ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic

Muhammad Abdul-Mageed[†] **AbdelRahim Elmadany**[†] **El Moatez Billah Nagoudi**[†]
Natural Language Processing Lab
The University of British Columbia
{muhammad.mageed,a.elmadany,moatez.nagoudi}@ubc.ca

Abstract

Pre-trained language models (LMs) are currently integral to many natural language processing systems. Although multilingual LMs were also introduced to serve many languages, these have limitations such as being costly at inference time and the size and diversity of non-English data involved in their pre-training. We remedy these issues for a collection of diverse Arabic varieties by introducing two powerful deep bidirectional transformer-based models, ARBERT and MARBERT. To evaluate our models, we also introduce ARLUE, a new benchmark for multi-dialectal Arabic language understanding evaluation. ARLUE is built using 42 datasets targeting six different task clusters, allowing us to offer a series of standardized experiments under rich conditions. When fine-tuned on ARLUE, our models collectively achieve new state-of-the-art results across the majority of tasks (37 out of 48 classification tasks, on the 42 datasets). Our best model acquires the highest ARLUE score (77.40) across all six task clusters, outperforming all other models including XLM-R_{Large} ($\sim$ 3.4$\times$ larger size). Our models are publicly available at https://github.com/UBC-NLP/marbert and ARLUE will be released through the same repository.

1 Introduction

Language models (LMs) exploiting self-supervised learning such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019a) have recently emerged as powerful transfer learning tools that help improve a very wide range of natural language processing (NLP) tasks. Multilingual LMs such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (XLM-R) (Conneau et al., 2020) have also been introduced, but are usually outperformed by monolingual models pre-trained with larger vocabulary and bigger language-specific datasets (Virtanen et al., 2019; Antoun et al., 2020; Dadas et al., 2020; de Vries et al., 2019; Le et al., 2020; Martin et al., 2020; Nguyen and Tuan Nguyen, 2020).

Since LMs are costly to pre-train, it is important to keep in mind the end goals they will serve once developed. For example, **(i)** in addition to their utility on 'standard' data, it is useful to endow them with ability to excel on wider real world settings such as in social media. Some existing LMs do not meet this need since they were trained on datasets that do not sufficiently capture the nuances of social media language (e.g., frequent use of abbreviations, emoticons, and hashtags; playful character repetitions; neologisms and informal language). It is also desirable to build models able to **(ii)** serve diverse communities (e.g., speakers of dialects of a given language), rather than focusing only on mainstream varieties. In addition, once created, models should be **(iii)** usable in energy efficient scenarios. This means that, for example, medium-to-large models with competitive performance should be preferred to large-to-mega models.

A related issue is **(iv)** how LMs are evaluated. Progress in NLP hinges on our ability to carry out meaningful comparisons across tasks, on carefully designed benchmarks. Although several benchmarks have been introduced to evaluate LMs, the majority of these are either exclusively in English (e.g., DecaNLP (McCann et al., 2018), GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019)) or use machine translation in their training splits (e.g., XTREME (Hu et al., 2020)). Again, useful as these benchmarks are, this circumvents our ability to measure progress in real-world settings (e.g., training and evaluation on native vs. translated data) for both cross-lingual NLP and in monolingual, non-English environments.

Context. Our objective is to showcase a scenario where we build LMs that meet *all* four needs listed above. That is, we describe novel LMs that (i) excel across domains, including social media, (ii) can serve diverse communities, and (iii) perform well compared to larger (more energy hungry) mod-

[†] All authors contributed equally.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 7088–7105
August 1–6, 2021.

els (iv) on a novel, standardized benchmark. We choose Arabic as the context for our work since it is a widely spoken language ($\sim$ 400M native speakers), with a large number of diverse dialects differing among themselves and from the standard variety, Modern Standard Arabic (MSA). Arabic is also covered by the popular mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), which provides us a setup for meaningful comparisons. That is, not only are we able to empirically measure monolingual vs. multilingual performance under robust conditions using our new benchmark, ARLUE, but we can also demonstrate how our base-sized models outperform (or at least are on par with) larger models (i.e., XLM-R_{Large}, which is $\sim 3.4\times$ larger than our models). In the context of our work, we also show how the currently best-performing model dedicated to Arabic, AraBERT (Antoun et al., 2020), suffers from a number of issues. These include (a) not making use of easily accessible data across domains and, more seriously, (b) limited ability to handle Arabic dialects and (c) narrow evaluation. We rectify all these limitations.

Our contributions. With our stated goals in mind, we introduce **ARBERT** and **MARBERT**, two Arabic-focused LMs exploiting large-to-massive diverse datasets. For evaluation, we also introduce a novel **AR**abic natural **L**anguage **U**nderstanding **E**valuation benchmark (**ARLUE**). ARLUE is composed of 42 different datasets, making it by far the largest and most diverse Arabic NLP benchmark we know of. We arrange ARLUE into six coherent cluster tasks and methodically evaluate on each independent dataset as well as each cluster task, ultimately reporting a single ARLUE score. Our models establish new state-of-the-art (SOTA) on the majority of tasks, across all cluster tasks. Our goal is for ARLUE to serve the critical need for measuring progress on Arabic, and facilitate evaluation of multilingual and Arabic LMs. To summarize, we offer the following contributions:

1. **We develop ARBERT and MARBERT**, two novel Arabic-specific Transformer LMs pre-trained on very large and diverse datasets to facilitate transfer learning on MSA as well as Arabic dialects.

2. **We introduce ARLUE**, a new benchmark developed by collecting and standardizing splits on 42 datasets across six different Arabic language understanding cluster tasks, thereby facilitating measurement of progress on Arabic and multilingual NLP.

3. We fine-tune our new powerful models on ARLUE and provide an extensive set of comparisons to available models. **Our models achieve new SOTA** on all task clusters in 37 out of 48 individual datasets and a SOTA *ARLUE score*.

The rest of the paper is organized as follows: In Section 2, we provide an overview of Arabic LMs. Section 3 describes our Arabic pre-tained models. We evaluate our models on downstream tasks in Section 4, and present our benchmark ARLUE and evaluation on it in Section 5. Section 6 is an overview of related work. We conclude in Section 7. We now introduce existing Arabic LMs.

2 Arabic LMs

The term *Arabic* refers to a collection of languages, language varieties, and dialects. The standard variety of Arabic is MSA, and there exists a large number of dialects that are usually defined at the level of the region or country (Abdul-Mageed et al., 2020a, 2021a,b). A number of Arabic LMs has been developed. The most notable among these is AraBERT (Antoun et al., 2020), which is trained with the same architecture as BERT (Devlin et al., 2019) and uses the $BERT_{Base}$ configuration. AraBERT is trained on 23GB of Arabic text, making $\sim$ 70M sentences and 3B words, from Arabic Wikipedia, the Open Source International dataset (OSIAN) (Zeroual et al., 2019) (3.5M news articles from 24 Arab countries), and 1.5B words Corpus from El-Khair (2016) (5M articles extracted from 10 news sources). Antoun et al. (2020) evaluate AraBERT on three Arabic downstream tasks. These are (1) sentiment analysis from six different datasets: HARD (Elnagar et al., 2018), ASTD (Nabil et al., 2015), ArsenTD-Lev (Baly et al., 2019), LABR (Aly and Atiya, 2013), and ArSaS (Elmadany et al., 2018). (2) NER, with the ANERcorp (Benajiba and Rosso, 2007), and (3) Arabic QA, on Arabic-SQuAD and ARCD (Mozannar et al., 2019) datasets. Another Arabic LM that was also introduced is ArabicBERT (Safaya et al., 2020), which is similarly based on BERT architecture. ArabicBERT was pre-trained on two datasets only, Arabic Wikipedia and

Arabic OSACAR (Suárez et al., 2019). Since both of these datasets are already included in AraBERT, and Arabic OSACAR[1] has significant duplicates, we compare to AraBERT only. GigaBERT (Lan et al., 2020), an Arabic and English LM designed with code-switching data in mind, was also introduced.[2]

3 Our Models

3.1 ARBERT

3.1.1 Training Data

We train ARBERT on 61GB of MSA text (6.5B tokens) from the following sources:

- **Books (Hindawi)**. We collect and pre-process 1,800 Arabic books from the public Arabic bookstore Hindawi.[3]
- **El-Khair**. This is a 5M news articles dataset from 10 major news sources covering eight Arab countries from El-Khair (2016).
- **Gigaword**. We use Arabic Gigaword 5th Edition from the Linguistic Data Consortium (LDC).[4] The dataset is a comprehensive archive of newswire text from multiple Arabic news sources.
- **OSCAR**. This is the MSA and Egyptian Arabic portion of the Open Super-large Crawled Almanach coRpus (Suárez et al., 2019),[5] a huge multilingual subset from Common Crawl[6] obtained using language identification and filtering.
- **OSIAN**. The Open Source International Arabic News Corpus (OSIAN) (Zeroual et al., 2019) consists of 3.5 million articles from 31 news sources in 24 Arab countries.
- **Wikipedia Arabic**. We download and use the December 2019 dump of Arabic Wikipedia. We use WikiExtractor[7] to extract articles and remove markup from the dump.

[1] https://oscar-corpus.com.

[2] Since GigaBERT is very recent, we could not compare to it. However, we note that our pre-training datasets are much larger (i.e., 15.6B tokens for MARBERT vs. 4.3B Arabic tokens for GigaBERT) and more diverse.

[3] https://www.hindawi.org/books/.

[4] https://catalog.ldc.upenn.edu/LDC2011T11.

[5] https://oscar-corpus.com/.

[6] https://commoncrawl.org.

