ArDia: Improving Arabic Dialectal Language Classification Using a Novel Dataset

Hossam Elsafty[1], **Bouthaina Abdou**[1], **Tobias Deußer**[1,2], **Maren Pielka**[1,2], **Christian Bauckhage**[1,2]
Rafet Sifa[1,2]

[1]University of Bonn, Bonn, Germany
[2]Fraunhofer IAIS, Sankt Augustin, Germany
hfelsafty@gmail.com, maren.pielka@iais.fraunhofer.de

Abstract

Despite Arabic being one of the most widely spoken languages, there is a scarcity of available dialectal Arabic data. In this paper, we address this challenge by proposing a novel approach to data collection through the main use of video captions from TikTok, and other resources such as dictionaries and articles, resulting in the creation of the ArDia dataset. To the best of our knowledge, the ArDia dataset is the largest labeled dialectal Arabic dataset, containing over 900,000 examples, each labeled with its respective dialect. We further leverage this dataset to pretrain transformer-based models, ArDiaBERT and ArDiaGPT. Due to a lack of research on the Arabic models, we present a comprehensive study of Arabic dialect identification using the ArDia dataset on the dialect identification task.

Introduction

Arabic is one of the most spoken languages in the world, with over 400 million native speakers. Modern Standard Arabic (MSA) is the standardized, formal variety used in education, media, and official communication, while dialectal Arabic refers to the diverse, informal spoken varieties used in daily life. These dialects vary significantly across regions and often lack standardization, which complicates natural language processing (NLP) tasks. Arabic dialects are mainly used as daily spoken language and in social media, where they appear in chats, comments, tweets, and TikTok videos. Due to the increasing content of Arabic online, the development of Arabic models has become increasingly important to solve a lot of Arabic NLP tasks.

The transformer-based language models have achieved an impressive step in solving many NLP tasks with great performance in many languages, but these models need a massive amount of data to understand the language and get a good performance, which is considered the main challenge for the Arabic language. Not only that, but also, most Arabic datasets are modern standard Arabic, so most of the Arabic models get poor performance when dealing with dialectal Arabic.

This paper aims to contribute to a potential solution to this challenge by gathering massive data on dialectal Arabic. We propose a new approach to data collection using TikTok video captions, and other resources such as dictionaries and articles. Also, it introduces a new dialectal Arabic dataset called the ArDia dataset. In addition to that, we present in this paper two new encoder-only models: ArDiaBERT(cosine), and ArDiaBERT(balanced). Moreover, we present two new decoder-only models: ArDiaGPT(cosine), and ArDiaGPT(balanced). Both encoder-only and decoder-only models are pre-trained on pure dialectal Arabic. The third contribution of our paper is performing a comprehensive study of the Arabic models on the dialect identification task to deeply understand the ability of the Arabic models with dialectal Arabic and to measure the strengths and weaknesses of each model.

Related Work

There is an ongoing effort to create different dialectal Arabic resources to allow building models for several natural language processing applications. We will discuss selected dialectal Arabic datasets that could be used in Arabic dialect identification.

NADI (Nuanced Arabic Dialect Identification) (Abdul-Mageed et al. 2023) : It is commonly used in dialect identification because of its high quality and because it provides an accurate label for the dialect at the region level, subregion level, and city level. The main resource for NADI to collect the data was Twitter. Regarding the amount of data, the NADI-2023 dataset contained a total of 23,400 tweets that included 18 Arabic dialects.

MADAR (Multi-Arabic Dialect Applications and Resources) (Bouamor, Hassan, and Habash 2019) : The MADAR shared task on fine-grained dialect was the first to target a large set of dialect labels at the city and country levels. The shared task consisted of two main subtasks: MADAR Travel Domain Dialects and MADAR Twitter User Dialect Identification. MADAR Twitter User Dialects Identification targeted the identification of dialects used by Twitter users across 21 Arab countries. Participants were required to develop models capable of recognizing and categorizing the dialects based on user-generated content on Twitter. For the dataset size, MADAR-26 was around 41600 as a

training set, 5200 as a validation set, and 5200 as a testing set. Regarding Madar-6, it has around 54000 as a training set and 6000 as a validation set.

DART (Dialectal ARabic Tweets) (Bosc, Cabrio, and Villata 2016) It is a large manually-annotated multi-dialect dataset of Arabic tweets. The dataset has about 25000 tweets that are annotated via crowdsourcing, and it is well-balanced over five main groups of Arabic dialects: Egyptian, Darija, Levantine, Gulf, and Iraqi.

IADD (integrated Arabic dialect identification dataset) (Zahir 2022) : It was created from the combination of subsets of five corpora: DART (Bosc, Cabrio, and Villata 2016), SHAMI (Kwaik et al. 2018), TSAC (Mdhaffar et al. 2017), PADIC (Meftouh et al. 2015), and AOC (Zaidan and Callison-Burch 2011) to support the task of automatic Arabic dialect detection. It contains 136,317 texts representing 5 regions: Darija, Levantine, Egyptian, Iraqi, and Gulf.

