ISSN: 1803-3814 (Printed), 2571-3701 (Online)
https://doi.org/10.13164/mendel.2023.1.055

Pre-training Two BERT-Like Models for Moroccan Dialect: MorRoBERTa and MorrBERT

Otman Moussaoui✉, **Yacine El Younoussi**

Information System and Software Engineering, National School of Applied Sciences, Abdelmalek Essaadi University, Morocco
otman.moussaoui@etu.uae.ac.ma✉, yacine.elyounoussi@uae.ac.ma

Abstract

This research article presents a comprehensive study on the pre-training of two language models, MorRoBERTa and MorrBERT, for the Moroccan Dialect, using the Masked Language Modeling (MLM) pre-training approach. The study details the various data collection and pre-processing steps involved in building a large corpus of over six million sentences and 71 billion tokens, sourced from social media platforms such as Facebook, Twitter, and YouTube. The pre-training process was carried out using the HuggingFace Transformers API, and the paper elaborates on the configurations and training methodologies of the models. The study concludes by demonstrating the high accuracy rates achieved by both MorRoBERTa and MorrBERT in multiple downstream tasks, indicating their potential effectiveness in natural language processing applications specific to the Moroccan Dialect.

Keywords: *Moroccan Dialect, BERT, RoBERTa, Natural Language Processing, Pre-trained, Machine Learning.*

Received: 27 May 2023
Accepted: 12 June 2023
Online: 21 June 2023
Published: 30 June 2023

1 Introduction

Natural Language Processing (NLP) has become a rapidly growing research field in recent years due to the emergence of deep learning models such as the Transformer [27] architecture. This neural network model was first introduced by [27] in 2017 and has revolutionized the field of NLP by providing a powerful way to process long sequences of text using attention mechanisms. The Transformer architecture has surpassed older models such as LSTMs (Long short-term memory networks) [17] and GRUs (Gated recurrent units) [10], and has served as a foundational architecture for a range of subsequent models that have achieved remarkable performance in various NLP tasks, including question-answering, language modeling, text classification, and others. These models include BERT (Bidirectional Encoder Representations from Transformers) [13], XLNet [31], and RoBERTa (Robustly Optimized BERT Pretraining Approach) [21], among others.

One of the key benefits of the Transformer architecture is that once pre-trained on a large corpus of text, the models can be fine-tuned for specific tasks using smaller, task-specific datasets. This approach has led to significant improvements in a range of NLP tasks, especially when large, well-curated training datasets are available. Despite the impressive progress in NLP enabled by the Transformer architecture, most of the pre-trained models are based on English or other widely spoken languages, and there is limited availability of models for other languages [22]. Moreover, most of the existing pre-trained models are multilingual, which means that they are trained on a collection of languages but do not account for the nuances and peculiarities of individual languages.

This lack of language-specific models is especially pronounced in the case of the Moroccan dialect, which has low resources and has received little attention from NLP researchers. The Moroccan dialect is spoken by more than 36 million people and is the primary communication tool in everyday life, while Modern Standard Arabic (MSA) is the official language used in Morocco [2]. The Moroccan dialect is also commonly used on social media platforms to express opinions and thoughts.

However, the Moroccan dialect poses significant challenges for NLP due to several factors. For one, the Moroccan dialect can be written using different scripts, including the Arabic alphabet, the Latin alphabet, or a combination of the two [23]. Additionally, the Moroccan dialect has its own unique vocabulary and syntax that differ from both MSA and other languages [26]. These factors make the Moroccan dialect an interesting challenge for NLP researchers to develop language-specific pre-trained models. To address this need, we propose to pre-train two BERT-like models for the Moroccan dialect. Specifically, we base the architecture of our models on the RoBERTa model and the original BERT model, which have achieved state-of-the-art performance in NLP tasks.