[7] https://github.com/attardi/wikiextractor.

Source	Size	#Tokens
Books (Hindawi)	650MB	72.5M
El-Khair	16GB	1.6B
Gigawords	10GB	1.1B
OSIAN	2.8GB	292.6M
OSCAR-MSA	31GB	3.4B
OSCAR-Egyptian	32MB	3.8M
Wiki	1.4GB	156.5M
Total	**61GB**	**6.5B**

Table 1: ARBERT 's pre-train resources.

We provide relevant size and token count statistics about the datasets in Table 1.

3.1.2 Training Procedure

Pre-processing. To prepare the raw data for pre-training, we perform light pre-processing. This helps retain a faithful representation of the naturally occurring text. We only remove diacritics and replace URLs, user mentions, and hashtags that may exist in any of the collections with the generic string tokens `URL`, `USER`, and `HASHTAG`, respectively. We do not perform any further pre-processing of the data before splitting the text off to wordPieces (Schuster and Nakajima, 2012). Multilingual models such as mBERT and XLM-R have 5K (out of 110K) and 14K (out of 250K) Arabic WordPieces, respectively, in their vocabularies. AraBERT employs a vocabulary of 60K (out of 64K).[8] For ARBERT, we use a larger vocabulary of 100K WordPieces. For tokenization, we use the WordPiece tokenizer (Wu et al., 2016) provided by Devlin et al. (2019).

Pre-training. For ARBERT, we follow Devlin et al. (2019)'s pre-training setup. To generate each training input sequence, we use the whole word masking, where 15% of the N input tokens are selected for replacement. These tokens are replaced 80% of the time with the [MASK] token, 10% with a random token, and 10% with the original token. We use the original implementation of BERT in the TensorFlow framework.[9] As mentioned, we use the same network architecture as $\text{BERT}_{\text{Base}}$: 12 layers, 768 hidden units, 12 heads, for a total of $\sim$ 163M parameters. We use a batch size of 256 sequences and a maximum sequence length of 128 tokens (256 sequences $\times$ 128 tokens = 32,768 tokens/batch) for 8M steps, which is approximately 42 epochs over the 6.5B tokens. For all our models, we use a learning rate of 1e$-$4.

[8] The empty 4K vocabulary bin is reserved for additional wordPieces, if needed.

[9] https://github.com/google-research/bert.

We pre-train the model on one Google Cloud TPU with eight cores (v2.8) from TensorFlow Research Cloud (TFRC).[10] Training took $\sim$ 16 days, for 42 epochs over all the tokens. Table 2 shows a comparison of ARBERT with mBERT, XLM-R, AraBERT, and MARBERT (see Section 3.2) in terms of data sources and size, vocabulary size, and model parameters.

3.2 MARBERT

As we pointed out in Sections 1 and 2, Arabic has a large number of diverse dialects. Most of these dialects are under-studied due to rarity of resources. Multilingual models such as mBERT and XLM-R are trained on mostly MSA data, which is also the case for AraBERT and ARBERT. As such, these models are not best suited for downstream tasks involving dialectal Arabic. To treat this issue, we use a large Twitter dataset to pre-train a new model, MARBERT, from scratch as we describe next.

3.2.1 Training data

To pre-train MARBERT, we randomly sample 1B Arabic tweets from a large in-house dataset of about 6B tweets. We only include tweets with at least three Arabic words, based on character string matching, regardless whether the tweet has non-Arabic string or not. That is, we do not remove non-Arabic so long as the tweet meets the three Arabic word criterion. The dataset makes up 128GB of text (15.6B tokens).

3.2.2 Training Procedure

Pre-processing. We employ the same pre-processing as ARBERT.

Pre-training. We use the same network architecture as $\text{BERT}_{\text{Base}}$, but *without* the next sentence prediction (NSP) objective since tweets are short.[11] We use the same vocabulary size (100K wordPieces) as ARBERT, and MARBERT also has $\sim$ 160M parameters. We train MARBERT for 17M steps ($\sim$ 36 epochs) with a batch size of 256 and a maximum sequence length of 128. Training took $\sim$ 40 days on one Google Cloud TPU (eight cores). We now present a comparison between our models and popular multilingual models as well as AraBERT.

[10] https://www.tensorflow.org/tfrc.

[11] It was also shown that NSP is *not* crucial for model performance (Liu et al., 2019a).

3.3 Model Comparison

Our models compare to mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) (base and large), and AraBERT (Antoun et al., 2020) in terms of training data size, vocabulary size, and overall model capacity as we summarize in Table 2. In terms of the actual Arabic variety involved, Devlin et al. (2019) train mBERT with Wikipedia Arabic data, which is MSA. XLM-R (Conneau et al., 2020) is trained on Common Crawl data, which likely involves a small amount of Arabic dialects. AraBERT is trained on MSA data only. ARBERT is trained on a large collection of MSA datasets. Unlike all other models, our MARBERT model is trained on Twitter data, which involves both MSA and diverse dialects. We now describe our fine-tuning setup.

3.4 Model Fine-Tuning

We evaluate our models by fine-tuning them on a wide range of tasks, which we thematically organize into six clusters: **(1)** sentiment analysis (SA), **(2)** social meaning (SM) (i.e., age and gender, dangerous and hateful speech, emotion, irony, and sarcasm), **(3)** topic classification (TC), **(4)** dialect identification (DI), **(5)** named entity recognition (NER), and **(6)** question answering (QA). For all classification tasks reported in this paper, we compare our models to four other models: mBERT, $\text{XLM-R}_{\text{Base}}$, $\text{XLM-R}_{\text{Large}}$, and AraBERT. We note that $\text{XLM-R}_{\text{Large}}$ is $\sim 3.4\times$ larger than any of our own models ($\sim$ 550M parameters vs. $\sim$ 160M). We offer two main types of evaluation: on ***(i) individual tasks***, which allows us to compare to other works on each individual dataset (48 classification tasks on 42 datasets), and ***(ii) ARLUE clusters*** (six task clusters).

For all reported experiments, we follow the same light pre-processing we use for pre-training. For all individual tasks and ARLUE task clusters, we fine-tune on the respective training splits for 25 epochs, identifying the best epoch on development data, and reporting on both development and test data.[12] We typically use the exact data splits provided by original authors of each dataset. Whenever no clear

[12] A minority of datasets came with no development split from source, and so we identify and report the best epoch only on test data for these. This allows us to compare all the models under the same conditions (25 epochs) and report a fair comparison to the respective original works. For *all* ARLUE cluster tasks, we identify the best epoch *exclusively* on our development sets (shown in Table 10).

Models	Training Data		Vocabulary		Configuration	
	Source	Tokns (ar/all)	Tok	Size (ar/all)	B / L	Param.
mBERT	Wiki.	153M/1.5B	WP	5K/110K	B	110M
XLM-R_B	CC	2.9B/295B	SP	14K/250K	B	270M
XLM-R_L	CC	2.9B/295B	SP	14K/250K	L	550M
AraBERT	3 sources	2.5B/2.5B	SP	60K/64K	B	135M
ARBERT	6 sources	6.2B/6.2B	WP	100K/100K	B	163M
MARBERT	Ara. Tweets	15.6B/15.6B	WP	100K/100K	B	163M

Table 2: Models compared. **B:** Base, **L:** Large, **CC:** Common Crawel, **SP:** SentencePiece, **WP:** WordPiece.

splits are available, or in cases where expensive cross-validation was used in source, we divide the data following a standard 80% training, 10% development, and 10% test split. For all experiments, whether on individual tasks or ARLUE task clusters, we use the Adam optimizer (**?**) with input sequence length of 256, a batch size of 32, and a learning rate of 2e−6. These values were identified in initial experiments based on development data of a few tasks.[13] We now introduce individual tasks.

4 Individual Downstream Tasks

4.1 Sentiment Analysis

Datasets. We fine-tune the language models on all publicly available SA datasets we could find in addition to those we acquired directly from authors. In total, we have the following 17 MSA and DA datasets: AJGT (Alomari et al., 2017), AraNET$_{Sent}$ (Abdul-Mageed et al., 2020b), AraSenTi-Tweet (Al-Twairesh et al., 2017), ArSarcasm$_{Sent}$ (Farha and Magdy, 2020), ArSAS (Elmadany et al., 2018), ArSenD-Lev (Baly et al., 2019), ASTD (Nabil et al., 2015), AWATIF (Abdul-Mageed and Diab, 2012), BBNS & SYTS (Salameh et al., 2015), CAMel$_{Sent}$ (Obeid et al., 2020), HARD (Elnagar et al., 2018), LABR (Aly and Atiya, 2013), Twitter$_{Abdullah}$ (Abdulla et al., 2013), Twitter$_{Saad}$,[14] and SemEval-2017 (Rosenthal et al., 2017). Details about the datasets and their splits are in Section A.1.

Baselines. We compare to the STOA listed in Table 3 and Table 4 captions. For all datasets with no baseline in Table 3, we consider AraBERT our baseline. Details about SA baselines are in Section A.2.

Results. To facilitate comparison to previous works with the appropriate evaluation metrics, we split our results into two tables: We show results in F_1^{PN} in Table 3 and F_1 in Table 4. We typically **bold** the best result on each dataset. ***Our models achieve best results in*** 13 ***out of the*** 17 ***classification tasks reported in the two tables combined***, while XLM-R (which is a much larger model) outperforms our models in the 4 remaining tasks. We also note that XLM-R acquires better results than AraBERT in the majority of tasks, a trend that continues for the rest of tasks. Results also clearly show that MARBERT is more powerful than than ARBERT. This is due to MARBERT's larger and more diverse pre-training data, especially that many of the SA datasets involve dialects and come from social media.

