Automated De-Identification of Arabic Medical Records

Veysel Kocaman
John Snow Labs Inc.

Youssef Mellah
John Snow Labs Inc.

Hasham Ul Haq
John Snow Labs Inc.

David Talby
John Snow Labs Inc.

Abstract

As Electronic Health Records (EHR) become ubiquitous in healthcare systems worldwide, including in Arabic-speaking countries, the dual imperative of safeguarding patient privacy and leveraging data for research and quality improvement grows. This paper presents a first-of-its-kind automated de-identification pipeline for medical text specifically tailored for the Arabic language. This includes accurate medical Named Entity Recognition (NER) for identifying personal information; data obfuscation models to replace sensitive entities with fake entities; and an implementation that natively scales to large datasets on commodity clusters.

This research makes two contributions. First, we adapt two existing NER architectures—BERT For Token Classification (BFTC) and BiLSTM-CNN-Char – to accommodate the unique syntactic and morphological characteristics of the Arabic language. Comparative analysis suggests that BFTC models outperform BiLSTM models, achieving higher F1 scores for both identifying and redacting personally identifiable information (PII) from Arabic medical texts. Second, we augment the deep learning models with a contextual parser engine to handle commonly missed entities. Experiments show that the combined pipeline demonstrates superior performance with micro F1 scores ranging from 0.94 to 0.98 on the test dataset, which is a translated version of the i2b2 2014 de-identification challenge, across 17 sensitive entities. This level of accuracy is in line with that achieved with manual de-identification by domain experts, suggesting that a fully automated and scalable process is now viable.

1 Introduction

Arabic is one major language that covers a large geographic and demographic portion of the world population with a high EHR adoption rate (Abdullah Alharbi, 2023). This means there is a high volume of both structured and unstructured digital data available that can be leveraged for different use cases. However, the data needs to be de-identified before being used for any research or development purpose.

De-identification of unstructured documents poses challenges due to various types of noise. Furthermore, every language has its own lexical rules, which makes it challenging to have a single model that can perform well across multiple languages. Therefore, there is a need to have models trained for different languages to get the best results. Usually, Named Entity Recognition (NER) models are used to extract sensitive information from the text which can then be de-identified (Uzuner et al., 2007). However, training NER models require labeled datasets, which are scarce and laborious to produce. In particular, the Arabic language has an extremely limited number of public datasets that can be leveraged.

The principal aim of this study is fourfold: Firstly, we introduce the first-of-its-kind medical Named Entity Recognition (NER) and De-identification models tailored specifically for the Arabic language, addressing a critical gap in the field. Secondly, we adapt existing NER architectures—BiLSTM-CNN-Char and BERT For Token Classification (BFTC)—to meet the unique syntactic and morphological requirements of the Arabic language. Thirdly, we implement a novel approach to overcome dataset limitations by translating a standard English dataset used in the 2014 i2b2 De-Identification challenge to Arabic using an entity-preservation technique. Fourthly, we employ a contextual parser engine to supplement weak entity extractions, thereby increasing the robustness of our models.

To train, evaluate, and compare these NER models, we use the Spark NLP for Healthcare library (Kocaman and Talby, 2021b), which offers both comprehensive NER support (Kocaman and Talby, 2022) and token embedding models for the Arabic

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 33–40

language. Importantly, this is not purely academic research; it's an applied study that has been engineered to be fully compatible and scalable with Apache Spark, making it immediately deployable in large-scale healthcare systems.

2 Related Work

The concept of automatic de-identification was first introduced into the Informatics for Integrating Biology and the Bedside (i2b2) project as explained by (Uzuner et al., 2007) and then expanded by (Stubbs et al., 2015), as an academic NLP challenge on automatically detecting PHI identifiers from medical records. These challenges have boosted research and development of Machine & Deep Learning algorithms for robust PHI identification.

Since then, there have been numerous studies to expand automatic de-identification to multiple languages. (Marimon et al., 2019) generated a dataset, and trained NER models for medical texts in Spanish language. (Catelli et al., 2020) applied similar techniques to Italian COVID-19 documents for de-identification.

