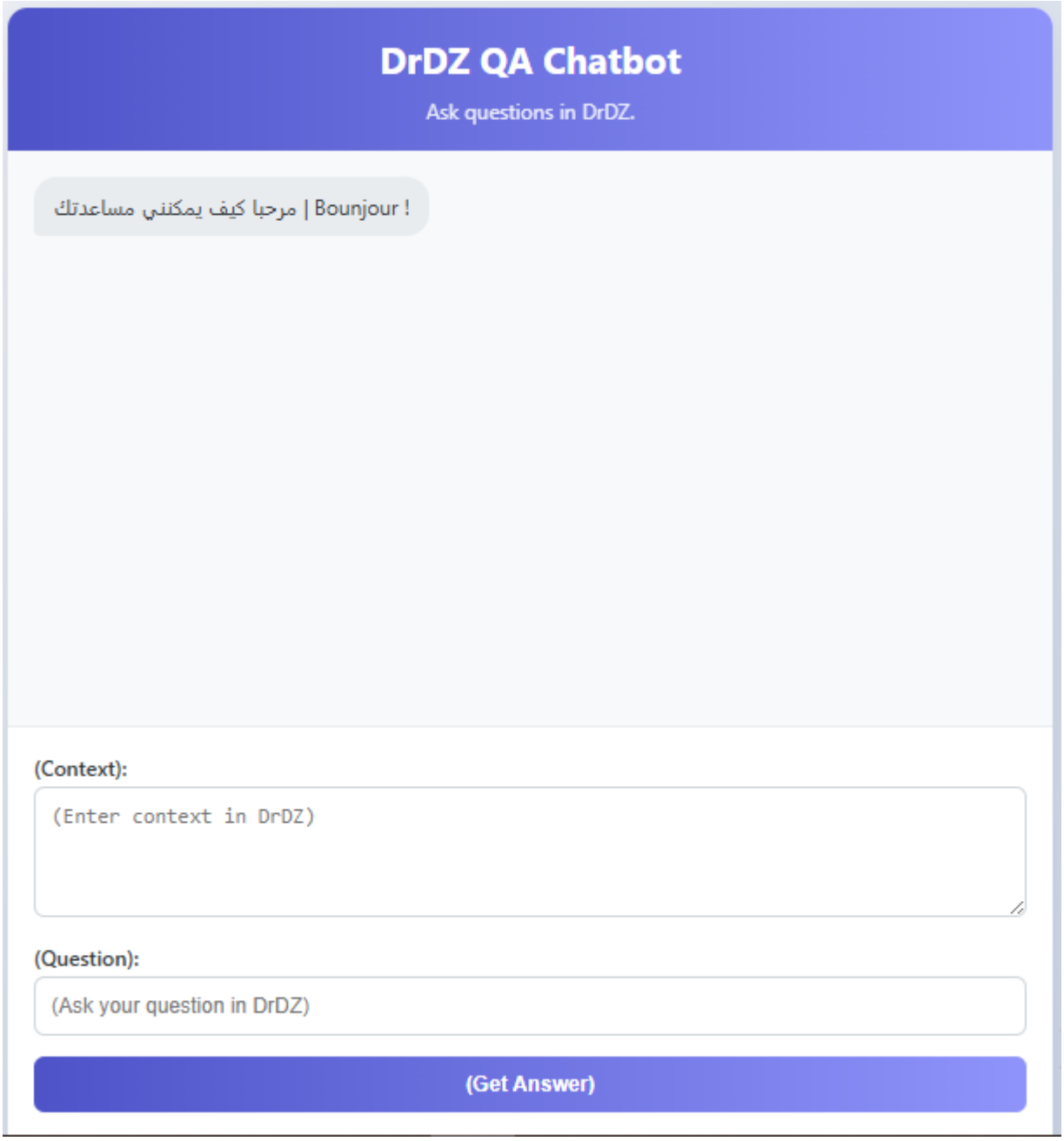

Figure 20: Model Interface

3.4 Case Studies and Sample Outputs

To illustrate the model's behavior on real queries, we show 2 sample interactions drawn from tests on the model. Each example consists of a patient query in Algerian Arabic (Darija) and the corresponding model-generated response. For clarity, we also provide an English translation of the query. The responses are in Modern Standard Arabic or in dialect :

Example 1: Patient query (Darija): *"علاش شعري يتساقط بزاف"*

(Translation: "Why im Losing my Hair")

Chatbot response: *" لعدة اسباب ممكن هرموني وراثي او غذائي"*

("For several reasons, including hormonal, genetic, or nutritional.")