Dataset (classes)	SOTA	mBERT	XLM-R_B	XLM-R_L	AraBERT	ARBERT	MARBERT
ArSAS (3)	92.00*	87.50	90.00	91.50	91.00	92.00	**93.00**
ASTD (3)	73.00*	67.00	60.67	67.67	72.00	76.50	**78.00**
SemEval (3)	69.00*	57.00	64.00	67.00	62.00	69.00	**71.00**
AraNET$_{Sent}$ (2)	76.20†	84.00	92.00	**93.00**	86.50	89.00	92.00
ArSarc$_{Sent}$ (3)	-	60.50	63.50	70.00	63.50	68.00	**71.50**
AraSenTi (3)	-	89.50	92.00	**93.50**	91.00	90.00	90.00
BBN (3)	-	55.50	69.50	72.00	70.00	76.50	**79.00**
SYTS (3)	-	67.00	78.00	76.50	75.50	**79.00**	76.50
Tw$_{Saad}$ (2)	-	79.00	95.00	95.00	81.00	90.00	**96.00**
SAMAR (5)	-	22.50	54.00	**57.00**	36.50	43.50	55.50
AWATIF (4)	-	60.50	63.50	68.50	66.50	71.50	**72.50**
Tw$_{Abdullah}$ (2)	-	81.50	91.00	92.00	89.50	91.50	**95.00**

Table 3: SA results (I) in F_1^{PN}. * Obeid et al. (2020); † Abdul-Mageed et al. (2020b). Default baseline is AraBERT.

Dataset (classes)	SOTA	mBERT	XLM-R_B	XLM-R_L	AraBERT	ARBERT	MARBERT
AJGT (2)	93.80	86.67	89.44	91.94	92.22	94.44	**96.11**
HARD (2)	96.20	95.54	95.74	95.96	95.89	96.12	**96.17**
ArsenTD-LEV (5)	59.40	50.50	55.25	**62.00**	56.13	61.38	60.38
LABR (2)	86.70	91.20	91.23	92.20	91.97	**92.51**	92.49
ASTD-B(2)	92.60	79.32	87.59	77.44	83.08	93.23	**96.24**

Table 4: SA results (II) in Acc. SOTA by Antoun et al. (2020).

4.2 Social Meaning Tasks

We collectively refer to a host of tasks as **social meaning**. These are age and gender detection; dangerous, hateful, and offensive speech detection; emotion detection; irony detection; and sarcasm detection. We now describe datasets we use for each of these tasks.

Datasets. For both age and gender, we use

[13] NER and QA are expetions, where we use sequence lengths of 128 and 384, respectively; a batch sizes of 16 for both; and a learning rate of 2e−6 and 3e−5, respectively.

[14] www.kaggle.com/mksaad/arabic-sentiment-twitter.

Task (classes)	SOTA	mBERT	XLM-R$_B$	XLM-R$_L$	AraBERT	ARBERT	MARBERT
Age (3)	51.42 ‡‡	56.35	59.73	53.60	57.72	58.95	**62.27**
Dangerous (2)	59.60 †	62.66	62.76	65.01	64.37	63.21	**67.53**
Emotion (8)	60.32 ‡‡	65.79	70.67	74.89	65.68	67.73	**75.83**
Gender (2)	65.30 ‡‡	68.06	71.00	71.14	67.75	69.86	**72.62**
Hate (2)	82.28**	72.81	71.33	79.31	78.89	83.02	**84.79**
Irony (2)	82.40 ‡	80.96	81.97	82.52	83.01	**85.59**	85.33
Offensive (2)	90.51*	84.25	85.26	88.28	86.57	90.38	**92.41**
Sarcasm (2)	46.60 ‡‡	68.20	66.76	69.23	72.23	75.04	**76.30**

Table 5: Results on social meaning tasks. F_1 score is the evaluation metric. * Hassan et al. (2020), ** Djandji et al. (2020), ‡ Zhang and Abdul-Mageed (2019a), † **?**, †† Farha and Magdy (2020), ‡‡ Abdul-Mageed et al. (2020b).

Arap-Tweet (Zaghouani and Charfi, 2018). We use AraDan (Alshehri et al., 2020) for dangerous speech. For offensive language and hate speech, we use the dataset released in the shared task (sub-tasks A and B) of offensive speech by Mubarak et al. (2020). We also use AraNET$_{Emo}$ (Abdul-Mageed et al., 2020b), IDAT@FIRE2019 (Ghanem et al., 2019), and ArSarcasm (Farha and Magdy, 2020) for emotion, irony and sarcasm, respectively. More information about these datasets and their splits is in Appendix B.1.

Baselines. Baselines for social meaning tasks are the SOTA listed in Table 5 caption. Details about each baseline is in Appendix B.2.

Results. As Table 5 shows, our models acquire best results on all eight tasks. Of these, MARBERT achieves best performance on seven tasks, while ARBERT is marginally better than MARBERT on one task (irony@FIRE2019). ***The sizeable gains MARBERT achieves reflects its superiority on social media tasks. On average, our models are 9.83 F_1 better than all previous SOTA.***

4.3 Topic Classification

Classifying documents by topic is a classical task that still has practical utility. We use four TC datasets, as follows:

Datasets. We fine-tune on Arabic News Text (ANT) (Chouigui et al., 2017) under three pretaining settings (*title only*, *text only*, and *title+text*.), Khaleej (Abbas et al., 2011), and OSAC (Saad and Ashour, 2010). Details about these datasets and the classes therein are in Appendix C.1.

Baselines. Since, to the best of our knowledge, there are no published results exploiting deep learning on TC, we consider AraBERT a strong baseline.

Results. As Table 6 shows, ***ARBERT acquires best results on both OSAC and Khaleej, and the title-only setting of ANT.*** AraBERT slightly outperforms our models on the text-only and title+text settings of ANT.

Dataset (classes)	mBERT	XLM-R$_B$	XLM-R$_L$	AraBERT	ARBERT	MARBERT
ANTText (5)	84.89	85.77	86.72	**88.17**	86.87	85.27
ANTTitle (5)	78.29	79.96	81.25	81.03	**81.70**	81.19
ANTText+Title (5)	84.67	86.21	86.96	**87.22**	87.21	85.60
Khallej (4)	92.81	91.87	93.56	93.83	**94.53**	93.63
OSAC (10)	96.84	97.15	**98.20**	97.03	97.50	97.23

Table 6: Performance on TC tasks. Our baseline is AraBERT.

Dataset (classes)	Task	SOTA	mBERT	XLM-R$_B$	XLM-R$_L$	AraBERT	ARBERT	MARBERT
ArSarc$_{Dia}$(5)	Regoin	-	43.81	41.71	41.83	47.54	**54.70**	51.27
MADAR (21)	Country	-	34.92	35.91	35.14	34.87	37.90	**40.40**
AOC (4)	Region	82.45*	77.27	77.34	78.77	79.20	81.09	**82.37**
AOC (3)	Region	78.81*	85.76	86.39	87.56	87.68	89.06	**90.85**
AOC (2)	Binary	87.23*	86.19	86.85	87.30	87.76	88.46	**88.59**
QADI (18)	Country	60.60†	66.57	77.00	82.73	72.23	88.63	**90.89**
NADI (21)	Country	26.78‡	13.32	16.36	17.17	17.46	22.56	**29.14**
NADI (100)	Province	06.06††	02.13	04.12	5.30	03.13	06.10	**06.28**

Table 7: DIA results in F_1. * Elaraby and Abdul-Mageed (2018), † Abdelali et al. (2020), ‡ El Mekki et al. (2020), †† Talafha et al. (2020). Default baseline is AraBERT.

4.4 Dialect Identification

Arabic dialect identification can be performed at different levels of granularity, including binary (i.e., MSA-DA), regional (e.g., *Gulf*, *Levantine*), country level (e.g., *Algeria*, *Morocco*), and recently province level (e.g., the Egyptian province of *Cairo*, the Saudi province of *Al-Madinah*) (Abdul-Mageed et al., 2020a, 2021b).

Datasets. We fine-tune our models on the following datasets: Arabic Online Commentary (AOC) (Zaidan and Callison-Burch, 2014), ArSarcasm$_{Dia}$ (Farha and Magdy, 2020),[15] MADAR (sub-task 2) (Bouamor et al., 2019), NADI-2020 (Abdul-Mageed et al., 2020a), and QADI (Abdelali et al., 2020). Details about these datasets are in Table D.1.

Baselines. Our baselines are marked in Table 7 caption. Details about the baselines are in Table D.2.

Results. As Table 7 shows, our models outperform all SOTA as well as the baseline AraBERT across all classification levels with sizeable margins. ***These results reflect the powerful and diverse dialectal representation of MARBERT, enabling it to serve wider communities***. Although ARBERT is developed mainly for MSA, it also outperforms all other models.

4.5 Named Entity Recognition

We fine-tune the models on five NER datasets.

Datasets. We use ACE03NW and ACE03BN (Mitchell et al., 2004), ACE04NW (Mitchell et al., 2004), ANERcorp (Benajiba and Rosso, 2007), and TW-NER (Darwish, 2013). Table E.1 shows the

[15] ArSarcasm$_{Dia}$ carries *regional* dialect labels.