An open access NLP dataset for Arabic dialects (Boujou et al. 2021) : This dataset was collected from Twitter and other social networks, which is labeled manually and consists of 50000 tweets in five dialects: Algerian, Lebanon, Morocco, Tunisian, and Egyptian. Moreover, this dataset was labeled for several applications: dialect detection, topic detection, and sentiment analysis.

Gulf Tweets (Mubarak, Chowdhury, and Alam 2022) : In this project, they collect a huge number of people who live in Gulf countries and collect their tweets to be used in NLP tasks. It contains over 108,000 tweets from different Gulf countries.

SDC (Saudi Dialect Corpus) and EDC (Egyptian Dialect Corpus) (Tarmom et al. 2020) : It was collected from social media platforms, such as Facebook and Twitter. Both of them were collected to train Saudi and Egyptian models. EDC contains 13,739 Egyptian examples, and SDC contains 14,891 Saudi examples.

DODa (Darija Open Dataset) (Outchakoucht and Es-Samaali 2021) : is an open-source project for the Moroccan dialect. With more than 10,000 entries, it is considered one of the largest open-source collaborative projects for Darija-English translation built for natural language processing purposes.

Multi-Parallel Corpus of North Levantine Arabic (Krubiński et al. 2023) : It is one of the largest Levantine dialect datasets, containing around 120,600 multi-parallel sentences in Levatine, English, French, German, Greek, Spanish, and MSA.

In addition to these efforts, we reference another dataset compilation work for Natural Language Inference (NLI) in Arabic, derived from SNLI, XNLI, and arNLI datasets and was utilized to explore informed pre-training methods for Arabic NLI.

Data Sources

It's important to note that a significant portion of Arabic text available on the internet is written in modern standard Arabic. However, for our research focused on dialectal Arabic, we need to seek out specific sources. To gather dialectal Arabic text, our initial methodology involved extracting data from two primary resources: online dictionaries and articles, but we noticed that the volume of the data collected from these two resources was insufficient for training a language model effectively. So, we have introduced a new approach to enhance our dataset. We now focus on obtaining captions from videos, particularly those on platforms like TikTok. By incorporating this new approach, we aim to diversify our data sources and capture a broader range of dialectal Arabic usage. This strategic shift allows us to explore linguistic variations in informal contexts, enriching models' understanding of dialectal Arabic beyond traditional written sources.

Our data collection process involves three key resources: an online dictionary, articles, and captioned videos.

Online Dictionaries

Online dictionaries play a crucial role in providing lexical information and linguistic insights. Two primary types of online dictionaries for dialectal Arabic exist: crowd-sourced dictionaries and book-based dictionaries. We will delve into both types in the following paragraphs:

Crowd-Sourced Dictionaries: These are web applications where users can contribute by adding new words and providing additional information such as dialect, examples of word usage, and word explanations. While this resource offers valuable linguistic insights into each dialect, it also presents drawbacks, including inaccurate labels, offensive examples, and random text. In our research, we work to minimize these drawbacks as much as possible, resulting in a cleaner dataset. We collect our data from two crowd-sourced dictionaries: Mo3jam[1] and 3amyah[2]

Dictionaries Based on Books: These dictionaries are designed to house digital versions of multiple dictionaries featuring various Arabic dialects translated into English. Their primary purpose is to facilitate word searches and aid in Arabic language learning for non-native speakers. While book-based dictionaries offer a wealth of dialectal Arabic data with English translations, they are not without their shortcomings, such as translation errors and potentially offensive examples. In our research, we have refined the data by removing as much error containing samples as possible to overcome these shortcomings. The main resources for dictionaries based on books are the following:

- **Living Arabic** [3]: Contains over 100,000 examples in all Arabic dialects.
- **Lisan Masry** [4]: Designed for learning the Egyptian dialect with examples and audio pronunciations. It includes around 1500 Egyptian examples with English translations.