Over the years, researchers have proposed multiple architectures aiming to achieve better performance. Initial approaches relied on hand-crafted features and lexical rules to extract required concepts from data. However, as token embedding models (Mikolov et al., 2013) advanced, other architectures started leveraging these embedding models. Among these, Bi-LSTM and Conditional Random Fields (CRF) based models (Huang et al., 2015) became notable for NER. More recently, attention-based models have been showing significantly better performance for sequence labeling tasks (Vaswani et al., 2023).

Regarding other efforts towards extracting medical terms from Arabic medical texts, several noteworthy studies have been conducted. (Nayel et al., 2023) explored deep learning techniques, including LSTM-CRF and BiLSTM-CRF models, for disease entity recognition in Arabic medical texts, achieving impressive precision, recall, and F1-scores. (Alanazi, 2017) introduced Bayesian Belief Networks (BBN) as an innovative approach to extracting various medical entities, demonstrating promising precision and recall for diseases and treatment methods.

In addition, (Abdelhay et al., 2023) tackled the challenges of implementing medical bots in Arabic with the introduction of the MAQA dataset, highlighting the effectiveness of Transformer models. (Hammoud et al., 2020) fine-tuned neural networks for medical entity recognition in Arabic medical texts, while (Hammoud et al., 2021) presented a novel dataset for disease classification, emphasizing the potential of pre-trained models. Finally, (Samy et al., 2012) compared strategies for medical term extraction, revealing the advantages of using Arabic equivalents of Latin prefixes and suffixes. These studies collectively advance the field of NER and medical term extraction in Arabic medical texts, offering a range of valuable approaches and insights.

Despite these advancements, it is crucial to note that there has been a notable absence of de-identification models or efforts explicitly targeting the Arabic language. This gap in the literature underscores the importance and timeliness of our study, which aims to address this void by introducing the first Arabic-specific medical NER and De-identification models.

3 Dataset Construction and Annotation

Training a named entity recognition model requires data to be annotated with named entities which is a laborious process. Instead of manually annotating an Arabic dataset, we took the standard 2014 i2b2 dataset (in English) (Stubbs et al., 2015) and translated it to Arabic using the Google translate API [1]. The i2b2 dataset is in CoNLL format, which means text is tokenized, and entities are identified using the IOB2 tagging scheme [2]. Since entities have fixed boundaries relative to the original text, translating the text naively would result in entity boundary mismatch.

For example, the name and age in the text *"Alan is a 30 year old male"* start at token 1 and 4, however, after translation, the name and age start at token 1 and 6 "آلان رجل يبلغ من العمر ٣٠ عامًا". This is because translation can change the entire structure of the text, consequently, making entity

[1] https://cloud.google.com/translate

[2] In Named Entity Recognition (NER), the IOB2 (Inside-Outside-Beginning) tagging scheme is a common way to annotate and identify entities in a text. In this scheme, each word in a sequence is tagged with one of the following prefixes: "B-" (Beginning): Indicates that the word is the start of a named entity. "I-" (Inside): Indicates that the word is inside a named entity, but is not the first word of the entity. "O" (Outside): Indicates that the word is not part of any named entity. These prefixes are then followed by the type of the entity, such as "PER" for person, "LOC" for location, "ORG" for organization, etc. This makes it easier to identify not just the entities in a sequence, but also their types and spans.

boundary mapping challenging. This problem is further exacerbated for entities spanning across multiple tokens as the number of tokens could also vary.

To solve this problem, we replace entities in the original (English) text with their types. For example, *"Alan is a 30 year old male"* would be converted to *"NAME is a AGE year old male"*. This way when the text is translated, we can search for the entity types by simple string matching, and replace them with Arabic values. For instance, "NAME" is replaced with an actual Arabic name "يوسف". In addition to solving the problem of preserving entity boundaries, this technique also helps to adapt the data to the new language, as entities, such as names, cities, addresses are native Arabic values.

