AdabNer: Arabic Digital Archive Books with Nested Entity Recognition

Aya Mourad[⋆] **Mustafa Jarrar**[1,2]

[⋆] Sorbonne Université, ISIR, Paris, France
[1] Hamad Bin Khalifa University, Qatar
[2] Birzeit University, Palestine

[⋆] **Corresponding:** aya.mourad@sorbonne-universite.fr

Abstract

Most studies on Arabic Named Entity Recognition (NER) have focused on news texts and social media posts, while the large and rich corpus of literary Arabic books has been underrepresented. We introduce AdabNer, the first large-scale nested NER dataset for Modern Standard Arabic (MSA) literary texts, comprising the first 6,000 words annotated from each of 138 books spanning ten literary genres, including history, biography, literary criticism, and travel literature, and covering works from the 1880s to the 2020s. The corpus comprises about 876K tokens, manually annotated using a nested 21 entity tag annotation scheme, yielding 78,530 entity mentions, 18.96% of which are nested. We fine-tuned five pre-trained Arabic BERT encoders in two settings: stratified and leave-book-out, achieving F_1 scores of 0.86 and 0.83 with AraBERTv2, respectively. We also evaluated five large language models through few-shot in-context learning, including open-source models and the closed-source Gemini 3 Pro, with Gemini 3 Pro achieving the highest LLM F_1 score of 0.59. Supervised results degraded under out-of-domain evaluation; however, joint multi-domain training reduced this gap to less than a 1% F_1 loss, demonstrating that domain-diverse training data is key to robust Arabic NER, though broader validation beyond the experiments reported is needed. AdabNer and its annotation guidelines are publicly available.[1]

1 Introduction

Natural language processing (NLP) involves several tasks such as speech recognition, text generation, sentiment analysis, and information extraction, among many others (Khurana et al., 2023; Singh, 2018). Named Entity Recognition (NER) is a subtask of information extraction (Li et al., 2020; Yadav and Bethard, 2018) that classifies entities into predefined categories like person names (PERS), geopolitical entities (GPE), organizations (ORG), location (LOC), monetary values (MONEY), percentages (PERCENT), etc. Such classification is helpful in question answering (Mollá et al., 2006), machine translation (Babych and Hartley, 2003), and text summarization (Mimoto et al., 2025).

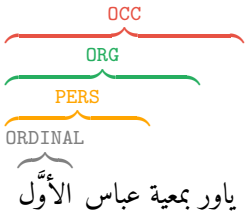
While a wide literature employs NER, its application has mainly focused on flat entity structures; however, real-world language contains nested entities, in which an entity mention can contain other entity instances (Ji et al., 2025). Additionally, NER has been explored extensively in high-resource languages (Krasadakis et al., 2024), such as English, while it remains not fully covered for Arabic (Qu et al., 2023). Compared to English NER, Arabic has its own difficulties (Shaalan, 2014), as it has a rich morphology that complicates entity boundary detection, and it lacks capitalization, which can help in other languages with identifying NER classes (Albahli, 2025). Despite growing research in Arabic NLP, available NER resources remain limited, and existing datasets such as AnerCorp (Benajiba et al., 2007) and Aqmar (Mohit et al., 2012) focus on flat NER derived from news and Wikipedia text. Although ACE2005 (Walker et al., 2006) was the first to introduce nested annotation in Arabic, it was developed on a small-scale news corpus. More recently, Wojood (Jarrar et al., 2022) was introduced as a large resource compiled from web-based reports, news, and social media. As such, models trained on these resources struggle to generalize to other domains (Liu et al., 2021), particularly literary books, which are rich in descriptive language and complex sentence structures (Bamman, 2020).

Arabic literary books, written in Modern Standard Arabic (MSA), including historiographies, biographies, travelogues, and geographies, employ their own distinctive linguistic and vocabulary. These books include historical groups of people, administrative structures, and political entities from different eras, ranging from the ancient

[1] https://doi.org/10.5281/zenodo.19468385

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 33382–33396

Such as the dialects of Tamim, Rabi'a, Mudar, Qays, Hudhayl, and Quda'a.



Aide-de-camp in the retinue of Abbas I.

Figure 1: Entity annotation structures: Example (a) flat entities (NORP), Example (b) nested entities (OCC, ORG, PERS, ORDINAL).

cities, e.g., Ptolemais Hermiou and Thebae, which existed during the Roman period, to old empires such as the Byzantine Empire. Moreover, literary books contain a complex structure of nested entities. Figure 1 presents an illustration of both cases in our corpus. Example (a) shows an instance of historical Arab tribes, resulting in a sequence of flat NORP entities. Example (b) illustrates a deeply nested structure: the full span denotes an occupation (OCC), nested with (ORG), person name (PERS), and ORDINAL within a single entity mention, a common attribute in literary Arabic, where a single entity represents a combination of different entity tags. Beyond the structural complexity of the entities span, philosophical, literary, and literary-critical writings often refer to authors, literary works, and scholars within the same sentence without prior context, making them more difficult to detect as entities. These challenges undercut the exploration of Arabic literary texts, but remain an essential domain for advancing Arabic NER research.

To address this gap, we introduce AdabNER, the first large-scale nested NER corpus derived from Arabic literary works. Our corpus incorporates the first 6,000 annotated words from each of 138 books spanning over 15 decades (1880s–2020s) across ten genres, including biographies, geography, history, literary criticism, literature, novels, philosophy, politics, social science, and travel literature. We construct AdabNER, a corpus of 876k tokens annotated with 21 entity types, using the Wojood nested annotation framework as a foundation, restoring the OntoNotes WORK_OF_ART and NORP categories to better capture literary works and introducing domain-specific adaptations to several category definitions. These refinements form the AdabNer annotation guidelines tailored to the literary domain. Our contribution can be summarized as follows:

- We built the first large-scale MSA Arabic NER dataset based on literary books to date, covering a wide scope of genres with complex nested structures.
- We fine-tuned five BERT-based models and employed in-context learning with 5 shots using five different multilingual large language models. Our experiments were conducted on two types of splits: leave-book-out and stratified splits of the books. The results show that AraBERTv2 achieved the highest micro-F_1 on both types of splits, with scores of 0.83 and 0.86, respectively.
- We performed out-of-domain evaluation against Wojood corpus and demonstrated that joint training on Wojood and AdabNer closes the domain gap to less than 1% micro-F_1 loss, indicating that robust nested Arabic NER requires multi-domain training.

The remaining sections of the paper are organized as follows: In Section 2 we overview the related work, Section 3 we present AdabNer corpus, the annotation process and challenges, the inter-annotator agreement metrics, and the corpus statistics. In Section 4, we present the experimental methodology of the fine-tuned NER models with the in-context learning along with the results and analysis. In Section 5, we present and analyze the out-of-domain and joint training performance, and we conclude in Section 6.

