Performance Analysis of Arabic Pre-Trained Models on Named Entity Recognition Task

Abdelhalim Hafedh Dahou, Mohamed Amine Cheragui, Ahmed Abdelali

GESIS – Leibniz-Institute for the Social Sciences, Cologne, Germany
Mathematics and Computer Science Department, Ahmed Draia University, Adrar, Algeria
Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar
abdelhalim.dahou@gesis.org
m_cheragui@univ-adrar.edu.dz
aabdelali@hbku.edu.qa

Abstract

Named Entity Recognition (NER) is a crucial task within natural language processing (NLP) that entails the identification and classification of entities such as person, organization and location. This study delves into NER specifically in the Arabic language, focusing on the Algerian dialect. While previous research in NER has primarily concentrated on Modern Standard Arabic (MSA), the advent of social media has prompted a need to address the variations found in different Arabic dialects. Moreover, given the notable achievements of Large-scale pre-trained models (PTMs) based on the BERT architecture, this paper aims to evaluate Arabic pre-trained models using an Algerian dataset that covers different domains and writing styles. Additionally, an error analysis is conducted to identify PTMs' limitations, and an investigation is carried out to assess the performance of trained MSA models on the Algerian dialect. The experimental results and subsequent analysis shed light on the complexities of NER in Arabic, offering valuable insights for future research endeavors.

1 Introduction

The expression named entities recognition (NER) has been used for the first time at the 6th edition of the Message Understanding Conference (MUC) in November 1995 (Grishman and Sundheim, 1996). The task of NER consisted in using SGML markers to identify entities in texts (names of persons, organizations, or places), temporal expressions, and numerical expressions ("currency" or "percentages"). Since then, NER has become a starting point and an important part of many applications in natural language processing (Ali et al., 2020), such as: Information Extraction (IE) (Kumar and Starly, 2022), Information Retrieval (IR) (Guo et al., 2009), Semantic Annotation (SA) (Li et al., 2022), Machine Translation (MT) (Babych and Hartley, 2003), Question Answering (QA) systems (Yadav and Bethard, 2018), Text Summarization (Aone, 1999) and Text Clustering (Nagrale et al., 2019).

The process of NER can be done according to three main approaches (Oudah and Shaalan, 2017) (Mansouri et al., 2008), (Gorinski et al., 2019): the symbolic or linguistic (rule-based) approach, where the main idea is to use linguistic knowledge (internal or external clues), dictionaries and gazetteers of proper names to establish a list of knowledge rules (called regular expressions or finite state transducers (Mesfar, 2007)). However, the principal inconvenience of this approach is that the rule-generation process is fastidious and time-consuming. The Machine Learning (ML) approach, is mainly based on a previously annotated corpus. where the recognition problem is converted into a classification problem and employs various ML models to solve it. A hybrid approach which combines the two previous approaches to boost the performance of the models developed have been tried as well.

In recent years, the deep learning approach has proven to be a very powerful for learning feature representations directly from datasets, achieving outstanding results. The approach can learn complex hidden representations without complex feature engineering and rich domain knowledge (Liu et al., 2022).

While the task for Latin scripted language is more advanced (Zhou and Chen, 2021), having features like capitalization gives a clue and differentiates between named entities and other words. Such feature is absent in languages like Arabic. The additional complexity of the task comes from the dialectal variations of Arabic.

In the literature, most of the works on NER in Arabic have been oriented towards the common version MSA, a variant that is both normalized and standardized. However, with the emergence of so-

Proceedings of Recent Advances in Natural Language Processing, pages 458–467
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_051

cial media (Facebook, tweeter, Youtube,. . . etc.) as a means of communication and also as a source of information. A huge amount of raw data generated every day, which represents a goldmine for many applications in NLP. Therefore, the research on NER has been oriented towards these variants of the Arabic language.

Dialectal Arabic is another form of Arabic language used in everyday' communications, and is generally spoken and written (social networks, advertisements, SMS, etc.). It varies not only from one Arab country to another, but also from one region to another within the same country. Thus, almost all Arab countries have their own dialects. Arabic dialectology generally distinguishes two main areas or families of dialects (Saadane et al., 2018), (Embarki, 2008), (John and Na'ama, 2019):

- The Eastern zone (Mashreq): including Egypt, Syria, and other Middle Eastern countries (Iraq, the Gulf States, Yemen, Oman, Jordan, etc.).
- The Western zone (North Africa): the Maghreb: which includes Algeria, Morocco, Tunisia, Libya, and Mauritania.

Various other granular classification were proposed in literature classify the dialects into five or more variants, namely Gulf, Nile Basin, Levant and Maghreb (Zaidan and Callison-Burch, 2011; Habash, 2010; Abdelali et al., 2021b) to even city level (Bouamor et al., 2018).

