International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-023-00498-2

REGULAR PAPER

DarijaBERT: a step forward in NLP for the written Moroccan dialect

Kamel Gaanoun[1] · Abdou Mohamed Naira[1,2] · Anass Allak[1,2] · Imade Benelallam[1,2]

Received: 7 February 2023 / Accepted: 12 December 2023

Abstract
The established performance of existing transformer-based language models, delivering state-of-the-art results on numerous downstream tasks, is noteworthy. However, these models often face limitations, being either confined to high-resource languages or designed with a multilingual focus. The availability of models dedicated to Arabic dialects is scarce, and even those that do exist primarily cater to dialects written in Arabic script. This study presents the first BERT models for Moroccan Arabic dialect, also known as Darija, called DarijaBERT, DarijaBERT-arabizi, and DarijaBERT-mix. These models are trained on the largest Arabic monodialectal corpus, supporting both Arabic and Latin character representations of the Moroccan dialect. Their performance is thoroughly evaluated and compared to existing multidialectal and multilingual models across four distinct downstream tasks, showcasing state-of-the-art results. The data collection methodology and pre-training process are described, and the Moroccan Topic Classification Dataset (MTCD) is introduced as the first dataset for topic classification in the Moroccan Arabic dialect. The pre-trained models and MTCD dataset are available to the scientific community.

Keywords BERT · Moroccan dialect · Arabic · NLP · Text mining · Low resource language

1 Introduction

The utilization of "transformers" [1] and language models (LMs) in scientific literature related to natural language processing (NLP) has witnessed a substantial increase in recent years, particularly highlighting the BERT model (Bidirectional Encoder Representations from Transformers) [2]. However, the widespread use of BERT and analogous models is predominantly limited to high-resource languages, including English (BERT), French (CamemBERT [3]), Spanish (BETO [4]), and a multilingual BERT model trained on Wikipedia across 104 languages. This discrepancy is attributed to the wealth of available data for these languages, a circumstance not mirrored in the case of low-resource languages.

The Arabic language stands out as a prime example of a low-resource language, lacking dedicated models, especially tailored to its diverse dialects. Presently, the landscape includes just three monodialectal BERT models for Arabic: SudaBERT [5], TunBERT [6], and DziriBERT [7]. Other models are either multidialectal or exclusively focused on modern standard Arabic (MSA), significantly distinct from dialectal Arabic (DA). Although there exist multidialectal models encompassing various dialects, incorporating certain aspects of the Moroccan dialect, the representation of the Moroccan dialect in existing language models remains constrained. The scarcity of models dedicated to the Moroccan dialect limits the available options, compelling researchers to predominantly rely on multilingual or multidialectal models for language understanding and processing tasks.

Existing multilingual models have a restricted focus on representing the Arabic language, let alone its dialects [8]. Whereas multidialectal models lack the specificity needed for an accurate representation of the Moroccan dialect, often leading to the loss of dialect-specific features. Moreover, there is no assurance of a satisfactory representation of dialectal vocabulary, given the substantial differences among Arabic dialects.

✉ Kamel Gaanoun
kgaanoun@insea.ac.ma

Abdou Mohamed Naira
nabdoumohamed@insea.ac.ma

Anass Allak
aallak@insea.ac.ma

Imade Benelallam
i.benelallam@insea.ac.ma

[1] SI2M Lab, INSEA, Rabat-Instituts, Rabat, Morocco

[2] AIOX LABS, rue Honain, Rabat, Morocco

Published online: 23 January 2024

In Morocco, MSA is the language used in official domains and taught in schools, while Darija, a blend of MSA, Amazigh, French, and Spanish, serves as the vernacular language widely spoken in everyday life. Previously solely spoken, Darija has recently acquired a written form due to the widespread use of social networks and increased access to technology. Nevertheless, owing to its recent emergence in written form, Darija lacks standardization in its written format and lacks established grammatical or syntactic rules.

Hence, implementing NLP applications that utilize text written in the Moroccan dialect necessitates a specialized BERT model. This paper presents three BERT models specifically developed for the dialect: DarijaBERT, DarijaBERT-arabizi, and DarijaBERT-mix, marking the initial implementation of BERT models exclusively dedicated to Darija, regardless of its written form in either Arabic or Latin script. Furthermore, the paper presents a benchmark of existing datasets containing Darija texts and introduces the first dataset specifically crafted for topic classification in this dialect.

The contributions of this paper are:

- Development of DarijaBERT, the first Moroccan dialect transformer language model, with three different variants:
  - DarijaBERT: Trained on Moroccan dialect written in Arabic letters
  - DarijaBERT-arabizi: Trained on Moroccan dialect written in Latin letters
  - DarijaBERT-mix: Trained on a larger dataset including both Arabic and Latin letters
- Introduction and release of MTCD, the first annotated dataset for Moroccan Arabic topic classification
- Fine-tuning and application of the models to:
  - Dialect identification
  - Sentiment analysis
  - Sarcasm automatic detection
  - Topic classification
- Release of DarijaBERT models on Github and Huggingface Hub.

The paper is organized in the following sections: Sect. 2 offers an overview of related work, Sect. 3 describes the approach used for data collection and models pre-training, Sect. 4 presents the evaluation process, Sects. 5 and 6 present and discuss the results, and the work is finally summarized in Sect. 7.

2 Related work

Arabic, as a language, displays diglossia, where MSA is employed in formal settings like communication, newspapers, and education, while the dialects dominate everyday life and social media. Designing NLP systems specifically for Arabic dialects poses challenges due to the limited availability of data and the intricate syntax and morphology of these dialects. Moroccan Arabic exhibits syntactic complexity due to the utilization of diverse sentence structures, as outlined by Meftouh et al. [9]. On the morphological dimension, the application of various affixes in Moroccan Arabic contributes to the formation of more intricate lexemes compared to MSA. An illustrative instance is the verb كتب (to write), which can be modified to كايكتب (he is writing) or غايكتب (he will write) through the introduction of affixes كا and غا in conjunction with يـ. In contrast, MSA represents these verbs as يكتب and سيكتب. Furthermore, the incorporation of a straightforward negation in Moroccan Arabic involves greater morphological complexity than its MSA counterpart. In the former, the verb transforms into ما كايكتبش (he is not writing) entailing the addition of four affixes, whereas in MSA, it is rendered as لا يكتب.

While the first BERT model was published in 2018, it was not until 2020 that the first MSA-specific models, such as AraBERT [10] and ArabicBERT [11], were introduced. The first dialect-specific models [5–7] were only published in 2021, three years after the release of the original BERT model.

AraBERT was the first model specifically designed for MSA, trained on a 23-GB text corpus consisting of approximately 3 billion words. It served as the reference model for MSA until the publication of ARBERT [12], a subsequent MSA-specific model that was trained on a larger corpus of 61 GB (6.5 billion tokens) and achieved state-of-the-art results on downstream MSA-related tasks. The authors of ARBERT later introduced a multidialectal model, called MARBERT [12], trained on a corpus of 128 GB (15.6 billion tokens) incorporating texts from various Arabic dialects. Although multilingual models like mBERT and XLM-RoBERTa [13] are available for other languages, the scarcity of monodialectal models for Arabic persists as a substantial concern.

