Exploring BERT Models for Part-of-Speech Tagging in the Algerian Dialect: A Comprehensive Study

Mohamed Amine Cheragui[1], **Abdelhalim Hafedh Dahou**[2] and **Amin Abdedaiem**[1]
[1] Mathematics and Computer Science Department Ahmed Draia University Adrar - Algeria
[2] GESIS-Leibniz-Institute for the Social Science Cologne - Germany
m_cheragui@univ-adrar.edu.dz
abdelhalim.dahou@gesis.org
aminabdedaiem@gmail.com

Abstract

Social media have given a new impetus to natural language processing, especially for Arabic, by orienting research towards varieties of languages called dialects, which are less prestigious linguistically than Modern Standard Arabic (MSA) but are becoming more and more important as informal communication channels through different platforms: emails, blogs, discussion forums and SMS, offering a fertile research area. Part-of-speech (POS) tagging holds significant importance in various natural language processing applications, particularly in languages with complex morphological characteristics like Arabic. While a substantial part of research has concentrated on POS tagging for MSA, studies on dialects are scarce due to limited linguistic resources. This paper aims to showcase our efforts in advancing a morphosyntactic tagger tailored for the Algerian dialect. We accomplish this through a series of experiments employing a pre-trained Arabic transformer model, fine-tuned on various writing styles of the Algerian dialect commonly encountered in social media and everyday communication. Our proposed model outperforms previous state-of-the-art models, achieving an accuracy rate of 87% for Dz writing style and 83% for Arabizi writing style.

1 Introduction

Recognizing the nature of a word in a context (classification of words according to their behaviour in language) is a non-trivial task in natural language processing (NLP). Indeed, making a machine capable of knowing the linguistic category of a word requires the implementation of sophisticated methods, in particular for ambiguous words, i.e. those that may belong to several morphosyntactic categories. Such automatic tools are called Part of Speech tagger.

POS tagging is a fundamental task for NLP, on which complex processes such as information extraction or machine translation, syntactic analysis, etc., are often based. By definition, POS tagging is a process that assigns a morpho-syntactic tag to each word in a text specifying, in particular, grammatical category, gender, number, tense, and mode (Nerabie et al., 2021).

Arabic language represents a real challenge in terms of POS tagging, mainly due to its particular morphological system, both rich and complex as a consequence of two linguistic phenomena which are inflection and derivation, which make the process of recognizing the parts of speech a tedious task (Habash and Rambow, 2005). Arabic is a language that is spoken by a population of about 428 million people [1] and extends over a huge geographical area from the Arabian Gulf to the Atlantic, spread over 22 countries. This geographical expansion has contributed to the emergence of several variants of the Arabic language called "aammiyya" dialect (colloquial Arabic) as opposed to fusha (literary Arabic). Although these dialects share some common characteristics, they differ on many linguistic levels from standard Arabic (Katz and Diab, 2011).

According to (Habash, 2010), we can enumerate 30 variants of Arabic dialects. The interest in this variant of the language, in despite of the difficulties it presents, in particular the lack of orthographic normalization and standardization, is due to its expansion in terms of use, especially in social media, offering a research field with many challenges.

In this paper, we outline our approach for the development of morphosyntactic POS tagging in the context of one of the most prevalent dialects found on social networks—the Algerian dialect. We achieve this by:

- Assessing and exploring several models based on the BERT architecture (AraBERT v0.2-base, AraBERT v0.2-Twitter-base, Dziribert,

[1] World Population Review. Arab Countries 2020. Washington, DC. https://worldpopulationreview.com/countries/arab-countries/. Accessed August 9, 2023

MARBERT and m-BERT).

- Tackle different Algerian writing styles including: Arabic letters (Dz), Latin characters (Arabizi), and code-switching.

- Addressing the research question of the performance achieved by models trained solely on MSA when tested in various writing styles of the Algerian dialect.

The paper is organized into six sections. Section 1 introduces the research problem, while Section 2 provides a comprehensive review of related works in the field of Arabic dialect POS tagging. Our contribution is detailed in Section 3, and Section 4 offers insights into the dataset employed across various stages of experimentation. The experimental results are deliberated upon in Section 5, and, in conclusion, Section 6 summarizes our findings and outlines a vision for future research endeavors.