The Algerian dialect, also known as Darija (”common language”), is spoken by 70% to 80% of the Algerian population (Saâdane, 2015) (of estimated 45 million people). When we speak about Algerian dialect, we must understand that it is a question of various sub-varieties of local dialect due to the geographical expansion of the country (2.382 million km²), because there is no unified Algerian dialect. There are therefore many varieties of Algerian dialect. It should be remembered that all these sub-varieties are heterogeneously influenced by other languages (e.g. Berber, French, Spanish, Turkish, Italian, etc.) (Harrat et al., 2016). Thus, we can distinguish Algiers dialect (mainly influenced by Berber and Turkish), Oranais dialect (influenced by Spanish), Constantinois dialect (influenced by Italian), Tlemçani dialect (influenced by Andalusian Arabic), etc.

In the context of NLP, the Algerian dialect constitutes a real challenge due to the multitude of constraints it presents, which are either inherited from standard Arabic, such as agglutination, and syntactic flexibility. Or they are due to the dialect itself, such as lack of normalization and standardization (it is common in Algerian dialect as the case of other dialects to find several orthographic transcriptions for the same), code-switching (a consequence of alternating two or more languages (or varieties of dialect) during the production of the same sentences or conversation). ARABIZI is a new spontaneous spelling variant of Algerian dialect, based particularly on Latin characters associated with numbers of special characters.

Such challenges motivated us to explore and focus more on this dialect in an attempt to investigate its particularities in the context of the new deep learning models. Our contributions in this paper can be summarized as follows:

- Answer the inquiry of whether training on the Modern Standard Arabic (MSA) corpus can yield favorable outcomes when testing on the Algerian dialect.
- Benchmark several Arabic pre-trained models and evaluated their performance on a publicly available Algerian dataset.
- Study the impact of using MSA dataset and its performance in reference to the Algerian dataset.
- Apply an error analysis on the best performing pre-trained model in order to figure the challenges and limitations of the model.

The remainder of this paper is organized as follows: the related work for NER in Arabic is presented in section 2. Section 3 gives some indications about Arabic pre-trained models. section 4 and 5 are devoted to experiments and results. The error analysis is described in section 6 and finally, conclusion and future works are presented in section 7.

2 Related Work

The first work on ANER was the TAGARAB system in 1998 (Maloney and Niv, 1998). Since then, many studies have followed covering different approaches: rule-based, machine learning or hybrid. In this section, we will divide the works into two categories: those on the Algerian dialect which are rare and the second category is the works on MSA adopting a Deep learning approach.

2.1 Algerian Dialect

According to our research, the problem with the Algerian dialect is the lack of resources to develop tools based on a machine learning or deep learning approach or even for evaluation (Harrat et al., 2014). For this reason, existing work in Algerian NER focuses more on building corpus (or dataset).

Touileb (2022), build NERDz, Algerian NER dataset. The corpus was an extension of NArabizi treebank (Touileb and Barnes, 2021), which contains initially 1500 sentences containing both Latin and Arabic characters (NERDz is a parallel corpus). statistically, NERDz contains 08 categories of entities, namely: PER for person name (467 entities); GPE for countries and cities (438 entities); ORG represents companies, organizations, and institutions (290 entities); NORP refers to nationalities, political beliefs, and religions (235 entities); EVT includes all types of cultural, political, and sports events (54 entities); LOC all geographical places (41 entities); PROD characterizes objects (23 entities); and MISC other entities with low occurrence in the dataset (18 entities). The author presented preliminary baseline results based on a neural architecture for NER that combines character-level CNN, word-level BiLSTM, and a CRF inference layer.

Adouane and Bernardy (2020), worked on a process that consists of mitigating the problem of the scarcity of labeled data for the Algerian dialect by the creation of a dataset for NER, and an investigation of the settings where it is beneficial to share representations learned between two or several tasks. For building the corpus, they used two corpus initially developed for Code-Switch Detection (CSD) (Adouane and Dobnik, 2017) and Sentiment Analysis (SA) (Adouane et al., 2020). The annotation was done manually by two native speakers, according to 06 predefined classes: person (PER), location (LOC), product (PRO), organization (ORG), and company (COM). They tagged the rest of named entity mentions like time and events as “other” (OTH) to distinguish them from non-named entities (OOO). In order to identify multi-word expressions as one named entity chunk, they use the IOB (Inside-Outside-Beginning) labeling scheme. For the Multi-task, the authors used an encoder-decoder architecture. However, here the encoders are shared between the tasks, while decoders are task-specific. For the experimentation, they proposed four scenarios, the first one NER alone (Macro F-score = 49.68%), the second one NER associated with CSD (Macro F-score = 48.65%), the third one NER associated with Spelling Normalisation and Correction (SPELL) (Macro F-score = 42.05%), and the fourth one NER associated with SA (Macro F-score = 34.60%).