SudaBERT, TunBERT, and DziriBERT are the only existing models in this sense for Sudan, Tunisia, and Algeria, respectively, among the 22 Arab countries. These models were trained on various sources of text data, including 7 million sequences from Twitter and Telegram public channels for SudaBERT, 500 thousand Tunisian social media comments for TunBERT, and 1 million tweets for DziriBERT. Currently, there is a lack of a dedicated model for the Moroccan dialect. The use of a multidialectal model such as MARBERT presents two significant limitations. Firstly, the support for multiple dialects results in reduced coverage of dialect-specific features and vocabulary. Secondly, the representation of the Moroccan dialect in the MARBERT dataset is expected to be inadequate, given that Twitter is more widely used in Gulf countries than in Morocco. Additionally, CAMeLBERT [14] offers three different models for MSA, dialectal, or classical Arabic, while Qarib [15] is another model that is based on both MSA and multidialectal corpora.

It is worth noting that, except for DziriBERT, which includes some text written in Latin characters (the proportion of which is unknown), all other models exclusively recognize dialects written in Arabic characters. This poses a disadvantage for these models, as they neglect a crucial form of the Arabic dialect commonly employed on social networks, known as Arabizi.

3 Methodology

3.1 Data collection

Having resources like Wikipedia and news articles for training BERT models is crucial for high-resource languages like English. Unfortunately, this is not the scenario for low-resource languages such as Darija. Furthermore, considering the distinct situations and contexts in which dialects are utilized compared to MSA is crucial when collecting data in Darija. This requires a careful consideration of the diversity and representativeness of the texts, while ensuring that a sufficient amount of data is collected to effectively train the models.

The written form of the Moroccan dialect is predominantly found on social media and the internet and can be written in either Arabic or Latin characters (referred to as Arabizi in the latter case, see Sect. 3.4.2). Hence, three sources were selected for this purpose: YouTube, Twitter, and "9essas",[1] a forum website specializing in stories written in Darija.[2] Following that, a collection method tailored to each of these sources was adopted.

For the 9essas website, the process involved straightforward scraping of the various stories from the website, given their exclusive use of Darija. The scraped text was divided into homogeneous sentences and through experimentation, it was established that sentence sequences of approximately 40 words were optimal. As a result of this process, a dataset, named "Stories" was formed, which consisted of approximately one million sequences.

For YouTube, comments were scraped from the most popular videos in Morocco as identified through Social Blade[3] and HypeAuditor.[4] A total of 46,106,073 comments were collected from 40 different channels. Channels primarily focused on music, religious content, or non-Moroccan content were excluded due to the predominantly song-related or mixed language nature of the comments in these channels. Subsequently, considering the mixture of comments in MSA, French, English, and other languages, comments written primarily in Latin characters were filtered out. Constructing the final YouTube dataset involved a two-step process. In the first step, a logistic regression classification model was developed using a dataset of 560,000 Darija comments randomly selected from Moroccan channels and 560,000 MSA sequences from the Sabanews dataset [16], this model was 70% accurate. To further enhance the model's performance, the logistic regression model was applied iteratively to the data, retaining only sequences with a prediction probability of at least 90%. The prediction probability, derived from the softmax output of the logistic regression model, represents the model's confidence in classifying a given sequence as belonging to the Moroccan dialect. Sequences that met or exceeded the 90% prediction probability threshold were considered highly likely to be in Darija and were included in the training dataset for subsequent iterations. This iterative process, often referred to as a self-learning method, allowed the model to learn from its own predictions and gradually improve its ability to accurately classify Darija comments. This method resulted in the retrieval of 4 million comments in Darija. For the remaining comments, with a probability below 90%, a rule-based learning approach was applied. The selection of keywords for this step involved excluding commonly used words in MSA to focus on capturing the unique features of the Moroccan dialect. Multiple preprocessing techniques were applied, including removing punctuation marks, filtering out short words, and excluding laughing interjections. From the preprocessed dataset, the most frequently used non-MSA words in the Stories dataset were identified. The top 350 keywords were selected, representing the most commonly used non-MSA words specific to the Darija dialect. This step resulted in the retrieval of an additional 2 million comments.

[1] https://9isas.modareb.info/.

[2] No copyright claims have been made by the anonymous writers or organizers, and we explicitly state that we do not assert any copyrights to the text. The collected data is exclusively utilized for training models and is not disseminated or shared in any manner.

[3] https://socialblade.com/.

[4] https://hypeauditor.com/.

In the second step, a sample of the total 6 million comments (4 million from the self-learning method and 2 million from the rule-based learning process) along with Stories and a sample of Twitter data were used to train DarijaBERT (see Sect. 3.4.1). Subsequently, DarijaBERT was utilized as a filtering mechanism for the initial dataset of 46,106,073 YouTube comments, aiming to obtain more refined predictions. In this filtering process, only comments identified as Darija, with a prediction threshold of at least 80%, were retained. Choosing this threshold aimed to ensure the inclusion of a broader spectrum of valid Darija comments while maintaining a high level of accuracy in their classification. Notably, owing to DarijaBERT's superior classification performance compared to logistic regression, it enables the extraction of more precise Darija sequences even at lower thresholds. As a result, a total of 5.5 million YouTube comments written in Darija were retrieved. To ensure the validity of the collected dataset, manual validation is performed on a subset of 1,000 comments. This subset consisted of five samples, each comprising 200 comments, and was evaluated by a native Darija speaker. The evaluation resulted in an average accuracy rate of 92%(± 3%).

The process of filtering comments written in Latin characters was distinct from filtering those written in Arabic script. In addition to Arabizi (Darija written in Latin letters), comments written in French were commonly included in this subset. To extract Arabizi comments, the LangDetect Python library was utilized to remove comments identified as French. The resulting dataset, referred to as Arabizi, consisted of 4.6 million comments.

For Twitter, The list of keywords from the Stories dataset was refined to create a more focused selection of 31 words specific to the Moroccan dialect[5]. This refinement aimed to improve the relevance and quality of the collected data by focusing on Moroccan-specific topics and keywords that are likely to capture Darija tweets. In contrast to YouTube comments, which require post-gathering filtering, on Twitter, Darija tweets were directly collected using the selected keywords, eliminating the need for additional filtering. The keyword selection process for the Twitter comments involved excluding common keywords shared between the Algerian and Moroccan dialects, as well as MSA, focusing on specific keywords unique to the Moroccan dialect. Indeed, the keyword selection process for the Twitter comments closely aligned with the tweet retrieval process. The observation was made that certain keywords predominantly retrieved Moroccan tweets, while others yielded a mix of Moroccan and Algerian tweets. This indicated that the selected keywords were not specific enough to the Moroccan dialect. Based on this observation, the decision was made to exclude keywords that were shared between the Moroccan and Algerian dialects[6]. Furthermore, Very short keywords[7] and ambiguous words were eliminated to ensure clarity and reliability of the dataset. To validate the collected data, a manual validation was conducted on a subset of 1,000 tweets, consisting of five samples of 200 tweets each. The tweets were evaluated by a native Darija speaker, resulting in an average accuracy rate of 94%(±2%). As a result of this methodology, a corpus of 3 million tweets written in Darija was successfully collected.