2 Related work

The Part of Speech tagging (POS) is a process that consists in assigning to each recognized entity a set of morphosyntactic features (Albared et al., 2011). This process has a crucial impact on the performance of several tools (Chunkers and Parsers, etc) and applications (Machine translation, Information retrieval, Text summarization, Sentiment analysis, etc) in NLP. For the MSA, POS tagging has been the subject of several works involving different approaches: rule-based, stochastic, and machine learning. However, for the Arabic dialect, research is scarce, due mainly to two factors: the lack of resources (corpus and tools: morphological analyzers, tokenizers, etc.), and there is no orthographic standards. The Dialectical Arabic (DA) POS tagging techniques follow two principal approaches. The first approach suggests using MSA resources and a few DA resources to create a POS tagger (Salloum and Habash, 2011) and the second intends to start from scratch.

Boujelbane et al. (Boujelbane et al., 2014), Retrained an MSA tagger which is the Stanford POS Tagger (Toutanvoa and Manning, 2000), using a corpus derived from a translation of the MSA Treebank into Tunisian Dialect, and adapt it to perform the tagging on the Tunisian dialect. The POS tagger set up achieved an accuracy of 78,5%.

Al-Sabbagh and Girju (Al-Sabbagh and Girju, 2012a), described a POS tagging based on the Brill's Transformation-Based Learning (Brill, 1994), for the Egyptian Dialect. For training and testing, the authors have built a golden corpus that contains 22,834 tweets, 423,691 tokens and 70,163 types. The tool obtained an F-measure score of 87.6%.

Baniata et al. (Baniata et al., 2018), presented a Bidirectional Long Short-Term Memory (Bi-LSTM)—Conditional Random Fields (CRF) segment-level Arabic Dialect POS tagger model for the Levantine Arabic (spoken variety of widely used in Jordan, Syria, Palestine and Lebanon) and Maghrebi (Morocco, Algeria and Tunisia), which will be integrated into the Multitask Neural Machine Translation (NMT) model. For the experimental part, they used the dataset described in (Darwish et al., 2018), which contains 350 tweets for four major Arabic dialects. Their POS tagger achieved an accuracy of 98% and 99% for the Levantine and Maghrebi dialect respectively.

Darwish et al. (Darwish et al., 2018), proposed a POS tagger for several Dialects (Egyptian, Levantine, Gulf, and Maghrebi), based on CRF. The authors have defined 03 features including clitic n-grams, clitic metatypes, and stem templates. For training and testing, a dataset covering all 04 dialects was built from 350 tweets for each dialect. For the results, the authors proposed 03 learning setups: the first one consists on treating each dialect alone, the model obtained the following results: 92.9% for Egyptian, 87.9% for Levantine, 87.8% for Gulf, and 88.3% for Maghrebi. In the second one, the dialects joined, the model gave the following results: 93.2% for Egyptian, 88.6% for Levantine, 87.2% for Gulf, and 87.7% for Maghrebi. In the third configuration, the dialects combined with the MSA, the model gave the following results: 93.4% for the Egyptian, 88.6% for the Levantine, 87.4% for the Gulf, and 87.6% for the Maghrebi.

Alharbi et al. (Alharbi et al., 2018), designed a Gulf Arabic (GA) POS taggers using two approaches: Support Vector Machine (SVM) classifier and BI-LSTM. For the SVM classifier, they defined 03 set features: Clitic features, Probabilistic features and Binary features. For the second Bi-LSTM classifier, the authors used Java Neural Network (JNN) toolkit for language modelling and POS tagging (Ling et al., 2015). The input of the network is a sequence of features: clitic, meta type, and/or stem template. For the experimental part, they used a gold annotated dataset which is built us-

ing gold segmented GA tweets taken from (Samih et al., 2017). Dataset consists of 343 Tweets with 6,844 tokens and 10,255 clitics. For the tag sets, they adopted the same one proposed by (Darwish et al., 2017) which composed of 18 tag sets. In addition, they added 04 others new tags for twitter specific data including: MENTION, URL, HASH, and EMOT. The two models SVM and Bi-LSTM obtained respectively an accuracy score of 85.96% and 91.2%.

Duh and Kirchhoff (Duh and Kirchhoff, 2005), built a Levantine and Egyptian POS tagger. They used the Buckwalter Morphological Analyzer designed for MSA, the LDC MSA Treebank corpus and some dialectal resources (the CallHome Egyptian Colloquial Arabic corpus ECA, the LDC Levantine Arabic corpus) in combination with unsupervised learning algorithms. The author's contribution consisted of bootstrap the Hidden Markov Models (HMM) tagger using POS information from the morphological analyzer. The developed tool obtained an accuracy of 70.88%.