Dataset	SOTA	mBERT	XLM-R_B	XLM-R_L	AraBERT	ARBERT	MARBERT
ANERcorp	88.77	86.78	87.24	**89.94**	89.13	84.38	80.64
ACE04NW	**91.47**	86.37	89.93	89.89	89.03	88.24	85.02
ACE03BN	94.92	91.23	53.97	85.41	91.94	**96.18**	79.05
ACE03NW	**91.20**	81.40	87.24	90.62	88.09	90.09	87.76
TW-NER	65.34	36.83	49.16	54.44	41.26	59.17	**66.67**

Table 8: NER results in F_1. SOTA by Khalifa and Shaalan (2019).

distribution of named entity classes across the five datasets.

Baseline. We compare our results with SOTA presented by Khalifa and Shaalan (2019) and follow them in focusing on person (PER), location (LOC) and organization (ORG) named entity labels while setting other labels to the unnamed entity (O). Details about Khalifa and Shaalan (2019) SOTA models are in Appendix E.2.

Results. As Table 8 shows, our models outperform SOTA on two out of the five NER datasets. We note that even though SOTA (Khalifa and Shaalan, 2019) employ a complex combination of CNNs and character-level LSTMs, which may explain their better results on two datasets, ***MARBERT still achieves highest performance on the social media dataset (TW-NER)***.

4.6 Question Answering

Datasets. We use ARCD (Mozannar et al., 2019) and the three *human* translated Arabic test sections of the XTREME benchmark (Hu et al., 2020): MLQA (Lewis et al., 2020), XQuAD (Artetxe et al., 2020), and TyDi QA (Artetxe et al., 2020). Details about these datasets are in Table F.1.

Baselines. We compare to Antoun et al. (2020) and consider their system a baseline on ARCD. We follow the same splits they used where we fine-tune on Arabic SQuAD (Mozannar et al., 2019) *and* 50% of ARCD and test on the remaining 50% of ARCD (ARCD-test). For all other experiments, we fine-tune on the Arabic *machine translated* SQuAD (AR-XTREME) from the *XTREME* multilingual benchmark (Hu et al., 2020) and test on the *human translated* test sets listed above. Our baselines in these is Hu et al. (2020)'s $mBERT_{Base}$ model on *gold* (human) data.

Results. As is standard, we report QA results in terms of both Exact Match (EM) and F_1. We find that results with ARBERT and MARBERT on QA are not competitive, a clear discrepancy from what we have observed thus far on other tasks. We hypothesize this is because the two models are pre-trained with a sequence length of only 128, which does not allow them to sufficiently capture both a question and its likely answer within the same sequence window during the pre-training.[16] To rectify this, we further pre-train the stronger model, MARBERT, on the same MSA data as ARBERT in addition to AraNews dataset (Nagoudi et al., 2020) (8.6GB), but with a bigger sequence length of 512 tokens for 40 epochs. We call this further pre-trained model **MARBERT-v2**, noting it has 29B tokens. As Table 9 shows, ***MARBERT-v2 acquires best performance on all but one test set***, where $XLM\text{-}R_{Large}$ marginally outperforms us (only in F_1).

5 ARLUE

5.1 ARLUE Categories

We concatenate the corresponding splits of the individual datasets to form ***ARLUE***, which is a conglomerate of task clusters. That is, we concatenate all training data from each group of tasks into a single TRAIN, all development into a single DEV, and all test into a single TEST. One exception is the social meaning tasks whose data we keep independent (see $\textbf{ARLUE}_{\textbf{SM}}$ below). Table 10 shows a summary of the ARLUE datasets.[17] We now briefly describe how we merge individual datasets into ARLUE.

$\textbf{ARLUE}_{\textbf{Senti}}$. To construct $ARLUE_{Senti}$, we collapse the labels *very negative* into *negative*, *very positive* into *positive*, and *objective* into *neutral*, and remove the *mixed* class. This gives us the 3 classes *negative*, *positive*, and *neutral* for $ARLUE_{Senti}$. Details are in Table A.1.

$\textbf{ARLUE}_{\textbf{SM}}$. We refer to the different social meaning datasets collectively as $ARLUE_{SM}$. We do not merge these datasets to preserve the conceptual coherence specific to each of the tasks. Details about individual datasets in $ARLUE_{SM}$ are in B.1.

$\textbf{ARLUE}_{\textbf{Topic}}$. We straightforwardly merge the TC datasets to form $ARLUE_{Topic}$, without modifying any class labels. Details of $ARLUE_{Topic}$ data are in Table C.1.

$\textbf{ARLUE}_{\textbf{Dia}}$. We construct three $ARLUE_{Dia}$ categories. Namely, we concatenate the AOC and $AraSarcasm_{Dia}$ MSA-DA classes to form $\boldsymbol{ARLUE_{Dia\text{-}B}}$ (binary) and the region level classes from the same two datasets to acquire $\boldsymbol{ARLUE_{Dia\text{-}R}}$ (4-classes, *region*). We then merge the country

[16] In addition, MARBERT is not trained on Wikipedia data from where some questions come.

[17] Again, $ARLUE_{SM}$ datasets are kept independent, but to provide a summary of all ARLUE datasets we collate the numbers in Table 10.

Dataset	SOTA		mBERT		XLM-R_B		XLM-R_L		AraBERT		ARBERT		MARBERT		MARBERT(v2)	
	EM	F_1	EM	F_1	EM	F_1	EM	F_1	EM	F_1	EM	F_1	EM	F_1	EM	F_1
ARCD-test*	30.10†	61.20†	29.63	60.93	30.20	59.55	32.05	64.77	30.20	62.30	30.34	63.89	21.65	54.06	**36.75**	**68.86**
ARCD-test	-	-	26.64	58.86	27.31	59.61	28.11	62.08	25.64	59.92	27.21	60.73	23.22	55.14	**29.63**	**63.05**
AR-MLQA	39.00‡	58.90 ‡	32.93	51.57	32.93	53.35	38.11	**60.00**	35.43	55.42	34.15	53.65	28.02	45.14	**39.23**	59.39
AR-XQuAD	54.20‡	71.00 ‡	48.66	66.26	45.88	64.91	51.85	72.19	51.60	68.79	49.92	67.90	41.09	58.46	**56.55**	**72.48**
AR-TyiDQA	39.00‡	58.90‡	46.36	64.02	39.41	60.99	44.41	67.06	44.19	64.39	46.80	66.94	38.98	57.51	**47.45**	**67.67**

Table 9: QA results. * Results on this test set are with models using the same training data as Antoun et al. (2020), while rest of rows report models trained with AR-XTREME (Hu et al., 2020). † Antoun et al. (2020); ‡ Hu et al. (2020).

Dataset	#Datasets	Task	TRAIN	DEV	TEST
$ARLUE_{Senti}$	17	SA	190.9K	6.5K	44.2K
$ARLUE_{SM}$*	8	SM	1.51M	162.5K	166.1K
$ARLUE_{Topic}$	5	TC	47.5K	5.9K	5.9K
$ARLUE_{Dia\text{-}B}$	2	DI	94.9K	10.8K	12.9K
$ARLUE_{Dia\text{-}R}$	2	DI	38.5K	4.5K	5.3K
$ARLUE_{Dia\text{-}C}$	3	DI	711.9K	31.5K	52.1K
$ARLUE_{NER}$†	5	NER	227.7K	44.1K	66.5K
$ARLUE_{QA}$‡	4	QA	101.6K	517	7.45K

Table 10: ARLUE categories across the different data splits. * Refer to Table B.1 for details about individual social meaning datasets in $ARLUE_{SM}$. † Statistics are at the token level. ‡ Number of question-answer pairs.

classes from the rest of datasets to get $\boldsymbol{ARLUE_{Dia\text{-}C}}$ (21-classes, *country*). Details are in Table D.1.

$ARLUE_{NER}$ & $ARLUE_{QA}$. We straightforwardly concatenate all corresponding splits from the different NER and QA datasets to form $\boldsymbol{ARLUE_{NER}}$ and $\boldsymbol{ARLUE_{QA}}$, respectively. Details of each of these task clusters data are in Tables E.1 and F.1, respectively.

5.2 Evaluation on ARLUE

We present results on each task cluster independently using the relevant metric for both the development split (Table 11) and test split (Table 12). Inspired by McCann et al. (2018) and Wang et al. (2018) who score NLP systems based on their performance on multiple datasets, we introduce an ***ARLUE score***. The ARLUE score is simply the macro-average of the different scores across all task clusters, weighting each task equally. Following Wang et al. (2018), for tasks with multiple metrics (e.g., accuracy and F_1), we use an unweighted average of the metrics as the score for the task when computing the overall macro-average. As Table 12 shows, ***our MARBERT-v2 model achieves the highest ARLUE score*** (77.40), followed by XLM-R_L (76.55) and ARBERT (76.07). We also note that in spite of its superiority on social data, MARBERT ranks top 4. This is due to MARBERT suffering on the QA tasks (due to its short input sequence length), and to a lesser extent on NER and TC.

6 Related Work

English and Multilingual LMs. Pre-trained LMs exploiting a self-supervised objective with masking such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b) have revolutionized NLP. Multilingual versions of these models such as mBERT and XLM-RoBERTa (Conneau et al., 2020) were also pre-trained. Other models with different objectives and/or architectures such as ALBERT (Lan et al., 2019), T5 (Raffel et al., 2020) and its multilingual version, mT5 (Xue et al., 2021), and GPT3 (Brown et al., 2020) were also introduced. More information about BERT-inspired LMs can be found in Rogers et al. (2020).