The original i2b2 Deid dataset provides two types of entity sets: Generic and Granular. The granular approach provides additional context that can be crucial for specific applications. For example, in a healthcare setting, knowing that a name refers to a "PATIENT" rather than just a "NAME" could be highly useful. Similarly, distinguishing between ZIP codes, cities, and countries can be very important in applications like location-based services or logistics. The generic approach is more broad and could be useful for general-purpose NER tasks where such granular distinctions are not necessary. It may also require less computational power and resources than the more detailed granular approach. Here is a sample list of mapping between generic and granular set of entities:

- NAME (PATIENT, DOCTOR, USERNAME)
- LOCATION (ROOM, DEPARTMENT, HOSPITAL, ORGANIZATION, STREET, CITY, STATE, COUNTRY, ZIP, OTHER)
- AGE
- DATE
- CONTACT (PHONE, FAX, EMAIL, URL, IPADDRESS)
- IDs (SOCIAL SECURITY NUMBER, MEDICAL RECORD NUMBER, HEALTH PLAN NUMBER, ACCOUNT NUMBER, LICENSE NUMBER, VEHICLE ID, DEVICE ID, BIOMETRIC ID, ID NUMBER)
- PROFESSION

Table 1 illustrates the difference between the generic and granular entity datasets. The details regarding the differences between entity sets, annotation schema, and annotation guidelines can be found at (Stubbs et al., 2015).

Chunk	Generic	Granular
2000 16	DATE	DATE
ليلى حسن	NAME	PATIENT
789	LOCATION	ZIP
جدة	LOCATION	CITY
54321	LOCATION	ZIP
المملكة العربية	LOCATION	CITY
السعودية	LOCATION	COUNTRY
النور	LOCATION	HOSPITAL
اميرة احمد	NAME	DOCTOR
ليلى	NAME	PATIENT
35	AGE	AGE

Table 1: Tokenized illustration of difference between generic and granular entities. In the "Generic" column, entities are tagged with broad, high-level categories. On the other hand, the "Granular" column takes entity recognition a step further by using more specific, detailed tags.

4 Architecture

4.1 Scalable NLP Pipeline

Our system leverages the capabilities of Spark NLP (Kocaman and Talby, 2021b), a widely-used open-source NLP library that excels in scalability for both training and inference tasks on any Apache Spark setup. The architecture allows for easy deployment either on a single machine or across a Spark cluster without requiring any modification to the code base. The de-identification process for Arabic text is realized through a multi-stage NLP pipeline, consisting of text pre-processing, deep learning models, contextual guidelines, and data masking techniques. The pipeline components can be seen at Figure 2.

4.1.1 Text Pre-Processing

The pipeline's initial phase involves multiple components such as a document assembler, sentence detector, token generator, and word embedding creator. These components are designed to prepare the data for identification and subsequent anonymization of Protected Health Information (PHI) tokens in Arabic.

At the outset, a document assembler is utilized to structure raw Arabic text, generating annota-