Darwish et al. (Darwish et al., 2020), built a multi-dialectal POS tagger (covering Egyptian, Levantine, Gulf, and Maghrebi dialects) based on two approaches: CRF classifier combined with linguistic features (stem templates and clitic metatypes), word clusters from a large unlabeled tweet corpus, and automatic dialect identification; while the second combines word-based and character-based representations in a deep neural network with stacked layers of convolutional and recurrent networks with a CRF output layer. They achieve a combined accuracy of 92.4% across all dialects, with per dialect results ranging between 90.2% and 95.4%.

Hamdi et al. (Hamdi et al., 2015), developed a POS tagger for the Tunisian dialect. Their idea was to convert Tunisian into an approximate form of MSA, called pseudo MSA, and use an existing MSA POS tagger. The output produced is then projected back on the Tunisian text. The system operates through a three steps process: firstly, conversion is performed using MAGEAD, a morphological analyzer/generator; secondly, disambiguation is carried out; and finally, POS tagging is accomplished using HMM. For the evaluation, they used a transcribed and annotated corpus of 805 sentences containing 10,746 tokens and 2,455 types. The system achieved an accuracy of 89%.

AlKhwiter and Al-Twairesh (AlKhwiter and Al-Twairesh, 2021), proposed two supervised POS taggers for both MSA and the Gulf Dialect that are developed based on two approaches including CRF and Bi-LSTM. For the experimentation, the authors built three annotated datasets named Mixed, MSA, and GLF containing respectively 3, 1000, and 1000 Arabic tweets. As a result for the Gulf Dialect, the CRF and Bi-LSTM achieved an accuracy of 90% and 95% respectively.

Inoue et al. (Inoue et al., 2022), proposed morphosyntactic tagging model for three Arabic dialects: Gulf, Egyptian and Levantine, based on Pre-trained Language Model (CAMeLBERT-Mix) with two variants Factored and Unfactored Tags. The authors report that they obtained an accuracy of 94.6% for the Egyptian, 97.9% for Gulf, and 94.0% for Levantine, using respectively, ARZTB, Gumar Corpus, and Curras Corpus.

Pasha et al. (Pasha et al., 2014), presented MADAMIRA, which is a combined version of previously developed tools: MADA (Habash et al., 2009) and AMIRA (Diab, 2009), based on SVM. It provides various functions, such as tokenization, POS tagging and phrase chunking. The tool was trained on the Penn Arabic Treebank corpus for MSA and the Egyptian Arabic Treebanks for the Egyptian dialect. The performance of MADAMIRA was evaluated through a blind test dataset, and achieved an accuracy rate of 92.4% for the Egyptian Dialect.

3 Contribution

As previously stated, POS tagging is a preprocessing phase and an essential block in numerous NLP applications that require the syntactic category for each text token. The related work section demonstrates that the Arabic language has less work than the other language owing to its highly inflectional structure. Furthermore, the majority of Arabic works in POS and cutting-edge POS taggers are dedicated to the MSA variant, which is the formal language used in journalism and government administrations. DA is the more casual Arabic version used in everyday life, got less attention by researchers due to its great complexity when compared to the MSA.

With the passage of time, DA grew more frequently utilized, particularly in social media, and MSA POS taggers struggled to acquire good results when applying for DA texts (Pasha et al., 2014). Our contribution focused on developing a dialec-

Table 1: Arabic Dialect Annotated (POS) Corpus.

Corpus	Dialect	Token	Annotation
YADAC (Al-Sabbagh and Girju, 2012b)	EGY	6 M	FST and Manually
ARZATB (Fashwan and Alansary, 2022)	EGY	475 K	CALIMA and Manually
NArabizi (Seddah et al., 2020)	ALG	19770	Manually
LATB (Maamouri et al., 2006)	LEV	26 K	/
Gumar (Khalifa et al., 2016)	GULF	112 M	MADAMIRA
Curras (Jarrar et al., 2014)	PAL	43 K	DIWAN and Manually
Baladi (Al-Haff et al., 2022)	LEB	9.6 K	Manually (AnnoSheet)
MOR (Al-Shargi et al., 2016)	MOR	64170	DIWAN
YEMS (Al-Shargi et al., 2016)	YEM	32445	DIWAN

tal POS tagger for one of the more broadly used Arabic dialects in social media, the Algerian dialect, using an Arabic pretrained model based on the BERT architecture and fine-tuned on different writing styles of the Algerian dialect found in social media and used in everyday life. Furthermore, this study will investigate whether a POS tagger trained on the Algerian dialect may outperform an MSA POS tagger.