Non-English LMs. Several models dedicated to individual languages other than English have been developed. These include AraBERT (Antoun et al., 2020) and ArabicBERT (Safaya et al., 2020) for Arabic, Bertje for Dutch (de Vries et al., 2019), CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020) for French, PhoBERT for Vietnamese (Nguyen and Tuan Nguyen, 2020), and the models presented by Virtanen et al. (2019) for Finnish, Dadas et al. (2020) for Polish, and Malmsten et al. (2020) for Swedish. Pyysalo et al. (2020) also create monolingual LMs for 42 languages exploiting Wikipedia data. Our models contributed to this growing work of dedicated LMs, and has the advantage of covering a wide range of dialects. Our MARBERT and MARBERT-v2 models are also trained with a massive scale social media dataset, endowing them with a remarkable ability for real-world downstream tasks.

NLP Benchmarks. In recent years, several NLP benchmarks were designed for comparative evaluation of pre-trained LMs. For English, McCann et al. (2018) introduced NLP Decathlon (DecaNLP) which combines 10 common NLP datasets/tasks. Wang et al. (2018) proposed GLUE, a popular benchmark for evaluating nine NLP tasks. Wang et al. (2019) also presented SuperGLUE, a more challenging benchmark than GLUE covering seven tasks. In the cross-lingual setting, Hu et al. (2020)

Dataset	mBERT	XLM-R_B	XLM-R_L	AraBERT	ARBERT	MARBERT	MARBERT (v2)
$ARLUE_{Senti}$ *	79.02 / 79.50	92.17 / 93.00	**93.18 / 94.00**	78.26 / 78.50	87.96 / 88.50	**93.30 / 94.00**	92.82 / 93.50
$ARLUE_{SM}$ †	66.84 / 61.76	69.18 / 64.12	68.79 / 64.20	67.63 / 62.11	69.12 / 64.23	**71.64 / 68.38**	70.43 / 66.26
$ARLUE_{Topic}$	91.10 / 91.67	91.57 / 92.24	**92.66 / 93.53**	92.42 / 93.17	91.06 / 92.23	90.48 / 92.01	91.52 / 92.50
$ARLUE_{Dia\text{-}B}$	87.83 / 87.50	88.20 / 87.93	88.92 / 88.57	89.30 / 89.06	89.53 / 89.23	89.80 / 89.50	**90.05 / 89.72**
$ARLUE_{Dia\text{-}R}$	86.45 / 85.89	86.00 / 85.46	86.97 / 86.54	87.30 / 86.92	88.85 / 88.49	**90.94 / 90.65**	90.04 / 89.67
$ARLUE_{Dia\text{-}C}$	41.08 / 32.03	40.59 / 31.75	39.73 / 31.51	37.90 / 30.41	42.51 / 34.26	43.54 / 34.25	**45.37 / 35.94**
$ARLUE_{NER}$	96.81 / 76.91	97.74 / 84.09	**97.97 / 85.56**	97.79 / 83.67	97.46 / 81.21	96.89 / 76.58	97.18 / 79.34
$ARLUE_{QA}$ ‡	32.30 / 51.14	32.30 / 52.43	35.18 / **58.08**	31.72 / 51.87	34.04 / 54.34	27.27 / 43.67	**37.14** / 57.93
Average	72.68 / 70.80	74.72 / 73.88	75.43 / 75.79	75.75 / 71.96	75.07 / 74.06	75.48 / 73.63	**76.82 / 75.61**
$ARLUE_{Score}$	71.74	74.30	75.34	72.38	74.56	74.56	**76.21**

Table 11: Performance of our models on the **DEV** splits of ARLUE. * Metric for $ARLUE_{Senti}$ is F_1^{PN}. † $ARLUE_{SM}$ results is the average score across the social meaning tasks described in Table B.2. ‡ Metric for $ARLUE_{QA}$ is Exact Match (EM) / F_1.

Dataset	mBERT	XLM-R_B	XLM-R_L	AraBERT	ARBERT	MARBERT	MARBERT (v2)
$ARLUE_{Senti}$ *	79.02 / 79.50	92.17 / 93.00	93.18 / **94.00**	78.26 / 78.50	87.96 / 88.50	**93.30 / 94.00**	**93.30 / 94.00**
$ARLUE_{SM}$ †	77.76 / 69.88	79.81 / 71.19	80.01 / 73.00	78.84 / 72.03	80.39 / 74.22	**82.35 / 77.13**	76.34 / **77.13**
$ARLUE_{Topic}$	90.88 / 92.12	90.90 / 91.81	**92.24 / 93.40**	92.15 / 92.97	90.81 / 92.65	89.67 / 90.97	90.07 / 91.54
$ARLUE_{Dia\text{-}B}$	85.52 / 84.88	86.54 / 85.98	87.82 / 87.17	87.74 / 87.21	88.31 / 87.74	**88.72 / 88.19**	88.47 / 87.87
$ARLUE_{Dia\text{-}R}$	86.45 / 85.89	86.00 / 85.46	86.97 / 86.54	87.30 / 86.92	88.85 / 88.49	**90.94 / 90.65**	90.04 / 89.67
$ARLUE_{Dia\text{-}C}$	42.80 / 35.23	42.67 / 35.40	41.94 / 34.98	39.71 / 33.56	44.44 / 36.87	45.89 / 37.69	**47.49 / 38.53**
$ARLUE_{NER}$	95.90 / 69.06	96.02 / 73.27	96.13 / **74.94**	96.76 / 76.19	97.00 / 76.83	96.38 / 71.93	**96.75** / 74.70
$ARLUE_{QA}$ ‡	34.34 / 55.74	34.62 / 56.67	39.37 / **63.12**	36.31 / 58.10	36.29 / 57.81	29.13 / 48.83	**40.47** / 62.09
Average	74.08 / 71.54	76.09 / 74.10	77.21 / 75.89	74.63 / 73.19	76.76 / 75.39	77.05 / 74.92	**77.87 / 76.94**
$ARLUE_{Score}$	72.81	75.09	76.55	73.91	76.07	75.99	**77.40**

Table 12: Performance of our models on the **TEST** splits of ARLUE (Acc / F_1). * Metric for $ARLUE_{Senti}$ is Acc/ F_1^{PN}. † $ARLUE_{SM}$ results is the average score across the social meaning tasks described in Table 5. ‡ Metric for $ARLUE_{QA}$ is Exact Match (EM) / F_1.

provide a Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark for the evaluation of cross-lingual transfer learning covering nine tasks for 40 languages (12 language families). ***ARLUE complements these benchmarking efforts, and is focused on Arabic and its dialects. ARLUE is also diverse (involves*** 42 ***datasets) and challenging (our best ARLUE score is at*** 77.40***).***

7 Conclusion

We presented our efforts to develop two powerful Transformer-based language models for Arabic. Our models are trained on large-to-massive datasets that cover different domains and text genres, including social media. By pre-training MARBERT and MARBERT-v2 on dialectal Arabic, we aim at enabling downstream NLP technologies that serve wider and more diverse communities. Our best models perform better than (or on par with) XLM-R_{Large} ($\sim 3.4\times$ larger than our models), and hence are more energy efficient at inference time. Our models are also significantly better than AraBERT, the currently best-performing Arabic pre-trained LM. We also introduced AraLU, a large and diverse benchmark for Arabic NLU composed of 42 datasets thematically organized into six main task clusters. ARLUE fills a critical gap in Arabic and multilingual NLP, and promises to help propel innovation and facilitate meaningful comparisons in the field. Our models are publicly available. We also plan to publicly release our ARLUE benchmark. In the future, we plan to explore self-training our language models as a way to improve performance following Khalifa et al. (2021). We also plan to investigate developing more energy efficient models.

Acknowledgements

We gratefully acknowledges support from the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, Canadian Foundation for Innovation, Compute Canada and UBC ARC-Sockeye (https://doi.org/10.14288/SOCKEYE). We also thank the Google TFRC program for providing us with free TPU access.

Ethical Considerations

Although our language models are pre-trained using datasets that were public at the time of collection, parts of these datasets might become private or get removed (e.g., tweets that are deleted by users). For this reason, we will not release or redistribute any of the pre-training datasets. Data coverage is another important consideration: Our datasets have wide coverage, and one of our contributions is offering models that can serve more diverse communities in better ways than existing models. However, our models may still carry biases that we have not tested for and hence we recommend they be used with caution. Finally, our models deliver better performance than larger-sized models and as such are more energy conserving. However, smaller models that can achieve simply 'good enough' results should also be desirable. This is part of our own future research, and the community at large is invited to develop novel methods that are more environment friendly.

References

Mourad Abbas, Kamel Smaïli, and Daoud Berkani. 2011. Evaluation of topic identification methods on arabic corpora. *JDIM*, 9(5):185–192.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2020. Arabic Dialect Identification in the Wild. *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28(1):20–37.

Muhammad Abdul-Mageed and Mona T Diab. 2012. AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. In *LREC*, volume 515, pages 3907–3914. Citeseer.

Muhammad Abdul-Mageed, Shady Elbassuoni, Jad Doughman, AbdelRahim Elmadany, El Moatez Billah Nagoudi, Yorgo Zoughby, Ahmad Shaher, Iskander Gaba, Ahmed Helal, and Mohammed El-Razzaz. 2021a. DiaLex: A benchmark for evaluating multidialectal Arabic word embeddings. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 11–20, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020a. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim Elmadany, Houda Bouamor, and Nizar Habash. 2021b. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Azadeh Hashemi, and El Moatez Billah Nagoudi. 2020b. AraNet: A deep learning toolkit for Arabic social media. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 16–23, Marseille, France. European Language Resource Association.

Nawaf Abdulla, N Mahyoub, M Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sentiment analysis: Corpus-based and lexicon-based. In *Proceedings of The IEEE conference on Applied Electrical Engineering and Computing Technologies (AEECT)*.