Original text	Deidentification with entity labels	Deidentification with special chars	Obfuscation with fakers
ملاحظات سريرية - مريضة الغدة الدرقية. التاريخ: 29 اكتوبر1995. اسم المريضة: سارة عبد الرحمن. العنوان: شارع النجاح، مبنى رقم 987، حي الأمل، القاهرة. الرمز البريدي: 67890. البلد: مصر. اسم المستشفى: مستشفى الأمل. اسم الطبيب: د. أحمد مصطفى. تفاصيل الحالة: المريضة سارة عبد الرحمن، البالغة من العمر 45 عامًا، تعاني من اضطرابات في الغدة الدرقية. تشكو من زيادة الوزن والتعب والاكتئاب.	ملاحظات سريرية - مريضة الغدة الدرقية. التاريخ: <DATE> <USERNAME>. اسم المريضة: <PATIENT>. العنوان: <LOCATION> مبنى رقم <ZIP> حي <CITY> <CITY>. الرمز البريدي: <ZIP>. البلد: <COUNTRY>. اسم المستشفى: مستشفى الأمل. اسم الطبيب: د. <DOCTOR>. تفاصيل الحالة: المريضة <DOCTOR> البالغة من العمر <AGE>عامًا، تعاني من اضطرابات في الغدة الدرقية. تشكو من زيادة الوزن والتعب والاكتئاب.	ملاحظات سريرية - مريضة الغدة الدرقية. التاريخ: [***]. اسم المريضة: [***] [***]. العنوان: [***] مبنى رقم [**] حي [***] [***]. الرمز البريدي: [***]. البلد: [***]. اسم المستشفى: مستشفى الأمل. اسم الطبيب: د. [***]. تفاصيل الحالة: المريضة [***] البالغة من العمر ***عامًا، تعاني من اضطرابات في الغدة .	ملاحظات سريرية - مريضة الغدة الدرقية. التاريخ: 30 اكتوبر1995. اسم المريضة: رجب زغلي . العنوان: الشارع الأزرق، مبنى رقم 799، حي الأمل،القصر . الرمز البريدي: 41019. البلد: أوكرانيا . اسم المستشفى: مستشفى الأمل. اسم الطبيب: د. إلهام محمود . تفاصيل الحالة: المريضة رأفت طاهر ، البالغة من العمر 57 عامًا، تعاني من اضطرابات في الغدة الدرقية. تشكو من زيادة الوزن والتعب والاكتئاب.
ملاحظات سريرية - مريض الربو. التاريخ: 13 أبريل 2000. اسم المريض: أحمد سليمان. العنوان: شارع السلام، مبنى رقم 555، حي الصفاء، الرياض. الرمز البريدي: 54321. البلد: المملكة العربية السعودية. اسم المستشفى: مستشفى الأمانة. اسم الطبيب: د. ريم الحمد. تفاصيل الحالة: المريض أحمد سليمان، البالغ من العمر 30 عامًا، يعاني من مرض الربو المزمن	ملاحظات سريرية - مريض الربو. التاريخ: <DATE>. اسم المريض: <PATIENT>. العنوان: شارع السلام، مبنى رقم <ZIP> حي الصفاء، <CITY>. الرمز البريدي: <ZIP> البلد: <COUNTRY> <CITY>. اسم المستشفى: <LOCATION>. اسم الطبيب: د. <DOCTOR>. تفاصيل الحالة: المريض أحمد سليمان، البالغ من العمر <AGE>عامًا، يعاني من مرض الربو المزمن.	ملاحظات سريرية - مريض الربو. التاريخ: [***]. اسم المريض: [***]. العنوان: شارع السلام، مبنى رقم [**] حي الصفاء، [****]. الرمز البريدي: [***]. البلد: [****] [***]. اسم المستشفى: [***]. اسم الطبيب: د. [***]. تفاصيل الحالة: المريض أحمد سليمان، البالغ من العمر **عامًا، يعاني من مرض الربو المزمن.	ملاحظات سريرية - مريض الربو. التاريخ: 13 مايو 2000. اسم المريض: إحسان . العنوان: شارع السلام، مبنى رقم 471، حي الصفاء، الأقصر . الرمز البريدي: 46763 . البلد: أوزبكستان . اسم المستشفى: مستشفى الأزرق. اسم الطبيب: د. حاتم غالية . تفاصيل الحالة: المريضأحمد سليمان، البالغ من العمر 38 عامًا، يعاني من مرض الربو المزمن.

Figure 1: Example of de-identifying a text in Arabic using masking and obfuscation.