3.1 Transformers and BERT Models

With the rise of the RNN model and its variations problems, recently developed techniques were proposed to overcome those limitations including the Transformer-based architecture built based on the attention mechanism (Vaswani et al., 2017). The Transformer-based model is known for their high performance in terms of learning contextualized text representation. BERT (stands for Bidirectional Encoder Representations from Transformers) is one of the most popular NLP models that utilizes a transformer at its core and which achieved state of the art performance on many NLP tasks including Classification, Question Answering, and NER Tagging when it was first introduced. Contextualized text or word representation means that the embeddings of a word is not static. That is, they depend on the context of words around it. So in a sentence like 'خويا روح نيشان، دوك تلقاه' 'my brother go forward, you will find it', and the other sentence 'نيشان راك صح عندك' 'you are right!', the two embeddings of the word 'نيشان' will be different, which in the first means "forward" and in the second sentence means 'right'. While directional models in the past like LSTM read the text input sequentially (left-to-right or right-to-left), the Transformer actually reads the entire sequence of words at once and thus is considered bidirectional.

The developed models for English such as BERT, DistilBERT (Sanh et al., 2019), BART (Lewis et al., 2020), were adopted and used in Arabic NLP showing remarkable performance. In this study, we will evaluate the performance of those Arabic pretrained models and take the one that achieves high performance. For instance, the AraBERT (Antoun et al., 2020) a pre-trained model on large MSA and dialects data from Wikipedia and Twitter, MARBERT trained on Maghrebi dialectes data representing coutries such as Algeria, Morocco, and Tunisia (Abdul-Mageed et al., 2021), DziriBERT trained on Algerian dialect (Abdaoui et al., 2022), mBERT (Pires et al., 2019) trained on the top 104 languages including Arabic and its dialects with the largest Wikipedia data. Table 2 presents a comparison between those Arabic pre-trained models in terms of size, dataset and vocab.

3.2 Fine-tuning Models

As mentioned previously, BERT is a big neural network architecture, with a huge number of parameters, that can range from 100 million to over 300 million. Given this complexity, training a BERT model from scratch, particularly on a small dataset, predisposes it to overfitting due to the disproportionate ratio of parameters to data points. Consequently, it is more effective to utilize a pre-trained BERT model, which has been subjected to rigorous training on a voluminous dataset, as an initial framework. Then further train the model on our relatively smaller dataset and this process is known as model fine-tuning. This mechanism can be done in three ways, the first is to train the entire architecture of the pre-trained model, the second consists of training some layers while freezing others, and the third one freezes the entire architecture and trains

Table 2: Used Arabic pre-trained models.

Model	Size (params)	DataSet (nwords)	Vocab size
AraBERT v0.2-base	136M	8.6 Billion	64K
AraBERT v0.2-Twitter-base	136M	8.6 Billion + 60 Million Multi-Dialect Tweets	/
Dziribert	124M	1 million tweets	50k
MARBERT	163M	6.2 Billion	100k
mBERT	110M	1.5 Billion	106k

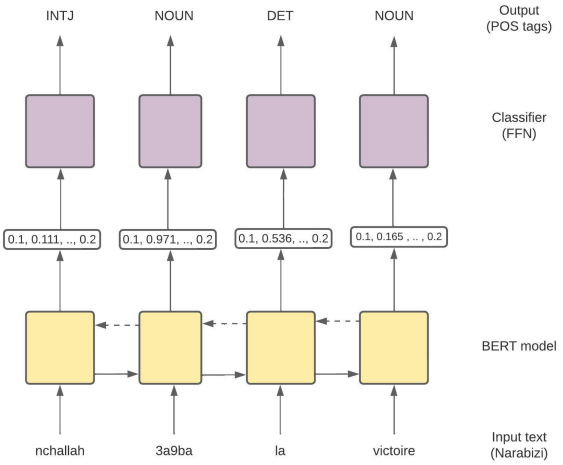

Figure 1: Model fine-tuning architecture.

just the classification layer. In our study, we train the entire pre-trained model on our dataset and feed the output to a softmax layer. In this case, the error is back-propagated through the entire architecture and the pre-trained weights of the model are updated based on the new dataset. Figure 1 describes the overall architecture of fine-tuning a bert model on POS tasks.