Dahou and Cheragui (2022), studied the impact of normalization and data augmentation on Algerian NER task, using 05 Arabic pre-trained models ARBERT, Arabert v0.2, DziriBERT, MARBERT, and mBERT. For that, they built a corpus based on Facebook’s comments, manually annotated according to 03 categories: person (578 entities), location (548 entities), and organization (186 entities). To evaluate the models, the authors set up 04 scenarios, the first one without normalization and data augmentation, in this case, the ARBERT model outperformed the other models with an F1 score of 84.4%. The second scenario is to use normalization, which enabled the DziriBERT model to get the highest F1 Score of 81.9%. The third scenario with data augmentation, where the Arabert v0.2 model yielded the best F1 score with 85.1%. The Arabert v0.2 model again obtained the best F1 Score with 86.2% in the last scenario combining normalization and data augmentation.

Dahou and Cheragui (2023a), presented DzNER, an Algerian dataset for NER, composed of more than 21,000 sentences (over 220,000 tokens) from Algerian Facebook pages and YouTube channels, the process of annotation is done manually by two professional annotators on the Algerian dialect, using the IOB2 scheme for three entities: PER which covers persons names, ORG that includes organizations, companies, institutions, political groups, and football clubs, and finally LOC that represents the geographical places. In order to evaluate the contribution and effectiveness of their corpus, the authors have carried out experiments to analyze the performance of pre-trained Arabic models which are: Arabert and DziriBERT. Where the training is done with DZNER and the test with NArabizi. The Arabert achieved a Macro F1 Score of 75.41% and DziriBERT obtained a Macro F1 Score of 74.69%.

(Dahou and Cheragui, 2023b) studied the impact of two phenomena, the first one was the segmentation and the second one was the use of Latin characters in the Algerian dialect. For this purpose, they pre-training 05 models: AraBERT, MARBERT, ARBERT, DziriBERT, and mBERT. For the experimentation, they use a novel annotated Algerian named entities recognition (DzNER) dataset. The results demonstrate that the ARBERT achieved the best results in Arabic characters with an F1 score of 0.819% on segmented dataset and 0.844% on unsegmented dataset, and the mBERT achieved the best results in Latin characters with an F1 score of 0.676

2.2 Modern Standard Arabic

Bazi and Laachfoubi (2019), introduced a neural network architecture based on bidirectional Long Short-Term Memory (LSTM) and Conditional Random Fields (CRF). The model gets two sources of information about words as input: pre-trained word embeddings and character-based representations and eliminated the need for any task-specific knowledge or feature engineering. For training and testing the authors used ANERcorp, their model achieved an F1 score of 90.6%.

Helwe and Elbassuoni (2019), adopted a semi-supervised co-training approach. Using of a small amount of labeled data, which is augmented with partially labeled data that is automatically generated from Wikipedia. The approach relies only on word embeddings as features and does not involve any additional feature engineering. For the test they used three different Arabic NER datasets: AQMAR, NEWS dataset, and TWEETS dataset, they obtained average F1 scores of 61.8%, 74.1%, and 59.2% respectively.

Ali and Tan (2019), employed a bidirectional encoder–decoder model for addressing the problem of ANER on the basis of recent work in deep learning, in which the encoder and decoder are bidirectional LSTMs. In addition to word-level embeddings, character-level embeddings are adopted, and they are combined via an embedding-level attention mechanism. The model can dynamically determine the information that must be utilized from a word - or character-level component through this attention mechanism. The authors run their experiments on the merged dataset (ANERcorp plus AQMAR). The model achieved a high F-score of 92, 01%.

Alkhatib and Shaalan (2020), proposed a hybrid mechanism based on a conventional neural network, followed by Bi-LSTM and CRF. The model was examined on ANERCorp and Kalimat Corpus. The overall results obtained for the categories: person, location, and organization, in terms of F-measure, are: 93.7%, 95.2%, and 95.3% respectively.

Al-Smadi et al. (2020), used a transfer learning with deep neural networks to build a Pooled-GRU model combined with the Multilingual Universal Sentence Encoder. The proposed model scored 90% with the F1 score, using WikiFANE Gold dataset.