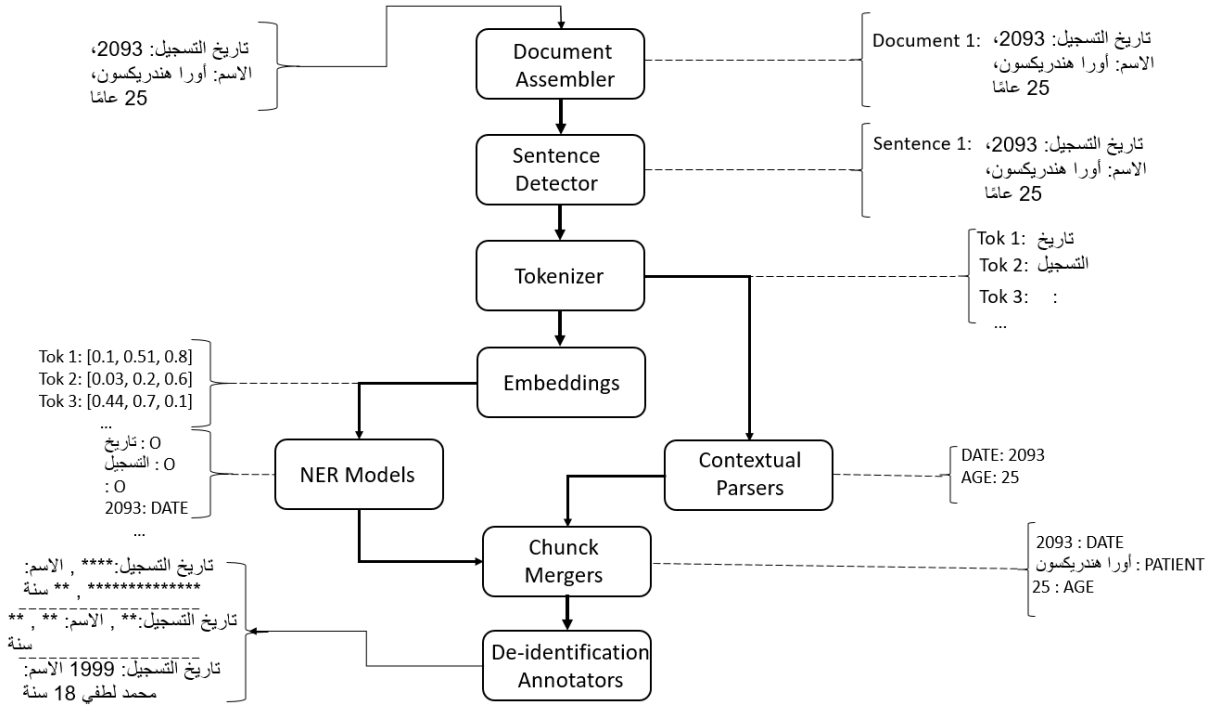

Figure 2: Full pipeline architecture

tions that can be processed further downstream. Following this, the pipeline employs a specialized deep-learning model (Schweter and Ahmed, 2019) optimized for Arabic clinical texts to perform sentence boundary detection. Rule-based techniques underperform in this context, owing to the unique grammar and punctuation in Arabic medical notes.

4.1.2 Named Entity Recognition

The core of the de-identification mechanism is the Named Entity Recognition (NER) model. It identifies PHI components like patients' names, healthcare providers, facilities, geographical locations, and specific identification numbers in Arabic text. The NER model is a crucial element as it minimizes data loss while recognizing PHI efficiently. For this, we employ a Bi-directional LSTM (BLSTM) architecture as detailed in (Kocaman and Talby, 2021a).

4.1.3 Enhancing NER with Contextual Rule Engine

While machine learning models excel in generalization, they might lack the granularity required for certain PHI identifiers. Therefore, a regular-expression-based rule engine is included in the pipeline to address this limitation. The rule engine, called Contextual Parser (CP), offers a set of adjustable parameters for prefix and suffix matching, enhancing the system's precision.

In the realm of de-identification, augmenting our NER models with CP rules offers a robust strategy for enhanced recognition and protection of Personal Health Information (PHI) elements. CP rules are linguistically tailored regulations that exploit the surrounding context of entities to optimize their detection accuracy. This is particularly useful for handling complex medical terminology, ambiguous entities, and cultural or geographical variations, especially in Arabic medical texts.

Rule Formulation: A collaborative effort between domain experts and translators allows us to design a set of CP rules that are specific to both the medical domain and the Arabic language. These rules address the unique linguistic complexities of medical texts, such as abbreviations, compound terms, and varying morphological patterns. Special attention is given to rules that target the identification of critical PHI elements like email addresses, dates, and identification numbers.

Entities Reinforced by CP Rules: The CP rules particularly bolster the NER model's ability to identify and protect a diverse array of entities. These include but are not limited to Social Security Numbers (SSN), Account Numbers (ACCOUNT), License Numbers (LICENSE), Ages (AGE), Phone Numbers (PHONE), ZIP Codes (ZIP), Medical Record Numbers (MEDICALRECORD), Emails (EMAIL), Dates (DATE), Driver's License Numbers (DLN), and Vehicle Identification Numbers (VIN).