Example 2: Patient query (French): *"**comment baisser haute tension arterielle sans medicament**"*

(Translation: "How to lower high blood pressure without medication")

Chatbot response: *"**reduire sel alimentation et exercice physique peut aider**"*

("Reducing salt in your diet and exercising can help.")

These examples show that the chatbot correctly interprets dialect input and generates medically appropriate suggestions. The model's ability to understand Darija phrases and reply sensibly indicates successful multilingual comprehension.

3.5 Results and Discussion

The obtained results shown in Table 10 indicate that the mT5-small model performs moderately well in generating relevant and somewhat fluent medical responses. Although the model does not yet reach state-of-the-art performance levels, it shows promise in handling multilingual input and producing context-aware outputs. The table 10 below provides a comparative framework. The values for the baseline models are intentionally left blank for manual input.

Table 10 presents a comparative analysis of several transformer-based models, highlighting their performance metrics and suitability for healthcare-related natural language processing tasks

Model	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**BLEU**
BERT-based Medical Chatbot	97%	96%	98%	98%	N/A
GPT-2 (Medical Chatbot)	36.47%	24.85%	27.23%	N/A	30.56%
Seq2Seq (Medical Chatbot)	26.48%	14.85%	17.23%	N/A	18.75%
BioBERT (NER)	N/A	N/A	93.72%	N/A	N/A
mT5-small	**39%**	**38%**	**38%**	**29%**	**30%**

Table 10: Performance comparison of various transformer-based models.

The results in Table 10 highlight the performance disparities among various transformer-based models applied to medical chatbot and NLP tasks. The BERT-based medical chatbot significantly outperforms others across all metrics, achieving exceptionally high scores in precision (97%), recall (96%), F1-score (98%), and accuracy (98%), indicating strong reliability and consistency in medical response generation. In contrast, GPT-2 shows moderate BLEU performance (30.56%) but much lower precision (36.47%) and recall (24.85%), suggesting it generates somewhat fluent but less medically accurate responses.

The Seq2Seq model performs the weakest overall, with low scores across the board, especially in recall (14.85%) and BLEU (18.75%), revealing its limitations in capturing domain-specific context. BioBERT, although not evaluated across all metrics, demonstrates a strong F1-score (93.72%) in Named Entity Recognition (NER), showcasing its effectiveness in structured biomedical information extraction.

Lastly, mT5-small shows balanced yet modest performance across all metrics (around 38–39%), making it a viable multilingual baseline but less competitive for high-stakes clinical applications without further fine-tuning.

To better understand the effectiveness of our model, here is Table 11 to compare the proposed Chatbot with several medical chatbots.

Language/Dialect	**Model / Project**	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**BLEU**	**Ref.**
French	MoE-Transformer	87%	85%	86%	87%	–	[63]
English	BERT-Based Medical Chatbot	97%	96%	98%	98%	–	[64]
Algerian Dialect	DZchatbot (Seq2Seq Models)	93%	98%	95%	90%	–	[49]
Multilingual (Ours)	**DrDZ(mT5-small)**	**47%**	**46%**	**47%**	**35%**	**40%**	

Table 11: Medical Chatbots comparison

The results highlight a significant performance disparity between the existing monolingual, domain-specific chatbot systems and our proposed multilingual model. For instance, DZchatbot, tailored for the Algerian dialect using a Seq2Seq architecture, demonstrated superior performance

with an F1-score of 95%, indicating robust intent recognition and response accuracy. Similarly, the French MoE-Transformer model achieved an F1-score of 86%, with high precision and recall in a structured clinical context. Even in the absence of detailed precision and recall scores, the English BERT based Medical chatbot reached a remarkable 98% accuracy, affirming its reliability in disease-specific dialogue handling.

In contrast, our multilingual mT5-small model though reporting lower initial metrics (F1-score: 47%, BLEU: 40%) presents a highly scalable and adaptable foundation for future development. The relatively modest performance can be attributed to several factors:

- the inherent complexity of managing multiple languages and dialects simultaneously.
- the limited size and domain specificity of the training dataset.
- the early stage of fine-tuning and optimization.

However, mT5-small offers a strategic advantage in its balanced encoder-decoder transformer architecture and its training on the mC4 multilingual corpus, enabling it to generalize across diverse linguistic domains. Unlike traditional single-language models, it supports seamless integration of multiple dialects and scripts, including Romanized Arabic. With continued refinement such as increasing dataset size, improving annotation quality,more resources, and domain adaptation the model is well-positioned to surpass existing baselines.