As shown in figure 1, our model is composed of an Arabic pre-trained BERT model and a simple linear layer. We can think of the BERT as an embedding layer and all we do is add a linear layer on top of these embeddings to predict the tag for each token in the input sequence. The yellow squares were the embeddings provided by the pretrained BERT model. All inputs are passed to BERT at the same time. The arrows between the BERT embeddings indicate how BERT does not calculate embeddings for each tokens individually, but the embeddings are actually based on the other tokens within the sequence which give us at the end a contextualized embedding. Finally, we fed the output of the pretrained BERT to the Linear layer of size (embedding_dim x n_outputs) and added a softmax layer on top to predict the POS Tagging like predicting noun, verb, or adjective.

4 Dataset

The dataset employed in this study, as delineated in (Touileb and Barnes, 2021), originates from the NArabizi treebank detailed by (Seddah et al., 2020). It encompasses a corpus of 1,300 Arabizi sentences sourced from an Algerian newspaper's web forum and an additional 200 sentences derived from song lyrics manually collated from various online platforms. Each sentence within this dataset is annotated across five distinct layers: tokenization, morphological analysis, code-switching identification, syntactic structure, and translation into French.
This dataset was further augmented to include two additional annotations for each token in the Arabizi sentences. The first enhancement involves the transliteration of each Arabizi token into the Arabic script, with the resultant dataset designated as 'DZ'. The second augmentation entails the transliteration of each Arabizi token into a code-switched script—either Arabic or Latin—depending on the token's origin, thus forming the code-switched dataset. As (Touileb and Barnes, 2021) assert, these annotations were meticulously conducted by bilingual native speakers of Algerian Arabic and French, adhering to standardized guidelines. Table 4, extracted from dataset' paper, exemplifies these stylistic variations within the dataset.
In the preceding sections, we outlined the focus of this study, which centers on evaluating the performance efficacy of a POS tagger specifically trained on the Algerian dialect. This investigation aims to ascertain whether such a dialect-specific POS tagger can surpass the performance of a MSA POS tagger. To facilitate a comprehensive and objective comparison, an MSA dataset, specifically curated for POS tagging, will be employed. For that, the

MSA dataset used is available in UD [2] (Zeman et al., 2020) and also has the same labels as the NArabizi dataset just with difference in terms of number of sentences and the average length. For the NArabizi, we have *19,770* tokens, *1,276* sentences with an average of *16.1* tokens. In MSA, we found *262,803* tokens, *8,664* sentences with an average of *42.3* tokens. Table 3 describe the distribution of POS tags in both datasets.

Table 3: Distribution of POS tags in both datasets in terms of numbers.

Category	NArabizi	UD (MSA)
NOUN	1981	10588
VERB	1819	3805
ADJ	624	4968
PROPN	552	1052
PRON	263	64
ADV	260	134
ADP	157	156
INTJ	120	5
DET	87	76
SCONJ	56	6
PART	36	36
PUNCT	36	18
CCONJ	32	95
NUM	9	994

5 Experiments and Evaluation

This section presents the experimental setup used, the experiments carried out as part of our research with a comparison between our best model and previous works.

5.1 Experimental setup

For the performance measures, the models are evaluated by calculating the precision, recall, Accuracy and F1-score of their output on the test dataset. Precision and recall are often used metrics to provide more accurate outcomes as well as to provide more information to the expert about the model's behavior, particularly in multi-class classification. To accelerate the training and testing phase, all of them were carried out using the Google Colab platform with a GPU Tesla P100-PCIE-16GB and the Hugging Face Transformers library (Wolf et al., 2020), was used in all our experiments. Using the test dataset, we fine-tuned the hyper-parameter to find the optimal configuration for each pre-trained model in order to achieve the best results. The hyper-parameter settings for each model are listed in table 5.

[2]Universal dependencies Corpus: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3226.