Nora Al-Twairesh, Hend Al-Khalifa, AbdulMalik Al-Salman, and Yousef Al-Ohali. 2017. Arasenti-tweet: A corpus for Arabic sentiment analysis of saudi tweets. *Procedia Computer Science*, 117:63–72.

Hassan Alhuzali, Muhammad Abdul-Mageed, and Lyle Ungar. 2018. Enabling deep learning of emotion with first-person seed expressions. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 25–35.

Khaled Mohammad Alomari, Hatem M ElSherif, and Khaled Shaalan. 2017. Arabic tweets sentimental analysis using machine learning. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 602–610. Springer.

Ali Alshehri, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2020. Understanding and detecting dangerous speech in social media. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 40–47, Marseille, France. European Language Resource Association.

Mohamed Aly and Amir Atiya. 2013. LABR: A Large Scale Arabic book Reviews Dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 494–498.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th*

Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pages 9–15.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Ramy Baly, Alaa Khaddaj, Hazem Hajj, Wassim El-Hajj, and Khaled Bashir Shaban. 2019. ArSentD-LEV: A multi-topic corpus for target-based sentiment analysis in Arabic levantine tweets. *arXiv preprint arXiv:1906.01830*.

Yassine Benajiba and Paolo Rosso. 2007. ANERsys 2.0: Conquering the NER Task for the Arabic Language by Combining the Maximum Entropy with POS-tag Information. In *IICAI*, pages 1814–1823.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2017. ANT Corpus : An Arabic News Text Collection for Textual Classification. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 135–142. IEEE.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, et al. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020. Pre-training Polish Transformer-based Language Models at Scale. *Artificial Intelligence and Soft Computing*.

Kareem Darwish. 2013. Named Entity Recognition using Cross-lingual Resources: Arabic as an Example. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1558–1567.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Marc Djandji, Fady Baly, Wissam Antoun, and Hazem Hajj. 2020. Multi-Task Learning using AraBert for Offensive Language Detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 97–101, Marseille, France. European Language Resource Association.

Ibrahim Abu El-Khair. 2016. 1.5 billion words Arabic Corpus. *arXiv preprint arXiv:1611.04033*.

Abdellah El Mekki, Ahmed Alami, Hamza Alami, Ahmed Khoumsi, and Ismail Berrada. 2020. Weighted combination of BERT and N-GRAM features for Nuanced Arabic Dialect Identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, Barcelona, Spain.

Mohamed Elaraby and Muhammad Abdul-Mageed. 2018. Deep models for Arabic dialect identification on benchmarked data. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 263–274, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

AbdelRahim Elmadany, Hamdy Mubarak, and Walid Magdy. 2018. ArSAS: An Arabic Speech-Act and Sentiment Corpus of Tweets. *OSACT*, 3:20.

Ashraf Elnagar, Yasmin S Khalifa, and Anas Einea. 2018. Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications. In *Intelligent Natural Language Processing: Trends and Applications*, pages 35–52. Springer.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Véronique Moriceau, and Paolo Rosso. 2019. IDAT@FIRE2019: Overview of the Track on Irony Detection in Arabic Tweets. . In *Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. In: CEUR-WS.org, Kolkata, India, December 12-15*.

Sabit Hassan, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, Ammar Rashed, and Shammur Absar Chowdhury. 2020. ALT Submission for OSACT Shared Task on Offensive Language Detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 61–65, Marseille, France. European Language Resource Association.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Fatemah Husain. 2020. OSACT4 Shared Task on Offensive Language Detection: Intensive Preprocessing-Based Approach. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 53–60, Marseille, France. European Language Resource Association.

Muhammad Khalifa, Muhammad Abdul-Mageed, and Khaled Shaalan. 2021. Self-training pre-trained language models for zero- and few-shot multi-dialectal Arabic sequence labeling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 769–782, Online. Association for Computational Linguistics.

Muhammad Khalifa and Khaled Shaalan. 2019. Character convolutions for Arabic Named Entity Recognition with Long Short-Term Memory Networks. *Computer Speech & Language*, 58:335–346.

Svetlana Kiritchenko, Saif Mohammad, and Mohammad Salameh. 2016. SemEval-2016 Task 7: Determining Sentiment Intensity of English and Arabic Phrases. In *Proceedings of the 10th international workshop on semantic evaluation (SEMEVAL-2016)*, pages 42–51.

Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. An Empirical Study of Pre-trained Transformers for Arabic Information Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4727–4734.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. pages 7315–7330.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Zihan Liu, Yan Xu, Genta Indra Winata, and Pascale Fung. 2019b. Incorporating Word and Subword Units in Unsupervised Machine. Translation Using Language Model Rescoring.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden–Making a Swedish BERT. *arXiv preprint arXiv:2007.01658*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model . In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The Natural Language Decathlon: Multitask Learning as Question Answering. *arXiv preprint arXiv:1806.08730*.

Alexis Mitchell, Stephanie Strassel, Mark Przybocki, J Davis, George Doddington, Ralph Grishman, and B Sundheim. 2004. Tides extraction (ACE) 2003 multilingual training data. *Linguistic Data Consortium, Philadelphia Web Download*.

S. Bravo-Marquez Mohammad, M. F. Salameh, and S. Kiritchenko. 2018. Semeval-2018 Task 1: Affect in Tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*. Association for Computational Linguistics.

Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem Hajj. 2019. Neural Arabic Question Answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, Florence, Italy. Association for Computational Linguistics.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of OSACT4 Arabic Offensive Language Detection Shared Task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52, Marseille, France. European Language Resource Association.

Mahmoud Nabil, Mohamed Aly, and Amir F Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Tariq Alhindi.

2020. Machine generation and detection of Arabic manipulated and fake news. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 69–84, Barcelona, Spain (Online). Association for Computational Linguistics.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7022–7032.

Sampo Pyysalo, Jenna Kanerva, Antti Virtanen, and Filip Ginter. 2020. WikiBERT models: deep transfer learning for many languages. *arXiv preprint arXiv:2006.01538*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21:1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Motaz K Saad and Wesam M Ashour. 2010. OSAC: Open Source Arabic Corpora . In *6th ArchEng Int. Symposiums, EEECS*, volume 10.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after Translation: A Case-Study on Arabic Social Media Posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, Denver, Colorado. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean Voice Search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructure. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein T Al-Natsheh. 2020. Multi-Dialect Arabic BERT for Country-Level Dialect Identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 111–118, Barcelona, Spain (Online). Association for Computational Linguistics.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. *arXiv preprint arXiv:1912.09582*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya

Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wajdi Zaghouani and Anis Charfi. 2018. 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. European Language Resources Association (ELRA).

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic Dialect Identification . *Computational Linguistics*, 40(1):171–202.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. OSIAN: Open Source International Arabic News Corpus - Preparation and Integration into the CLARIN-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182, Florence, Italy. Association for Computational Linguistics.

Chiyu Zhang and Muhammad Abdul-Mageed. 2019a. Multi-task bidirectional transformer representations for irony detection. *CoRR*.

Chiyu Zhang and Muhammad Abdul-Mageed. 2019b. No Army, No Navy: BERT Semi-Supervised Learning of Arabic Dialects. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 279–284, Florence, Italy. Association for Computational Linguistics.

Appendices

A Sentiment Analysis

A.1 SA Datasets

- **AJGT.** The Arabic Jordanian General Tweets (AJGT) dataset (Alomari et al., 2017) covers MSA and Jordanian Arabic, with 900 *positive* and 900 *negative* posts.

- **AraNET$_{\text{Sent}}$**. Abdul-Mageed et al. (2020b) collect 15 datasets in both MSA and dialects from Abdul-Mageed and Diab (2012) (AWATIF), Abdul-Mageed et al. (2014) (SAMAR), Abdulla et al. (2013); Nabil et al. (2015); Kiritchenko et al. (2016); Aly and Atiya (2013); Salameh et al. (2015); Rosenthal et al. (2017); Alomari et al. (2017); Mohammad et al. (2018), and Baly et al. (2019). These datasets carry both **binary** (*negative* and *positive*) and **three-way** (*negative*, *neutral*, and *positive*) labels, but Abdul-Mageed et al. (2020b) map them into binary sentiment only.

- **AraSenTi-Tweet**. This comprises 17,573 gold labeled MSA and Saudi Arabic tweets by Al-Twairesh et al. (2017).

- **ArSarcasm$_{\text{Sent}}$** This sarcasm dataset is labeled with sentiment tags by Farha and Magdy (2020) who extract it from ASTD (Nabil et al., 2015) (10,547 tweets) and SemEval-2017 Task 4 (Rosenthal et al., 2017) (8,075 tweets).

- **ArSAS**. This Arabic Speech Act and Sentiment (ArSAS) corpus (Elmadany et al., 2018) consists of tweets annotated with sentiment tags.

- **ArSenD-Lev**. The Arabic Sentiment Twitter Dataset for LEVantine dialect (ArSenD-Lev) (Baly et al., 2019) has 4,000 tweets retrieved from the Levant region.

- **ASTD.** This is a collection of 10,006 Egyptian tweets by Nabil et al. (2015).

- **AWATIF**. This is an MSA dataset from newswire, Wikipedia, and web fora introduced by Abdul-Mageed and Diab (2012).

- **BBNS & SYTS**. The **BBN** blog posts **S**entiment (BBNS) and **S**yria **T**weets **S**entiment (SYTS) are introduced by Salameh et al. (2015).

- **CAMel$_{\text{Sent}}$**. Obeid et al. (2020) merge training and development data from ArSAS (Elmadany et al., 2018), ASTD (Nabil et al., 2015), SemEval (Rosenthal et al., 2017), and ArSenTD (Al-Twairesh et al., 2017) to create a new training dataset ($\sim$ 24K) and evaluate on the independent test sets from each of these sources.