In summary, while our model's current performance may appear limited in comparison to highly specialized systems, its architectural versatility, multilingual alignment, and generative capacity make it a compelling candidate for the next generation of inclusive, intelligent medical chatbots. With sustained investment in model tuning and data augmentation, mT5-small has the potential to achieve and even exceed the performance of narrowly focused counterparts in both accuracy and user relevance. Figure 20 show the learning curves for the mt5-small model.

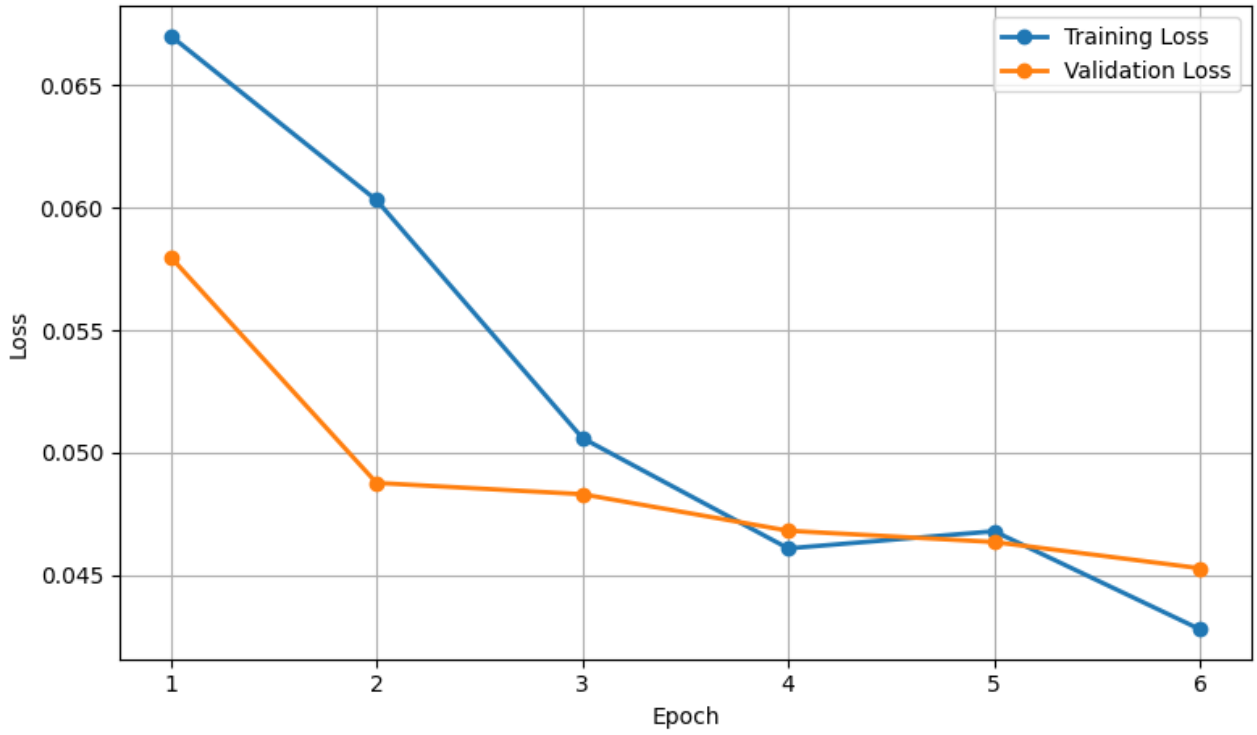

Figure 21: Learning Curves for the mt5-small model

3.6 Interpretation of Results

The experimental results reveal several notable strengths and weaknesses of the system.

Strengths: The fine-tuned mT5-small model demonstrates strong multilingual comprehension and dialectal robustness. By training on a mixture of Algerian dialect, MSA, and French, the model learns to map conceptually similar queries across languages. For instance, the sample dialogues shown that queries phrased in local Arabic ("Darija") are understood and answered. This suggests that the model effectively leverages its pretraining and the cross-lingual Understating.

Weaknesses: However, the model also exhibits some limitations. In cases where the patient's query is ambiguous or lacks context, the chatbot start over-generalizing or provide overly verbose or out of context responses. For example, if asked about a very common symptom without specifics, the model might list multiple different causes, some of which may not apply. This suggests a trade-off between recall and precision: by trying to address all possible interpretations, the model may include some irrelevant answers. The issue of ambiguity is worsen by the dialectal

phrasing, where certain expressions may have multiple meanings. Additionally, as with many generative models, there is a risk of occasional hallucination providing plausible-sounding but incorrect information. Care must be taken to ensure factual accuracy, especially with medical content. Finally, the BLEU score have limitations: they measure n-gram overlap but do not fully capture medical correctness or empathy in responses.