2 Related Work

2.1 MSA NER Corpora

In this section, we summarize the MSA NER corpora. Mostly, these corpora are relatively small and focused on news sources with only a few entity types. For example, OntoNotes 5 (Weischedel et al., 2013) is one of these few datasets, which contains a comprehensive flat NER with approximately 300K tokens drawn from MSA news media

sources and annotated with 18 entity types. One other example is ANERCorp (Benajiba et al., 2007), which consists of 150k tokens compiled from MSA news media sources and includes annotations for only four entity types: person, organization, location, and miscellaneous. On the other hand, a group of studies focused on domains other than news, such as the classical Arabic CANERCorpus (Salah and Zakaria, 2018), which comprises 258k tokens from Hadith documents, and Aqmar (Mohit et al., 2012), which comprises 47k tokens from Arabic Wikipedia, annotated with open-ended entity types across four domains: history, technology, science, and sports. While nested NER has been vastly explored in English, it still lags in Arabic. To illustrate, only two existing datasets include nested spans, such as the Multilingual ACE2005 corpus (Walker et al., 2006), which contains 30k mentions, five coarse-grained entity types (PERS, GPE, LOC, ORG, FAC), and 35 fine-grained subtypes. However, ACE2005 is expensive and limited to a single domain of media articles collected 20 years ago. The second source is the Wojood corpus (Jarrar et al., 2022), which represents an Arabic nested NER, with 550k tokens covering both MSA and dialects. It is annotated with 21 types, adapting 17 OntoNotes classes with four additional types Occupation (OCC), WEBSITE, UNIT, and Currency (CURR). Wojood corpus consists of sentences retrieved from online articles, Birzeit University digital archive, and dialectal text from Palestinian and Lebanese social media. Wojood was later extended to $\text{Wojood}_{\text{Fine}}$ (Liqreina et al., 2023), introducing 30 fine-grained sub-entity types, and to news specialized $\text{Wojood}_{\text{Gaza}}$ (Jarrar et al., 2024), and further complemented by Konooz (Hamad et al., 2025) which consists of 777K tokens covering 10 domains in 16 Arabic dialects; however, these resources remain confined to web-based reports, news, social media, and dialectal text.

2.2 Literary Book NER

Considerable progress in named entity recognition for different languages has been accomplished; however, literary NER in Arabic is surprisingly nonexistent in both fiction and non-fiction contexts. NER for fictional literary texts in high-resource languages has attracted researchers in recent years. One of the earliest works that inspired this research is by Bamman et al. (Bamman et al., 2019). In this work, the authors developed LitBank in English, with 210,532 annotated tokens across 100 novels, providing 13,912 annotations for six ACE entity categories, with 13.8% nested entities. Studies in European languages demonstrated equal accomplishment. In literary German-language processing, the Deutsches Roman Corpus (DROC) (Krug et al., 2017) contains 90 samples of German novels written from the 17th to the 20th centuries with over 50,000 manually labeled character mentions and their corresponding coreferences. For literary work in French, BookNLP-fr (Mélanie-Becquet et al., 2024) extended the BookNLP project to French and labeled the same set of NER categories, with an additional TIME tag in BookNLP's ACE schema. Additionally, European languages are represented by the ELTeC corpus (Frontini et al., 2020), which provides seven entities annotations across 100 novels per language for French, Portuguese, and Dutch, their work is complemented by PPORTAL_ner's (Silva and Moro, 2024) 25 Portuguese works and KIND's (Paccosi and Aprosio, 2022) 86 Italian fiction chapters. On the other hand, book-length NER in non-fiction books represents an even larger chasm across all languages. Chinese historical texts are well covered in CHisIEC (Tang et al., 2024), and Korean cultural heritage is comprehensively annotated in KoCHET (Kim et al., 2022). All other major languages, such as Arabic, lack non-fiction book-length NER resources.

Among Arabic NER datasets, most available corpora have focused on newswire texts and Wikipedia, leaving the rich tradition of Arabic literary books completely unrepresented. Moreover, this includes not only novels, poetry, and dramatic texts but also a whole body of non-fictional literature, including classical historiography, biographical lexicography, travelogues, and philosophical tracts. To fill this gap, we present the first Arabic nested NER dataset based on books, containing the first 6,000 words annotated from each of 138 books spanning 15 decades (1880s-2020s) across 10 genres.

3 AdabNer Corpus

We built our corpus from MSA non-fiction works from the Hindawi Library[2], the largest freely accessible Arabic digital library. It contains over 3,000 books published over a period of more than 150 years. To ensure the diversity of non-fiction discourse, we selected books from ten genres, including history and biography, to cover historical events and figures, as well as geography and travel liter-

[2] https://www.hindawi.org/

Genre	1880s	1890s	1900s	1910s	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s	2010s	2020s	Total
Biographies	–	417	278	597	496	284	781	806	767	719	689	1033	509	608	511	8495
Geography	–	776	–	512	928	–	1221	395	1611	364	–	705	495	707	445	8159
History	1065	829	818	1332	641	1265	1120	821	1047	794	549	613	682	915	726	13217
Literary Criticism	–	745	382	455	886	1377	688	339	263	235	862	898	560	546	688	8924
Literature	381	696	748	256	278	388	174	903	501	620	303	778	616	669	289	7600
Novels	251	176	400	650	429	265	431	252	318	262	252	640	236	555	362	5479
Philosophy	388	–	306	111	724	555	547	453	563	295	178	924	186	165	699	6094
Politics	–	–	–	–	294	587	517	618	473	723	583	521	504	727	–	5547
Social Sciences	583	464	468	211	438	849	537	134	504	481	416	330	309	455	287	6466
Travel Literature	784	374	330	706	566	873	776	522	377	354	786	555	298	694	554	8549
Total	3452	4477	3730	4830	5680	6443	6792	5243	6424	4847	4618	6997	4395	6041	4561	**78530**

Table 1: AdabNer entity counts by genre and decade.

ature, which describe geographical locations and narrate authors' travels. For intellectual discourse, we selected philosophy, literary criticism, and literature, which feature Arabic literary essays and critical discussions. For analytical discourse, we selected politics and the social sciences to explore the social and political aspects of Arab societies, as well as novels, which enrich the corpus with the narrative discourse and plot structures. However, we excluded Arabic poetry and fiction genres, such as children's fiction, science fiction, and detective fiction. These genres have different syntactic structures, metaphorical entity references, and flexible word order that would require specialized annotation guidelines beyond the scope of AdabNer. Within each genre, books were grouped by decade of publication from the 1880s to the 2020s, and one book from each genre in each decade was randomly selected, with the first 6,000 words of each book annotated after manually omitting the book summary and overview, starting from the first chapter of each book. We segmented each book using spaCy[3] and separated special punctuation markers used in literary books from the word tokens, assigning them as a new row with O tags. The resulting output was saved in CSV format, containing four columns: `book ID`, `sentence ID`, `token`, and `NER tag`.

Our annotation framework follows the Wojood nested annotation methodology. We adapt the 18 OntoNotes categories together with the three additional classes introduced by Wojood (`OCC`, `CURR`, and `UNIT`), and omit the `WEBSITE` tag due to its relative scarcity in our dataset, resulting in 21 categories in total. While using a model trained on Wojood as a starting point for annotation, we refined the tags definitions to better align with the literary domain. For `NORP`, we followed the OntoNotes definition, restricting the tag to proper nouns representing ethnic, religious, and political groups. Wojood defines `NORP` more broadly as any group of people, including occupational plurals (e.g., workers, kings, deputies). We excluded these common noun forms because literary narrative frequently uses them as general descriptions rather than references to specific named groups, and including them would introduce systematic annotation noise. For `WORK_OF_ART`, we restored it as a distinct category following OntoNotes rather than merging it with `PRODUCT` as in Wojood, in which book titles, poems, and artistic references are more appropriate than commercial products for reflecting literary works. For example, in literary criticism texts, `WORK_OF_ART` comprises 5.8% of entities compared to only 0.2% `PRODUCT`. A description of all classes, with examples, is provided in Table 9, and detailed class definitions and span conventions are presented in the AdabNER guidelines.