[1]https://en.mo3jam.com/

[2]https://3amyah.com/

[3]https://livingarabic.com

[4]https://www.lisaanmasry.org/

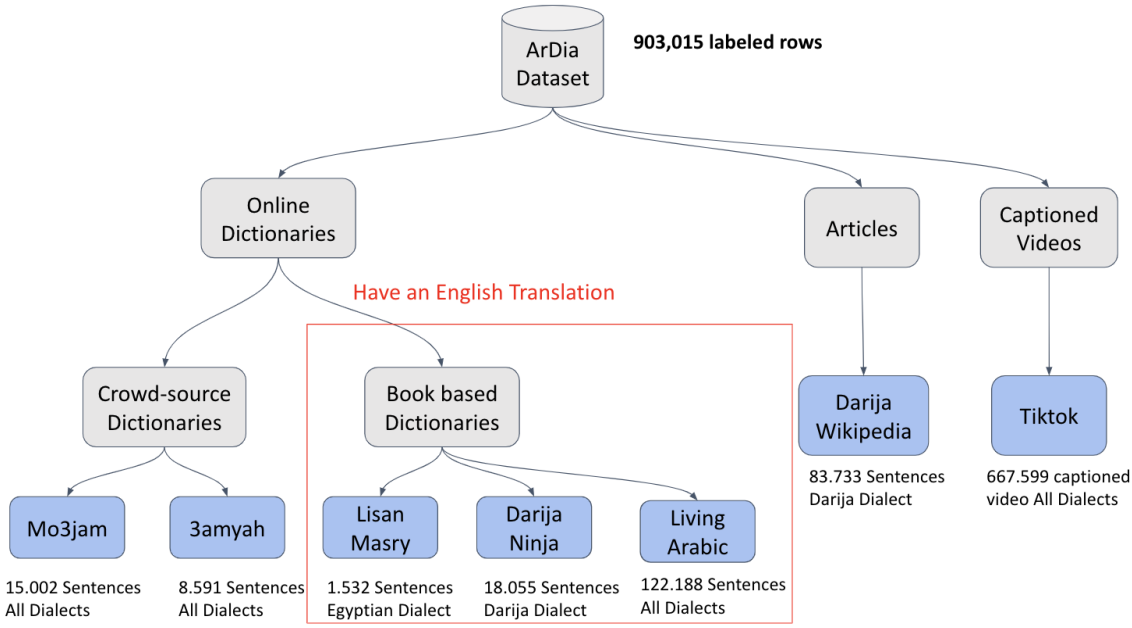

Figure 1: The Resources of the ArDia datasets

- **Darija Ninja** [5]: Designed for searching words in the Darija dialect with examples and translations. It boasts over 18,000 examples with their English translations.

Articles

One of the most important resources in datasets is articles because they have long sentences on different topics. Also, it provides clean data with correct punctuation. The only drawback of getting the text from articles was mixing dialectal Arabic with modern standard Arabic in many places. In our dataset, we get the article data from Darija Wikipedia[6]: which contains over 10,000 articles with more than 80,000 sentences in the Darija dialect.

Videos Caption

Nowadays, videos serve as a valuable resource for collecting natural and spontaneous language usage in dialectal Arabic. By leveraging videos from platforms like TikTok, YouTube, and others, we gain access to a wealth of user-generated content that reflects real-life communication scenarios in different topics like educational tutorials, entertainment, marketing materials, vlogs, or storytelling. In our case, dialectal Arabic data is a spoken language, so video captions are considered the most valuable type of data in our dataset. In addition to that, the text of the video will be pure dialectal Arabic without mixing MSA. This type of data is dependent on the audio recognition quality to get high-quality data, and based on our own observations, we found TikTok provides high quality text generation from Arabic content. Although TikTok contains hundreds of millions of Arabic videos, due to our limited computational resources, we collect around 2,600,000 videos from 7,500 users, which yields around 650,000 captions.

[5]https://derja.ninja//

[6]https://ary.wikipedia.org/

Data Collection Methodology

In this section, we will show the processes that we follow to collect the data from the resources. The data in each resource has a different format, which makes the process of collecting the data different from one resource to another. To ensure ethical compliance and mitigate potential negative outcomes, we anonymized all collected data, employed robust documentation practices, and implemented access controls to safeguard the data. We will go through the process for each type of resource in the following sections.

Collection Crowd-Sourced Dictionaries Data

The process of collecting data from crowd-sourced dictionaries involved web scraping techniques to extract the data from the dictionaries' websites. Our initial step involves navigating to the dictionary URL and starting the scraping process.

The scraping process starts with extracting all available data on the first page of the dictionary, which includes the examples and their corresponding dialect labels. After that, we navigate to the next page of the dictionary and repeat the same procedure. This iterative process continues until reaching the last page of the dictionary.

Once collected, the data is exported to a CSV file, ensuring it is thoroughly documented and anonymized for subsequent model training.

Resource Name	Resource Type	Egyptian Count	Levantine Count	Gulf Count	Iraqi Count	Yemeni Count	Darija Count	Total Count
3amaya	Crowdsourced Dict.	1.024	2.706	4.180	4.618	606	1.868	15.002
Mo3jam	Crowdsourced Dict.	559	1.213	2.862	485	270	3.202	8.591
Living Arabic	Book-based Dict.	30.055	43.696	14.888	**18.984**	2.569	11.996	122.188
Darja Ninja	Book-based Dict.	0	0	0	0	0	18.055	18.055
Lisan Masry	Book-based Dict.	1.532	0	0	0	0	0	1.532
Wikipedia Darija	Articles	0	0	0	0	0	**83.733**	83.733
TikTok Captions	Video Captions	**368.119**	**162.188**	**113.235**	14.365	**9.660**	32	667.599
Total	-	401.289	209.803	135.165	38.452	13.105	118.886	903.015

Table 1: Dataset Counts Overview

Collection Book-based Dictionaries Data

Book-based dictionary websites provide searching for words only. So we collect dictionary data in 3 phases:

Phase 1: Prepare the words to be searched To collect the data from the website, we need to use a dictionary search. So we collect the most frequent 60,000 words in English and Arabic and merge the words into a CSV file to be used for the search.