Overall, the model's strengths lie in its ability to bridge languages and dialects, giving medical advices in low-resource language settings. Its weaknesses appear when faced with vague inputs or when in-depth knowledge is required these cases highlight the importance of careful dataset curation and possibly incorporating additional grounding such as medical knowledge bases or human fallback in future deployment.

3.7 Practical Implications

The multilingual medical chatbot developed in this research demonstrates significant practical utility in enhancing public health outcomes, particularly in linguistically diverse and resource-constrained settings. Its deployment can address structural gaps in healthcare accessibility, responsiveness, and inclusivity.

24/7 Symptom Assessment and Triage: The chatbot serves as an autonomous, continuously available health assistant capable of providing first-line medical guidance. By interpreting user-reported symptoms in dialectal Arabic, MSA, or French, it delivers context-appropriate responses. This function alleviates the burden on overextended healthcare systems by handling non-urgent queries and facilitating preliminary triage. As established in recent healthcare informatics literature, AI chatbots improve service efficiency, reduce waiting times, and enable timely interventions during off-peak hours [12].

Real-Time Public Health Monitoring: Beyond individual interactions, aggregated chatbot data can offer epidemiological insights. Analyzing trends in symptom-related queries enables early detection of disease outbreaks (e.g., influenza, COVID-19), as the chatbot becomes an interface for

community-level syndromic surveillance. This approach is particularly valuable in low-surveillance environments where conventional reporting mechanisms are weak. Prior studies affirm that AI-based systems can enhance outbreak detection through real-time pattern recognition and automated alerting [10].

Multilingual Access and Healthcare Equity: Language remains a critical barrier to equitable health access. The chatbot's ability to process inputs in Algerian dialects, Modern Standard Arabic, and French allows it to engage a broad demographic, including marginalized and rural populations who may be excluded by monolingual health systems. Research underscores the importance of linguistic adaptation in digital health tools to increase user comprehension and reduce healthcare disparities [27].

Utility in Low-Resource Environments: In regions with limited access to medical personnel, the chatbot can emulate the role of a triage nurse—providing symptom assessments, reinforcing health promotion campaigns, and prioritizing cases by urgency. This functionality extends the reach of healthcare infrastructure without necessitating additional workforce. Empirical evidence supports the efficacy of conversational agents in reducing the workload on providers and enhancing patient engagement through scalable, automated interfaces [10].

3.8 Safety, Diagnostic Accuracy and Usability Framework

Safety and Appropriateness: We must ensure the chatbot never gives harmful advice. Following prior work on medical chatbot, we would implement guardrails and verify that responses comply with clinical guidelines. A safety evaluation would include testing the chatbot on adversarial or emergency questions to ensure it defers to a human when needed. We also plan to adopt established guidelines to systematically score safety.

Diagnostic Accuracy: This measures the correctness of the chatbot's medical reasoning. We can frame this as a multi-class classification on symptom-to-diagnosis mapping. For example, given a question describing symptoms, the chatbot's answer implies a diagnosis, we can compute recall/sensitivity on these implied diagnoses against ground-truth. More generally, we will track standard metrics (accuracy, F1) on the medical QA tasks to quantify its clinical correctness.

Patient Usability: We measure how easily patients can use the chatbot. This follows ISO 9241-11 notions of effectiveness, efficiency, and satisfaction. We conduct user studies where participants try to get health advice via the chatbot: we record task success rates, time to answer, and user satisfaction (through surveys). We also check for accessibility and acceptance. High usability means users can effectively get understandable, helpful answers in their native dialect without undue effort.

3.9 Conclusion

Chapter III has presented a thorough evaluation of the mT5-small–based multilingual medical chatbot. While the current results remain modest with Precision, Recall, and F1-score all averaging around 47%, and overall accuracy at 35% they still reflect meaningful progress, particularly within the constraints of limited resources and under-resourced language informal varieties. The BLEU score of 40% is especially noteworthy, indicating that the model is beginning to capture some of the surface-level fluency and lexical coherence needed in a dialogue setting.