3.1 Annotation Process

Phase I: We started the annotation process by using a model finetuned on Wojood nested NER dataset, which provided us with initial NER annotation layer that served as a starting point for manual refinement.
Phase II: The annotation team composed of three annotators: the main author and two annotators holding master's degrees in computer science, each with prior experience in Arabic NLP. Annotators were trained according to established Wojood guidelines (Jarrar et al., 2022) and a set of updated, continuously refined annotation guidelines adapted to the literary domain and developed throughout the project. These refinements were guided by discussions and observations during weekly meetings to ensure a consistent interpretation of the 21 entity types, resulting in AdabNer guidelines.

[3] https://spacy.io

Entity Type (Total)			
PERS	15495	WORK_OF_ART	1927
GPE	13883	FAC	1794
NORP	7065	EVENT	1315
OCC	6499	LANGUAGE	1058
DATE	6495	UNIT	761
CARDINAL	5321	QUANTITY	758
LOC	5047	CURR	742
ORG	3759	MONEY	718
ORDINAL	2987	LAW	296
TIME	2143	PERCENT	247
		PRODUCT	220
Overall: 78530			

Table 2: AdabNer entity type distribution.

Phase III: We then trained a BERT-based multi-label model meant for predicting nested named entities (see Section 4), after the last round of manual annotation. The resulting trained model was then used to reannotate the corpus. The model predictions were presented to the annotators for systematic review. The annotators compared the model-generated labels against their original annotations. This step enabled identifying and correcting missing or incomplete annotations, thereby improving the overall quality and coverage of the corpus.

3.2 Annotation Challenges

We encountered several challenges during the annotation phase. First, the literary texts from the period of colonization, historically toward the end of the nineteenth century and into the mid-twentieth century, posed a unique difficulty. The main challenge originated from the entities' non-standard spatio-temporal nature during that period, which made it difficult for annotators to identify the appropriate entity tags for such references. For instance, annotators frequently had difficulty annotating regions such as the Trucial Coast (مشيخات ساحل الصلح) or the Aden Protectorate (محمية عدن) and choosing the correct entity type for such vague references, between geographic locations (LOC) and administrative proto-states (GPE). Annotators were asked to research whether such references should be labeled as LOC or GPE to accurately reflect the geopolitical realities of the nineteenth and twentieth centuries. The second source of complexity arises from the transliteration of non-Arab entities, which lack standardized spelling in Arabic and make it difficult to determine their meaning. Moreover, the Ottoman-era context poses distinct annotation challenges due to lexical and temporal complexity. Many terms used in books of this period originate from Ottoman

Entity Type	TP	FN	FP	κ	F_1
CARDINAL	396	10	21	0.974	0.962
CURR	54	1	1	0.985	0.982
DATE	484	32	70	0.963	0.905
EVENT	84	16	15	0.924	0.844
FAC	189	35	49	0.905	0.818
GPE	1105	42	59	0.965	0.956
LANGUAGE	172	5	11	0.961	0.956
LAW	8	1	1	0.988	0.889
LOC	518	57	76	0.948	0.886
MONEY	48	3	4	0.968	0.932
NORP	576	26	47	0.957	0.940
OCC	454	28	85	0.925	0.889
ORDINAL	277	5	5	0.983	0.982
ORG	216	34	57	0.920	0.826
PERCENT	1	0	0	1.000	1.000
PERS	1315	52	61	0.981	0.959
PRODUCT	10	0	1	0.990	0.952
QUANTITY	102	4	5	0.990	0.958
TIME	150	9	22	0.941	0.906
UNIT	102	3	1	0.986	0.981
WORK_OF_ART	117	11	21	0.928	0.880
Overall	**6378**	**374**	**612**	**0.938**	**0.923**

Table 3: Entity IAA metrics.

	Genre	TP	FN	FP	κ	F_1
	Biographies	799	40	59	0.962	0.942
	Geography	443	35	43	0.936	0.919
	History	1391	91	142	0.949	0.923
	Literary Criticism	625	25	49	0.951	0.944
	Literature	533	23	31	0.982	0.952
	Novels	406	49	60	0.901	0.881
	Philosophy	228	12	24	0.956	0.927
	Politics	585	25	62	0.947	0.931
	Social Science	382	22	45	0.924	0.919
	Travel Literature	986	52	97	0.939	0.930

Table 4: IAA metrics by genre.

Turkish and Persian administrative and military systems, which added another layer of difficulty for annotators when annotating titles such as Jokhdar (جوخدار), Janissary (الينكجرية), and Sipahi (إسباهية). These terms were difficult for the annotators, as they did not know the exact meanings of these words, making it essential to differentiate them from royal titles and classify them as occupations (OCC) as markers of socio-military status. Also, temporal annotation proved equally challenging due to the concurrent use of multiple dating systems. The books often contain Coptic years (e.g., للشهداء ١٦٦٠) and Coptic months (e.g., توت). Annotators were asked to identify these references as valid DATE expressions rather than skipping such labels, which reflect that period of time. Accordingly, we updated the annotation schemes to consistently represent multiple temporal frameworks.

3.3 Corpus Statistics

Our corpus comprises 26,162 sentences from 138 books, with an average of 189 sentences per book. The mean sentence length is 33.51 words, annotated with the 21 nested NER tags. Our corpus contains 78,530 entities, of which 14,082 are nested, yielding a global nesting rate of 18.96%. We provide detailed statistics for the corpus in the Table 1, grouped by genre and decade. The history genre has the most annotations and novels the fewest. Temporally, entity counts peak in the 1930s–1960s period. Missing values for some genre-decade combinations come from the fact that there are no corresponding books in the Hindawi library. We also present the detailed overall statistics of entity types covered across the corpus in Table 2.

3.4 Inter-Annotator Agreement

To assess the consistency of our annotations, we employed Cohen's Kappa (Cohen, 1960), a widely used metric for measuring inter-annotator agreement (IAA) (Hripcsak and Rothschild, 2005). For this purpose, we randomly sampled approximately 5% from each genre within each decade (a total of 83k tokens) and had them annotated by all annotators. Researchers do not recommend using solely kappa (κ) for evaluating named entity annotations (Campillos-Llanos et al., 2021) because of the unbalanced nature of the task, where the number of tokens tagged as "O" is much higher than other labels, which inflates κ scores and overestimates IAA. Instead, researchers recommend using the F_1 score, as it more appropriately reflects agreement in NER tasks. Nevertheless, in our research, we evaluate the IAA agreement using the κ score for each class and the F_1 score. The results presented in Table 3 show high agreement across all entity types, where at the individual entity type level, we have achieved an overall κ of 0.938 and an F_1 score of 0.923. The agreement scores per genre are shown in Table 4, which demonstrates very good overall agreement across all book genres. All genres have κ agreement above 0.92 except for novels, which achieve 0.901, most probably due to the frequent use of ambiguous references, which makes annotating this genre a challenging task for the annotators.