In summary, the incorporation of CP rules into a de-identification process enhances the capabilities of our NER models, making them highly adaptable and effective in identifying a broad range of PHIs. Our model now proficiently identifies and protects the aforementioned entities, demonstrating the efficacy of our approach in safeguarding patient information in Arabic medical texts.

This multi-dimensional approach, combining data-driven deep learning with domain-specific linguistic rules, showcases the flexibility and robustness of our NER models. It not only fortifies our system against privacy intrusion but also aligns it with data protection laws.

4.1.4 Chunk Merger

Subsequent to the identification of PHI chunks by machine learning models and rule-based methods, the pipeline consolidates these identifications to optimize overall accuracy. The system assigns priorities to each type of entity, allowing for customization depending on use-cases.

4.1.5 Masking or Obfuscation

In the final stage, the system performs the actual deidentification and obfuscation. This involves masking or substituting PHI elements with dummy data while preserving the overall structure and format of the documents.

Accurate NER is the first step towards de-identifying a text - the next step is to redact the information. This can be achieved by applying either masking or obfuscation. Masking essentially replaces the identified entities with either their entity type or asterisks. These asterisks can either be of fixed character length for all the identified entities, or of the same length as the entity chunk being replaced; we found the later option to be helpful while de-identifying pdf and image documents, as it minimizes any changes to the original document layout.

Obfuscation involves replacing PHI with surrogate values that are semantically, and linguistically correct. For example, names are replaced with random names, similarly, dates are replaced with randomized dates within an offset window. Although obfuscation appears to be the better de-identification strategy as it obfuscates the entire text, making it harder to re-identify (even when an entity is missed by the NER model), there are some inherent challenges while maintaining data integrity. For example, multiple occurrences of names, addresses, and dates should be replaced with similar values throughout the document to maintain data integrity. The Spark NLP for Healthcare library already has built-in methods to track entities for consistent obfuscation.

Figure 1 illustrates text de-identified using masking and obfuscation.

5 Experimentation & Analysis

Two different NER architectures are trained and evaluated on a standard 80-20 split, and their performance is evaluated based on the model architecture and the embeddings used while training. The first model is based on a Bi-LSTM architecture as explained in (Kocaman and Talby, 2021a). This Bi-LSTM model is versatile and can be paired with virtually any token embedding model. In our experiments, we use this architecture with GLoVe (Pennington et al., 2014) and BERT (Devlin et al., 2019) embeddings. The GLoVe embeddings are

trained on the Arabic common crawl dataset [3] [4]. For Arabic BERT embeddings, we utilize models pre-trained on an Arabic dataset; AraBERT (Antoun et al., 2021), and CamelBERT (Inoue et al., 2021). The second model architecture is based solely on BERT, upon which we train end-to-end BERT For Token Classification (BFTC) models.

In terms of model architecture, the BFTC models outperformed Bi-LSTM based models on both datasets as explained in Table 2 and 3. The Bi-LSTM model trained with GLoVe, AraBERT, and CamelBERT embeddings achieved macro F1 score of **0.9378**, **0.9372**, **0.9590** on the generic entity dataset, and **0.9386**, **0.9178**, **0.9369** on the granular entity dataset. In comparison, the BFTC models achieved 1-2% higher F1 scores.

In addition to the named entities in our training dataset, most documents contain certain rule-based entities like unique organizational/national identifiers. Extracting such information does not necessarily require re-training the model, as most of these identifiers have a fixed format, and can be easily extracted using regular expressions. Therefore, we include a regular expression engine in the final pipeline that is fully customizable as explained in section 4.1.3. Figure 2 illustrates a complete end-to-end pipeline with all the components.

6 Conclusion

In conclusion, this study successfully presents a groundbreaking advancement in healthcare data privacy and research for Arabic-speaking communities by introducing the first medical Named Entity Recognition (NER) and De-identification models tailored specifically for the Arabic language. Through the adaptation of existing architectures—BiLSTM-CNN-Char and BERT For Token Classification (BFTC)—we were able to accommodate the unique linguistic features of Arabic. Furthermore, our novel entity-preservation technique was pivotal in overcoming the challenges associated with limited datasets, enabling the translation of a standard English dataset into Arabic for training and evaluation.