To pre-train our models, we use a dataset consisting of over six million sequences of YouTube comments, Facebook comments, and tweets. We then compare the performance of our models with existing Moroccan dialect models and multilingual models, including multilingual BERT [13], XLM-R [11], CamelBERT [18], and MARBERT [5] using both publicly available datasets for the Moroccan dialect and our own manually an-

notated data. Our primary contributions in this work are two-fold. First, we present two pre-trained BERT-like models, MorRoBERTa and MorrBert, specifically designed for the Moroccan dialect. These models are pre-trained on a large dataset of more than six million sequences consisting of comments from social media platforms such as YouTube, Facebook, and Twitter.

We use the pre-trained models to perform downstream natural language understanding tasks, such as sentiment analysis, dialect identification, and language classification as used by Moroccans. Our results show that both MorRoBERTa and MorrBert are effective for sentiment analysis, dialect identification, and language classification tasks on Moroccan datasets, and can achieve comparable or better performance than other state-of-the-art models in the field. Second, we compare the performance of our models to that of existing Moroccan dialect and multilingual models using varying data sizes and varying classes. Our experiments show that MorRoBERTa and MorrBert consistently outperform other models, indicating their effectiveness in handling the unique challenges posed by the Moroccan dialect.

Overall, our work provides a valuable contribution to the field of natural language processing, especially for low-resource languages like the Moroccan dialect. Our pre-trained models can be fine-tuned for specific tasks such as sentiment analysis, named entity recognition, and machine translation, among others. Furthermore, our experiments highlight the importance of developing language-specific models rather than relying on multilingual models, especially for languages with distinct characteristics and limited resources. Our work opens up avenues for future research in the development of language-specific models for other low-resource languages.

2 Related Work

The development of pre-trained language models based on the Transformer architecture has revolutionized the field of NLP in recent years. One of the most successful examples is the BERT model proposed by [13], which employs a multi-layer Transformer-encoder architecture and two pre-training tasks, MLM, and Next Sentence Prediction (NSP), to learn contextualized word representations. BERT has achieved state-of-the-art performance in a wide range of NLP tasks, including sentiment analysis, question answering, and natural language inference. In response to the success of BERT, many BERT-based models with varying goals have been proposed in the literature. For instance, RoBERTa [21] optimizes BERT's pre-training methods by removing the NSP pre-training task and significantly increasing the size of the training corpus and batch size. DistilBERT [24] and ALBERT [20] reduce the model size and complexity of BERT by using knowledge distillation and parameter sharing techniques, respectively.

Apart from these models, several pre-trained models have been developed specifically for low-resource languages, including XLM [12], which pre-trains a shared multilingual model on parallel data from 100 languages, and mBERT [13], which pre-trains a single model on a large corpus of Wikipedia articles from 104 languages. These models have been shown to improve the performance of downstream NLP tasks in low-resource languages, although they may not capture the nuances and peculiarities of individual languages.

One of the major challenges in developing NLP models for Arabic dialects is the lack of high-quality, large-scale labeled datasets for training and evaluation. This challenge is particularly relevant for the Moroccan dialect, which has unique phonetic and grammatical features that distinguish it from other Arabic dialects. To overcome this challenge, researchers have proposed various pre-trained models for the Moroccan dialect, such as CamelBERT [18], MARBERT [5], QARiB [4] and AraBERT [7], which are based on the BERT architecture and pre-trained on multidialectal corpora. However, these models have limitations due to the relatively low representation of the Moroccan dialect in the training corpus, which can result in reduced performance and accuracy when applied to specific dialectal variations.

In contrast, DarijaBERT, the monodialectal Moroccan model developed by [14], was trained specifically on a Moroccan Arabic dialect corpus, which allows for more accurate modeling of the linguistic features and nuances of the Moroccan dialect. The model also supports both Arabic and Latin character representations of the dialect.

In this work, we propose to pre-train two BERT-like models, MorRoBERTa, and MorrBERT, specifically designed for the Moroccan dialect. We use a large corpus of social media data to pre-train our models using MLM and compare their performance to existing Moroccan dialect and multilingual models, as well as to our own manually annotated data. Our experiments demonstrate that MorRoBERTa and MorrBERT are comparable to or better than the performance of other models and achieve state-of-the-art performance in downstream NLP tasks for the Moroccan dialect, highlighting the importance of developing language-specific models for low-resource languages like the Moroccan dialect.