4 Experimental Methodology

4.1 Data Splitting

We split our dataset into three parts for training (80%), validation (10%), and testing (10%). We use two strategies: stratified-book split, in which we split the books so that each book is proportionally distributed across all sets, and leave-book-out split, in which each set (train, test, and val) contains entirely disjoint books.

4.2 Multilabel Bert-based Model for Nested NER

We solved nested NER as a multilabel classification task (Figure 2). Our model architecture consists of a BERT-based encoder followed by a dropout layer (p=0.1) and a fully connected classifier with sigmoid activation. We use a sigmoid function to estimate independent probabilities for each BIO label across all entity types, enabling multilabel tagging in which tokens can receive multiple entity annotations, yielding a label space of 43 labels: for each of the 21 entity types, a `B-` and `I-` tag, plus a single `O` tag. This single-head design enables cross-entity feature sharing, allowing the model to learn correlations between co-occurring entity types, for instance, the relationship between `OCC` and `ORG` in nested occupational names. The multilabel formulation naturally handles nested entities, the majority of our 18.96% nesting rate, and allows tokens predictions from multiple entity types, however the same type nesting layers presents a special challenge. When nested entities share the same type (e.g., `ORG` within `ORG`), the model cannot distinguish the boundaries using standard `BIO` labels alone. Thus, we apply a layering approach where we append numerical indices to only the repeated layered labels within the same token position (e.g., `B-ORG`, $\texttt{B-ORG}_2$), treating each suffix as an additional class in the same classifier. Note that layered entities represent less than 1% of total entities (Table 8), with a maximum depth of 3 of only two instances.

We then conducted experiments and evaluated several Arabic BERT-based models to compare multiple Arabic BERT variants, including AraBERTv1 and AraBERTv2 (Antoun et al., 2020), CAMeLBERT-MSA (Inoue et al., 2021), and ARBERT and ARBERTv2 (Abdul-Mageed et al., 2021) to identify the most effective encoder for literary Arabic NER. We fine-tuned all models for 50 epochs each, with early stopping based on the validation macro-F_1 score, using a patience of 5 epochs. We use the AdamW optimizer (Loshchilov and Hutter, 2017), an exponential learning rate scheduler, focal loss (Lin et al., 2017) with $\alpha = 0.75$ and $\gamma = 1.0$ to handle class imbalance, and a dropout

Model	Leave-Book-Out				Stratified-Book			
	P	R	F_1	Macro F_1	P	R	F_1	Macro F_1
BERT-based								
AraBERTv1	0.77±0.010	0.79±0.004	0.78±0.004	0.74±0.022	0.80±0.007	0.82±0.003	0.81±0.005	0.78±0.011
AraBERTv2	**0.82**±0.014	**0.85**±0.004	**0.83**±0.006	**0.81**±0.007	**0.85**±0.003	**0.87**±0.003	**0.86**±0.003	**0.83**±0.006
CAMeLBERT	0.77±0.005	0.77±0.003	0.77±0.002	0.74±0.002	0.79±0.006	0.81±0.006	0.80±0.006	0.78±0.008
ArBERT	0.78±0.010	0.81±0.001	0.79±0.004	0.77±0.009	0.76±0.011	0.80±0.006	0.79±0.008	0.75±0.015
ARBERTv2	0.79±0.019	0.83±0.011	0.81±0.005	0.78±0.006	0.82±0.007	0.85±0.002	0.83±0.004	0.81±0.006
In-Context Learning								
gemini-3-pro	0.55±0.009	0.64±0.020	0.59±0.007	0.57±0.005	0.59±0.007	0.49±0.040	0.53±0.026	0.51±0.023
Qwen3-235B	0.30±0.012	0.53±0.003	0.38±0.009	0.35±0.021	0.32±0.009	0.50±0.008	0.39±0.007	0.35±0.012
Qwen2.5-72B-Instruct	0.30±0.020	0.53±0.004	0.38±0.017	0.35±0.026	0.29±0.012	0.46±0.004	0.35±0.010	0.33±0.015
aya-expanse-32b	0.29±0.020	0.48±0.011	0.36±0.017	0.30±0.034	0.32±0.013	0.44±0.013	0.37±0.006	0.32±0.009
c4ai-command-r-v01	0.22±0.026	0.43±0.007	0.29±0.025	0.25±0.020	0.22±0.008	0.38±0.007	0.27±0.004	0.24±0.003

Table 5: Evaluation results with micro precision (P), micro recall (R), micro F_1, and Macro F_1. Results are reported as mean across 3 different runs with three different seeds, mean±std.

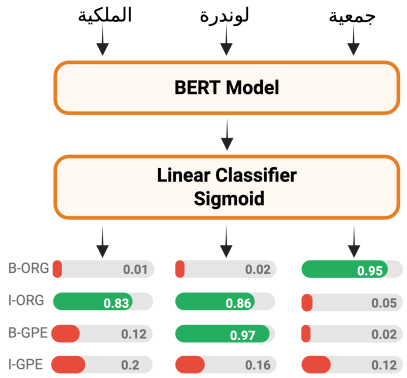


Figure 2: Multi-label token classification architecture for nested NER. BERT Model refers to one of five pretrained models we are using. The sigmoid produces independent probabilities for each label to predict the overlapping entities at the token level.[4]

of 0.1. We set the maximum sequence length to 512, the batch size to 16, and the learning rate to $\eta = 6e{-}5$. At inference, we set the probability threshold to 0.5. All models are implemented using PyTorch[5] with the HuggingFace Transformers library[6]. In general, all models converged around epoch 25 trained on 3 NVIDIA RTX A6000 GPUs, each with 48 GB of memory.

4.3 In-context Learning

To evaluate the inherent capabilities of large language models (LLMs) for Arabic nested named entity recognition, we conducted few-shot in-context learning with our 21 entity categories on five multilingual large models spanning both open-source and closed-source families: Qwen 2.5-72B-Instruct (Team, 2024) and Qwen3-235B-A22B-Instruct is included to assess the impact of model scale on Arabic NER performance under few-shot prompting, c4ai-command-r-v01 [7] which focuses on multilingual reasoning , and aya-expanse-32b model (Dang et al., 2024) represents a moderate-scale (32B) multilingual model designed for broad language coverage and gemini-3-pro (Google, 2026) to evaluate closed-source model capabilities on Arabic literary NER. These LLMs were selected for their generalization performance across multiple languages, their support for the Arabic language in particular, and their efficiency in structured predictions via prompt-based instructions. For open-source models, we run our tests using the vLLM inference engine (Kwon et al., 2023) with tensor parallelism (TP = 8) on NVIDIA A100-80GB GPUs with 90% memory usage and a temperature of 0. Each prompt includes the system prompt shown in the Figure 3, the entity list description provided in the Table 9, and user prompt with the five shots (k = 5) randomly selected via stratified sampling from the training set to maximize entity-type coverage by iterating over a shuffled set of entity categories. We instruct the model to produce entity annotations in JSON format containing the text and its corresponding type.