Phase 2: Scrape the content of the dictionary: After collecting the searching words, we start to search the dictionary word by word for the most frequent words and store the search result words, Arabic examples, corresponding English translations, and the dialect label. In the end, we store the data as a parallel dataset from dialectal Arabic to English with a dialect label.

Phase 3: Translate English examples to modern standard Arabic: In the last phase, we generate a different parallel dataset from dialectal Arabic to modern standard Arabic using the English examples. To do that, we used the Google Translation API to translate the English examples to modern standard Arabic. Throughout this process, we ensured the data was curated to remove artifacts, offensive content, and inaccuracies.

Collection TikTok's Video Captions

Unlike other resources, the data is unlabeled. So our task in this approach was to find a way to label the captions of the videos and then scrape the captions of the videos.

The main challenge in labeling the videos is its large volume which needs a lot of time and effort. To address this, we labeled the user's dialect and inferred that all of their videos share the same dialect. However, this approach introduces potential noise, as some users may post videos in multiple dialects. To mitigate this issue, we carefully selected users based on the following criteria:

- The user is not a fake account.
- The user has a public account.
- The user speaks only one dialect.
- The content of the user is based on talking content, not dancing or reacting content.
- The user enables the caption on their videos.

Collecting the TikTok data contains 3 phases. We will explain each phase in detail as follows:

Phase 1: Collecting and Labeling Arabic Accounts We decided to collect and label the users manually by browsing TikTok using different accounts for each dialect to leverage TikTok's Algorithm to show videos from this respective dialect. As an example, we created an Egyptian account that only interacts with Egyptian content and follows only the people who speak in the Egyptian dialect. another account for the Gulf dialect, and the same thing for other dialects. and we follow the account that satisfies the criteria. After manually collecting a suitable amount of users in the following list, we scrape the list and label all of them based on the used account (Gulf, Egyptian, Darija, Yemeni, and Iraqi accounts). The result of this phase is a list of accounts labeled with the dialect. To preserve privacy, no personal identifiers were collected, and accounts were labeled solely by dialect.

Phase 2: Scrape the users' videos The main goal of this phase is to scrape the users' videos and store them in a new dataset. Also, the dialect label of the video will be the same dialect as the user. This approach was complemented by careful selection to minimize noise from multi-dialect users.

Phase 3: Scrape the videos' caption In this phase, we will scrape the text captions of the videos if they are available. By the end of this phase, we had created a large dataset of captions labeled with the dialect. To further enhance the dataset's quality, we documented the process and ensured that all data were anonymized before inclusion in the final dataset.

Data Analysis

In this section, we will analyze the ArDia dataset to understand its details and extract valuable insights.

Table 2 shows some examples from the dataset which include the Arabic text, the dialect of the text, and the data source of the text.

Dataset Size

Table 1 shows the counts of the dataset in each dialect. We noticed that the data in the crowded source dictionary resources (3amaya and Mo3jam) are quite small compared to the other resources. On the other hand, the book-based

Text	Dialect	Source
اليوم طلبت من مطعم تك وريش شوف شوف المتعه وعلمني اي شيخ من عندهم سعره ٥ ريال ونص الاكل حقهم مره لذيذ التكه الرومي كانت مره رهيبه وريش الغنم لازم تطلبونها من عندهم كان موجود بحي العنود بالدمام الحمد لله احس اني شبعت يبيلها بيبسي دايت جربوها علموني وين راح تندمون	**Gulf**	**TikTok**
نحمّل في فلوس بش نشري سبادي آديداس	**Darija**	**Darja Ninja**
لا تِجْرَح عَوَاطْفَه بْهِيكِي (بْهِيچِي) كَلِمات	**Iraqi**	**Living Arabic**

Table 2: Three Examples From ArDia Dataset

dictionary resources are quite large compared to the crowd-sourced dictionary. In addition to that, book-based dictionary resources provide an English translation of the examples. Moreover, the table shows that the TikTok captions resource provides us with a massive amount of data in all dialects except Darija because the speech recognition of TikTok does not support this dialect.

We could also notice that the TikTok data not only provides us with massive data but also gives us acceptable data in rare dialects like Yemeni and Iraqi, which will help us design models to understand these dialects. The final size of our dataset is around 900,000 labeled texts, which is considered the largest labeled dialectal Arabic dataset.

Resource Name	Resource Type	AVG. Number of Words Per Sample
3amaya	Crowdsourced Dict.	7.14
Mo3jam	Crowdsourced Dict.	11.141
Living Arabic	Book-based Dict.	4.16
Darja Ninja	Book-based Dict.	5.21
Lisan Masry	Book-based Dict.	4.86
Wikipedia Darija	Article	149.37
TikTok	Video Captions	153.27

Table 3: Average Number of Words Per Sample in Different Arabic Resources

Number of Words in Different Resources

The length of text data plays a crucial role in training models to comprehend languages effectively. One limitation faced with many dialectal Arabic datasets is the relatively short text length, often due to the character limit imposed by platforms like Twitter, which is a primary source for such data.