Alsaaran and Alrabiah (2021), proposed a deep learning-based model by fine-tuning BERT model to recognize and classify Arabic-named entities. The pre-trained BERT context embeddings were used as input features to a Bidirectional Gated Recurrent Unit (BGRU) and were fine-tuned using two annotated ANER datasets. For the experimentation, they set up two scenarios, the first using ANERCorp dataset and obtained F1 score of 92.28%. The second merged ANERCorp and AQMAR dataset and achieved an F1 score of 90.68%,

Al-Qurishi and Souissi (2021), proposed an effective model for ANER. The architecture of this model consists of three layers: a transformer-based language model layer, a fully connected layer, and the last layer is a conditional random field(CRF). For the test, the model achieved an F1-macro score of 89.6% on the ANERCorp and 88.5% on the AQMAR datasets.

Boudjellal et al. (2021), presented ABioNER a BERT-based model to identify biomedical named entities in the Arabic text data (specifically disease and treatment named entities) that investigates the effectiveness of pretraining a monolingual BERT model with a small-scale biomedical dataset on enhancing the model understanding of Arabic biomedical text. The model performance was compared with two state-of-the-art models (AraBERT and multilingual BERT cased), and it outperformed both models with 85% F1 Score.

Shaker et al. (2023), proposed long short-term memory (LSTM) units and Gated Recurrent Units (GRU) for building the NER model in the Arabic language. For the experimentation, they built a new dataset in seven different fields (Geography, History, Medical, Sport, Technology, News, and Cooking). The entities' names were labeled in nine categories: Person (PER), Location (LOC), geopolitical (GEO), time (TIM), profession (PRO), organization (ORG), disease (DIS), geography (GEO), and miscellaneous (MISC). The tests show that the LSTM model achieved better accuracy than the GRU model, 80.24% and 77.78% respectively.

3 Arabic Pre-Trained Models

Pre-trained language models, including BERT (Devlin et al., 2018a) and RoBERTa (Liu et al., 2019), have demonstrated significant success across a wide range of NLP tasks in various languages. Arabic NLP has witnessed substantial advancements with the development of dedicated pre-trained language models, achieving state-of-the-art outcomes in both MSA and DA as shown in table 1. However, selecting the most suitable model is challenging due to differences in design decisions and hyper-parameters, such as data size, language variant, tokenization, vocabulary size, and number of training steps. Despite fine-tuning being the common approach to choosing the best-performing pre-trained model for a specific task, the reasons behind the superior performance of one model over another and the impact of design choices remain unclear. This study aims to address this question specifically for the Arabic NER task. We selected the following models based on their popularity and coverage for both MSA and DA.

- **AraBERT** (Antoun et al., 2020) is a BERT pre-trained model was trained on around 77GB of Arabic text (8B words) that included Wikipedia Arabic dump, OSCAR corpus (Ortiz Suárez et al., 2020), OSIAN Corpus (Zeroual et al., 2019), Abu El-Khair Corpus (El-khair, 2016) and a large collection from Assafir newspaper articles.

- **MARBERT** (Abdul-Mageed et al., 2021) A large pre-trained model trained and released by the UBC NLP team. The model used a collection of over 1B tweets 128GB of text (15.6B tokens) in combination with 61GB of MSA text (6.5B tokens) from publicly available collections.

- **mBERT** (Devlin et al., 2018b) A Pre-trained model from Google that was built on Wikipedia top 104 languages using a masked language modeling (MLM) objective. Even though this model is not purely trained for Arabic. It's coverage for Arabic is decent as it ranks on the top languages.

- **QARiB** (Abdelali et al., 2021a) The model was pre-trained on Arabic Gigaword Fourth Edition, Abu El-Khair Corpus (El-khair, 2016), Open Subtitles (Lison and Tiedemann, 2016) in addition to 440M unique tweets. This made a total of 14B tokens.

Model	**Params**	**N. Words**	**Vocab. size**
AraBERT	136M	8.6B	64K
MARBERT	163M	6.2B	100k
mBERT	110M	1.5B	106k
QARiB	110M	14B	64k

Table 1: The selected Arabic pre-trained models.

To evaluate the models listed in table 1, we conducted fine-tuning on our datasets and assessed their performance under various scenarios based on the proposed contributions in the introduction. The final architecture utilized consists of an Arabic pre-trained BERT model combined with a straightforward linear layer. Conceptually, the Arabic pre-trained model functions as an embedding layer. We simply augment this with a linear layer to predict the tag for each token in the input sequence. All inputs are simultaneously processed by the pre-trained model, generating individual embeddings for each token. These embeddings are contextually influenced by the other tokens within the sequence, resulting in contextualized embeddings. Subsequently, we passed the output of the pre-trained model to the Linear layer. To predict NER tagging, such as identifying a person, organization, or location, we incorporated a softmax layer on top.

4 Experimental Setup

This section details the experimental setup used in our research. In our experiments, we investigated the performance and limitations of the Arabic pre-trained model in the NER task.