3 Pre-training Process

3.1 Data Description

In this section, we describe our proposed approaches for large data collection for Moroccan Dialect.

Data Collection

To train BERT-like models, it is necessary to amass significant amounts of raw text data and preprocess it accordingly. For our Moroccan Dialect project, we identified three sources from which to compile large datasets:

Table 1: Dataset used to train our models.

Source	Number of Words	Number of Sentences	Size
Facebook	26,559,210	3,191,281	144 MB
YouTube	43,901,655	2,785,302	347 MB
Twitter	869,184	51,479	6.61 MB

Facebook, Twitter, and YouTube. Consequently, our pre-training corpus primarily consists of informal language comments, which are commonly used by social media users to express their opinions. However, the aim of collecting this data is to capture as many vocabularies and texts used on social media as possible. We were able to gather over six million sentences written and shared by Moroccan users, resulting in a final dataset that comprised more than 71 billion tokens, equivalent to approximately 700 MB of text data.

We used different collection approaches for each source, as described below:

- Twitter: We used the Twitter API to collect data for research by searching for Moroccan-related hashtags such as #Maroc, #أسود_الأطلس, #bdarija, #dahk, #فضايح, etc. and extracting all tweets with a Moroccan location setting. We collected around 1.1 million tweets, and after filtering and removing duplicates, we ended up with 65,000 tweets.
- Facebook: To collect data from Facebook, we selected pages with a large number of members and publications related to Moroccan contexts and users, and we used the "Facebook Graph API" to import the data. We collected nearly 5 million comments between 11/09/18 and 18/11/18. The total number of comments collected is equal to 157.7M. We cleaned the comments by deleting duplicates and empty ones.
- YouTube: We utilized the YouTube API version 3.0 to extract comments from different channels. We selected 33 channels based on their popularity, identified by Hypeauditor [3] (most commented and followers), and used SocialReaper[1] to scrape the comments. In total, we collected 2,785,302 comments from 27,000 different videos.

Table 1 contains statistical information for each source that was utilized to construct our dataset.

Pre-processing

In the realm of data analysis and machine learning, the preprocessing steps applied to sequences are crucial in ensuring accurate and consistent results [15]. In an effort to standardize the data, the dataset underwent a series of modifications. Firstly, any hashtags, URLs, and email addresses were removed from the sequences in order to eliminate any noise or distractions from the data.

[1]https://github.com/ScriptSmith/socialreaper

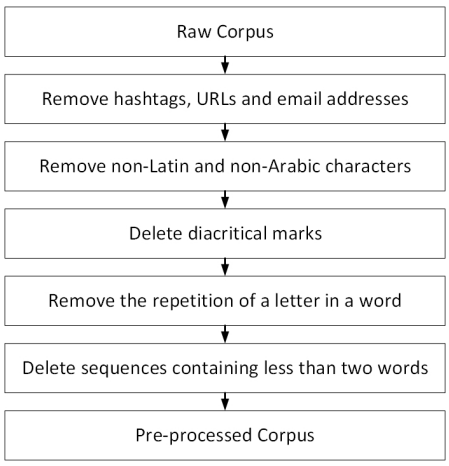

Figure 1: Data preprocessing process

Additionally, the repetition of a letter in a word was removed, as this can often be indicative of emphasis in social media communication rather than a significant feature of the text. Diacritics marks were also removed to ensure consistency and prevent potential misinterpretations of the data.

Furthermore, non-Latin and non-Arabic characters were removed from the sequences to create a more uniform dataset. This was done in order to reduce the variability of the data and allow for a more accurate comparison between different sequences.

Finally, only sequences with at least two words were retained. This was done in order to eliminate any overly simplistic or ambiguous sequences, which could potentially skew the results. By applying these preprocessing steps, the dataset was standardized and prepared for further analysis and machine learning applications.