3.2 Data preprocessing

The preprocessing steps applied to the sequences aimed to preserve their similarity to the original sequences while ensuring data standardization. Specifically, the Arabizi dataset underwent the following modifications: Repeated characters were consolidated into a single instance, hashtags, user mentions, and URLs were replaced with the tokens "HASHTAG," "USER," and "URL," respectively. Only sequences with a minimum of three Latin words were retained and all Latin words were converted to lowercase. A similar methodology was applied to the dataset containing Arabic letters, with the exception that only sequences containing a minimum of two Arabic words were retained, and the Tatweel character (letter elongation) and diacritics were removed (as illustrated in Table 1). Although sequences may contain non-Arabic words, they were maintained in order to preserve the overall meaning of the sequences. Attempts to train a model using a dataset that excluded non-Arabic words led to suboptimal performance and were consequently discarded. The evaluation of the training process output involved native Darija speakers participating in MASK filling tasks on diverse Darija sentences. The model configured in this manner encountered challenges in delivering contextually accurate and semantically meaningful completions for masked words in a majority of instances.

[5] كاتشوف، كيضحك، زوينة، كتبكي، مزيان، داكشي، كيشوف، كتشوف، واخا، كيزيدو، دابا ديال، الجلاخة، تبوگيصة، مكلخ، حشومة، منبقاوش، شلاهبية، تخربيق، كايدوي، برهوش، كاندوي يسيفطوه، يصيفطوه، السماسرية، مكاينش، مزيانين، الفقصة، زوينين، سيمانة، الدراري **English translation** you see, he's laughing, beautiful, she's crying/you're crying, well, that, he's watching, she's watching/you're watching, okay, they're adding, now, of, disgusting person, beauty, stupid, shame, we don't stay anymore/we won't continue to, thugs, gibberish, he speaks, little kid, I speak, they send him, they send him, the commercial intermediaries, there is no, good, frustration, beautiful, a week, the children/the boys.

[6] Some Algerian/Moroccan words: كارطون، ماقراتش، ندير

[7] Some short keywords: زم، فك، رشّ، هزّ

Table 1 Examples of preprocessing steps

Raw text	Preprocessed text
مشا ليا التيـــلفووون	مشا ليا التيلفون
أنا كـلّما قاليا شي واحد نسكت عاد كاندوينّيـــــت	أنا كلما قاليا شي واحد نسكت عاد كاندوي نيــت

3.3 Pre-training setup

The three models employed the same architecture, which was based on the BERT-base architecture [2]. The architecture consisted of 12 encoder blocks, 768 hidden units, and 12 attention heads. Whole-word masking was applied during the training process with a 15% replacement probability. Tokens selected for replacement were substituted with the [MASK] token in 80% of cases, a random token in 10% of cases, and the original token in the remaining 10%. The batch size was set to 512, and the maximum sequence length was fixed at 128. It is worth noting that in the conducted experiments, only the masked language modeling (MLM) task was utilized, omitting the Next Sentence Prediction (NSP) task. This decision was based on compelling evidence from recent studies [17–19] illustrating the minimal impact of the NSP task on model performance. Also, given the independent nature of the majority of the sequences in the dataset, incorporation of the NSP task was deemed unnecessary.

A WordPiece tokenizer [20] was utilized to generate 80k, 110k, and 160k tokens for DarijaBERT, DarijaBERT-arabizi, and DarijaBERT-mix, respectively. The models were trained using Google's Tensorflow Research Cloud (TRC[8]) TPU v3.8 with the aid of the HuggingFace trainer. The learning rate for all models is set to $1e^{-4}$. The hyperparameter settings were chosen in accordance with the recommendations from the original BERT paper to ensure consistency with established practices and achieve optimal performance. The batch size was based on available memory capacity. Additionally, considering the nature of our data, which comprises short sequences like comments and tweets, we set the sequence length to 128.

3.4 Models

3.4.1 DarijaBERT

The first model developed was DarijaBERT, trained specifically on sequences composed of Arabic letters. The training dataset for this model was created using a sample of 1 million tweets and 1 million YouTube comments obtained during the initial filtering step (Sect. 3.1), along with all sequences from the Stories dataset. This resulted in a dataset consisting of 3 million sequences with a total size of 691 MB (as depicted in Fig. 1). A WordPiece tokenizer was trained on the dataset producing an 80k-token vocabulary. The training process lasted 49 h, encompassing 234,800 steps or 40 epochs, resulting in a model with 147 million parameters. The model is publicly available on the HuggingFace models' hub.[9]

3.4.2 DarijaBERT-arabizi

Arabizi [21–23], also known as Latinized Arabic [24, 25] or Arabic chat Alphabet [26, 27], is a form of writing Arabic that uses Latin letters and Arabic numbers. It emerged in the early 1990 s as a means of communication among young people in Arab countries who were using new technology that did not support the Arabic alphabet. Latin scripts were used as a solution due to the similarity in pronunciation between some Arabic and Latin characters, along with the use of Arabic digits that resemble specific Arabic characters. This mode of communication is still prevalent today, particularly on social networks. Table 2 contains a few samples of Arabizi sentences and their equivalents in Arabic letters along with English translation. Therefore, the Arabizi dataset, consisting of 5 million Darija sequences written in Latin letters (287 MB), was used to create a dedicated model named DarijaBERT-arabizi (see Fig. 2). Initially, the model was trained using a vocabulary of 80k tokens; however, the results of mask filling tests were not satisfactory. The model was unable to predict the tokens that give meaning to the test sentences, as confirmed by experiments conducted by two native speakers of Darija. This could be due to the various styles of writing that the same word can take in Arabizi, as shown in Table 3. Therefore, the vocabulary size was expanded to 110k and satisfactory results were obtained. The model was trained for 60 h, 364,280 steps equivalent to 40 epochs. The model has 170 million parameters in total and is publicly available on the HuggingFace models' hub.[10]

3.4.3 DarijaBERT-mix

To support both Arabic and Arabizi writing styles, DarijaBERT-mix was developed. The model was trained on a total of around 14 million sequences, comprising 9.5 million sequences in Arabic letters and 4.6 million in Arabizi from the Arabizi dataset.

[8] https://sites.research.google/trc/about/.

[9] https://huggingface.co/SI2M-Lab/DarijaBERT.

[10] https://huggingface.co/SI2M-Lab/DarijaBERT-arabizi.