4.1 Dataset

We conducted experiments using two Arabic datasets: the DzNER corpus (Dahou and Cheragui, 2023a)[1], designed for the Algerian dialect NER task and encompassing various domains such as Sports, Travel, Electronics, and Politics. This corpus comprises 220k tokens with 18,387 entities annotated with organization (ORG), person (PER), and location (LOC) tags. The training set accounts for 80% of the total tokens, while the remaining portion is allocated for testing. For MSA NER, we utilized the ANERcorp dataset (Benajiba et al., 2007) using the splitting provided by CAMeL Lab

[1] DzNER Corpus in Github

(Obeid et al., 2020). ANERcorp consists of 316 articles selected from different newspapers to create a diverse corpus, totaling 150k tokens, with 11% of them representing named entities (NEs). The training split comprises 125,102 tokens, and the test split contains 25,008 tokens, all labeled with organization (ORG), person (PER), location (LOC), and miscellaneous (MISC) tags. In our study, we focused exclusively on the three primary entities: person, organization, and location. To accommodate ANERcorp, we replaced the MISC label with the label O. Figure 1 details the overall distribution of the entities in both datasets. Table **??** illustrates the distribution of entities in the training and testing splits for both datasets.

	DzNER		**ANERCorp**	
Entities	**Train**	**Test**	**Train**	**Test**
Person	6189	2204	2721	858
Location	5077	1315	3776	668
Organization	3740	1185	1576	450

Table 2: Statistics of the evaluation datasets.

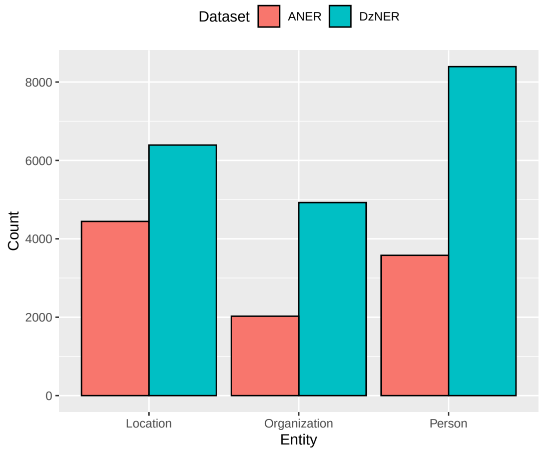


Figure 1: Distribution of NER categories in DzNER and ANERCorp.

4.2 Metrics

The metrics employed in this study include precision, recall, and F1-score. These metrics were selected to evaluate the model's performance in predicting the entity tag.
Precision gauges the ratio of true positives among the instances predicted as positive.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Recall assesses the ratio of true positives correctly identified.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

The F1-score represents the harmonic mean of precision and recall. It provides a measure of the balance between precision and recall, with values ranging from 0 to 1. Higher values indicate superior performance.

$$F1 = \frac{2 \times (precision \times recall)}{precision + recall} \quad (3)$$

4.3 Hyper-parameters

The finetuning and testing processes took place on the Google Colab platform, making use of a Tesla P100 - 16GB GPU. To achieve superior results, we fine-tuned the hyper-parameters by leveraging the test subset of the DzNER dataset. We employed the Adam optimizer (Kingma and Ba, 2014), setting the learning rate to 5×10^{-5}, with a batch size of 16, and a seed of 42 for six epochs. Throughout all our experiments, we utilized the Huggingface Transformers library (Wolf et al., 2020).

5 Results and Discussion

We carried out a battery of experiments in the following order:

5.1 Evaluating DzNER Performance on ANERCorp

We finetuned the selected pre-trained models using the training part of ANERCorp and evaluated both test sets of ANERCorp and DzNER. Table **??** shows the results. It is clear that the DzNER did not perform well on the MSA content. This stress the challenges of dealing with dialectal content and how much models trained only on MSA will underperform, eventhough the original pre-trained models were already exposed to such dialectal content.

5.2 Evaluating ANERCorp Performance on DzNER

The objective of this experiment is to benchmark MSA dataset and its performance when evaluated on dialectal content. Despite that both are Arabic text, the lack of standard orthography and the extensive code-switching in the dialectal content present a major challenge as detailed in section 1. The results in Table **??** similarly to experiment 5.1; the finetuned models performed sub-optimally on the MSA dataset. It is worth to note that the numbers are slightly better than finetuning only on MSA. This indicate that the dialecatal content subsumes the MSA in such task. While most of the MSA features are captured in the dialectal dataset. Extensive code-switching and unstandarized writing is typically absent in MSA.