As highlighted in Table 3, resources like crowd-sourced and book-based dictionaries typically contain fewer than 12 words per entry. Contrastingly, resources such as Wikipedia Darija, and TikTok exhibit significantly larger text lengths, averaging around 150 words per entry. This abundance of text on platforms like Wikipedia Darija, and TikTok presents a valuable opportunity for training models to gain a deeper understanding of dialectal Arabic, owing to the richness and diversity of content available.

Data Preprocessing

Preprocessing is one of the most important steps that will improve the performance of the models. We use the same preprocessing used in the Arabic research, such as in (Antoun, Baly, and Hajj 2021) and (Zahir 2022). We will explain the preprocessing steps as follows:

Letters normalization: The main goal of this step is to standardize the writing style by using only one form of the letter. For example, replace all forms of the letter "Alef" "ا آ أ إ" with the standard form "ا". Another example is replacing the letter "ة" with "ه".

Diacritics removal: text often includes diacritics, which are small marks added to letters to indicate vowels or the

word pronunciation. In many natural language processing tasks, diacritics are removed because they can add complexity without significantly affecting the meaning, and most Arabic speakers do not write the diacritic part of the words.

Remove filler noise: Remove filler noise such as "aaaa", or ﻻﻻﻻﻻ in Arabic. This filler noise mainly appears in the videos' captions. it is important to filter out these elements to get clean data.

ArDia Models

In this section, we discuss the methodology that we used for continued pretraining transformer-based models for Arabic dialect classification. For the pretraining process, we aim to enhance the model performance without losing previously learned knowledge. The methodology was inspired by (Ibrahim et al. 2024), which compares different pretraining methods to find the best one. As a summary, Table 4 shows a comparison between all ArDia models.

Model Name	Pretrained Data	Learning Scheduling
ArDiaGPT (cosine)	• ArDia dataset. • Open source datasets.	Cosine with hard restart scheduling with a warmup
ArDiaBERT (cosine)	• ArDia dataset. • Open source datasets.	Cosine with hard restart scheduling with a warmup
ArDiaGPT (balanced)	• ArDia dataset. • Open source datasets. • Duplicated Iraqi and Yemeni data.	Linear scheduling with a warmup
ArDiaBERT (balanced)	• ArDia dataset. • Open source datasets. • Duplicated Iraqi and Yemeni data.	Linear scheduling with a warmup

Table 4: Comparison of ArDia Models

We continued pretrained the encoder-only model ArDiaBERT(cosine)[7] and ArDiaBERT(balanced) [8] based on MARBERT(Abdul-Mageed, Elmadany, and Nagoudi 2021) and the decoder-only model ArDiaGPT(cosine)[9] and ArDiaGPT(balanced)[10] based on ArbianGPT0.3(Koubaa et al. 2024). We used two approaches as follows:

1. **Improved Continued Pretraining Approach: ArDiaBERT(cosine), ArDiaGPT(cosine)**: According to (Ibrahim et al. 2024), further pretraining a model with a new dataset without losing the old data requires integrating 5% of the previously used data into the new dataset. we continued pretrained the based Arabic models on the ArDia dataset and supplemented it with open-source dialectal Arabic datasets that are commonly used in existing dialectal Arabic models. We applied the cosine with a hard restart scheduling strategy with a warmup to manage the learning rate during pretraining. We named the models ArDiaBERT(cosine) and ArDiaGPT(cosine).
2. **Rare Dialect Balancing Approach: ArDiaBERT(balanced), ArDiaGPT(balanced)**: To further improve the performance of our models, especially on rare dialects (Yemeni and Iraqi), the data for these dialects was duplicated (3 times duplication in Iraqi data and 8 times duplication in Yemeni data) to balance it with other dialects. Next, the models were then continued pretrained using a linear scheduling strategy with a warmup. Consequently, we have named these models ArDiaBERT(balanced) and ArDiaGPT(balanced).

Experiments Setup

In this section, we will show the software and hardware requirements for our experiments. Regarding the hardware, we used four NVIDIA A100 GPUs.
For ArDiaBERT models, we follow the MARBERT(Abdul-Mageed, Elmadany, and Nagoudi 2021) configuration that has 12 encoder blocks, 768 hidden dimensions, 12 attention heads, and a maximum sequence length of 512. We ran the experiments for 3 epochs with a 3 x 10^{-5} learning rate, and the optimizer was AdamW. For continued pretraining, we used AutoModelForMLM, which helped us use the Arabic model to perform the masked language model task. Moreover, we chose to use the MARBERT model as a base model and its tokenizer since it was pretrained on dialectal Arabic and it is state-of-the-art in the dialect identification task.