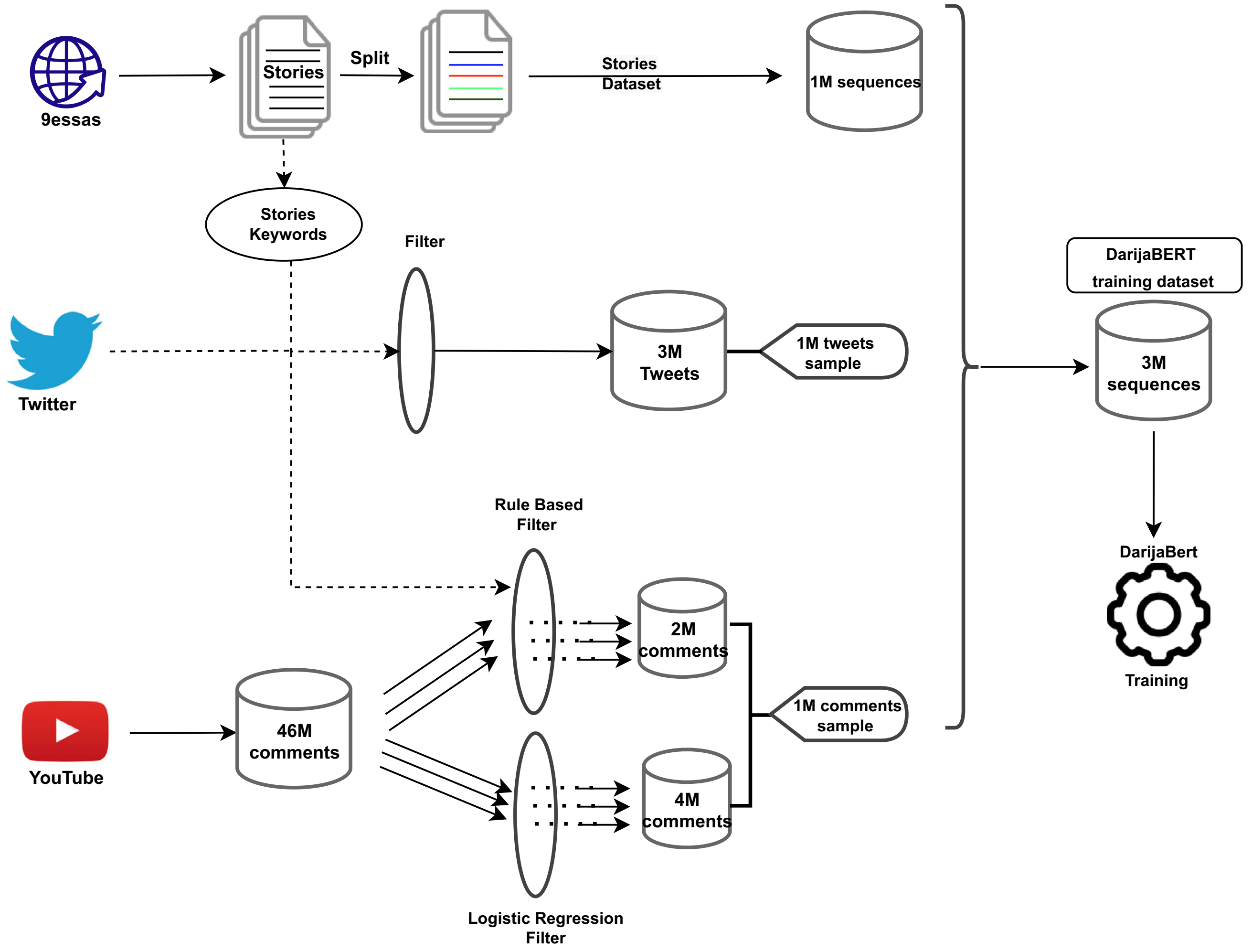


Fig. 1 Data collection process for DarijaBERT training dataset

Table 2 Examples of Arabizi sequences

Moroccan dialect	Arabizi	English translation
مشيتالبارحللسوق وشريتالتفاح	Mchit lbare7 lssou9 w chrit ttffa7	I went to the market yesterday and bought some apples
بابا شرا ليا لعبة زوينة	Baba chra lia lo3ba zwina	My father bought me a nice toy
هداك الولد مأدب	8adak lweld m2ddeb	This boy is well educated

The 9.5 million sequences were obtained using DarijaBERT instead of logistic regression to filter out the Youtube comments not used for DarijaBERT (see Fig. 2). The DarijaBERT model trained on a smaller volume of data (containing the initial 3 million sequences) was used to produce higher quality data for training DarijaBERT-mix. Built on a vocabulary of 160k tokens, this model underwent training for approximately 96 h, covering 510,000 steps or about 10 epochs. The number of parameters in this model is 209 million. DarijaBERT-mix is also available on the HuggingFace models' hub.[11]

[11] https://huggingface.co/SI2M-Lab/DarijaBERT-mix.

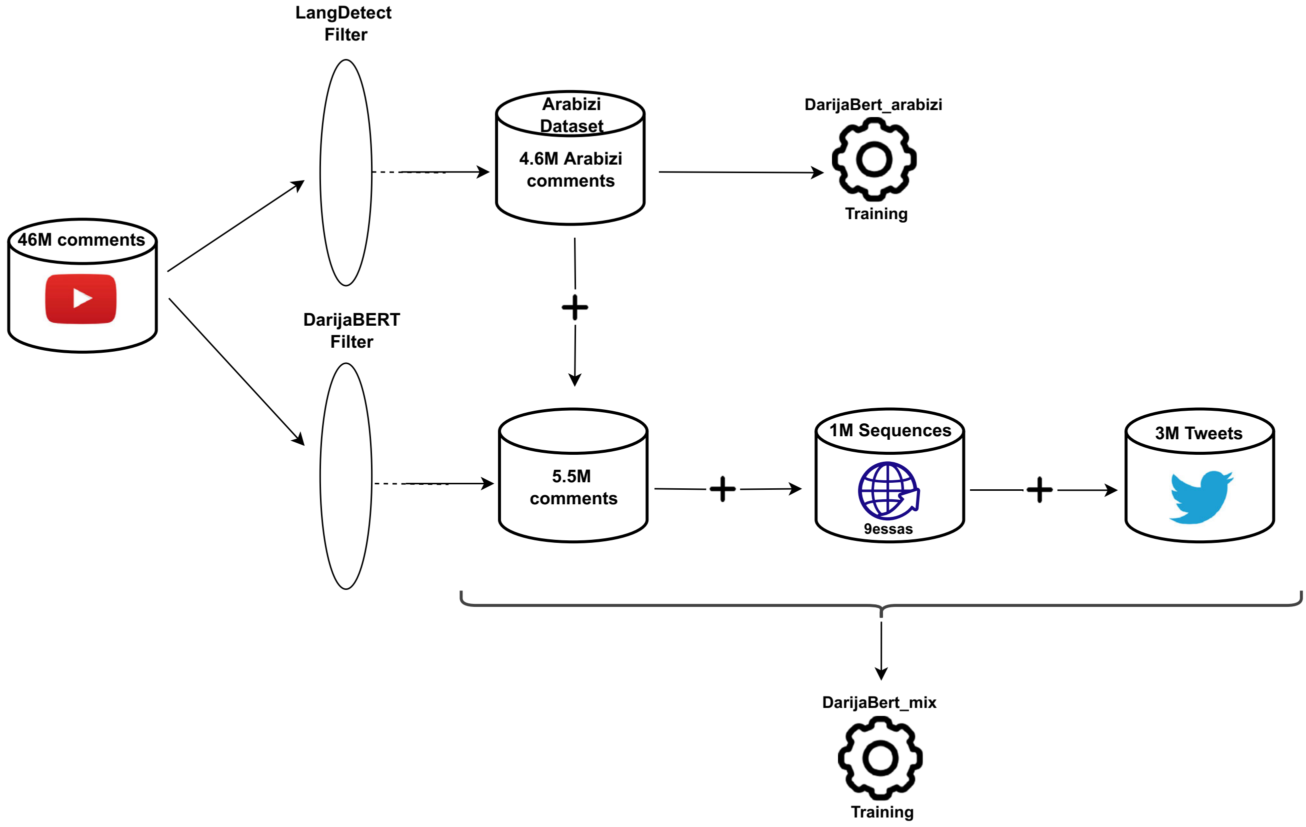


Fig. 2 DarijaBERT-arabizi and DarijaBERT-mix datasets creation process

Table 3 Examples of different spellings in Arabizi Darija

Arabizi spellings	- tkhebia, tkhbia, tkhabia - Geltha lik, goltha lek, gltha lk
CODA*[28] equivalent	- تْخَبْيَة, تْخْبْيَة, تْخْبِيَة - قْلْتْهَا لِيك, قُلْتْهَا لِيك, قْلْتْهَا لِيك
MSA	-إختِباء -قُلتُها لك
English translation	- Hiding DarijaBert - I told you