	ANER		**DzNER**	
Model	**ANER**	**DzNER**	**ANER**	**DzNER**
AraBERT	0.850	0.639	0.779	0.855
MARBERT	0.827	0.615	0.643	0.827
mBERT	0.776	0.372	0.545	0.790
QARiB	0.820	0.570	0.708	0.828

Table 3: Results of the evaluation cross-datasets using different pre-trained models using micro F1 score. The upper row represents the training data, and the second row represents the testing data.

5.3 Evaluation on Combined Data

Another set of experiments where we attempted to explore whether combining the datasets would have any impact or not on the evaluation. The goal is to see if the Algerian dialect will benifit from the existance of the MSA in the training data or the inverse. After combining both ANERCorp and DzNER training datasets, we evaluated the new finetined models using the test sets of ANERCorp and DzNER separately. Table ?? shows the results of the evaluation. It is clear that the differences are very marginal and not significant as shown in Figure 2. The results are a good indication that both datasets are disjoint and the features present in both are not redundant.

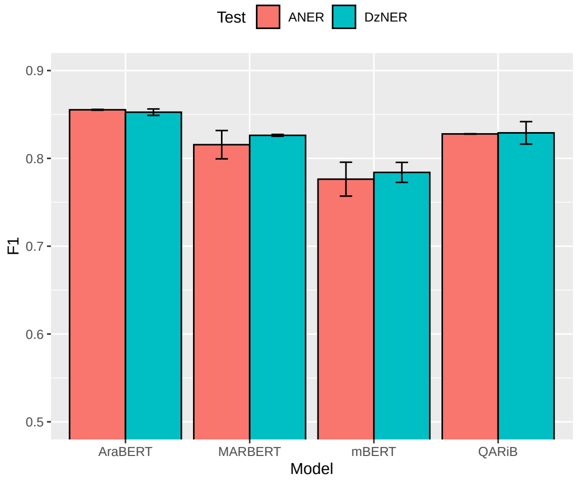


Figure 2: Performance of models per dataset.

Model	**ANER**	**DzNER**
AraBERT	0.8557	0.8552
MARBERT	0.8042	0.8255
mBERT	0.7627	0.7921
QARiB	0.8277	0.8381

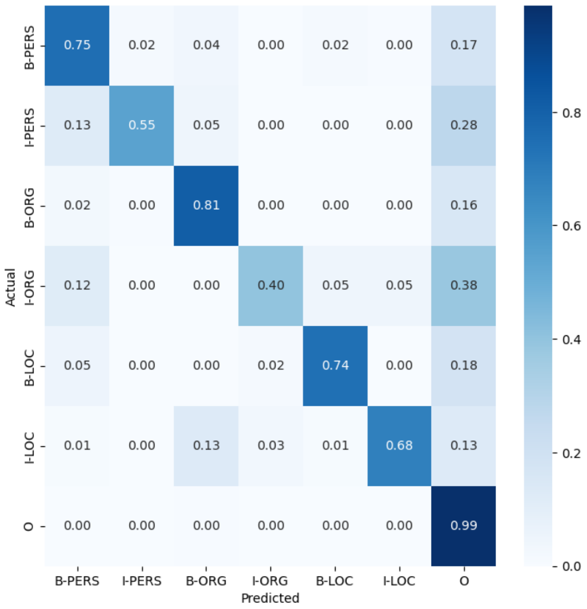
Table 4: Results of the data combination using different pre-trained models using micro F1 score.

6 Error Analysis

For further investigation, we selected the best performing model AraBERT to probe and examine the shortfall of such class of models. For such task, we inspected the errors on DzNER. Figure 3 shows the confusion matrix for the results of evaluating DzNER on model finetuned with the training set from the same dataset. It is clear that the majority of the errors are caused by not detecting PERS, ORG and LOC respectively on the order of error severity. Looking deeper into the issue, we selected 100 samples among the errors resulted from the classification. We noted that the bulk of these errors are caused by lack of spelling standards such as the case of “ المووغريب، العرق، بانقلاداش “ which are misspellings for “المغرب، العراق، بنغلاديش ” respectively. Such cases represents over 13% of the errors. While another large set of errors are caused by transliteration, this is mostly when using foreign or entities in another language but transcribing them in Arabic. Cases such as “ڨوقل، الجيري، لألجي “ that represents “ Google, Alger, Algerie “ respectively. Such category of errors represent another 21% among the errors. Errors such missing capitalization in Latin transcribed entities is very common as well. This is the case for “bougara, paris, and zanzibar” that supposed to be transcribed with capitals as “ Bougara, Paris, Zanzibar “. Such issues highlight the challenges dealing with dialectal content that is present in this dataset and similar ones.