These initial results underscore both the challenges and potential of transformer-based NLP systems applied to dialectal and multilingual healthcare queries. Limitations such as domain ambiguity, informal spelling, and lack of annotated corpora likely contributed to the moderate classification performance. However, the generative capabilities, especially for code-switched input and dialectal queries, show early signs of promise.

Most importantly, this chapter illustrates that the architecture and data pipeline developed for this project are operational and scalable. With better annotated datasets, more computational resources, and further domain adaptation, we expect performance especially on linguistic fluency and medical relevance—to improve significantly in future iterations. Thus, while the current model represents a foundational step, it holds the potential for impactful public health applications in multilingual environments.

.

General Conclusion

This thesis presents a comprehensive advancement in the domain of multilingual and dialect-aware medical Chatbot. We proposed, developed, and evaluated an end-to-end question-answering medical chatbot based on the mT5-small model, tailored for use in Algerian Arabic dialects, Modern Standard Arabic (MSA), and French. The work bridges the gap between state-of-the-art NLP architectures and the realities of under-resourced, multilingual populations.

The first major contribution is the design and deployment of a multilingual encoder–decoder pipeline powered by the mT5-small architecture. By adapting this generative model to the medical domain, we demonstrated its capacity to produce fluent, and medically coherent responses in multiple languages. Our unified pipeline performs intent recognition, contextual understanding and answer generation within a single architecture.

Second, this thesis addresses the challenge of dialectal Arabic by introducing a curated dataset specific to Algerian variants, alongside preprocessing steps designed to normalize orthographic and morphological variation. This dialect-sensitive adaptation significantly enhances the model's understanding of informal expressions and regional language patterns, a feature often overlooked in prior multilingual NLP studies.

Third, we implemented a fully functional end-to-end chatbot system, integrating data collection, preprocessing, annotation, fine-tuning, and evaluation components. This practical framework demonstrates real-time deployment feasibility and can be readily adapted to other low-resource settings.

Finally, this work contributes new baselines and training strategies for dialectal Arabic and low-resource languages in the medical NLP space. By creating tools and data that reflect real-world linguistic diversity, our approach fosters greater inclusivity and paves the way for deploying intelligent medical systems in underserved regions.

Perspectives

To enhance the performance and robustness of transformer-based models in healthcare applications, several strategies can be employed. **Data expansion** through collaborations with healthcare providers or crowdsourcing can significantly improve model generalization, particularly for rare symptoms or unique linguistic expressions. This scaling of the training corpus also facilitates the effective deployment of larger transformer variants such as *mT5-base* or *mT5-large*, which require substantial data to perform optimally. Complementing this, **improved dialect detection** via a dedicated identification module in the preprocessing pipeline can enhance language-specific normalization, ensuring more accurate downstream processing in multilingual or dialect-rich settings. Furthermore, **domain-specific adaptation**, such as continued pretraining on biomedical literature or clinical notes, can refine the model's understanding of complex medical terminology and improve precision in specialized contexts. Integrating **structured medical knowledge bases**, such as ontologies or clinical decision support platforms, can strengthen the factual accuracy of generated responses and mitigate hallucinations. Finally, leveraging **semi-supervised learning techniques**, including self-training or pseudo-labeling on large unlabeled datasets, offers a promising path to boost model performance without incurring the high cost of manual annotation. Together, these approaches form a comprehensive strategy for building more reliable, scalable, and domain-adapted AI systems in healthcare.

Bibliography

[1] M. A. Hasham, V. Parshuram, and G. R. Bandre, "Social media in healthcare: Transforming patient engagement, education, and research dissemination," *Int. J. Acad. Med.*, vol. 11, no. 1, pp. 4–9, Jan.–Mar. 2025, doi: 10.4103/ijam.ijam_48_24.

[2] M. Abdelhay, A. Mohammed, and H. A. Hefny, "Deep learning for Arabic healthcare: MedicalBot," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 71, Apr. 2023, doi: 10.1007/s13278-023-01077-w.

[3] R. Mohammad, O. S. Alkhnbashi, and M. Hammoudeh, "Optimizing large language models for Arabic healthcare communication: A focus on patient-centered NLP applications," *Big Data Cogn. Comput.*, vol. 8, no. 11, p. 157, Nov. 2024, doi: 10.3390/bdcc8110157.

[4] A. Babu and S. B. Boddu, "BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding," *Explor. Res. Clin. Soc. Pharm.*, vol. 13, p. 100419, Feb. 2024, doi: 10.1016/j.rcsop.2024.100419.