- **HARD**. The Hotel Arabic Reviews Dataset (HARD) (Elnagar et al., 2018) consists of 93,700 MSA and dialect hotel reviews.

- **LABR**. The Large Arabic Book Review Corpus (Aly and Atiya, 2013) has 63,257 book reviews from Goodreads,[18] each rated with a 1-5 stars scale.

- **Twitter$_{\text{Abdullah}}$.**[19] This is a dataset of 2,000 MSA and Jordanian Arabic tweets manually labeled by Abdulla et al. (2013).

- **Twitter$_{\text{Saad}}$.** This dataset is collected using an emoji lexicon by Moatez Saad in 2019 and is available on Kaggle.[20]

- **SemEval-2017**. This is the SemEval-2017 sentiment analysis in Arabic Twitter task datasetby Rosenthal et al. (2017).

A.2 SA Baselines

For SA, we compare to the following STOA:

- **Antoun et al. (2020).** We compare to best results reported by the authors on five SA datasets: HARD, balanced ASTD (which we refer to as ASTD-B), ArSenTD-Lev, AJGT, and the unbalanced positive and negative classes for LABR. They split each dataset into 80/20 for Train/Test, respectively, and report in accuracy using the best epoch identified on test data. For a valid comparison, we follow their data splits and evaluation set up.

- **Obeid et al. (2020).** They fine-tune mBERT and AraBERT on the merged CAMel$_{\text{sent}}$

[18] www.goodreads.com.

[19] For ease of reference, we assign a name to this and other unnamed datasets.

[20] www.kaggle.com/mksaad/arabic-sentiment-twitter-corpus.

Dataset (classes)	Classes	TRAIN	DEV	TEST
AJGT (2)	{neg, pos}	1.4K	-	361
$\text{AraNET}_{\text{Sent}}$ (2)	{neg, pos}	100.5K	14.3K	11.8K
AraSenTi-Tweet (3)	{neg, neut, pos}	11.1K	1.4K	1.4K
$\text{ArSar}_{\text{Sent}}$ (3)	{neg, neut, pos}	8.4K	-	2.1K
ArSAS (3)	{neg, neut, pos}	24.8K	-	3.7K
ArSenD-LEV (5)	{neg, neut, pos, neg^+, pos^+}	3.2K	-	801
ASTD (3)	{neg, neut, pos}	24.8K	-	664
ASTD-B (2)	{neg, pos}	1.1K	-	267
AWATIF (4)	{neg, neut, obj, pos }	2.3K	288	284
BBN (3)	{neg, neut, pos}	960	125	116
HARD (2)	{neg, pos}	84.5K	-	21.1K
LABR (2)	{neg, pos}	13.2K	-	3.3K
SAMAR (5)	{mix, neg, neut, obj, pos}	2.5K	310	316
SemEval (3)	{neg, neut, pos}	24.8K	-	6.1K
SYTS (3)	{neg, neut, pos}	960	202	199
$\text{Tw}_{\text{Abdullah}}$ (2)	{neg, pos}	1.6K	202	190
Tw_{Saad} (2)	{neg, pos}	47K	5.8K	5.8K
$\text{ARLUE}_{\text{Senti}}$ (3)	{neg, pos, neut}	190.9K	6.5K	44.2K

Table A.1: Sentiment analysis datasets. **neg**$^+$: "very negative"; **pos**$^+$: "very positive". We construct $\text{ARLUE}_{\text{Senti}}$ by merging the different datasets and collapsing, or removing, the less frequent classes (details in text).

datasets and report in $F_1{}^{PN}$, which is the macro F_1 score over the positive and negative classes only (while neglecting the neutral class).

- **Abdul-Mageed et al. (2020b).** They fine-tune mBERT on the $\text{AraNET}_{\text{Sent}}$ data and report results in F_1 score on test data.

A.3 SA Evaluation on DEV

Table A.2 shows results of SA on DEV for datasets where there is a development split.

Dataset (classes)	mBERT	XLM-R_{B}	XLM-R_{L}	AraBERT	ARBERT	MARBERT
$\text{AraNET}_{\text{Sent}}$(2)	84.00	92.00	**93.00**	86.50	89.00	92.00
AraSenTi(3)	93.00	93.50	**95.00**	91.50	92.00	93.50
BBN(3)	68.00	75.00	77.00	70.00	**79.50**	78.50
SYTS(3)	62.00	**80.50**	66.00	65.00	69.00	72.50
$\text{Twitter}_{\text{Saad}}$(2)	80.00	95.50	95.50	81.50	90.00	**96.00**
SAMAR(5)	26.00	54.50	61.00	42.50	50.50	**62.50**
AWATIF(4)	63.50	62.00	67.50	65.00	70.50	**72.00**
$\text{Twitter}_{\text{Abdullah}}$(2)	87.50	91.00	95.50	92.50	**99.00**	97.00

Table A.2: SA results (F_1) on DEV.

B Social Meaning

B.1 SM Tasks & Datasets

- **Age and Gender**. For both age and gender, we use the *Arap-Tweet* dataset (Zaghouani and Charfi, 2018), which covers 17 different countries from 11 Arab regions. We follow the 80-10-10 data split of AraNet (Abdul-Mageed et al., 2020b).

- **Dangerous Speech.** We use the dangerous speech *AraDang* dataset from Alshehri et al. (2020), which is composed of tweets manually labeled with *dangerous* and *safe* tags.

Task	Dataset (classes)	Classes	TRAIN	DEV	TEST
Age	Arap-Tweet (3)	$\{\leq 24$ yrs, $25-34$ yrs, ≥ 35 yrs $\}$	1.3M	160.7K	160.7K
Dangerous	AraDang (2)	{dangerous, not-dangerous}	3.5K	616	664
Emotion	$\text{AraNET}_{\text{Emo}}$ (8)	{ang, anticip, disg, fear, joy, sad, surp, trust}	190K	911	942
Gender	Arap-Tweet (2)	{female, male}	1.3M	160.7K	160.7K
Hate Speech	HS@OSACT (2)	{hate, not-hate}	10K	1K	2K
Irony	FIRE2019 (2)	{irony, not-irony}	3.6K	-	404
Offensive	OFF@OSACT (2)	{offensive, not-offensive}	10K	1K	2K
Sarcasm	AraSarcasm (2)	{sarcasm, not-sarcasm}	8.4K	-	2.1K

Table B.1: Social Meaning datasets.

- **Offensive Language and Hate Speech**. We use manually labeled data from the shared task of offensive speech (Mubarak et al., 2020).[21] The shared task is divided into two sub-tasks: **sub-task A**: detecting if a tweet is *offensive* or *not-offensive*, and **sub-task B**: detecting if a tweet is *hate-speech* or *not-hate-speech*.

- **Emotion.** We use the *AraNet$_{emo}$* dataset from Abdul-Mageed et al. (2020b), which is created by merging two datasets from Alhuzali et al. (2018).

- **Irony**. We use the irony identification dataset for Arabic tweets released by IDAT@FIRE2019 shared task (Ghanem et al., 2019), following Abdul-Mageed et al. (2020b) data splits.

- **Sarcasm**. We use the *ArSarcasm* dataset developed by Farha and Magdy (2020).

More details about these datasets are in Table B.1.

B.2 SM Baselines

- **Age and Gender**. We compare to AraNET Abdul-Mageed et al. (2020b) age and gender models, trained by fine-tuning mBERT. The authors report 51.42 and 65.30 F_1 on age and gender, respectively.

- **Dangerous Speech**. We compare to Alshehri et al. (2020), who report a best of 59.60 F_1 on test with an mBERT model fined-tuned on emotion data.

- **Emotion**. We compare to Abdul-Mageed et al. (2020b), who acquire 60.32 F_1 on test with a fine-tuned mBERT.

- **Hate Speech**. The best results on the offensive and hate speech shared task (Mubarak et al., 2020) are at 95 F_1 score and are reported by Husain (2020), who employ heavy

[21] http://edinburghnlp.inf.ed.ac.uk/workshops/OSACT4.

Task (classes)	mBERT	XLM-R$_B$	XLM-R$_L$	AraBERT	ARBERT	MARBERT
Age (3)	56.33	59.70	53.63	57.67	58.60	**62.19**
Dangerous (2)	67.35	65.09	69.95	67.73	68.58	**75.50**
Emotion (8)	61.34	72.09	72.78	65.46	68.05	**75.18**
Gender (2)	68.06	71.10	71.23	67.61	69.97	**72.81**
Hate (2)	75.91	76.56	78.00	72.09	75.01	**82.91**
Irony (2)	81.08	83.12	81.29	79.12	84.83	**86.77**
Offensive (2)	84.04	85.26	86.72	87.21	88.77	**91.68**

Table B.2: SM results in F_1 on DEV.

feature engineering with SVMs. Since our focus is on methods exploiting language models, we compare to Djandji et al. (2020) who rank second in the shared task with a fine-tuned AraBERT (83.41 F_1 on test).

- **Irony**. We compare to Zhang and Abdul-Mageed (2019a) who fine-tune mBERT on the irony task, with an auxiliary author profiling task, and report 82.4 F_1 on test.
- **Offensive Language**. We compare to the best results on the offensive sub-task (Mubarak et al., 2020) reported by Hassan et al. (2020). They propose an ensemble of SVMs, CNN-BiLSTM, and mBERT with majority voting and acquire 90.51 F_1.
- **Sarcasm**. We compare to Farha and Magdy (2020) who train a BiLSTM model using the AraSarcasm dataset, reporting 46.00 F_1 score.