4.4 Experimental Setup

For each model, both BERT-based and in-context learning, we report the average over three runs, each using a different seed. We report the mean micro

[5] https://pytorch.org/

[6] https://huggingface.co/transformers/

[6] In this example, جمعية (Association) is tagged as B-ORG (0.95) and لوندرة (London) as I-ORG (0.86) and B-GPE (0.97), capturing the nested structure where Royal London Association (جمعية لوندرة الملكية) is an organization located in London (لوندرة).

[7] https://huggingface.co/CohereLabs/c4ai-command-r-v01

System Prompt

You are an expert in Named Entity Recognition (NER) systems for Arabic text. Your task is to identify and extract all named entities from the given Arabic text. This is a Nested NER task. A single token can belong to multiple entities simultaneously.
Entity Types: *{entity_list}*
Output Format:
- Return a JSON array of entities
- Each entity must have "type" and "text" fields
- If no entities found, return an empty array: []

Figure 3: System prompt template for the in-context learning.

precision, recall, and F_1, and the macro F_1, along with their standard deviations across the three runs, on the test set for each model under the two splitting strategies. We apply in-context learning to both strategies to ensure methodological parity by evaluating on the same test sets. To evaluate entity-level detection, we use exact span matching, in which an entity prediction is counted as correct only when both the entity boundaries and the entity type exactly match the ground truth. The source code of the experiments is available on Github[8].

4.5 Results and Discussion

Table 5 presents the evaluation of the nested Arabic NER for both the supervised BERT-based encoders and in-context learning using LLMs. Apparently, BERT-based models perform exceptionally well compared to in-context learning experiments, with AraBERTv2 achieving the strongest metrics (highlighted in bold). Interestingly, all BERT-based models showed improved results when using the stratified-book split, indicating that this type of splitting allows the models to learn the authors' style and vocabulary use within the same book. In comparison, under the leave-book-out setting, all BERT-based models still perform well with only a slight deterioration, e.g., around 2 to 3 percent drop for the AraBERTv2 model, suggesting sizeable cross-book generalization (see Section A.2 for detailed analysis).

At the extreme end, in-context learning using open-source models demonstrates systematic underperformance relative to the BERT-based models, with Qwen3-235B-A22B-Instruct, the largest-scale model with 235B parameters, achieving the best micro F_1, followed by aya-expanse-32b on the stratified split. Qwen3-235B, despite being substantially larger than the 72B variant, shows improvements on the stratified set and similar performance on leave-book-out. The closed-source model results, on the other hand, represented by Gemini 3 Pro outperform all open-source models with an F_1 score of 0.59 on the stratified split, yet still fall short of the fine-tuned AraBERTv2 baseline, which achieves an F_1 score of 0.86. This underperformance is expected due to the following limitations: LLMs are better in generative tasks rather than token level NER classification task (Lu et al., 2025), the complexity of the nested task and 21 categories, making it difficult for the in-context learning to learn these classes (Kim et al., 2024), the literary domain of AdabNer, which is likely underrepresented in the LLM pretrained datasets, and the only five-shot prompting setup, which seems to be insufficient and constraining the in-context learning model's ability to generalize (Xiao et al., 2025). Finally, we note that the in-context learning models exhibit higher recall than precision. This indicates that LLMs tend to over-predict entities, including false positives, and fail to accurately detect the entity boundary within the nested annotation structure.

[8] https://github.com/aiamourad/AdabNER

Train · Test	P	R	F_1	Macro F_1
In-Domain				
AdabNer · AdabNer	0.85	0.88	0.86	0.84
Wojood · Wojood	0.92	0.93	0.92	0.83
Out-of-Domain				
AdabNer · Wojood	0.66	0.66	0.66	0.55
Wojood · AdabNer	0.61	0.58	0.59	0.45

Table 6: Out-of-domain evaluation between AdabNer and Wojood.

5 Out-of-Domain Evaluation

In this section, we discuss the out-of-domain evaluation between our corpus AdabNer and Wojood. Aqmar (Mohit et al., 2012) and ANERCorp (Benajiba et al., 2007) are the most popular benchmarks in MSA NER; however, performing a direct comparison with AdabNer is challenging due to differences in annotation schemes. Both datasets adopt flat NER labeling and differ in NER tag types. ANERCorp is limited to only four types, and Aqmar includes open-ended entity tags. A fair comparison would require either simplifying AdabNer by removing its nested structure and reducing the sets to match Aqmar and ANERCorp, or extending both data sources with more complex annota-

tion schemes. However, both approaches would introduce inconsistencies by either removing AdabNer's main contribution, which is nested NER, or introducing inconsistencies and annotation noise. Thus, the out-of-domain evaluation will be limited to Wojood, which shares 20 nested named entity types with AdabNer.

We first mapped AdabNer entity types to the Wojood tags by dropping the WEBSITE tag from Wojood and merging WORK_OF_ART and PRODUCT entities into a single PRODUCT entity type. Then, we selected the best-performing model on AdabNer, AraBERTv2, and retrained it on the stratified split using a new random seed. For Wojood, we trained the same model architecture and evaluated its performance on the Wojood nested dataset. In Table 6, we show the evaluation results for the out-of-domain, where the model trained on AdabNer was tested on Wojood, and vice versa. We also include the in-domain evaluation results for comparison. As expected, in-domain performance results in better metrics than out-of-domain. Similar domain degradation is reported by (Hamad et al., 2025). We quantify the domain shift by employing the Maximum Mean Discrepancy (MMD) analysis (Table 12) at the entity level. The highest distributional discrepancies are observed for LAW (0.445) and PERCENT (0.480), followed by EVENT (0.299), ORG (0.282), and LANGUAGE (0.278). At the other end, TIME (0.152) and PERS (0.161) exhibit the lowest shift. These MMD scores strongly align with the entity-level transfer performance reported in Table 13.

The performance degradation is mainly due to differences in the lexical and temporal scales of the corpora being tested. AdabNer is derived from literary works, where entities peak in books written in the 1930s–1960s (Table 1), and includes ancient units of measurement, currencies, and colonial occupational titles. These entities have minimal lexical overlap with Wojood's news and social media content, and thus, show an asymmetric transfer behavior between both domains, as clearly observed in MONEY and OCC. Moreover, the use of variation in writing style in both corpora increases the domain disparity. The literary book entities detection varies with the author's style of writing, specifically for context-dependent entities (Table 11), whereas news text follows a uniform and standard pattern. The poor performance of contextual entities such as EVENT, FAC, and LOC, and the weak transfer of numerical entities, reflects this fact. Moreover, dialectical inclusion in Wojood introduces informal entity mentions and code switching words that are absent from AdabNer, and thus widens the gap between the two domains. A detailed analysis is provided in Section A.4.

Test	P	R	F_1	Macro F_1
Joint Training				
AdabNer	0.847±.008	0.859±.004	0.853±.006	0.839±.005
Wojood	0.916±.002	0.927±.002	0.921±.002	0.841±.006

Table 7: Joint multi-domain training using AraBERTv2 performed with AdabNer and Wojood training sets.