[5] S. Rekik, M. Elamine, and L. Belguith, *A Medical Chatbot for Tunisian Dialect using a Rule-Based and Machine Learning Approach*, pp. 1–7, Dec. 2023, doi: 10.1109/AICCSA59173.2023.10479249.

[6] C. Blanc, A. Bailly, É. Francis, T. Guillotin, F. Jamal, B. Wakim, and P. Roy, "FlauBERT vs. CamemBERT: Understanding patient's answers by a French medical chatbot," *Artif. Intell. Med.*, vol. 127, p. 102264, May 2022, doi: 10.1016/j.artmed.2022.102264.

[7] A. Jerfy, O. Selden, and R. Balkrishnan, “The Growing Impact of Natural Language Processing in Healthcare and Public Health,” *Inquiry*, vol. 61, 2024, Art. no. 469580241290095.

[8] L. Xue *et al.*, “mT5: A massively multilingual pre-trained text-to-text transformer,” *arXiv* preprint arXiv:2010.11934, 2020.

[9] I. García-Ferrero *et al*., “Medical mT5: An open-source multilingual text-to-text LLM for the medical domain,” *CoRR*, vol. 01, 2024.

[10] O. Baclic *et al*., “Challenges and opportunities for public health made possible by advances in natural language processing,” *Can. Commun. Dis. Rep.*, vol. 46, no. 6, pp. 161–168, Jun. 2020.

[11] M. J. Paul *et al*., “Social media mining for public health monitoring and surveillance,” in *Proc. Pacific Symp. Biocomput.* 2016, pp. 468–479.

[12] M. Clark and S. Bailey, “Chatbots in Health Care: Connecting Patients to Information,” CADTH Horizon Scan Report EH0122, Jan. 2024.

[13] M. Albirini, *Modern Arabic Sociolinguistics: Diglossia, Variation, Codeswitching, Attitudes and Identity*. Routledge, 2016.

[14] M. A. Al-Sobh, A. H. Abu-Melhim, and N. A. Bani-Hani, “Diglossia as a result of language variation in Arabic: Possible solutions in light of language planning,” *Journal of Language Teaching and Research*, vol. 6, no. 2, pp. 274–279, Mar. 201

[15] H. Saadane and N. Habash, "A conventional orthography for Algerian Arabic," in *Proc. 1st Workshop on Arabic Natural Language Processing (ANLP)*, Beijing, China, Jul. 2015, pp. 69–79.

[16] M. O. A. Hegazi, Y. Al-Dossari, A. Al-Yahy, A. Al-Sumari, and A. Hilal, "Preprocessing Arabic text on social media," *Heliyon*, vol. 7, no. 2, e06191, Feb. 2021

[17] A. Abdaoui, M. Berrimi, M. Oussalah, and A. Moussaoui, "DziriBERT: a pre-trained language model for the Algerian dialect," *arXiv preprint arXiv:2109.12346*, Dec. 2022

[18] S. Bougrine, A. Chorana, A. Lakhdari, and H. Cherroun, "Toward a web-based speech corpus for Algerian Arabic dialectal varieties," in *Proceedings of the Third Arabic Natural Language Processing Workshop*, Valencia, Spain, Apr. 2017

[19] AltexSoft, "Named Entity Recognition: What Is It and How Does It Work?," *AltexSoft Blog*.

[20] X. Qu, Y. Gu, Q. Xia, Z. Li, Z. Wang, and B. Huai, "A survey on Arabic named entity recognition: Past, recent advances, and future trends," *arXiv preprint arXiv:2302.03512*, Feb. 2023.

[21] N. Manchev, "Building a Named Entity Recognition Model Using a BiLSTM-CRF Network," *Domino Data Lab Blog*, Jul. 13, 2023.

[22] A. H. Dahou and M. A. Cheragui, "Named entity recognition for Algerian Arabic dialect in social media," in *Proceedings of the 12th International Conference on Information Systems and Advanced Technologies (ICISAT 2022)*, Lecture Notes in Networks and Systems, vol. 525, Springer, Cham, 2023, pp. 135–145.

[23] S. Touileb, "NERDz: A Preliminary Dataset of Named Entities for Algerian," in *Proc. 2nd Conf. Asia-Pacific Chapter of the Association for Computational Linguistics and 12th Int. Joint Conf. Natural Language Processing (Vol. 2: Short Papers)*, Online Only, Nov. 2022, pp. 95–101.