B.3 SM Evaluation on DEV

Table B.2 shows results of the social meaning tasks on development splits.

C Topic Classification

C.1 TC Datasets

- **Arabic News Text.** Chouigui et al. (2017) build the Arabic news text (ANT) dataset from transcribed Tunisian radio broadcasts.
- **Khaleej**. Abbas et al. (2011) created the Khaleej from Gulf Arabic websites.
- **OSAC.** Saad and Ashour (2010) collect OSAC from news articles.

Dataset (classes)	Classes	TRAIN	DEV	TEST
ANT (5)	{C, E, I, ME, S, T}	25.2K	3.2K	3.2K
Khallej (4)	{E, I, LOC, S}	4.6K	570	570
OSAC (10)	{E, F, H, HIST, L, R, RLG, SPS, S, STR}	18K	2.2K	2.2K
ARLUE$_{\text{Topic}}$ (16)	{*all classes*}	47.7K	5.9K	5.9K

Table C.1: TC datasets. **C**: culture, **E**: economy, **F**: family, **H**: health, **HIST**: history, **I**: international news, **L**: law, **LOC**, local news, **ME**: middle east, **R**: recipes, **RLG**: religion, **SPS**: space, **S**: sports, **STR**: stories, **T**: technology.

Dataset (classes)	mBERT	XLM-R$_B$	XLM-R$_L$	AraBERT	ARBERT	MARBERT
ANTText (5)	85.04	86.74	87.41	**87.98**	87.06	85.80
ANTTitle (5)	79.46	80.77	82.04	**83.56**	81.10	82.36
ANTText+Title (5)	87.24	86.36	88.45	**88.76**	87.27	85.99
Khallej (4)	94.48	95.32	96.09	95.65	96.16	**96.31**
OSAC (10)	97.87	97.75	97.61	**97.94**	97.56	97.66

Table C.2: TC results tasks (F_1) on DEV.

C.2 TC Evaluation on DEV

Results of TC tasks on DEV data are in Table C.2.

D Dialect Identification

D.1 DIA Datasets

We introduce each dataset briefly here and provide a description summary of all datasets in Table D.1.

- **Arabic Online Commentary (AOC).** This is a repository of 3M Arabic comments on online news (Zaidan and Callison-Burch, 2014). It is labeled with MSA and three **regional** dialects (*Egyptian*, *Gulf*, and *Levantine*).
- **ArSarcasm$_{\text{Dia}}$**. This dataset is developed by Farha and Magdy (2020) for sarcasm detection but also carries **regional** dialect labels from the set {*Egyptian, Gulf, Levantine, Maghrebi*}.
- **MADAR**. Sub-task 2 of the MADAR shared task (Bouamor et al., 2019)[22] is focused on user-level dialect identification with manually-curated **country** labels (n=21).
- **NADI-2020.** The first Nuanced Arabic Dialect Identification shared task (NADI 2020) (Abdul-Mageed et al., 2020a)[23] targets **country** level (n=21) as well as **province** level (n=100) dialects.
- **QADI**. The QCRI Arabic Dialect Identification (QADI) dataset (Abdelali et al., 2020) is labeled at the **country** level (n=18).

Details of the datasets are in Table D.1.

D.2 DIA Baselines

- **Elaraby and Abdul-Mageed (2018)** report three levels of classification on AOC data: **(1) MSA vs. DA** (87.23 accuracy), **(2) regional** (i.e., *Egyptian*, *Gulf*, and *Levantine*) (87.81 accuracy), and **(3) MSA, Egyptian, Gulf, and**

[22] https://camel.abudhabi.nyu.edu/madar-shared-task-2019/.

[23] https://github.com/UBC-NLP/nadi.

Task (classes)	Dataset	Classes	TRAIN	DEV	TEST
AOC (2)	Binary	{DA, MSA}	86.5K	10.8K	10.8K
AOC (3)	Region	{Egypt, Gulf, Levnt}	35.7K	4.5K	4.5K
AOC (4)	Region	{Egypt, Gulf, Levnt, MSA}	86.5K	10.8K	10.8K
$\text{ArSarcasm}_{\text{Dia}}$ (5)	Regoin	{Egypt, Gulf, Levnt, Magreb, MSA}	8.4K	-	2.1K
MADAR-TL (21)	Country	{*Multiple countries**}	193.1K	26.6K	44K
NADI (21)	Country	{*Multiple countries**}	2.1K	5K	5K
QADI (18)	Country	{*Multiple countries*†}	497.8K	-	3.5K
$\text{ARLUE}_{\text{Dia-B}}$ (2)	Binary	{DA, MSA}	94.9K	10.8K	12.9K
$\text{ARLUE}_{\text{Dia-R}}$ (4)	Region	{Egypt, Gulf, Levnt, Magreb}	38.5K	4.5K	5.3K
$\text{ARLUE}_{\text{Dia-C}}$ (21)	Country	{*Multiple countries**}	711.9K	31.5K	52.1K

Table D.1: Dialect datasets. * All Arab countries except Comoros. † All Arab countries except Comoros, Djibouti, Mauritania, and Somalia.

Dataset (classes)	Task	mBERT	XLM-R_{B}	XLM-R_{L}	AraBERT	ARBERT	MARBERT
MADAR(21)	Country	33.75	34.54	33.28	33.47	39.24	**40.61**
AOC(4)	Regoin	80.07	78.97	79.55	80.85	81.96	**83.56**
AOC(3)	Regoin	87.07	86.80	88.21	88.46	89.57	**91.56**
AOC(2)	Binary	87.89	87.63	88.38	88.76	89.32	**89.66**
NADI(21)	Country	14.49	17.30	18.62	16.18	23.73	**26.40**
NADI(100)	Province	02.32	03.91	4.00	03.04	**06.05**	05.23

Table D.2: DIA results on DEV in F_1.

Levantine (accuracy of 82.45). Their best results are based on BiLSTM.

- **Abdelali et al. (2020)** fine-tune AraBERT on the QADI dataset. They report 60.6 F_1.
- **Zhang and Abdul-Mageed (2019b)** developed the top ranked system in MADAR subtask 2, with 48.76 accuracy and 34.87 F_1 at tweet level.
- **Talafha et al. (2020)** developed NADI subtask 1 (**country level**) winning system, an ensemble of fine-tuned AraBERT (26.78 F_1).
- **El Mekki et al. (2020)** developed NADI subtask 2 (**province level**) winning system using a combination of word and character n-grams to fine-tune AraBERT (6.08 F_1).
- **AraBERT.** For $\text{ArSarcasm}_{\text{Dia}}$, where no dialect id system was previously developed, we consider a fine-tuned AraBERT a baseline.

D.3 DIA Evaluation on DEV

Table D.2 shows results of the dialect identification tasks on development splits.

E Named Entity Recognition

E.1 NER datasets

Table E.1 and Table E.2 show the data splits across our NER datasets, and the results of all our models on the development splits.

Dataset	Tokens	Train	DEV	Test
ANERcorp	150.2K	95.5K	24.8K	29.9K
ACE03BN	15.6K	11.6K	2K	2K
ACE03NW	27K	21.3K	2.7K	3K
ACE04BN	70.5K	56.5K	7K	7K
TW-NER	74.8K	42.9K	7.4K	24.5K
$\text{ARLUE}_{\text{NER}}$	338.3K	227.7K	44.1K	66.5K

Table E.1: Distribution of the Arabic NER datasets.

Dataset (classes)	mBERT	XLM-R_{B}	XLM-R_{L}	AraBERT	ARBERT	MARBERT
ANERcorp	86.20	87.24	89.64	**90.24**	83.24	80.86
ACE03NW	80.57	88.21	**90.49**	89.76	88.17	85.02
ACE03BN	80.35	80.36	83.39	81.05	**90.91**	79.05
ACE04NW	87.21	90.08	**91.94**	89.70	89.33	86.80
TW-NER	52.60	73.61	**77.70**	**73.61**	70.78	67.39

Table E.2: NER results (F_1) on DEV.

E.2 NER Baselines

Khalifa and Shaalan (2019) apply CNNs and BiLSTMs and report F_1 scores on test sets, as follows: 88.77 (ANERcorp), 91.47 (ACE03NW), 94.92 (ACE03BN), 91.20 (ACE04NW), and 65.34 (Twitter). We use their exact data splits.

F Question Answering Datasets

- **ARCD.** Mozannar et al. (2019) use crowdsourcing to develop the Arabic Reading Comprehension Dataset. We use the same ARCD data splits used by Antoun et al. (2020).
- **MLQA.** This MultiLingual Question Answering benchmark is proposed by Lewis et al. (2020). It consists of over 5K extractive question-answer instances in SQuAD format in seven languages, including Arabic.
- **XQuAD.** This Cross-lingual Question Answering Dataset Artetxe et al. (2020) consists of 1,190 question-answer pairs and 240 paragraphs from SQuAD v1.1 (Rajpurkar et al., 2016) translated into ten languages (including Arabic) by professional translators.
- **TyDi QA.** The TyDi QA dataset Artetxe et al. (2020) is manually curated and covers 11 languages (including Arabic). We focus on the "Gold" passage task only.

Dataset	TRAIN	DEV	TEST
AR-XTREME	86.7K (MT)	-	-
ARCD	-	-	1.4K (H)
AR-MLQA	-	517 (HT)	5.3K (HT)
AR-XQuAD	-	-	1.2K (HT)
AR-TyDi-QA	14.8K (H)	-	921 (H)
ARLUE_{QA}	101.6K	517	11.6K

Table F.1: Multilingual & Arabic QA datasets. **H**: Human Created. **HT**: Human Translated. **MT**: Machine Translated.