5.2 Results and discussion

This study comprises two experimental series, the first series of experiments looks at the performance of each Arabic pre-trained model indicated above on the Algerian dataset. The second trial series investigates how much performance can be obtained by applying MSA-specific models to adapt to Algerian dialect.

5.2.1 Experimental series 1

We ran three separate trials in this experimental series to assess the performance of each model on the dataset. The experiments are as follows: first, focus on the Dz writing style, which only uses Arabic letters; second, on the Arabizi style, which utilizes Latin characters; and third, on the code-switched style, which combines Latin and Arabic characters. Table 6 provides the results for each model on the three writing styles in terms of accuracy and F1 score. Finally, we compare our best-obtained results to previously published research in table 7.

The findings shown in table 6 demonstrated that the models performed well on the three writing styles of the Algerian dialect. The AraBERT twitter model obtained the highest results for the first style of writing that employs only Arabic words (Dz), with an F1 score of 84.6%, followed by the AraBERT base and DziriBERT models. This accomplishment is due to the variety of text sources and the volume of MSA and dialectal data encountered in the pre-training phase, which allows the model to represent the majority of Arabic words while avoiding out-of-vocabulary words. In the Arabizi writing style, the DziriBERT model exhibited superior performance, attaining the highest F1-score of 79.5%. Following closely behind was the mBERT model, which was trained across multiple languages, including French, predominantly used by the Algerian community especially in the Arabizi writing style. This multilingual training contributed to its commendable results. Even though DziriBERT's vocabulary is limited and it has seen less text in the pre-training phase compared to the other models, this demonstrates that pre-training a model for one dialect on a small training set

Table 4: Examples of the writing styles exist in the dataset.

Arabizi	ycombati la misere li las9at fina welat kiste
Arabic transliteration (Dz)	يكومباطي لا ميزار لي لسقت فينا ولات كيست
Code-switched transliteration	يكومباطي la misère لي لسقت فينا ولات kyste
English translation	He fights the misery that sticks to us and which has become a cyst

Table 5: Hyper-parameters values for each used model.

Models	Epochs	learning rate	warmup steps	seed
AraBERT v0.2-base	20	5e-5	42	42
AraBERT v0.2-Twitter-base	20	5e-5	42	42
Dziribert	15	5e-5	0	42
MARBERT	15	3e-5	42	42
mBERT	15	5e-5	42	666

Table 6: Performance results on Algerian dataset for the three writing styles.

	Dz		Arabizi		Code-switching	
Model	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**
AraBERT base v0.2	0.871	0.845	0.801	0.764	0.893	0.873
AraBERT v02-twitter	0.867	**0.846**	0.795	0.752	0.892	0.869
MARBERT	0.861	0.834	0.795	0.757	0.892	0.864
DziriBERT	0.866	0.840	0.831	**0.795**	0.895	**0.875**
mBERT	0.841	0.812	0.807	0.773	0.888	0.863

may provide better results than pre-training a multi-dialectal model on considerably larger data. In contrast to MARBERT, which underwent training on a substantially larger corpus encompassing diverse Arabic dialects, DziriBERT exhibited consistent superiority in performance. In the final phase of evaluating writing styles (Code-switching), DziriBERT once again demonstrated its excellence, achieved an impressive F1 score of 87.5%. This exceptional performance is attributed to DziriBERT's training on a relatively modest yet substantial corpus of Algerian text, comprising both Arabic and Latin characters.

As seen in Table 7, our models performed the best in terms of accuracy on both Dz and Arabizi writing styles. In this comparison, we compared our models' results to those of (Touileb and Barnes, 2021), who fine-tuned the multilingual BERT on their own data, (Seddah et al., 2020), who used a feature-based alVWTagger, and (Muller et al., 2020), who use mBERT and the StanfordNLP tagger.

5.2.2 Experimental series 2

The experiments are the same as in the previous series, except this time we train all the models on the MSA dataset and test them on the Algerian dataset with the three writing styles. Table 8 shows the accuracy and F1 score results for each model on the three writing styles.

Table 8 demonstrated that the models performed poorly on the three writing styles of the Algerian dialect as compared to the results obtained when the models were trained on the Algerian dataset. With 43%, 20%, and 49.1% F1 scores in Dz, Arabizi, and code-switched, respectively, DziriBERT and MARBERT outperformed the other models in the three writing styles. We may support this with the pre-training corpus for both models, which are trained only on Arabic dialects for MARBERT and Algerian dialects for DziriBERT. Furthermore, as compared to the Arabizi style, both models behaved well in the Dz and code-switched styles.