The preprocessing process is shown in Fig. 1.

Tokenization

Each model employs a tokenizer that adheres to its original research paper to maintain conformity. For the MorRoBERTa model, a byte-level BPE tokenizer is used [25]. This tokenizer merges byte-based characters based on their frequency of occurrence until the desired vocabulary size is achieved.

In contrast, the MorrBERT model employs a WordPiece tokenizer [30], which is similar to the BPE tokenizer. The WordPiece tokenizer creates subwords of characters that are likely to appear in the training data and adds them to the vocabulary. For further details on the parameters, please refer to Table 2.

3.2 Pre-training Models

Our team developed two models, MorRoBERTa and MorrBERT, using the same corpus presented in last Section. Both models were trained using MLM pre-training and a vocabulary size of 52K subword tokens. We utilized HuggingFace Transformers API [29] to build the models, and they were executed on the

Table 2: Displays the configurations for the MorRoBERTa and MorrBERT models.

	MorRoBERTa	MorrBERT
train_batch_size	128	64
steps	565,980	565,980
vocab_size	52K	52K
Vocab Tokenizer	Byte-level BPE	WordPiece
Model type	roberta	bert
hidden_layers	6	12
num_parameters	83,504,416	125,977,344

HPC-MARWAN [1] platform using GPU cards for accelerated performance.

MorRoBERTa is a smaller version of the RoBERTa-base[21] model with 6 layers, 12 attention heads, 768 hidden dimensions, and a maximum sequence length of 512. During training, the batch size was set to 128, and the model was trained for a total of 565,980 steps. The training process took nearly 92 hours to complete 12 epochs across the full training set.

Similarly, MorrBERT was configured exactly like the BERTBASE [13] model, with 12 layers, 12 attention heads, and a batch size of 64. The model was also trained for a total of 565,980 steps, but the training process took nearly 120 hours to complete 12 epochs across the full training set.

Table 2 provides a summary of the different configurations for MorRoBERTa and MorrBERT, highlighting the varying specifications and training times required for these models. Our findings underscore the importance of selecting the appropriate configuration and training methodology for neural models in natural language processing, and the computational resources required to achieve high accuracy and robustness in these applications.

4 Evaluation Tasks and Test Data

In order to evaluate the effectiveness of our models, we rely on both publicly available datasets for the Moroccan dialect and our own manually annotated data. Our assessment focuses on three main tasks: sentiment analysis, dialect identification, and language classification as used by Moroccans.

The Sentiment Analysis task, also referred to as Polarity Detection, is a type of classification task that involves analyzing a given text and assigning it a sentiment polarity label [8]. The primary goal of this task is to determine whether a piece of text expresses a positive, negative, or neutral sentiment. To evaluate the effectiveness of sentiment analysis models, we utilized two Moroccan dialect datasets, namely MAC [16] and MSDA [9].

The MAC dataset consists of two class, namely polarity (positive, negative, neutral, and mixed) and language of the tweets (Standard Arabic or Dialectal Arabic). On the other hand, the MSDA dataset comprises three class: Arabic dialect (Algerian, Lebanese, Moroccan, Tunisian, and Egyptian), topic (other, politics, health, social, sport, and economics), and sentiment analysis (positive, negative, and neutral). For more information about these datasets, please refer to Table 3.

Arabic dialect identification involves recognizing and distinguishing the various spoken dialects of the Arabic language. As Arabic is spoken across several countries in the Middle East and North Africa, there are significant differences in the way people pronounce words, use vocabulary, and apply grammar rules across dialects. Arabic dialect identification can be done at different levels of detail, including binary (distinguishing between Modern Standard Arabic and dialects), regional (such as Gulf, Iraqi, Levantine, Egyptian, and North African dialects), and country-specific (for example, Moroccan, Algerian, Saudi dialects, etc.) [6].