4 Evaluation

DarijaBERT was fine-tuned on four downstream tasks: dialect identification (DI), sentiment analysis (SA), sarcasm automatic detection (SAD), and topic classification (TC). The latter was achieved using the MTCD dataset (see Sect. 4.3). Results were compared with those from six additional models supporting Arabic either fully or partially, as shown in Table 4.

Except for CAMeLBERT-DA, exclusively trained on Arabic dialects, existing models underwent training on a multi-lingual corpus, involving MSA and DA, as observed in XLM-RoBERTa and mBERT, or have been fully trained on MSA alone, as demonstrated in the case of AraBERT. Furthermore, some models, such as MarBERT and Qarib, have undergone training on a combination of MSA and DA.

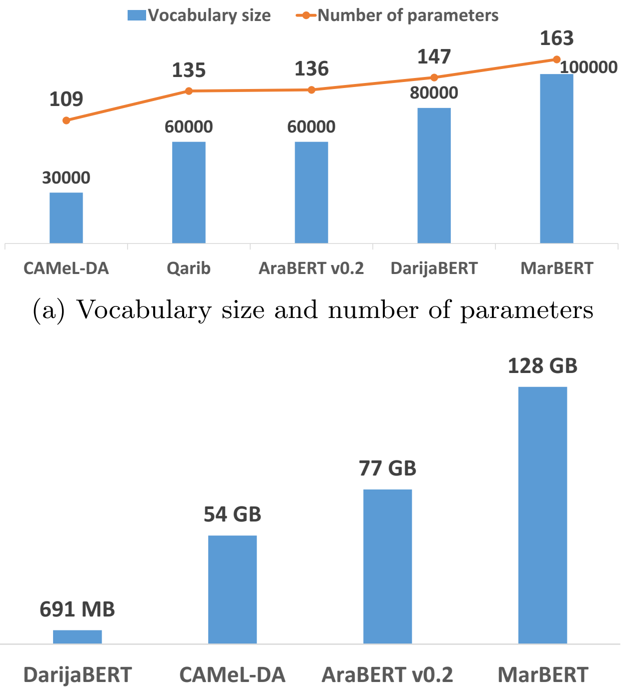
The size of the vocabulary was found to be correlated with the number of parameters in the model. This relationship was observed to be linear, with the number of parameters ranging from 109 million for CAMeLBERT-DA to 163 million for MarBERT, corresponding to vocabulary sizes of 30,000 and 100,000 tokens, respectively (as shown in Fig. 3a). Among the models, MarBERT used the largest dataset of 128 GB, while DarijaBERT used the smallest dataset of 691 MB. The remaining models used datasets larger than 50 GB (as shown in Fig. 3b).

Comprehensive gold standard datasets for the Moroccan dialect are scarce. Existing datasets focusing on Darija are relatively limited compared to datasets available for other dialects, like Egyptian or Gulf dialects. Moreover, these Darija datasets are not consistently accessible in the public domain. To facilitate the evaluation of our models on downstream tasks, a curated list of available datasets that incorporate Darija is compiled and presented in Table 5.

In order to accurately evaluate models on downstream tasks related to the Moroccan dialect, only publicly available

Table 4 BERT models comparison

Model	Arabic composition	Vocabulary size (ar/all)	# Tokens (ar/all)	Data size	#Params	#Steps
XLM-RoBERTa-Base [13]	Partially (MSA+DA)	14K/250K	2.9B/295B	2.5TB	278 M	–
mBERT-uncased [2]	Partially (MSA)	5K/110K	153 M/1.5B	–	167 M	–
AraBERTv0.2 [10]	Fully (MSA)	60K/64K	2.5B/2.5B	77GB	136 M	3 M
CAMeLBERT-DA [14]	Fully (DA)	30K/30K	5.8B/5.8B	54GB	109 M	1 M
Qarib [15]	Fully (MSA+DA)	64K/64K	14B/14B	–	135 M	10 M
MarBERT [12]	Fully (MSA+DA)	100K/100K	15.6B/15.6B	128GB	163 M	17 M
DarijaBERT	**Fully (DA)**	**80K/80K**	**100 M/100 M**	**691MB**	**147 M**	**235K**
DarijaBERT-mix	**Fully (DA)**	**160K/160K**	**–**	**1.7GB**	**209 M**	**510K**



(a) Vocabulary size and number of parameters

(b) Data size

Fig. 3 Comparison of Arabic BERT models statistics

datasets that contain a substantial proportion of Darija-labeled sequences were selected. Datasets with a high number of inaccurately labeled Darija sequences or those dominated by modern standard Arabic were excluded. Additionally, a translation of the Bible into the Moroccan dialect was excluded due to its strong religious bias and the potential lack of text diversity and generalizability. Based on these criteria, MADAR, MSTD, and MSDA were selected as evaluation datasets for their public accessibility and substantial number of Darija-labeled sequences

During the evaluation process, cross-validation was employed, with three folds for the larger datasets (MADAR and MTCD) and five folds for the smaller datasets (MSTD and MSDA). To address the issue of unbalanced distributions, a stratified approach was implemented. Throughout all experiments, the default parameters of the Transformers library trainer class were utilized and the number of epochs was set to 1. In line with previous studies [2], the [CLS] token of the last hidden layer was used as the representation of the sequences, followed by the integration of a linear layer for classification purposes. The evaluations were performed using an NVIDIA P100 GPU, except for the sentiment analysis (SA) and sarcasm automatic detection (SAD) tasks, which were conducted using an NVIDIA T4 GPU. To ensure comparability, the same fine-tuning process was applied to all the compared models.

4.1 Dialect identification

This task was conducted on two datasets:

- **MSDA dialect detection dataset** [46]: approximately 50k social media posts in different Arabic dialects. It is transformed into a binary dataset, i.e., Moroccan dialect vs. other dialects.
- **MADAR** [30]: approximately 111k sequences in 25 Arabic dialects. The dataset was transformed into a binary formatformat. One class represents Moroccan dialect, while the other encompasses all remaining dialects.

Table 6 reveals the imbalance in both datasets, featuring 16% and 13% of Darija sequences, respectively.