To evaluate whether it is possible to close the gap between the AdabNer and Wojood with multi-domain training, we performed joint training across 3 different seeds using AraBERTv2 on the combined datasets and evaluated it on the test set of each dataset. The results, as shown in Table 7, demonstrate that it is, in fact, possible to close the domain gap with a loss of less than 1% in F_1 score compared to the in-domain results for both datasets. These results, while limited to a single out-of-domain pair, provide initial evidence that multi-domain training can close the domain gap in nested Arabic NER.

6 Conclusion

We introduce AdabNer, the first Arabic NER corpus that fully benefits from existing MSA literary works. AdabNer consists of 138 books spanning 15 decades with diverse genres. The corpus comprises about 876K tokens, annotated with 21 entity types using a nested annotation scheme. Evaluations with fine-tuned Arabic BERT models and in-context learning under stratified and leave-book-out settings reveal exceptional performance of the BERT-based models, especially for AraBERTv2. However, these results degraded under out-of-domain evaluation against Wojood, the only available MSA heavy nested NER benchmark, but joint training on AdabNer and Wojood closes this gap to less than 1% F_1 loss on both datasets, which highlights the need for domain-specific or multi-domain training to achieve robust Arabic NER. In the future, we plan to extend AdabNer to include additional genres, specifically fictions, perform entity linking across books and knowledge base graphs, and annotate coreference chains within AdabNer. We also plan to study advanced prompting strategies, such as chain-of-thought and iterative refinement with human-in-the-loop, and fine-tune open-source LLMs on our AdabNer corpus.

7 Ethical Consideration

Books under the Hindawi license have been authorized by the Hindawi library for academic and non-commercial research purposes only.

8 Limitations

A number of considerations related to limitations and ethics are relevant to our work:

- Our models perform nested named entity recognition in literary books and can be applied to information extraction tasks; however, despite being trained on diverse books across genres, they perform poorly on the news source.
- Although we report high agreement scores between annotators, the annotation schema itself may need to be modified when adapting to different domains.
- WOJOOD is the only available MSA nested dataset, and the analysis of the out-of-domain evaluation is thus limited.
- Our baseline model uses a layering strategy to handle same-type nested entities with shallow nesting depth (<1% of entities) in our corpus. We intend to investigate span-based or hypergraph models for deeper nesting in future work.

Acknowledgments

Aya Mourad would like to express her deep gratitude to her advisor, Prof. Abdenour Hadid, for his invaluable supervision, guidance, and support throughout this research project.

This project is co-funded by the European Union's Horizon Europe research and innovation programme Cofund SOUND.AI under the Marie Skłodowska-Curie Grant Agreement No 101168090.

We gratefully acknowledge the support of SCAI (Sorbonne Center for Artificial Intelligence). This work was performed using HPC resources from GENCI-IDRIS (Grant 2025-AD011016170). It was additionally supported by the Google Cloud Research Credits program (Award GCP19980904), which provided the computational resources for the Gemini 3 Pro and Qwen3-235B experiments. We acknowledge Christophe Khalil for his role as primary annotator and for his insights in refining the annotation guidelines. We would like to thank Shaymaa for introducing the initial WOJOOD annotation guidelines at the early phase of the project.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and 1 others. 2021. 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)*, pages 7088–7105.

Saleh Albahli. 2025. An advanced natural language processing framework for arabic named entity recognition: A novel approach to handling morphological richness and nested entities. *Applied Sciences*, 15(6):3073.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.

David Bamman. 2020. Litbank: Born-literary natural language processing. *Computational Humanites, Debates in Digital Humanities (2020, preprint)*.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144.

Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 143–153. Springer.

Leonardo Campillos-Llanos, Ana Valverde-Mateos, Adrián Capllonch-Carrión, and Antonio Moreno-Sandoval. 2021. A clinical trials corpus annotated with umls entities to enhance the access to evidence-based medicine. *BMC medical informatics and decision making*, 21(1):69.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and 1 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Niama El Khbir, Urchade Zaratiana, Nadi Tomeh, and Thierry Charnois. 2023. Lipn at wojoodner shared task: A span-based approach for flat and nested arabic named entity recognition. In *Proceedings of ArabicNLP 2023*, pages 789–796.

Francesca Frontini, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. Named entity recognition for distant reading in eltec. In *CLARIN Annual Conference 2020*.

Google. 2026. Gemini 3 pro preview. Large language model. Accessed: 2026-03-04.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773.

Nagham Hamad, Mohammed Khalilia, and Mustafa Jarrar. 2025. Konooz: Multi-domain multi-dialect corpus for named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7316–7331.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

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. *arXiv preprint arXiv:2103.06678*.

Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, Nagham Hamad, and 1 others. 2023. Wojoodner 2023: The first arabic named entity recognition shared task. In *Proceedings of ArabicNLP 2023*, pages 748–758.

Mustafa Jarrar, Nagham Hamad, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Wojoodner 2024: The second arabic named entity recognition shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 847–857.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. Wojood: Nested arabic named entity corpus and recognition using bert. *arXiv preprint arXiv:2205.09651*.

Lixia Ji, Yiping Dang, Yunlong Du, Wenzhao Gao, and Han Zhang. 2025. Nested named entity recognition: A survey of latest research. *Expert Systems*, 42(7):e70052.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744.

Gyeongmin Kim, Jinsung Kim, Junyoung Son, and Heuiseok Lim. 2022. Kochet: A korean cultural heritage corpus for entity-related tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3496–3505.

Hongjin Kim, Jai-Eun Kim, and Harksoo Kim. 2024. Exploring nested named entity recognition with large language models: Methods, challenges, and insights. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8670.

Panteleimon Krasadakis, Evangelos Sakkopoulos, and Vassilios S Verykios. 2024. A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. *Electronics*, 13(3):648.

Markus Krug, Lukas Weimer, Isabella Reger, Luisa Macharowsky, Stephan Feldhaus, Frank Puppe, and Fotis Jannidis. 2017. Description of a corpus of character references in german novels-droc [deutsches roman corpus].

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Haneen Liqreina, Mustafa Jarrar, Mohammed Khalilia, Ahmed El-Shangiti, and Muhammad Abdul-Mageed. 2023. Arabic fine-grained entity recognition. In *Proceedings of ArabicNLP 2023*, pages 310–323.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13452–13460.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Qiuhao Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. 2025. Large language models struggle in token-level clinical named entity recognition. In *AMIA Annual Symposium Proceedings*, volume 2024, page 748.

Frédérique Mélanie-Becquet, Jean Barré, Olga Clément Seminck, Clément Plancq, Marco Naguib, Martial Pastor, and Thierry Poibeau. 2024. Booknlp-fr, the french versant of booknlp. a tailored pipeline for 19th and 20th century french literature. In *Conference on Computational Literary Studies (CCLS 2024)*.

Tomoaki Mimoto, Kentaro Kita, Yuta Gempei, Takamasa Isohara, Shinsaku Kiyomoto, and Toshiaki Tanaka. 2025. Cyber threat intelligence report summarization with named entity recognition. In *International Workshop on Security*, pages 468–485. Springer.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A Smith. 2012. Recall-oriented learning of named entities in arabic wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173.

Diego Mollá, Menno Van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian language technology workshop 2006*, pages 51–58.