To perform the dialect identification task, two distinct datasets are commonly used. The first is MSDA [9] Arabic Dialect Detection for Social Media Posts, which is a labeled dataset with around 50K sentences. The second is IADD [32] which stands for Integrated Arabic Dialect Identification Dataset. This dataset consists of three categories: region (MGH, LEV, EGY, IRQ, GLF, or general), country (MAR, TUN, DZ, EGY, IRQ, SYR, JOR, PSE, LBN), and data source (PADIC, DART, AOC, SHAMI, or TSAC). To simplify the task, we transformed the IADD dataset into a binary format where one label indicates the Moroccan dialect, while the other label represents all other dialects. Table 4 provides additional details about these datasets.

Language classification is the process of automatically identifying the language of a given text [28]. This task is an important part of natural language processing, as it is often necessary to know the language of a text in order to properly analyze or process it. In our language classification task, we utilized our own dataset of Facebook comments, which we manually annotated with seven different labels, including Dialect in Latin Alphabet (DAL), Dialect in Arabic Alphabet (DAA), Classical Arabic (ARC), French (FRN), English (ANG), Spanish (ESP), and Others (AUT). Table 5 provides a detailed description of the dataset, including the content and label descriptions.

5 Results and Discussion

We conducted experiments for each of the tasks described earlier by fine-tuning the MorrBert and MorRoBERTa models, and comparing their performance with that of other models in the field, such as mBERT, XLM-R, CamelBERT-mix [18], MARBERT, and DBERT-mix [14]. The models' performance was evaluated using the F1 and accuracy measures.

All models were fine-tuned for four epochs using a batch size of 16 and the Adam optimizer [19]. The default values were maintained for other hyperparameters. We split the data into 80% for training and 20% for testing, using a random stratified split. We con-

Table 3: An overview of the evaluation datasets for the sentiment analysis task.

Dataset	Size	Number of Class	Number of Labels	Positive polarity labels	Negative polarity labels	Neutral labels	Domain
MAC	1.71 Mo	2	4	10,671	2,057	17,272	Twitter
Sentiment Anaysis MSDA	6.53 Mo	1	3	6,792	15,385	30,033	Twitter

Table 4: An overview of the evaluation datasets for the dialect identification task.

Dataset	Size	Number of Class	Number of Labels	Moroccan polarity labels	Total Dialect	Domain
IADD	24.6 Mo	3	10	7,213	135,804	Varied
Dialect MSDA	6.62 Mo	1	5	6,792	49,306	Twitter

Table 5: An overview of the evaluation dataset for the language classification task.

Size	Number of Labels	Total	Domain	Labels	Description	Number of Comments
7.3 Mo	7	75,509	Facebook	DAL	Comment in Dialect Latin Alphabet	28,143
				DAA	Comment in Dialect Arabic Alphabet	27,012
				ARC	Comment in Classic Arabic	13,324
				FRN	Comment in French	5,300
				ANG	Comment in English	1,172
				ESP	Comment in Spanish	251
				AUT	If a comment meets any of the following criteria: other languages, only Facebook usernames, mixed Arabic/Latin characters, named entities, only numbers, or ambiguous	30

ducted these experiments using GPUs in Google Colaboratory[2].

The experimental results, comparing the performance of MorrBert and MorRoBERTa models to other models across various downstream tasks, are presented in Tables 6, 7, and 8.

5.1 Sentiment Analysis

The results of the sentiment analysis task reveal the impressive performance of MorrBert and MorRoBERTa in terms of accuracy and F1 scores on both the MAC and MSDA datasets, as displayed in Table 6. MorRoBERTa achieved F1 scores of 74.44% and 78.07% for the MAC and MSDA datasets, respectively, while MorrBert achieved F1 scores of 75.13% and 76.50% for the same datasets. These scores are comparable to or even surpass the performance of other models in the field, such as CamelBERT-mix, MARBERT, and XLM-R.

Interestingly, we found that the DBERT-mix model, which is designed specifically for the Moroccan dialect of Arabic, did not perform as well as expected. This may suggest that models trained on more general varieties of Arabic are better suited for sentiment analysis tasks on these datasets.