[24] M. ALMutairi, L. AlKulaib, M. Aktas, S. Alsalamah, and C.-T. Lu, "Synthetic Arabic Medical Dialogues Using Advanced Multi-Agent LLM Techniques," in *Proc. 2nd Arabic Natural Language Processing Conf. (ArabicNLP 2024)*, Bangkok, Thailand, Aug. 2024, pp. 11–26. doi: 10.18653/v1/2024.arabicnlp-1.2.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Proc. 31st Conf. Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, CA, USA, Dec. 2017, pp. 6000–6010, doi: https://doi.org/10.5555/3295222.3295349.

[26] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, Aug. 2021, pp. 7088–7105.

[27] H. Al Shamsi, A. G. Almutairi, S. Al Mashrafi, and T. Al Kalbani, "Implications of language barriers for healthcare: A systematic review," *Oman Medical Journal*, vol. 35, no. 2, pp. e122, Apr. 2020.

[28] K. Xu, Y. Cheng, W. Hou, Q. Tan, and W. Li, "Reasoning Like a Doctor: Improving Medical Dialogue Systems via Diagnostic Reasoning Process Alignment," *arXiv preprint arXiv:2406.13934*, Jun. 2024. [Online]. Available: https://arxiv.org/abs/2406.13934.

[29] Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, 2020.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, Jun. 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423

[31] BigScience, "BLOOM: A 176B-Parameter Multilingual Open-Science Language Model," *BigScience*, Apr. 2023.

[32] Li and Qin, "Anonymizing and Sharing Medical Text Records," *IS Research*, 2017.

[33] Dasigi and Diab, "CODACT: Identifying Orthographic Variants in Dialectal Arabic," *IJCNLP*, 2011.

[34] Obeid *et al.*, "CAMeL Tools: An Open-Source Python Toolkit for Arabic NLP," *LREC*, 2020.

[35] Xinyu *et al.*, "Named Entity Recognition on Medical Data using Pre-Trained LLM," CS 6301 Project, 2024.

[36] Dowling *et al.*, “From MLOps to ML Systems with Feature/Training/Inference Pipelines,” *Hopsworks Blog*, 2023.

[37] Canadian Agency for Drugs and Technologies in Health (CADTH), "Chatbots in Health Care: Connecting Patients to Information," *CADTH Issues in Emerging Health Technologies*, no. 201, Mar. 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK602381/

[38] F. Branda, M. Stella, C. Ceccarelli, F. Cabitza, G. Ceccarelli, A. Maruotti, M. Ciccozzi, and F. Scarpa, "The Role of AI-Based Chatbots in Public Health Emergencies: A Narrative Review," *Future Internet*, vol. 17, no. 4, p. 145, Apr. 2025, doi: 10.3390/fi17040145.

[39] G. M. S. Himel, M. S. Hasan, U. S. Salsabil, and M. M. Islam, "MedLingua: A conceptual framework for a multilingual medical conversational agent," *MethodsX*, vol. 12, p. 102614, Feb. 2024, doi: 10.1016/j.mex.2024.102614.

[40] Google Developers, "Classification: Accuracy, recall, precision, and related metrics," *Machine Learning Crash Course*, Jun. 2024. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall

[41] M. A. Al-Saqqa, A. A. Al-Ayyoub, and Y. Jararweh, "Cross-Lingual Transfer Learning for Arabic Task-Oriented Dialogue Systems Using Multilingual Transformer Model mT5," *Mathematics*, vol. 10, no. 5, p. 746, Mar. 2022, doi: 10.3390/math10050746.

[42] S. Harrat, K. Meftouh, M. Abbas, and K. Smaili, "Building Resources for Algerian Arabic Dialects," presented at the 1st International Workshop on Social Media Processing for Arabic (SMPA), 2014. [Online]. Available: https://inria.hal.science/hal-01066989v1/document

[43] H. Saleh, A. AlMohimeed, R. Hassan, M. M. Ibrahim, S. H. Alsamhi, M. R. Hassan, and S. Mostafa, "Advancing Arabic dialect detection with hybrid stacked transformer models," *Front. Hum. Neurosci.*, vol. 19, Feb. 2025, Art. no. 1498297, doi: 10.3389/fnhum.2025.1498297.