7 Conclusion

In this study, we conducted a series of experiments to investigate NER performance in the context of Arabic, with a specific focus on the Algerian dialect. Our findings shed light on the challenges and limitations of existing Arabic pre-trained models trained on MSA and DA when applied to dialectal content. The experiments comparing the performance of ANERCorp on DzNER and vice versa revealed the difficulties posed by the lack of stan-



Actual \ Predicted	B-PERS	I-PERS	B-ORG	I-ORG	B-LOC	I-LOC	O
B-PERS	0.75	0.02	0.04	0.00	0.02	0.00	0.17
I-PERS	0.13	0.55	0.05	0.00	0.00	0.00	0.28
B-ORG	0.02	0.00	0.81	0.00	0.00	0.00	0.16
I-ORG	0.12	0.00	0.00	0.40	0.05	0.05	0.38
B-LOC	0.05	0.00	0.00	0.02	0.74	0.00	0.18
I-LOC	0.01	0.00	0.13	0.03	0.01	0.68	0.13
O	0.00	0.00	0.00	0.00	0.00	0.00	0.99

Figure 3: Confusion matrix for the results of evaluating DzNER on AraBERT model finetuned with DzNER train set.

dardized orthography and extensive code-switching in dialectal content. While the fine tuned models showed slightly improved results on the MSA dataset, the dialectal content encompassed MSA features, highlighting the dominance of dialectal data in this task. The combination of the ANERCorp and DzNER datasets did not significantly impact the evaluation results, indicating that the datasets offer non-redundant features and are disjoint from each other. The error analysis, conducted using the best performing model AraBERT, identified common sources of errors in dialectal content, such as spelling variations, transliteration issues, and missing capitalization in latin transcribed entities. These findings emphasize the challenges associated with dialectal content and the need to address spelling variations and non-standardized writing in dialectal Arabic. Future research will focus on: (i) refining NER models to better handle dialectal Arabic; (ii) explore strategies to expand these resources and improve performance in dialectal contexts; and (iii) investigate joint training NER with other auxiliary tasks such as part of speech tagging. Both tasks can mutually benefit from each other and share useful knowledge.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021a. Pre-training bert on arabic tweets: Practical considerations.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021b. QADI: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 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, Online. Association for Computational Linguistics.

Wafia Adouane and Jean-Philippe Bernardy. 2020. When is multi-task learning beneficial for low-resource noisy code-switched user-generated algerian texts? In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 17–25.

Wafia Adouane and Simon Dobnik. 2017. Identification of languages in algerian arabic multilingual documents. In *Proceedings of the third Arabic natural language processing workshop*, pages 1–8.

Wafia Adouane, Samia Touileb, and Jean-Philippe Bernardy. 2020. Identifying sentiments in algerian code-switched user-generated comments. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2698–2705.

Muhammad Al-Qurishi and Riad Souissi. 2021. Arabic named entity recognition using transformer-based-crf model. In *International Conference on Natural Language and Speech Processing*.

Mohammad Al-Smadi, Sa'ad A. Al-Zboon, Yaser Jararweh, and Patrick Juola. 2020. Transfer learning for arabic named entity recognition with deep neural networks. *IEEE Access*, 8:37736–37745.

Brahim Ait Ben Ali, Soukaina Mihi, Ismail El Bazi, and Nabil Laachfoubi. 2020. A recent survey of arabic named entity recognition on social media. *Rev. d'Intelligence Artif.*, 34:125–135.

Mohammed NA Ali and Guanzheng Tan. 2019. Bidirectional encoder–decoder model for arabic named entity recognition. *Arabian Journal for Science and Engineering*, 44:9693–9701.

Manar Alkhatib and Khaled Shaalan. 2020. Boosting arabic named entity recognition transliteration with deep learning. In *The thirty-third international flairs conference*.

Norah Alsaaran and Maha Alrabiah. 2021. Arabic named entity recognition: A bert-bgru approach. *Comput. Mater. Contin*, 68:471–485.

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

Chinatsu Aone. 1999. A trainable summarizer with knowledge acquired from robust nlp techniques. *Advances in automatic text summarization*, pages 71–80.

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*.

Ismail El Bazi and Nabil Laachfoubi. 2019. Arabic named entity recognition using deep learning approach. *International Journal of Electrical and Computer Engineering (IJECE)*.

Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *Computational Linguistics and Intelligent Text Processing: 8th International Conference, CICLing 2007, Mexico City, Mexico, February 18-24, 2007. Proceedings 8*, pages 143–153. Springer.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nada Boudjellal, Huaping Zhang, Asif Khan, Arshad Ahmad, Rashid Naseem, Jianyun Shang, and Lin Dai. 2021. Abioner: a bert-based model for arabic biomedical named-entity recognition. *Complexity*, 2021:1–6.