6 Conclusion

In this study, we assessed POS tagging for the Algerian dialect using transformer-based pre-trained

Table 7: Comparison of our best model with previous works in terms of accuracy.

Model	Dz	Arabizi
Touileb et al. (Touileb and Barnes, 2021)	82.5	76.3
Seddah et al. (Seddah et al., 2020)	80.4	-
Muller et al. (Muller et al., 2020)	81.6	-
Our model	**0.871**	**0.831**

Table 8: Performance results on Algerian test set with the use of MSA dataset as a training set.

	Dz		Arabizi		Code-switching	
Model	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**
AraBERT base v0.2	0.443	0.409	0.199	0.101	0.515	0.462
AraBERT v02-twitter	0.444	0.410	0.171	0.107	0.521	0.471
MARBERT	0.450	0.417	0.226	**0.200**	0.540	0.487
DziriBERT	0.462	**0.430**	0.223	0.196	0.538	**0.491**
mBERT	0.437	0.408	0.181	0.152	0.504	0.464

models. Our research was organized to evaluate these models' performance across diverse writing styles of the Algerian dialect, with the primary objective of ascertaining how effectively models trained on MSA text can deal with the Arabic dialects. As a results, DziriBERT consistently achieved the highest F1 scores across all the writing styles, showcasing its adaptability and robustness in handling these variations of the Algerian dialect. Our model outperformed previous works in accuracy for Dz and Arabizi styles. Moreover, when models trained on MSA were tested on Algerian data, performance dipped, but DziriBERT and MARBERT maintained strong results, especially in Dz and code-switched styles due to the amount of Algerian dialect data seen during the pre-training phase. Overall, this highlights the importance of tailoring models to dialects due to significant differences. DziriBERT excelled, even with a small training corpus, offering promise for dialect-specific language tasks. Future research will explore POS tagging's impact in other tasks like named entity recognition, machine translation, and segmentation.

References

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2022. Dziribert: a pre-trained language model for the algerian dialect. arXiv:2109.12346v3.

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.

Karim Al-Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + baladi: Towards a Levantine corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 769–778, Marseille, France. European Language Resources Association.

Rania Al-Sabbagh and Roxana Girju. 2012a. A supervised POS tagger for written Arabic social networking corpora. In *Proceedings of KONVENS 2012*, pages 39–52. ÖGAI.

Rania Al-Sabbagh and Roxana Girju. 2012b. YADAC: Yet another dialectal Arabic corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2882–2889, Istanbul, Turkey. European Language Resources Association (ELRA).

Faisal Al-Shargi, Aidan Kaplan, Ramy Eskander, Nizar Habash, and Owen Rambow. 2016. Morphologically annotated corpora and morphological analyzers for Moroccan and sanaani yemeni Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1300–1306, Portorož, Slovenia. European Language Resources Association (ELRA).

Mohammed Albared, Nazlia Omar, and Mohd. Juzaiddin Ab Aziz. 2011. Developing a competitive hmm arabic pos tagger using small training corpora. In *Intelligent Information and Database Systems*, pages 288–296, Berlin, Heidelberg. Springer Berlin Heidelberg.

Randah Alharbi, Walid Magdy, Kareem Darwish, Ahmed AbdelAli, and Hamdy Mubarak. 2018. Part-of-speech tagging for Arabic Gulf dialect using Bi-LSTM. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wasan AlKhwiter and Nora Al-Twairesh. 2021. Part-of-speech tagging for arabic tweets using crf and bi-lstm. *Computer Speech Language*, 65:101138.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Laith H Baniata, Seyoung Park, and Seong-Bae Park. 2018. A multitask-based neural machine translation model with part-of-speech tags integration for arabic dialects. *Applied Sciences*, 8(12):2502.

Rahma Boujelbane, Mariem Ellouze, Frédéric Béchet, and Lamia Belguith. 2014. De l'arabe standard vers l'arabe dialectal : projection de corpus et ressources linguistiques en vue du traitement automatique de l'oral dans les médias tunisiens [from Modern Standard Arabic to Tunisian dialect: corpus projection and linguistic resources towards the automatic processing of speech in the Tunisian media]. *Traitement Automatique des Langues*, 55(2):73–96.