Teresa Paccosi and Alessio Palmero Aprosio. 2022. Kind: an italian multi-domain dataset for named entity recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 501–507.

Xiaoye Qu, Yingjie Gu, Qingrong Xia, Zechang Li, Zhefeng Wang, and Baoxing Huai. 2023. A survey on arabic named entity recognition: Past, recent advances, and future trends. *IEEE Transactions on Knowledge and Data Engineering*, 36(3):943–959.

Ramzi Esmail Salah and Lailatul Qadri Binti Zakaria. 2018. Building the classical arabic named entity recognition corpus (canercorpus). In *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pages 1–8. IEEE.

Khaled Shaalan. 2014. A survey of arabic named entity recognition and classification. *Computational Linguistics*, 40(2):469–510.

Mariana O Silva and Mirella M Moro. 2024. Pportal_ner: An annotated corpus of portuguese literary entities. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12927–12937.

Sonit Singh. 2018. Natural language processing for information extraction. *arXiv preprint arXiv:1807.02383*.

Xuemei Tang, Zekun Deng, Qi Su, Hao Yang, and Jun Wang. 2024. Chisiec: an information extraction corpus for ancient chinese history. *arXiv preprint arXiv:2403.15088*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *(No Title)*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0. LDC2013T19, Web Download, Philadelphia: Linguistic Data Consortium.

Yuhui Xiao, Jianjian Zou, and Qun Yang. 2025. Advancing few-shot named entity recognition with large language model. *Applied Sciences*, 15(7):3838.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th international conference on computational linguistics*, pages 2145–2158.

A Appendix

A.1 Corpus

Layered Entity	Count	Percentage
$\texttt{LOC}_2$	164	0.21%
$\texttt{ORG}_2$	102	0.13%
$\texttt{OCC}_2$	35	0.04%
$\texttt{FAC}_2$	22	0.03%
$\texttt{LAW}_2$	7	0.01%
$\texttt{ORG}_3$	2	0.00%
$\texttt{EVENT}_2$	2	0.00%
Total Entities	**78,530**	**100%**

Table 8: Distribution of layered nested entity instances of the same class.

In Table 8, we show the statistics of layered entities of the same type, where nested LOC within LOC has the highest count, followed by ORG and OCC. The maximum depth of layered entities of the same type is three entities which occurs in only two instances and is therefore negligible. In Table 9 , we list all the entities classes annotated in this study with the description and arabic examples.

A.2 Model Performance Analysis

We evaluated our proposed architecture, using AraBERTv2 as the Bert-based model, on the Wojood corpus. When evaluated on the test set, our

Label	Description & Examples
CARDINAL	Numerals. واحد، اثنان، 100
CURR	Currency names or symbols. دولار، جنيه، يورو
DATE	Dates: days, months, years, ages. يوليو 1925، أمس
EVENT	Named events: battles, wars, sports. الربيع العربي، كأس العالم
FAC	Man-made structures: buildings, bridges, streets. برج القاهرة، الجامع الأزهر
GPE	Geopolitical entities: countries, cities. مصر، القاهرة، فرنسا
LANGUAGE	Any named language. العربية، الإنجليزية
LAW	Named legal documents. الدستور الفرنسي
LOC	Non-GPE locations: rivers, mountains, seas. النيل، البحر الأحمر، جبل سيناء
MONEY	Monetary values with unit. جنيه 50 دولار، 100
NORP	Nationalities, religious, and political groups. مصري، مسلم، جمهوريون
OCC	Professional titles. مهندس، طبيب، أستاذ
ORDINAL	First, second, etc. الأول، الثاني
ORG	Companies, agencies, institutions. الجزيرة، اليونسكو، جوجل
PERCENT	Percentage values including "%". 99%، خمسون بالمئة
PERS	Names of people. Excludes honorifics (Mr., Mrs., Dr.). محمد علي باشا، حسين بك
PRODUCT	Products: vehicles, foods, etc. آيفون، مرسيدس، ب٢٩
QUANTITY	All kinds of measurements. كم، 50 كجم، 100
TIME	Times smaller than 24 hours. 4:00مساءً، هذا الصباح
UNIT	Unit names or symbols. كيلومتر، كيلوغرام
WORK_OF_ART	Titles of books, artifacts, etc. مبادئ الأخلاق ، صولجان صلاح الدين

Table 9: Entity types with descriptions and examples

model achieves a micro F_1 score of 0.92, a precision of 0.92, and a recall of 0.93. According to the reported evaluation metrics for nested NER in the Wojood shared task (Jarrar et al., 2023), the proposed architecture's performance is comparable to that of the top-5 models (Table 10), performing similar to a span-based suggested model implemented by the LIPN team (El Khbir et al., 2023) with micro F_1 score of 0.92 and outperforms the metrics reported by (Jarrar et al., 2022), which treat the task as a multitask NER tagging problem. Our results demonstrate that the multilabel formulation achieves competitive performance while maintaining architectural simplicity. We adopt the multilabel formulation for its computational efficiency on long literary sentences and aim to explore biaffine and hypergraph models in the future.

Rank	Team	F1	P	R	Model Type
1	Elyadata	93.73	93.99	93.48	DiffusionNER
2	UM6P & UL	93.03	92.46	93.61	BERT-based Multi-task
3	AlexU-AIC	92.61	92.10	93.13	MRC + Seq. Labeling
4	LIPN	92.45	92.31	92.59	Span-based
5	Ours	92.30	91.83	92.77	Multilabel
6	Wojood	88.40	87.72	89.09	Multi-task

Table 10: Model performance comparison on Wojood-NER 2023 shared task.

Entity Type	Leave-book-out		Stratified		ΔF_1
	F_1	Sup.	F_1	Sup.	
Context-Dependent					
EVENT	0.72	131	0.71	207	+0.01
FAC	0.63	232	0.75	268	–0.12
LAW	0.69	47	0.65	45	+0.04
LOC	0.68	555	0.78	777	–0.10
ORG	0.75	576	0.77	640	–0.02
WORK_OF_ART	0.71	312	0.81	337	–0.10
Context-Independent					
CARDINAL	0.88	694	0.90	887	–0.02
CURR	0.97	128	0.92	91	+0.05
DATE	0.84	848	0.87	1041	–0.03
GPE	0.86	1446	0.90	2126	–0.04
LANGUAGE	0.91	94	0.89	172	+0.02
MONEY	0.93	123	0.91	92	+0.02
NORP	0.85	1091	0.84	1072	+0.01
OCC	0.77	941	0.79	1045	–0.02
ORDINAL	0.94	473	0.95	470	–0.01
PERS	0.88	2711	0.91	2371	–0.03
PERCENT	1.00	9	0.78	26	+0.22
QUANTITY	0.86	82	0.92	103	–0.06
TIME	0.87	333	0.90	319	–0.03
UNIT	0.91	83	0.91	105	0.00
Poor Performance (Both Splits)					
PRODUCT	0.45	27	0.55	36	–0.10
Micro Average	**0.83**	**10,961**	**0.86**	**12,272**	**–0.03**
Macro Average	**0.81**	–	**0.83**	–	**–0.02**

Table 11: AraBertv2 entity classification mean F_1 across 3 seeds grouped by context dependence across two splitting strategies: leave-book-out and stratified split.Sup. = support (number of gold entities in the test set).