Abdelhalim Hafedh Dahou and Mohamed Amine Cheragui. 2022. Impact of normalization and data augmentation in ner for algerian arabic dialect. In *Modelling and Implementation of Complex Systems: Proceedings of the 7th International Symposium, MISC 2022, Mostaganem, Algeria, October 30-31, 2022*, pages 249–262. Springer.

Abdelhalim Hafedh Dahou and Mohamed Amine Cheragui. 2023a. Dzner: A large algerian named entity recognition dataset. *Natural Language Processing Journal*, page 100005.

Abdelhalim Hafedh Dahou and Mohamed Amine Cheragui. 2023b. Named entity recognition for algerian arabic dialect in social media. In *12th International Conference on Information Systems and Advanced Technologies "ICISAT 2022" Intelligent Information, Data Science and Decision Support System*, pages 135–145. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Ibrahim Abu El-khair. 2016. 1.5 billion words arabic corpus.

Mohamed Embarki. 2008. Les dialectes arabes modernes: état et nouvelles perspectives pour la classification géo-sociologique. *Arabic*, pages 583–604.

Philip John Gorinski, Honghan Wu, Claire Grover, Richard Tobin, Conn Talbot, Heather C. Whalley, Cathie L. M. Sudlow, William Whiteley, and Beatrice Alex. 2019. Named entity recognition for electronic health records: A comparison of rule-based and machine learning approaches. *ArXiv*, abs/1903.03985.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 267–274, New York, NY, USA. Association for Computing Machinery.

Nizar Y Habash. 2010. Introduction to arabic natural language processing. *Synthesis lectures on human language technologies*, 3(1):1–187.

Salima Harrat, Karima Meftouh, Mourad Abbas, Khaled-Walid Hidouci, and Kamel Smaili. 2016. An algerian dialect: Study and resources. *International Journal of Advanced Computer Science and Applications*, 7(3).

Salima Harrat, Karima Meftouh, Mourad Abbas, and Kamel Smaili. 2014. Building resources for algerian arabic dialects. In *15th Annual Conference of the International Communication Association Interspeech*.

Chadi Helwe and Shady Elbassuoni. 2019. Arabic named entity recognition via deep co-learning. *Artificial Intelligence Review*, 52:197–215.

Huehnergard John and Pat-El Na'ama. 2019. The semitic languages.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Aman Kumar and Binil Starly. 2022. "fabner": information extraction from manufacturing process science domain literature using named entity recognition. *Journal of Intelligent Manufacturing*.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li. 2022. Chinese named entity recognition: The state of the art. *Neurocomputing*, 473:37–53.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

John Maloney and Michael Niv. 1998. Tagarab: a fast, accurate arabic name recognizer using high-precision morphological analysis. In *Computational approaches to semitic languages*.

Alireza Mansouri, Lilly Suriani Affendey, and Ali Mamat. 2008. Named entity recognition approaches. *International Journal of Computer Science and Network Security*, pages 339–344.

Slim Mesfar. 2007. Named entity recognition for arabic using syntactic grammars. In *Natural Language Processing and Information Systems, 12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007, Paris, France, June 27-29, 2007, Proceedings*, volume 4592 of *Lecture Notes in Computer Science*, pages 305–316. Springer.

Deepali Nagrale, Vaibhav Khatavkar, and Parag Kulkarni. 2019. Document theme extraction using named-entity recognition. In *Computing, Communication and Signal Processing*, pages 499–509, Singapore. Springer Singapore.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Mai Oudah and Khaled Shaalan. 2017. Nera 2.0: Improving coverage and performance of rule-based named entity recognition for arabic. *Natural Language Engineering*, 23(3):441–472.

Houda Saadane, Hosni Seffih, Christian Fluhr, Khalid Choukri, and Nasredine Semmar. 2018. Automatic identification of maghreb dialects using a dictionary-based approach. In *International Conference on Language Resources and Evaluation*.

Houda Saâdane. 2015. *Le traitement automatique de l'arabe dialectalisé: aspects méthodologiques et algorithmiques*. Ph.D. thesis, Université Grenoble Alpes.

Alaa Shaker, Alaa Aldarf, and Igor Bessmertny. 2023. Using lstm and gru with a new dataset for named entity recognition in the arabic language. *arXiv preprint arXiv:2304.03399*.

Samia Touileb. 2022. Nerdz: A preliminary dataset of named entities for algerian. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 95–101.

Samia Touileb and Jeremy Barnes. 2021. The interplay between language similarity and script on a novel multi-layer algerian dialect corpus. *arXiv preprint arXiv:2105.07400*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. OSIAN: Open source international Arabic news corpus - preparation and integration into the CLARIN-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182, Florence, Italy. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2021. Learning from noisy labels for entity-centric information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.