Eric Brill. 1994. Some advances in transformation-based part of speech tagging. arXiv:cmp-lg/9406010.

Kareem Darwish, Mohammed Attia, Hamdy Mubarak, Younes Samih, Ahmed Abdelali, Lluís Màrquez, Mohamed Eldesouki, and Laura Kallmeyer. 2020. Effective multi-dialectal arabic pos tagging. *Natural Language Engineering*, 26(6):677–690.

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, and Mohamed Eldesouki. 2017. Arabic POS tagging: Don't abandon feature engineering just yet. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 130–137, Valencia, Spain. Association for Computational Linguistics.

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer. 2018. Multi-dialect Arabic POS tagging: A CRF approach. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mona Diab. 2009. Second generation amira tools for arabic processing: Fast and robust tokenization, pos tagging, and base phrase chunking. In *2nd international conference on Arabic language resources and tools*, volume 110, page 198.

Kevin Duh and Katrin Kirchhoff. 2005. POS tagging of dialectal Arabic: A minimally supervised approach. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 55–62, Ann Arbor, Michigan. Association for Computational Linguistics.

Amany Fashwan and Sameh Alansary. 2022. Developing a tag-set and extracting the morphological lexicons to build a morphological analyzer for Egyptian Arabic. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 142–160, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan Claypool Publishers.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan. Association for Computational Linguistics.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. Mada+ tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt*, volume 41, page 62.

Ahmed Hamdi, Alexis Nasr, Nizar Habash, and Núria Gala. 2015. POS-tagging of Tunisian dialect using Standard Arabic resources and tools. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 59–68, Beijing, China. Association for Computational Linguistics.

Go Inoue, Salam Khalifa, and Nizar Habash. 2022. Morphosyntactic tagging with pre-trained language models for Arabic and its dialects. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1708–1719, Dublin, Ireland. Association for Computational Linguistics.

Mustafa Jarrar, Nizar Habash, Diyam Akra, and Nasser Zalmout. 2014. Building a corpus for palestinian Arabic: a preliminary study. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 18–27, Doha, Qatar. Association for Computational Linguistics.

Graham Katz and Mona Diab. 2011. Introduction to the special issue on arabic computational linguistics. *ACM Transactions on Asian Language Information Processing*, 10(1).

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of Gulf Arabic. In *Proceedings of the Tenth International*

Conference on Language Resources and Evaluation (LREC'16), pages 4282–4289, Portorož, Slovenia. European Language Resources Association (ELRA).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramón Fermandez, Silvio Amir, Luís Marujo, and Tiago Luís. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530, Lisbon, Portugal. Association for Computational Linguistics.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, Mona Diab, Nizar Habash, Owen Rambow, and Dalila Tabessi. 2006. Developing and using a pilot dialectal Arabic treebank. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Benjamin Muller, Benoit Sagot, and Djamé Seddah. 2020. Can multilingual language models transfer to an unseen dialect? a case study on north african arabizi. *arXiv preprint arXiv:2005.00318*.

Abdul Munem Nerabie, Manar AlKhatib, Sujith Samuel Mathew, May El Barachi, and Farhad Oroumchian. 2021. The impact of arabic part of speech tagging on sentiment analysis: A new corpus and deep learning approach. *Procedia Computer Science*, 184:148–155. The 12th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 4th International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101, Reykjavik, Iceland. European Language Resources Association (ELRA).

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Wael Salloum and Nizar Habash. 2011. Dialectal to Standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21, Edinburgh, Scotland. Association for Computational Linguistics.

Younes Samih, Mohamed Eldesouki, Mohammed Attia, Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak, and Laura Kallmeyer. 2017. Learning from relatives: Unified dialectal Arabic segmentation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 432–441, Vancouver, Canada. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content North-African Arabizi treebank: Tackling hell. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139–1150, Online. Association for Computational Linguistics.

Samia Touileb and Jeremy Barnes. 2021. The interplay between language similarity and script on a novel multi-layer Algerian dialect corpus. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3700–3712, Online. Association for Computational Linguistics.

Kristina Toutanvoa and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 63–70, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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, Online. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Željko Agic, Lars Ahren-

berg, Chika Kennedy Ajede, Gabriele Aleksandraviciute, Lene Antonsen, et al. 2020. Universal dependencies 2.6. *LINDAT/CLARIAHCZ digital library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University. url: http://hdl. handle. net/11234/1-3226*.