[2]https://colab.research.google.com/

5.2 Dialect Identification

Table 7 provides a comprehensive overview of the accuracy and F1 scores for dialect identification. Our results show that MorRoBERTa and MorRBERT perform similarly on the MSDA dataset, with MorRoBERTa achieving slightly higher scores in both F-1 and accuracy measures. However, on the IADD dataset, both models significantly outperform the other models, with MorRoBERTa achieving the highest scores in both measures. DarijaBERT-mix performs worse than the other models on both datasets, while MARBERT and CamelBERT-mix perform well, especially on the IADD dataset.

5.3 Moroccan Language Classification

In the language classification task, the findings from Table 8 reveal that MorrBert and MorRoBERTa outperformed other models in terms of both accuracy and F1 score. MorrBert achieved an impressive F1 score of 91.06%, while MorRoBERTa achieved a notable F1 score of 90.33%. Although DBERT-mix, MARBERT, CamelBERT-mix, mBERT, and XLM-R also demonstrated high accuracy and F1 scores, their performance was slightly lower compared to MorrBert and MorRoBERTa.

Overall, the experiments showed that fine-tuned MorrBert and MorRoBERTa models are effective for sentiment analysis, dialect identification, and language

Table 6: Sentiment Analysis accuracy and F1-Macro scores on the 10k MAC and MSDA Dataset.

	MAC		MSDA	
Model	Acc.	F1	Acc.	F1
CamelBERT-mix	83.8	80.1	**85.2**	77.9
DBERT-mix	77.4	72.2	83.9	75.9
MARBERT	**85.0**	**82.2**	84.8	77.6
mBERT	74.1	67.8	83.2	75.0
XLM-R	75.1	69.1	83.4	75.8
MorrBERT	79.2	75.1	84.4	76.5
MorRoBERTa	78.9	74.4	85.1	**78.1**

Table 7: Dialect Identification accuracy and F1-Macro scores on the 10k MSDA and IADD Datasets.

	MSDA		IADD	
Model	Acc.	F1	Acc.	F1
CamelBERT-mix	**81.4**	75.8	99.4	96.9
DBERT-mix	76.1	70.9	98.9	93.9
MARBERT	82.7	**78.0**	99.5	97.4
mBERT	76.4	70.8	99.1	94.8
XLM-R	69.0	62.7	98.2	90.4
MorrBERT	75.1	67.0	99.2	95.3
MorRoBERTa	78.2	72.8	**99.6**	**98.6**

classification tasks on Moroccan datasets, and can achieve comparable or better performance than other state-of-the-art models in the field. Additionally, models specifically designed for Arabic dialects such as MARBERT and CamelBERT-mix also show promising results.

6 Conclusion

This article outlines our proposed process for creating a large Moroccan Dialect dataset as well as our approach for pre-training two BERT-like models. Our aim was to compile a dataset with over six million sentences. written and shared by Moroccan users on three different platforms: Facebook, Twitter, and YouTube. In the end, we obtained a dataset of more than 71 billion tokens, which we standardized using a series of preprocessing steps. We used two different tokenizers - a Byte-level BPE tokenizer for MorRoBERTa and a WordPiece tokenizer for MorrBERT - and both models were trained using MLM pre-training. To build the models, we utilized the HuggingFace Transformers API. and we executed them on the HPC-MARWAN platform using GPU cards for accelerated performance.

The models exhibited different specifications and training times, highlighting the importance of selecting the appropriate configuration for a model based on the task at hand and the computational resources available for training. Through our pre-training process, we developed two robust models for Moroccan Dialect, and we hope that this work will contribute to the growing body of research on NLP for low-resource languages.

Our models are publicly available for research purposes via the Hugging Face repository[3].

[3]https://huggingface.co/otmangi

Table 8: Language Classification accuracy and F1-Macro scores on our Dataset of Facebook Comments.