For ArDiaGPT models, we follow the ArabianGPT0.3B (Koubaa et al. 2024) configuration which is based on the GPT2 architecture with 24 decoder blocks, a vocabulary size of 64,000, and a context window size of 1024 tokens. We ran the experiment for 3 epochs with a 3 x 10^{-5} learning rate, and the optimizer was AdamW. For the continued pretraining process, we used AutoModelForCasualLM, which enabled us to train the model with the next word prediction task. Moreover, we chose ArabianGPT0.3B to be the base model, and we used its tokenizer because it was pretrained on dialectal Arabic.

For the dialect identification experiments, We used the ArDia dataset to fine-tune the models on 80% of the dataset on dialect identification and evaluate the models on 20% of the ArDia dataset. We ran the experiments for 10 epochs with a 5 x 10^{-5} learning rate, batch size of 32, a linear scheduler with a warmup, and an optimizer was AdamW.Also, we used AutoModelForSequenceClassification in Pytorch, which adds the classification layer on top of the language model to be used in classification tasks. To make the robust

[7]https://huggingface.co/HossamElsafty/ArDiaBERTv2_cosine
[8]https://huggingface.co/HossamElsafty/ArDiaBERTv2_scaleup
[9]https://huggingface.co/HossamElsafty/ArbianGPT_cosine
[10]https://huggingface.co/HossamElsafty/ArDiaGPT_Scaleup

Model Name	Pretrained Data	Model Type	Darija F1	Gulf F1	Iraqi F1	Yemeni F1	Levantine F1	Egyptian F1	Macro F1
AraBERT(Antoun, Baly, and Hajj 2020)	MSA	BERT	93.94	86.26	74.71	71.13	88.38	94.83	84.87
ARBERT(Abdul-Mageed, Elmadany, and Nagoudi 2021)	MSA	BERT	94.08	86.17	75.07	71.86	88.22	94.88	85.05
AraGPT(Antoun, Baly, and Hajj 2021)	MSA	GPT2	92.38	84.27	71.41	67.63	86.69	94.27	82.78
ArabianGPT0.1B(Koubaa et al. 2024)	DA/MSA	GPT2	93.43	85.17	73.38	69.24	87.72	94.61	83.93
ArabianGPT0.3B(Koubaa et al. 2024)	DA/MSA	GPT2	94.37	86.23	75.70	72.78	88.72	94.96	85.46
ArabianGPT0.8B(Koubaa et al. 2024)	DA/MSA	GPT2	92.94	85.00	73.01	69.06	87.39	94.57	83.66
CAMeLBERT(Inoue et al. 2021)	DA	BERT	94.25	86.63	75.68	73.34	88.72	95.01	85.61
QARiB(Abdelali et al. 2021)	DA	BERT	94.60	87.35	76.40	73.56	88.99	95.22	86.02
MARBERT(Abdul-Mageed, Elmadany, and Nagoudi 2021)	DA	BERT	94.70	87.42	77.20	75.11	89.04	95.29	86.46
ArDiaBERT(balanced)	DA	BERT	94.53	87.41	**77.31**	**75.45**	89.04	95.25	86.50
ArDiaBERT (cosine)	DA	BERT	94.61	**87.61**	77.16	75.08	89.10	**95.32**	86.48
ArDiaGPT(balanced)	DA	GPT2	94.67	87.12	76.74	75.18	89.24	95.09	86.34
ArDiaGPT(cosine)	DA	GPT2	**94.86**	87.45	**77.31**	74.95	**89.48**	95.26	**86.55**

Table 5: Evaluation of Arabic models on dialect identification on the ArDia dataset using a classification layer. MSA stands for modern standard Arabic, and DA stands for dialectal Arabic.

evaluation result, we performed the fine-tuning 5 times with a different seed for each model and calculated the average F1 score of the 5 experiments as the F1 score.

Results

The result of our research focuses on showing the performance of the Arabic models on the dialect identification task using the ArDia dataset.

We present our evaluation using the ArDia dataset in Table 5. We explain our observation using the following points:

Comparing the models based on the pretrained data: According to Table. 5, the best model that is pretrained on modern standard Arabic is ARBERT, which is less than all BERT models that are pretrained on dialectal Arabic (QARiB, CAMeLBERT, ArDiaBERT, and MARBERT). The same observation is applied to the GPT model since the ArabianGPT models that were pretrained on modern standard Arabic and dialectal Arabic outperformed the AraGPT2 model that was pretrained on modern standard Arabic. According to this observation, the models that were pretrained on dialectal Arabic outperformed the models that were pretrained on modern standard Arabic.

Comparing the models based on the model type If we excluded ArDia models, we can tell from Table. 5 that BERT-based models outperformed GPT-based models on dialect identification task. We could notice it in models which are pretrained on modern standard Arabic AraBERT and ARBERT outperformed AraGPT. Also, CAMeLBERT, QARiB, and MARBERT outperformed ArabianGPT models. Regarding, ArDia models, we successfully continued pretrained GPT-based model which outperformed all Arabic models on ArDia dataset.