4.2 Sentiment analysis and sarcasm automatic detection

The model underwent fine-tuning to detect sentiment polarity and sarcasm in texts written in the Moroccan dialect, making predictions on MSTD (Moroccan Sentiment Twitter Dataset) [40]. This dataset comprises 12k tweets, labeled as negative, objective, positive, or sarcastic. To facilitate the analysis, two data subsets were created, one with sentiment

Table 5 Arabic datasets benchmark

Dataset[a]/Author name	Content	Public	#Moroccan dialect sequences/tokens
ArSarcasm [29]	10,547 tweets in Egyptian, Gulf, Levantine, and Maghrebi	Yes	Unknown
MADAR [30]	265K sentences in 25 Arab cities dialects	Yes	20k sequences
NADI [31]	30,957 Tweets from 21 Arab countries	Yes	1.5k sequences
QADI [32]	540k tweets from 18 Arab countries	Yes	13k sequences
Arap-tweet [33]	2.4M multidialectal tweets from 16 countries	No	Unknown
Al-Shargi et al. [34]	96.5K words Morphologically Annotated Corpus for Moroccan and Sanaani Yemeni Arabic	No	64k words
Darwish et al. [35]	Bible translated in Darija	Yes	Unknown
Samih et al. [36]	223k tokens from Darija and MSA blog posts	No	76k tokens
Voss et al. [37]	corpus of tweets of Moroccan dialect written in Roman script	No	Unknown
Laoudi et al. [38]	1836 Hespress news website comments	No	1.8k sequences
Maghfour et al. [39]	10k Facebook comments labeled for sentiment analysis	No	3.5k sequences
MSTD [40]	12k Darija tweets	Yes	12k sequences
Refaee et al. [41]	8,868 Multidialectal twitter corpus annotated for subjectivity and sentiment analysis	Yes	Unknown
MSAC [42]	2000 tweets/comments for sentiment analysis	Yes	2k sequences
el2017sentiment et al. [43]	700 Tweets geolocated in Morocco	No	Unknown
Habbat et al. [44]	25 146 Moroccan tweets	No	Unknown
TEAD [45]	6 million Arabic tweets in Maghrebi, Egypt,Gulf,Levantine	Yes	Unknown
MSDA [46]	+50K Tweets in 5 dialects: Algeria, Egypt, Lebanon, Tunisia and Morocco	Yes	10k sequences

[*a*]Available datasets at the time of writing this paper

Table 6 MSDA and MADAR datasets content description

	MSDA	MADAR
Darija	7169	13,871
Other	38,351	97,035
Total	45,520	110,906

Table 7 MSTD dataset sentiment labels distribution

Label	No. of sequences
Negative	2667
Objective	6220
Positive	732
Total	9619

Table 8 MSTD dataset sarcasm labels distribution

Label	No. of sequences
Sarcastic	2176
Non sarcastic	9619
Total	10,895

labels and the other with a binary label for sarcasm (refer to Tables 7 and 8 for the content description).

4.3 Topic classification

Initially, an attempt was made to classify topics using the MSDA dataset, however, the resulting dataset was insufficiently small. Hence, a new dataset was created specifically for topic classification in the Moroccan dialect and was named the Moroccan Topic Classification Dataset (MTCD). This dataset is the first of its kind for the Moroccan dialect and has been made open source[12] to encourage further research in this area. The MTCD was curated by collecting comments from four Moroccan YouTube channels covering topics such as Gaming, Cooking, Sports, and General. Due to a shortage of comments from the Sports channel, additional comments from two popular Moroccan soccer teams' Facebook pages (Raja and Wydad) were included. The DarijaBERT model was employed to filter comments in the Moroccan dialect, retaining only those with a prediction probability exceeding 80%. This filtering process resulted in a total of 64k comments, as indicated in Table 9. A Darija native speaker evaluated a sample of these comments, and the validation yielded an accuracy rate of 94% in identifying Darija comments.

Table 9 MTCD dataset content description

Topic	No. of comments
Gaming	14,000
Cooking	10,000
Sports	20,000
General	20,000
Total	64,000

5 Results

The results of the evaluation of DarijaBERT for each of the downstream tasks are presented and analyzed in this section. The performance of DarijaBERT is compared to the six models specified in Table 4. The comparison is not performed with DarijaBERT-arabizi as the evaluation datasets do not include Arabizi, and the models being compared only support text written in the Arabic script. The findings are reported in terms of accuracy and the appropriate F1 score for the respective downstream task. The highest scores for each metric are emphasized in bold.

5.1 Dialect identification

Results indicate that DarijaBERT displays superior performance in accuracy and $\mathrm{F1}_{Darija}$ scores when applied to the MSDA and MADAR datasets, as shown in Table 10. DarijaBERT-mix surpasses DarijaBERT by 1.09 points for MSDA and 0.39 points for MADAR in $\mathrm{F1}_{Darija}$ scores. MARBERT closely trails behind the DarijaBERT models, being 1.75 and 7.44 points behind DarijaBERT-mix for MSDA and MADAR, respectively. The other models show similar results for MADAR, while for MSDA the Arabic language models have an advantage over the multilingual models. Regarding $\mathrm{F1}_{Darija}$, DarijaBERT-mix performs on average 9.6 points better than all other models (excluding DarijaBERT) and 7 points better than the Arabic language models (excluding DarijaBERT). The observed improvement in performance of DarijaBERT-mix over DarijaBERT can be attributed to its extensive training dataset, which includes 9.5 million sequences compared to DarijaBERT's 3 million. The larger dataset equips DarijaBERT-mix with a more profound understanding of Moroccan dialect-specific patterns and nuances, leading to enhanced proficiency in dialect identification tasks.

5.2 Sentiment analysis and sarcasm automatic detection

Table 11 summarizes the sentiment analysis accuracy and $\mathrm{F1}_{PN}$ scores (positive and negative labels), as well as the accuracy and $\mathrm{F1}_{Sar}$ scores (for sarcastic label).

[12] https://github.com/AIOXLABS/DBert.

Table 10 Dialect identification results

Model	MSDA[a]		MADAR	
	Acc.	$F1_{Darija}$	Acc.	$F1_{Darija}$
XLM-RoBERTa-base	87.55	53.07	93.40	76.07
mBERT	89.07	59.37	93.83	77.74
AraBERT	90.90	67.49	93.44	77.57
CAMeLBERT-DA	91.31	70.10	93.59	78.86
Qarib	91.49	71.38	92.53	77.43
MarBERT	92.90	76.25	92.62	77.70
DarijaBERT	93.07	76.91	95.81	84.75
DarijaBERT-mix	93.43	78.00	95.93	85.14

[a] Significant at the 5% level. p-value = 0.000228. Note: ANOVA tests are exclusively conducted for results derived from five folds experiments. For experiments involving three folds, there is an insufficient number of estimation points to perform robust statistical tests

Table 11 Sentiment analysis and sarcasm automatic detection results

Model	SA[a]		SAD	
	Acc.	$F1_{PN}$	Acc.	$F1_{Sar}$
XLM-RoBERTa-base	67.05	21.80	81.55	0
mBERT	67.10	17.33	81.87	18.01
AraBERT	69.90	31.50	82.37	27.42
CAMeLBERT-DA	72.31	45.75	82.61	31.82
Qarib	73.96	47.60	82.60	33.10
MarBERT	73.95	50.82	82.61	**35.96**
DarijaBERT	**74.78**	**50.85**	**82.80**	33.17
DarijaBERT-mix	71.50	26.20	81.90	11.80