Our comparative analyses demonstrated that BERT For Token Classification models outperformed Bi-LSTM models, achieving higher F1 scores in both the identification and redaction of personally identifiable information (PII) in Arabic medical texts. The contextual parser engine deployed in our study further enhanced the robustness of our models.

Significantly, this work is more than just an academic endeavor; it is an applied study with tools that are ready to be deployed at scale using Apache Spark. As a seminal contribution, this research not only provides essential tools for the safe and efficient handling of Arabic medical records but also lays a foundation for future studies, opening up avenues for the adaptation of NER and De-identification techniques to other underrepresented languages.

7 Limitations

Following are some of the limitations of the solution that may affect its generalizability and reliability, and need to be studied further for improvements:

7.1 Dataset quality and Diversity

The translation of English to Arabic (achieved through the Google Translate API), may not be able to completely take into account the detailed linguistic diversity and medical terminology in this domain. This could result in inaccurate data from a translated dataset that would affect the performance of NER models. Moreover, since there are differences in grammatical structures between the languages, direct substitution of masked chunks with Arabic texts may produce syntactic and contextual ambiguities. The division of entities and their classifications may be affected by these ambiguities. In translation errors, noise, and inconsistencies in the dataset could be introduced that might affect model performance.

7.2 Limited Vocabulary and Language Nuances

Arabic, which may be difficult for the NER models to read accurately, is a diverse language with different dialects and nuances. In the field of medicine, there are further difficulties to be encountered with domain-specific jargon and terminology. The model's performance may be hindered by the fact that it does not have an effective ability to deal with uncommon and distinct domain terms which could result in erroneous negative findings or misclassification.

[3] https://commoncrawl.org/

[4] https://fasttext.cc/docs/en/crawl-vectors.html

Model	AGE	CNTC	DATE	ID	LOC	NAME	PRO	GEND	Macro	Micro
Bi-LSTM (GLoVe CC)	0.9870	0.9799	0.9870	0.8358	0.9413	**0.9648**	**0.9210**	0.8863	0.9378	0.9572
Bi-LSTM (AraBERT-base)	0.9727	0.9696	0.9734	0.8450	0.8675	0.8784	0.8071	0.8869	0.9372	0.9505
Bi-LSTM (CamelBERT)	**0.9885**	0.9666	0.9757	0.8656	0.8975	0.9111	0.8675	0.9096	0.9590	0.9712
BFTC (AraBERT-base)	0.9854	**0.9852**	**0.9901**	**0.9467**	0.9225	0.9425	0.8622	0.9507	0.9600	0.9800
BFTC (CamelBERT)	0.9830	0.9828	0.9899	0.9333	**0.9494**	0.9624	0.8601	**0.9556**	**0.9700**	**0.9800**

Table 2: F1 scores on the generic entity dataset (CNTC: Contact, LOC: Location, PRO: Profession, GEND: Gender).

Entity	1	2	3	4	5
ZIP	0.9756	0.9580	0.9566	0.9483	0.9510
USER	1.0000	1.0000	1.0000	0.9557	1.0000
STR	0.9856	0.9841	0.9836	0.9186	0.9824
GEND	0.8850	0.8895	0.8918	0.9508	0.9262
PRO	0.9113	0.8284	0.8780	0.8498	0.8676
PH	0.9268	0.9135	0.8918	0.9352	0.9558
PAT	0.8711	0.7786	0.7898	0.8054	0.8134
ORG	0.8283	0.6046	0.7469	0.7376	0.8571
MR	0.9714	0.8571	0.7441	0.9230	1.0000
ID	0.9630	0.9629	0.9629	0.9718	0.9390
HOSP	0.8319	0.8081	0.8766	0.8969	0.9363
EMAIL	0.9782	0.9955	1.0000	1.0000	1.0000
DOC	0.9392	0.8951	0.9199	0.9345	0.9314
DATE	0.9876	0.9775	0.9768	0.9903	0.9922
CNTR	0.9461	0.8650	0.8750	0.9038	0.9362
CITY	0.9756	0.8788	0.8953	0.9400	0.9641
AGE	0.9799	0.9755	0.9879	0.9854	0.9830
Macro	0.9386	0.9178	0.9369	0.9400	0.9100
Micro	0.9434	0.9419	0.9547	0.9800	0.9800