Comparing ArDia models with their baseline: Based on Table. 5, ArDiaBERT models were contained pretrained based on MARBERT. Although ArDiaBERT models have better F1 scores than MARBERT, it isn't considered a significant improvement. ArDiaBERT(balanced) and ArDiaBERT(cosine) improved their performance by 0.04% and 0.02% in F1 scores respectively.

Regarding ArDiaGPT models which were contained pretrained based on ArbianGPT0.3, they have a significant improvement by increasing the F1 scores by 1.09% in ArDiaGPT(cosine) and 0.88% in ArDiaGPT(balanced).

Comparing models based on dialects' scores: We show in the Table. 5 the details of each dialect score, which could help us deeply understand the models' performance. According to that, ArDiaGPT(cosine) achieved the best F1 score in Darija, Iraqi, and Levantine dialects. Moreover, ArDiaBERT(balanced) achieved the best F1 score in rare dialects like Yemeni and Iraqi dialects. ArDiaBERT(cosine) achieved the best F1 score in the Gulf and Egyptian dialects.

Overall comparison: Based on our evaluation in Table 5, ArDia models outperformed the Arabic models on the dialect identification task. The ArDiaGPT(cosine) achieved the best performance compared with other ArDia models.

Conclusion

In conclusion, this paper has made notable advancements in natural language processing for dialectal Arabic through several key contributions. By introducing a new approach to collecting dialectal Arabic data using TikTok videos' captions, we have successfully built the largest labeled dialectal Arabic dataset. This dataset is named the "ArDia" dataset and contains more than 900,000 examples. Also, it provides a comprehensive resource for dialectal Arabic, including less common dialects such as Yemeni and Iraqi, which are often underrepresented in other data sources.

TikTok has proven to be an invaluable resource for gathering diverse dialectal data, especially for dialects not commonly found in traditional datasets. This novel approach has allowed us to capture a wide range of linguistic variations and nuances. In addition to that, it provides high-quality and long text, which will be a key point in getting more robust language models.

Another major contribution of this research paper is the development of new transformer-based models: ArDiaBERT and ArDiaGPT. These models have been continued pretrained on pure dialectal Arabic data, which will help the model have a good performance at the dialectal Arabic iden-

tification task.

To ensure the reproducibility of findings, we have made the ArDia dataset, pretrained models, and training configurations available under a responsible release policy. The computational environment, hyperparameters, and data processing scripts are fully documented to support transparency and enable replication of our results.

A possible ethical concern in the context of our data collection method is the use of data from TikTok without explicit consent of the users. We do not consider this a major issue, as all samples are publicly available online, and we are not violating any of TikTok's terms of use.

Outlook

Future research could focus on leveraging TikTok as a valuable resource for collecting extensive dialectal Arabic data. With advanced computational capabilities, a robust pipeline could be developed to harvest large-scale data from TikTok, enriching datasets, capturing a wider range of dialectal variations, and significantly improving the robustness and adaptability of language models.

Due to computational limitations, this study utilized relatively small BERT-based and GPT-based models. Future research should explore the capabilities of larger and more powerful models, such as JAIS (Sengupta et al. 2023), ACEGPT (Huang et al. 2023), and Llama (Touvron et al. 2023). Pretraining these models on large-scale data derived from video captions has the potential to enhance their performance across various Arabic language tasks, including toxicity detection, machine translation, emotion analysis, and paraphrasing.

By ensuring ethical research practices through anonymization, access control, and comprehensive documentation, this study lays the foundation for future advancements in Arabic NLP while promoting transparency, accountability, and reproducibility in the research community.

Furthermore, we are planning to make the ArDia dataset publicly available. Upon release, it will adhere to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, ensuring that the data is discoverable through persistent identifiers, accessible via a recognized repository with clear licensing terms, interoperable with standardized formats and ontologies, and reusable with comprehensive documentation for future research.

Acknowledgments

This research has been partially funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence.

We would like to thank the University of Bonn for allowing us to access the university's HPC cluster to perform our experiments.

References

Abdelali, A.; Hassan, S.; Mubarak, H.; Darwish, K.; and Samih, Y. 2021. Pre-Training BERT on Arabic Tweets: Practical Considerations.

Abdul-Mageed, M.; Elmadany, A.; and Nagoudi, E. M. B. 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)*, 7088–7105. Online: Association for Computational Linguistics.

Abdul-Mageed, M.; Elmadany, A.; Zhang, C.; Nagoudi, E. M. B.; Bouamor, H.; and Habash, N. 2023. NADI 2023: The Fourth Nuanced Arabic Dialect Identification Shared Task. *arXiv preprint arXiv:2310.16117*.

Antoun, W.; Baly, F.; and Hajj, H. 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, 9.

Antoun, W.; Baly, F.; and Hajj, H. 2021. AraGPT2: Pre-Trained Transformer for Arabic Language Generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 196–207. Kyiv, Ukraine (Virtual): Association for Computational Linguistics.