[a] Significant at the 5% level. p-value $< 2e^{-16}$. Note: ANOVA tests are exclusively conducted for results derived from five folds experiments. For experiments involving three folds, there is an insufficient number of estimation points to perform robust statistical tests

Table 12 Topic classification results

Model	MTCD	
	Acc.	$F1_{macro}$
XLM-RoBERTa-base	82.61	82.78
mBERT	84.38	84.48
AraBERT	83.67	83.92
CAMeLBERT-DA	87.18	87.37
Qarib	88.52	88.68
MarBERT	89.10	89.26
DarijaBERT	90.52	90.77
DarijaBERT-mix	90.60	90.90

Lower scores than the DI downstream scores were reported, which can be attributed to the complexity of downstream tasks related to sentiment analysis. Additionally, the F1 scores for sarcasm detection were observed to be lower compared to those in sentiment analysis. It also turns out that models trained on Arabic dialects perform better than other models. Specifically, AraBERT trained on MSA shows lower scores than models supporting DA (CAMeLBERT-DA, Qarib, MarBERT, and DarijaBERT). DarijaBERT is the best-performing model in terms of $F1_{PN}$ for SA, outperforming Arabic models by 6.9 points on average. Regarding sarcasm detection, the accuracy scores are consistently high and comparable, given the highly unbalanced nature of the dataset. However, aside from the multilingual models and AraBERT, the scores at the $F1_{Sar}$ level exhibit relative similarity. MarBERT outperforms other models by an average of 4.6 points, slightly ahead of DarijaBERT. Conversely, DarijaBERT-mix falls short in sentiment analysis tasks, despite its success in dialect identification. Indeed, it slightly outperforms the multilingual models for SA and has the second-lowest SAR score. The performance difference between DarijaBERT-mix and DarijaBERT in sentiment analysis can be ascribed to their unique characteristics and training data. DarijaBERT-mix, akin to a multilingual model, encompasses Moroccan Arabic in both Arabic and Arabizi scripts, which broadens linguistic patterns but may introduce noise and variability absent in monolingual DarijaBERT. Indeed, sentiment analysis relies on grasping language-specific emotional nuances, and the inclusion of mixed scripts and training data variations in DarijaBERT-mix can complicate the capture of sentiment-related features effectively.

5.3 Topic classification

The accuracy and macro-F1 scores of the different models are shown in Table 12. Multilingual models achieve relatively higher scores than the other downstream tasks. On one hand, mBERT even outperforms AraBERT by 0.6 points in terms of $F1_{macro}$. On the other hand, AraBERT is in the second-last position, exceeding only XLM-RoBERTa. DarijaBERT models are more effective than all the other models in both metrics. Indeed DarijaBERT-mix exceeds MarBERT by 1.64 points and is on average 4.8 points better than the rest of the models (excluding DarijaBERT).

6 Discussion

DarijaBERT illustrates that a monodialectal model can outperform multilingual or multidialectal models, even with a smaller dataset.This is evidenced by MarBERT's training dataset being approximately 190 times larger than DarijaBERT's. Despite this, better results are obtained for Moroccan dialect identification, sentiment analysis, and topic classification. Additionally, results for automatic sarcasm detection are ranked second, despite the complexity of these two tasks. Contrary to expectations, more data for DarijaBERT-mix does not produce better results than DarijaBERT on SA and SAD tasks. We hypothesize, that the reason is the balance and diversity of the data utilized for

DarijaBERT, obtained by combining three different sources with equal shares of sequences.

The vocabulary size is also an essential parameter for the model's performance. When compared to models for the Arabic language, DarijaBERT uses 80k tokens for training, which is only surpassed by MarBERT's 100k tokens. However, tests were conducted with both higher and smaller vocabulary sizes (60k, 120k, and 160k) with no substantial improvement in performance was observed. This leads to the conclusion that a larger vocabulary size does not always translate to higher performance, highlighting the need to determine the optimal value for this parameter. It is also important to note that the lack of dedicated datasets for Arabizi Darija is a barrier to developing effective systems. Furthermore, the substantial vocabulary size utilized in DarijaBERT-arabizi (110k) underscores the importance of addressing Arabizi preprocessing to enhance the model's performance. Arabizi text poses challenges due to its informal nature and lack of strict rules. One of the key issues that need to be addressed by the language model is the wide range of word forms that convey the same meaning. To tackle this, it is crucial to employ appropriate tokenization techniques that can effectively handle data sparsity and capture the variations in word forms.

7 Conclusion

This paper introduces pioneering BERT models tailored to the Moroccan Arabic dialect, referred to as Darija. The open-sourced models cover both written forms of Darija, encompassing DarijaBERT for text in Arabic letters, DarijaBERT-arabizi for text in Latin letters, and DarijaBERT-mix, trained on both character types. These models outperformed other multidialectal models (trained on approximately a 190 times larger corpus) in Moroccan dialect identification, sentiment analysis, and topic classification tasks. As a result, researchers will be able to create novel processing systems for the Moroccan dialect. The research uses diverse data from YouTube, Twitter, and stories in Moroccan Arabic, employing iterative self-learning to enhance models for real-world Darija language nuances. Furthermore, MTCD (Moroccan Topic Classification Dataset) is made publically available, marking the first dataset for Darija topic classification encompassing four distinct topic classes. It is planed to extend the evaluations to other downstream tasks, create Arabizi-specific datasets and improve the models through text preprocessing. In summary, DarijaBERT models excel in various tasks but face key limitations: the blending of dialects in DarijaBERT-mix may impact certain tasks negatively, the absence of Arabizi benchmark datasets poses challenges for evaluation, and the scarcity of Moroccan dialect ground truth data limits comprehensive comparisons. Overcoming these limitations is vital for advancing Darija model development.

Acknowledgements We thank the Google TRC program for giving us access to their TPUs Cloud.

Author Contributions KG, AMN, AA and Imade Benelallam contributed to this paper with equal efforts. KG, AMN, AA particpated to the data collection and cleaning. KG was responsible for the development and implementation of the models on GCP. All authors discussed and reviewed the results. KG was responsible for writing the paper. IB supervised the whole work.