Table 3: F1 scores on the granular entity dataset. Numbers in the columns refer to the following models: 1: Bi-LSTM (GLoVe CC), 2: Bi-LSTM (AraBERT-base), 3: Bi-LSTM (CamelBERT), 4: BFTC (AraBERT-base), 5: BFTC (CamelBERT) (USER: UserName, GEND: Gender, PRO: Profession, PH: Phone, PAT: PATIENT, ORG: Organization, MR: Medical Record, HOSP: Hospital, DOC: Doctor, CNTR: Country).

7.3 Privacy and Ethical Considerations

For patients' privacy and to comply with laws and regulations, de-identification of medical data is necessary. However, limitations may exist even in the case of state-of-the-art de-identification pipelines. It should be noted that the automated de-identification process does not guarantee absolute confidentiality, and manual verification by healthcare professionals may still be needed to ensure the correct erasure of sensitive information. Careful consideration has to be given to the ethical consequences of false positives and false negatives in de-identification.

7.4 Performance Evaluation Metrics

The metrics of precision, recall, and F1 score are widely applied for evaluating NER model's performance, but they may lack a full understanding of the actual world impact of false positives and false negatives in healthcare contexts. In order to provide a more comprehensive assessment of model efficiency, it would be useful to develop domain-specific evaluation metrics that account for the criticality of different types of entities in medical documents.

References

Mohammed Abdelhay, Ammar Mohammed, and Hesham A Hefny. 2023. Deep learning for arabic healthcare: Medicalbot. *Social Network Analysis and Mining*, 13(1):71.

Raed Abdullah Alharbi. 2023. Adoption of electronic health records in saudi arabia hospitals: Knowledge and usage. *Journal of King Saud University - Science*, 35(2):102470.

Saad Alanazi. 2017. *A named entity recognition system applied to Arabic text in the medical domain*. Ph.D. thesis, Staffordshire University.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. Arabert: Transformer-based model for arabic language understanding.

Rosario Catelli, Francesco Gargiulo, Valentina Casola, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2020. Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set. *Applied Soft Computing*, 97:106779.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Jaafar Hammoud, Natalia Dobrenko, and Natalia Gusarova. 2020. Named entity recognition and information extraction for arabic medical text. In *Multi Conference on Computer Science and Information Systems, MCCSIS*, pages 121–127.

Jaafar Hammoud, Aleksandra Vatian, Natalia Dobrenko, Nikolai Vedernikov, Anatoly Shalyto, and Natalia Gusarova. 2021. New arabic medical dataset for diseases classification. In *Intelligent Data Engineering and Automated Learning–IDEAL 2021: 22nd International Conference, IDEAL 2021, Manchester, UK,*

November 25–27, 2021, Proceedings 22, pages 196–203. Springer.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models.

Veysel Kocaman and David Talby. 2021a. Biomedical named entity recognition at scale. In *International Conference on Pattern Recognition*, pages 635–646. Springer.

Veysel Kocaman and David Talby. 2021b. Spark nlp: natural language understanding at scale. *Software Impacts*, 8:100058.

Veysel Kocaman and David Talby. 2022. Accurate clinical and biomedical named entity recognition at scale. *Software Impacts*, 13:100373.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *IberLEF@ SEPLN*, pages 618–638.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Hamada Nayel, Nourhan Marzouk, and Ahmed Elsawy. 2023. Named entity recognition for arabic medical texts using deep learning models. In *2023 Intelligent Methods, Systems, and Applications (IMSA)*, pages 281–285. IEEE.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Doaa Samy, Antonio Moreno-Sandoval, Conchi Bueno-Díaz, Marta Garrote Salazar, and José María Guirao. 2012. Medical term extraction in an arabic medical corpus. In *LREC*, pages 640–645.

Stefan Schweter and Sajawel Ahmed. 2019. Deep-eos: General-purpose neural networks for sentence boundary detection. In *KONVENS*.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.