Bosc, T.; Cabrio, E.; and Villata, S. 2016. DART: A dataset of arguments and their relations on Twitter. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, 1258–1263.

Bouamor, H.; Hassan, S.; and Habash, N. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, 199–207.

Boujou, E.; Chataoui, H.; Mekki, A. E.; Benjelloun, S.; Chairi, I.; and Berrada, I. 2021. An open access NLP dataset for Arabic dialects: Data collection, labeling, and model construction. *arXiv preprint arXiv:2102.11000*.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2025-01-04.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Huang, H.; Yu, F.; Zhu, J.; Sun, X.; Cheng, H.; Song, D.; Chen, Z.; Alharthi, A.; An, B.; Liu, Z.; et al. 2023. AceGPT, Localizing Large Language Models in Arabic. *arXiv preprint arXiv:2309.12053*.

Ibrahim, A.; Thérien, B.; Gupta, K.; Richter, M. L.; Anthony, Q.; Lesort, T.; Belilovsky, E.; and Rish, I. 2024. Simple and scalable strategies to continually pre-train large language models. *arXiv preprint arXiv:2403.08763*.

Inoue, G.; Alhafni, B.; Baimukan, N.; Bouamor, H.; and Habash, N. 2021. The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Online): Association for Computational Linguistics.

Koubaa, A.; Ammar, A.; Ghouti, L.; Najar, O.; and Sibaee, S. 2024. ArabianGPT: Native Arabic GPT-based Large Language. *arXiv preprint arXiv:2402.15313*.

Krubiński, M.; Sellat, H.; Saleh, S.; Pospíšil, A.; Zemánek, P.; and Pecina, P. 2023. Multi-Parallel Corpus of North Levantine Arabic. In Sawaf, H.; El-Beltagy, S.; Zaghouani, W.; Magdy, W.; Abdelali, A.; Tomeh, N.; Abu Farha, I.; Habash, N.; Khalifa, S.; Keleg, A.; Haddad, H.; Zitouni, I.; Mrini, K.; and Almatham, R., eds., *Proceedings of Arabic-NLP 2023*, 411–417. Singapore (Hybrid): Association for Computational Linguistics.

Kwaik, K. A.; Saad, M.; Chatzikyriakidis, S.; and Dobnik, S. 2018. Shami: A corpus of levantine arabic dialects. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Mdhaffar, S.; Bougares, F.; Esteve, Y.; and Hadrich-Belguith, L. 2017. Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In *Third Arabic natural language processing workshop (WANLP)*, 55–61.

Meftouh, K.; Harrat, S.; Jamoussi, S.; Abbas, M.; and Smaili, K. 2015. Machine translation experiments on PADIC: A parallel Arabic dialect corpus. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*, 26–34.

Mubarak, H.; Chowdhury, S. A.; and Alam, F. 2022. Arabgend: Gender analysis and inference on arabic twitter. *arXiv preprint arXiv:2203.00271*.

Outchakoucht, A.; and Es-Samaali, H. 2021. Moroccan dialect-darija-open dataset. *arXiv preprint arXiv:2103.09687*.

Sengupta, N.; Sahu, S. K.; Jia, B.; Katipomu, S.; Li, H.; Koto, F.; Afzal, O. M.; Kamboj, S.; Pandit, O.; Pal, R.; et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Tarmom, T.; Teahan, W.; Atwell, E.; and Alsalka, M. A. 2020. Compression versus traditional machine learning classifiers to detect code-switching in varieties and dialects: Arabic as a case study. *Natural Language Engineering*, 26(6): 663–676.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Zahir, J. 2022. IADD: An integrated Arabic dialect identification dataset. *Data in Brief*, 40: 107777.

Zaidan, O.; and Callison-Burch, C. 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*, 37–41.

Ethics Checklist

(a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see section "Data Collection Methodology"

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see section "Data Analysis"

(e) Did you describe the limitations of your work? No, but since this is a dataset paper, it should be clear that the main goal is to introduce a novel data source and not to significantly improve upon the state of the art.

(f) Did you discuss any potential negative societal impacts of your work? No, because we did not see any obvious negative societal impact our work could have.

(g) Did you discuss any potential misuse of your work? No, because we did not see any possible directions of misuse.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see section "Data Collection Methodology"

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

(j) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes

(k) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see section "Experiments Setup"

(l) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, due to constraints on time and computing resources.

(m) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see section "Experiments Setup"

(n) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see section "Results"

(o) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No, because we did not consider this very relevant, as it is not a critical application.

(p) If your work uses existing assets, did you cite the creators? Yes, see sections "Related Work" and "Data Sources"

(q) Did you mention the license of the assets? Not explicitly, but we provide the links under which those are available. (see section "Data Sources")

(r) Did you include any new assets in the supplemental material or as a URL? Yes

(s) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, we did not consider this relevant as it is all publicly available data that we are re-using.

(t) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No, we did not consider this relevant as it is all publicly available data.

(u) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes, See sections "Outlook"

(v) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Not yet, but we are planning to do that upon publication of this work.