Declarations

Conflicts of interest The authors state that they have no conflicts of interest.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: 31st NIPS, pp. 6000–6010 (2017)
2. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019). https://doi.org/10.18653/v1/n19-1423
3. Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B.: CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, pp. 7203–7219 (2020). https://doi.org/10.18653/v1/2020.acl-main.645
4. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
5. Elgezouli, M., Elmadani, K.N., Saeed, M.: Sudabert: pre-trained encoder representation for Sudanese Arabic dialect. In: 2020 ICCCEEE, pp. 1–4 (2021). https://doi.org/10.1109/ICCCEEE49695.2021.9429651
6. Messaoudi, A., Cheikhrouhou, A., Haddad, H., Ferchichi, N., BenHajhmida, M., Korched, A., Naski, M., Ghriss, F., Kerkeni, A.: Tunbert: Pretrained contextualized text representation for Tunisian dialect. In: Intelligent Systems and Pattern Recognition, Cham, pp. 278–290 (2022)
7. Abdaoui, A., Berrimi, M., Oussalah, M., Moussaoui, A.: Dziribert: pre-trained language model for the Algerian dialect. arXiv preprint arXiv:2109.12346 (2021)
8. Slim, A., Melouah, A., Faghihi, U., Sahib, K.: Improving neural machine translation for low resource Algerian dialect by transductive transfer learning strategy. Arab. J. Sci. Eng. **47**, 1–8 (2022)
9. Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., Smaili, K.: Machine translation experiments on PADIC: a parallel Arabic DIalect corpus. In: Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, Shanghai, China, pp. 26–34 (2015). https://aclanthology.org/Y15-1004
10. Antoun, W., Baly, F., Hajj, H.: 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, Marseille, France, pp. 9–15 (2020)

11. Safaya, A., Abdullatif, M., Yuret, D.: KUISAIL at SemEval-2020 : BERT-CNN for offensive speech identification in social media. In: 40th SemEval, pp. 2054–2059. ICCL, Barcelona (online) (2020)
12. Abdul-Mageed, M., Elmadany, A., Nagoudi, E.M.B.: 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), Online, pp. 7088–7105 (2021). https://doi.org/10.18653/v1/2021.acl-long.551
13. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, pp. 8440–8451 (2020). https://doi.org/10.18653/v1/2020.acl-main.747
14. Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., Habash, N.: The interplay of variant, size, and task type in Arabic pre-trained language models. In: Workshop on Arabic Natural Language Processing (2021)
15. Abdelali, A., Hassan, S., Mubarak, H., Darwish, K., Samih, Y.: Pre-training bert on Arabic tweets: practical considerations. arXiv preprint arXiv:2102.10684 (2021)
16. El-Khair, I.A.: 1.5b words Arabic corpus. preprint arXiv:1611.04033 (2016)
17. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: generalized autoregressive pretraining for language understanding. Adv. Neural Inf. Process. Syst. **32** (2019)
18. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
19. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: a lite BERT for self-supervised learning of language representations. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020 (2020)
20. Schuster, M., Nakajima, K.: Japanese and Korean voice search. In: 2012 IEEE ICASSP, pp. 5149–5152 (2012). IEEE
21. Bianchi, R.M.: Glocal Arabic online: the case of 3arabizi. SSLLT **2**(4), 483–503 (2012)
22. Yaghan, M.A.: "Arabizi": a contemporary style of Arabic slang. Design Issues **24**(2), 39–52 (2008)
23. Alghamdi, H., Petraki, E.: Arabizi in Saudi Arabia: a deviant form of language or simply a form of expression? Soc. Sci. **7**(9), 155 (2018)
24. Aboelezz, M.: 'we are young. we are trendy. buy our product!': The use of Latinized Arabic in printed edited magazines in Egypt. UAJSS (9), 47–72 (2012)
25. Palfreyman, D., Khalil, M.A.: "A funky language for teenz to use": representing gulf Arabic in instant messaging. J. Comput. Med. Commun. **9**(1), 917 (2003)
26. Mostafa, L.: A survey of automated tools for translating Arab chat alphabet into Arabic language. Am. Acad. Sch. Res. J. **4**(3), 44–50 (2012)
27. Elmahdy, M., Gruhn, R., Abdennadher, S., Minker, W.: Rapid phonetic transcription using everyday life natural chat alphabet orthography for dialectal Arabic speech recognition. In: 2011 IEEE ICASSP, pp. 4936–4939 (2011). IEEE
28. Habash, N., Eryani, F., Khalifa, S., Rambow, O., Abdulrahim, D., Erdmann, A., Faraj, R., Zaghouani, W., Bouamor, H., Zalmout, N., : Unified guidelines and resources for arabic dialect orthography. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
29. Abu Farha, I., Magdy, W.: From Arabic sentiment analysis to sarcasm detection: the ArSarcasm dataset. In: 4th OSACT, Marseille, France, pp. 32–39 (2020)
30. Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A.,: The madar arabic dialect corpus and lexicon. In: LREC (2018)
31. Abdul-Mageed, M., Zhang, C., Bouamor, H., Habash, N.: NADI 2020: The first Nuanced Arabic dialect identification shared task. In: Proceedings of the Fifth WANLP, pp. 97–110 (2020)
32. Abdelali, A., Mubarak, H., Samih, Y., Hassan, S., Darwish, K.: Qadi: Arabic dialect identification in the wild. In: Workshop on Arabic Natural Language Processing (2021)
33. Zaghouani, W., Charfi, A.: Arap-tweet: a large multi-dialect Twitter corpus for gender, age and language variety identification. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan (2018)
34. Al-Shargi, F., Kaplan, A., Eskander, R., Habash, N., Rambow, O.: Morphologically annotated corpora and morphological analyzers for Moroccan and Sanaani Yemeni Arabic. In: 10th LREC 2016 (2016)
35. Darwish, K., Abdelali, A., Mubarak, H., Samih, Y., Attia, M.: Diacritization of Moroccan and Tunisian Arabic dialects: A CRF approach. OSACT **3**, 62 (2018)
36. Samih, Y., Maier, W.: An Arabic-Moroccan Darija code-switched corpus. In: Proceedings of LREC'16, pp. 4170–4175 (2016)
37. Voss, C., Tratz, S., Laoudi, J., Briesch, D.: Finding Romanized Arabic dialect in code-mixed tweets. In: Proceedings of LREC'14, pp. 2249–2253 (2014)
38. Laoudi, J., Bonial, C., Donatelli, L., Tratz, S., Voss, C.: Towards a computational lexicon for Moroccan darija: Words, idioms, and constructions. In: Proceedings of LAW-MWE-CxG-2018, pp. 74–85 (2018)
39. Maghfour, M., Elouardighi, A.: Standard and dialectal Arabic text classification for sentiment analysis. In: ICMDE, pp. 282–291 (2018). Springer
40. Mihi, S., Ait, B., El, I., Arezki, S., Laachfoubi, N.: Mstd: Moroccan sentiment twitter dataset. Int. J. Adv. Comput. Sci. Appl **11**(10), 363–372 (2020)
41. Refaee, E., Rieser, V.: An Arabic twitter corpus for subjectivity and sentiment analysis. In: LREC, pp. 2268–2273 (2014)
42. Oussous, A., Benjelloun, F.-Z., Lahcen, A.A., Belfkih, S.: Asa: A framework for Arabic sentiment analysis. J. Inf. Sci. **46**(4), 544–559 (2020)
43. El Abdouli, A., Hassouni, L., Anoun, H.: Sentiment analysis of Moroccan tweets using naive bayes algorithm. IJCSIS **15**(12) (2017)
44. Habbat, N., Anoun, H., Hassouni, L.: Topic modeling and sentiment analysis with LDA and NMF on Moroccan tweets. In: The Proceedings of the Third ICSCA, pp. 147–161 (2020). Springer
45. Abdellaoui, H., Zrigui, M.: Using tweets and emojis to build tead: an Arabic dataset for sentiment analysis. Computaci'on y Sistemas **22**(3) (2018)
46. Boujou, E., Chataoui, H., Mekki, A.E., Benjelloun, S., Chairi, I., Berrada, I.: An open access nlp dataset for arabic dialects: data collection, labeling, and model construction. preprint arXiv:2102.11000 (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.