	Our DATA	
Model	Acc.	F1
CamelBERT-mix	93.9	87.7
DBERT-mix	94.0	87.7
MARBERT	94.1	88.0
mBERT	93.5	88.0
XLM-R	93.1	87.1
MorrBERT	**95.2**	**91.1**
MorRoBERTa	94.9	90.3

Acknowledgement: This research was supported through computational resources of HPC-MARWAN (www.marwan.ma/hpc) provided by the National Center for Scientific and Technical Research (CNRST), Rabat, Morocco.

References

[1] High performance computing (hpc). https://www.marwan.ma/index.php/en/services/hpc [Retrieved May 17, 2022].

[2] La constitution, edition 2011. http://www.sgg.gov.ma/Portals/0/constitution/constitution_2011_Fr.pdf [Retrieved April 18, 2023].

[3] Top most-commented youtube channels in morocco — hypeauditor. https://hypeauditor.com/top-youtube-all-morocco/most-commented/ [Retrieved April 15, 2023].

[4] Abdelali, A., Hassan, S., Mubarak, H., Darwish, K., and Samih, Y. Pre-training bert on arabic tweets: Practical considerations. *arXiv preprint arXiv:2102.10684* (2021).

[5] Abdul-Mageed, M., Elmadany, A., and Nagoudi, E. M. B. Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785* (2020).

[6] Abdul-Mageed, M., Zhang, C., Elmadany, A., Bouamor, H., and Habash, N. Nadi 2021: The second nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2103.08466* (2021).

[7] Antoun, W., Baly, F., and Hajj, H. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104* (2020).

[8] Bhatia, S., Sharma, M., and Bhatia, K. K. *Sentiment Analysis and Mining of Opinions*. Springer International Publishing, Cham, 2018, pp. 503–523.

[9] Boujou, E., Chataoui, H., Mekki, A. E., Benjelloun, S., Chairi, I., and Berrada, I. An open access nlp dataset for arabic dialects: Data collection, labeling, and model construction. *arXiv preprint arXiv:2102.11000* (2021).

[10] Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[11] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[12] Conneau, A., and Lample, G. Cross-lingual language model pretraining. *Advances in neural information processing systems 32* (2019).

[13] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] Gaanoun, K., Naira, A. M., Allak, A., and Benelallam, I. Darijabert: a step forward in nlp for the written moroccan dialect. *ResearchSquare* (2023). https://doi.org/10.21203/rs.3.rs-2560653/v1.

[15] Gani, M. O., Ayyasamy, R. K., Sangodiah, A., and Fui, Y. T. Ustw vs. stw: A comparative analysis for exam question classification based on bloom's taxonomy. *MENDEL 28*, 2 (2022), 25–40.

[16] Garouani, M., and Kharroubi, J. Mac: an open and free moroccan arabic corpus for sentiment analysis. In *The Proceedings of the International Conference on Smart City Applications* (2021), Springer, pp. 849–858.

[17] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[18] Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., and Habash, N. The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678* (2021).

[19] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[20] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[21] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[22] Ruder, S. Why you should do nlp beyond english. https://ruder.io/nlp-beyond-english [Retrieved April 18, 2023].

[23] Samih, Y., and Maier, W. An arabic-moroccan darija code-switched corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (2016), pp. 4170–4175.

[24] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[25] Sennrich, R., Haddow, B., and Birch, A. Das hunderttage-stadion: Entstehungsgeschichte des bad nauheimer kunsteisstadions unter colonel paul r. knight. *Acl* (2016), 1715–1725.

[26] Tachicart, R., Bouzoubaa, K., and Jaafar, H. Lexical differences and similarities between moroccan dialect and arabic. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)* (2016), IEEE, pp. 331–337.

[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

[28] Wijaya, M. C. The classification of documents in malay and indonesian using the naive bayesian method uses words and phrases as a training set. *MENDEL 26*, 2 (2020), 23–28.

[29] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (2020), pp. 38–45.

[30] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

[31] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems 32* (2019).

[32] Zahir, J. Iadd: An integrated arabic dialect identification dataset. *Data in Brief 40* (2022), 107777.