Further, we disaggregate the evaluation results by entity type, e.g., context- and non-context-based classes. Table 11 shows that numerical- and formulaic-based entities (CARDINAL, ORDINAL, MONEY, CURR, UNIT, and TIME) demonstrated stable metrics with no remarkable changes between the leave-out-book and stratified book settings, mainly because these kinds of classes are not dependent on the context in which they appear with minimal ΔF_1 (typically ≤0.03), as these entities follow con-

sistent surface patterns regardless of authorial style. This could also be linked to the fact that the numerical entities had uniform distributions across the two settings. The 22% improvement in PERCENT under leave-book-out (F_1= 1.00 vs. 0.78) is a statistical artifact: the leave-book-out test set contains only 9 instances, and the model achieves perfect detection across all three seeds, whereas the stratified split with 26 instances yields higher variance. Similarly, the 6% drop in QUANTITY reflects lexical diversity in historical measurement units across different time periods. High-frequency entities such as PERS with 2711 instances and GPE with 1446 instances also remain relatively stable (Δ=0.03 and Δ= 0.04), benefiting from sufficient training signal across diverse books. In comparison, the context-dependent entities showed a decline in performance under the leave-out-book setting compared to the stratified-book; however, this drop was moderate. PRODUCT remains the weakest entity type in both settings (0.55 and 0.45). This shows that the dataset contains diverse contexts, enabling the model to generalize well to unseen books. At the same time, it may be valuable to further enrich the distribution of context-dependent entities with additional data to enhance the model's ability to generalize to new linguistic contexts. This will essentially be the focus of the future work.

A.3 MMD Analysis

To quantify the distributional divergence between the WOJOOD corpus and ADABNER, we employ Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), which compares the distribution of the dataset by first projecting the data into Reproducing Kernel Hilbert Space (RKHS) and then estimates the distance between two probability distributions in Hilbert space.

For each entity type, we extract contextual embeddings using AraBERTv2 (Antoun et al., 2020) and compute full-span representations by averaging subword embeddings across the entire entity mention. We then calculate the empirical MMD *per entity* using a Radial Basis Function (RBF) kernel. Given samples $X_e = \{x_1, \ldots, x_m\}$ from the literary domain ADABNER and $Y_e = \{y_1, \ldots, y_n\}$ from news domain WOJOOD for entity class e, MMD estimate is defined as follows:

$$\text{MMD}_e(X_e, Y_e) = \left\| \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \phi(y_j) \right\|_{\mathcal{H}} \quad (1)$$

Entity	MMD	Shift
PERCENT	0.480	High
LAW	0.445	High
EVENT	0.299	Moderate
ORG	0.282	Moderate
OCC	0.280	Moderate
LANGUAGE	0.278	Moderate
LOC	0.276	Moderate
CURR	0.256	Moderate
MONEY	0.254	Moderate
DATE	0.250	Moderate
GPE	0.245	Moderate
FAC	0.237	Moderate
PRODUCT	0.219	Moderate
NORP	0.204	Moderate
QUANTITY	0.193	Low
CARDINAL	0.187	Low
UNIT	0.179	Low
ORDINAL	0.164	Low
PERS	0.161	Low
TIME	0.152	Low
Mean	**0.252**	**Moderate**

Table 12: Domain shift analysis (MMD) between WOJOOD and ADABNER across entity types. Shift levels: High (> 0.35), Moderate (0.20–0.35), Low (< 0.20).

where $\phi : \mathcal{X} \rightarrow \mathcal{H}$ is the nonlinear projection to feature representation in RKHS, and $e \in \{\text{PERS}, \text{LOC}, \text{ORG}, \ldots\}$. We use the RBF kernel for the mapping.

A.4 Out-of-Domain Evaluation Analysis

In Table 13, we group the performance at the entity level based on transfer performance, and in Table 12, we report the MMD score per entity type. The entities showing strong transfer in both directions are CURR, GPE, and PERS. Although CURR and GPE show moderate MMD scores, they still perform well due to the class density in both corpora. However, when WOJOOD is transferred to ADABNER, performance degrades, likely due to the time span of ADABNER, which reflects historical context and lexical divergence, with patterns different from those found in news. Occupation class (OCC) (MMD score 0.28) shows the most degradation when WOJOOD is applied to ADABNER, which is likely due to occupations related to the colonization period, where lexically the occupation titles are derived from Turkish and Persian words; thus, WOJOOD performs poorly on this entity. The same applies to ORG, DATE, and UNIT, which show moderate MMD and perform poorly, likely due to lexical differences reflecting the 1930s and 1960s. However, ADABNER still performs better on these entities because it overlaps

Entity Type	AdabNer →Wojood	Wojood →AdabNer	Δ
Strong Transfer ($F_1 \geq 0.70$ ***in both directions)***			
CURR	0.88	0.73	-0.15
GPE	0.83	0.77	-0.06
PERS	0.83	0.72	-0.11
Moderate Transfer ($F_1 \geq 0.50$ ***in both directions)***			
CARDINAL	0.41	0.56	+0.15
DATE	0.58	0.62	+0.04
OCC	0.76	0.53	-0.23
ORDINAL	0.53	0.56	+0.03
ORG	0.68	0.51	-0.17
QUANTITY	0.57	0.59	+0.02
TIME	0.42	0.57	+0.15
UNIT	0.74	0.66	-0.08
Asymmetric Transfer			
EVENT	0.64	0.43	-0.21
MONEY	0.75	0.43	-0.32
PERCENT	0.36	0.88	+0.52
Weak Transfer ($F_1 < 0.50$ ***in both directions)***			
FAC	0.44	0.44	0.00
LANGUAGE	0.31	0.06	-0.25
LOC	0.44	0.47	+0.03
NORP	0.18	0.44	+0.26
Failed Transfer ($F_1 = 0$ ***in one direction)***			
LAW	0.33	0.00	-0.33
PRODUCT	0.32	0.00	-0.32
Not Evaluated			
WEBSITE	–	–	–
Micro Average	**0.66**	**0.59**	**-0.07**

Table 13: Out-of-domain F_1 scores for AraBERTv2. AdabNer→Wojood: trained on AdabNer, evaluated on Wojood; Wojood→AdabNer: vice versa. Δ = (Wojood→AdabNer) − (AdabNer→Wojood).

to some extent with the Wojood corpus, which includes books from the 2000s to the 2020s. For weak and asymmetric transfer, context-dependent entities such as EVENT, FAC, and LOC perform poorly. This performance difference could be due to differences in writing style between books and news (all with Moderate MMD scores ranging from 0.22 to 0.3). As for NORP, in this class we restrict the entity to groups of people, unlike Wojood, which includes all entities indicating plural forms of occupations, such as workers, kings, and deputies, among others. The metrics reflect this fact: when using Wojood to predict AdabNer's NORP entity type, the model performs well, but degrades in the opposite direction. This is because NORP in Wojood intersects with AdabNer, while AdabNer does not fully cover all NORP types present in Wojood . LAW, with a high MMD score, and PRODUCT (moderate score) fail badly in both directions, which shows domain sensitivity in their MMD scores.