[44] D. P. Jeong, S. Garg, Z. C. Lipton, and M. Oberst, "Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress?," *arXiv preprint arXiv:2411.04118*, Nov. 2024. [Online]. Available: https://arxiv.org/abs/2411.04118

[45] Z. Shi, F. Tonolini, N. Aletras, E. Yilmaz, G. Kazai, and Y. Jiao, "Rethinking Semi-Supervised Learning with Language Models," in *Findings of the Association for Computational Linguistics: ACL 2023*, Jul. 2023, pp. 5614–5634. [Online]. Available: https://aclanthology.org/2023.findings-acl.347.pdf

[46] Q. Bao, L. Ni, and J. Liu, "HHH: An Online Medical Chatbot System based on Knowledge Graph and Hierarchical Bi-Directional Attention," *arXiv preprint arXiv:2002.03140*, Feb. 2020. [Online]. Available: https://arxiv.org/abs/2002.03140

[47] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. (online manuscript), Jan. 12, 2025.

[48] "Natural Language Processing (NLP): what it is and how it works?," *Future Processing Blog*, 29 Mar. 2021.

[49] A. Boulesnane, Y. Saidi, O. Kamel, M. M. Bouhamed, and R. Mennour, "DZchatbot: A Medical Assistant Chatbot in the Algerian Arabic Dialect using Seq2Seq Model," in *Proc. 4th Int. Conf. Pattern Analysis and Intelligent Systems (PAIS)*, Oum El Bouaghi, Algeria, Oct. 12–13, 2022, pp. 1–8, doi: 10.1109/PAIS56586.2022.9946867.

[50] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007. [Online]. Available: https://nlp.cs.nyu.edu/sekine/papers/li07.pdf

[50] A. Dey and B. S. Prukayastha, "Named Entity Recognition using Gazetteer Method and N-gram Technique for an Inflectional Language: A Hybrid Approach," *Int. J. Comput. Appl.*, vol.

84, no. 9, pp. 31–35, Dec. 2013.

[52] S. Thenmalar, J. Balaji, and T. V. Geetha, "Semi-supervised Bootstrapping Approach for Named Entity Recognition," *arXiv preprint arXiv:1511.06833*, Nov. 2015. [Online]. Available: https://arxiv.org/abs/1511.06833

[53] H. Fehri, S. Dardour, and K. Haddar, "ARmed question answering system," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 21, Jun. 10 2022, Art. e7054, doi: 10.1002/cpe.7054

[54] P. Nakov, L. Màrquez, A. Moschitti, and H. Mubarak, "Arabic community question answering," *Natural Language Engineering*, vol. 25, no. 1, pp. 5–41, Jan. 2019, doi: 10.1017/S1351324918000426.

[55] A. Alasmari, S. Alhumoud, and W. Alshammari, "AraMed: Arabic Medical Question Answering using Pretrained Transformer Language Models," in *Proc. 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) @ LREC-COLING 2024*, Torino, Italy, May 2024, pp. 50–56. [Online]. Available: https://aclanthology.org/2024.osact-1.6.pdf

[56] Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models Are Unsupervised Multitask Learners," *OpenAI Blog*, Feb. 14 2019.

[57] D. Boughareb, R. Boughareb, S. Hallaci, M.R.E Boukherouba, and H. Seridi. 2025. Addressing the Complexity of Dialectal Arabic: An Enhanced Encoder-Decoder Ensemble Approach for Optimized Sentiment Analysis. ACM Trans. Asian Low-Resour. Lang. Inf. Process. Just Accepted (May 2025). https://doi.org/10.1145/3735972.

[58] I. Guellil, F. Azouaou, F. Benali, A.-E. Hachani, and H. Saadane, "Approche Hybride pour la translitération de l'arabizi algérien : une étude préliminaire," in *Proc. 25e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Rennes, France, May 2018.

[59] W. Antoun, F. Baly, and H. Hajj, "AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding," *arXiv preprint arXiv:2012.15516*, Dec. 2020. [Online]. Available: https://arxiv.org/abs/2012.15516

[60] BigScience, "BLOOMZ: A Multilingual Instruction-Tuned Version of BLOOM," *BigScience*, 2023. [Online]. Available: https://huggingface.co/bigscience/bloomz

[61] P. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proc. 58th Annual Meeting of the Assoc. Comput. Linguistics (ACL 2020)*, Jul. 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747

[62] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT: A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," in *Proc. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing at NeurIPS 2019*, Dec. 2019.

[63] T.-D. Le, P. Jouvet, and R. Noumeir, "Improving Transformer Performance for French Clinical Notes Classification Using Mixture of Experts on a Limited Dataset," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 13, pp. 261–274, 2025, doi: 10.1109/JTEHM.2025.3576570.

[64] Arun Babu and Sekhar Babu Boddu, "BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding," *Exploratory Research in Clinical and Social Pharmacy*, vol. 13, art. 100419, Feb. 15, 2024, doi: 10.1016/j.rcsop.2024.100419