ORCA: A Challenging Benchmark for Arabic Language Understanding

AbdelRahim Elmadany[1,⋆] **El Moatez Billah Nagoudi**[1,⋆] **Muhammad Abdul-Mageed**[1,2,⋆]
[1] Deep Learning & Natural Language Processing Group, The University of British Columbia
[2]Department of Natural Language Processing & Department of Machine Learning, MBZUAI
{a.elmadany,moatez.nagoudi,muhammad.mageed}@ubc.ca

Abstract

Due to the crucial role pretrained language models play in modern NLP, several benchmarks have been proposed to evaluate their performance. In spite of these efforts, no public benchmark of diverse nature currently exists for evaluating Arabic NLU. This makes it challenging to measure progress for both Arabic and multilingual language models. This challenge is compounded by the fact that any benchmark targeting Arabic needs to take into account the fact that Arabic is not a single language but rather a collection of languages and language varieties. In this work, we introduce a publicly available benchmark for Arabic language understanding evaluation dubbed ORCA. It is carefully constructed to cover diverse Arabic varieties and a wide range of challenging Arabic understanding tasks exploiting 60 different datasets (across seven NLU task clusters). To measure current progress in Arabic NLU, we use ORCA to offer a comprehensive comparison between 18 multilingual and Arabic language models. We also provide a public leaderboard with a unified single-number evaluation metric (*ORCA score*) to facilitate future research.[1]

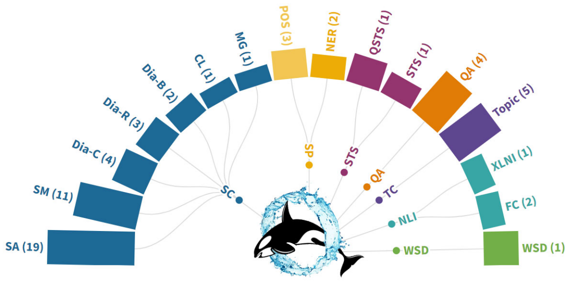


Figure 1: ORCA task clusters and datasets taxonomy. The task clusters are **SC:** Sentence Classification. **SP:** Structured Predictions. **TC:** Topic Classification. **STS:** Semantic Textual Similarity. **NLI:** Natural Language Inference. **QA:** Question-Answering. **WSD:** Word sense disambiguation. The value in parentheses is the number of datasets in each task cluster.

1 Introduction

Pretrained language models (PLMs) (Devlin et al., 2019; Liu et al., 2019b; Lan et al., 2019; Zhang et al., 2019; Sanh et al., 2019; Radford et al., 2019; Dai et al., 2019; Clark et al., 2020; Lewis et al., 2020a; Zaheer et al., 2020; Brown et al., 2020; Beltagy et al., 2020; Zhang et al., 2020a; Conneau et al., 2020a; Kitaev et al., 2020; Zhang et al., 2020b; He et al., 2021; Le et al., 2021; Raffel et al., 2022; Chung et al., 2022; Chowdhery et al., 2022) have become a core component of the natural language understanding (NLU) pipeline, making it all the more important to evaluate their performance under standardized conditions. For this reason, several English-based benchmarks such as GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), SyntaxGym (Gauthier et al., 2020), Evaluation Harness (Gao et al., 2021), GEM (Gehrmann et al., 2021), NL-Augmenter (Dhole et al., 2021), Dynabench (Kiela et al., 2021), MMLU (Hendrycks et al., 2021), NATURAL INSTRUCTIONS (Mishra et al., 2022), BIG-bench (Srivastava et al., 2022), and multilingual benchmarks such as XTREME (Hu et al., 2020), XGLUE (Liang et al., 2020), and MASAKHANE (Nekoto et al., 2020) have been introduced. Benchmarks for a few other languages also followed, including FLUE (Le et al., 2020) for French, CLUE (Xu et al., 2020) for Chinese, IndoNLU (Wilie et al., 2020) for Indonesian, KorNLI and KorSTS (Ham et al., 2020) for Korean, and JGLUE(Kurihara et al., 2022). This leaves behind the majority of the world's languages, and relies on multilingual benchmarks which often have limited coverage of dialects and naturally-occurring (rather than machine translated) text. This motivates us to introduce a benchmark for Arabic. One

[1]https://orca.dlnlp.ai/.
⋆All authors contributed equally.

Findings of the Association for Computational Linguistics: ACL 2023, pages 9559–9586
July 9-14, 2023

other reason that lends importance to our work is that Arabic is a rich collection of languages with both standard and dialectal varieties and more than 400M native speaker population.

To the best of our knowledge, there have only been two attempts to provide Arabic NLU evaluation benchmarks. These are ARLUE (Abdul-Mageed et al., 2021) and ALUE (Seelawi et al., 2021). Although useful, both of these have major limitations: ALUE has *modest coverage* (only eight datasets covering only three task clusters) and ARLUE involves datasets that are not publicly available. Our goal is to rectify these limitations by introducing ORCA, which expands task coverage using *fully* public datasets, while also offering an accessible benchmark with a public leaderboard and processing tools as well as wide geographical and linguistic coverage. ORCA exploits 60 different datasets, making it by far the most extensive benchmark for Arabic NLU and among the most extensive for any language. We present detailed analyses of the data comprising ORCA and evaluate a wide range of available pretrained language models (PLMs) on it, thus offering strong baselines for future comparisons.

In summary, we offer the following contributions: **(1)** We introduce ORCA, an extensive and diverse benchmark for Arabic NLU. ORCA is a collection of 60 datasets arranged into *seven task clusters*, namely: sentence classification, text classification, structured prediction, semantic similarity, natural language inference (NLI), question-answering (QA), and word sense disambiguation (WSD). **(2)** We provide a comprehensive comparison of the performance of publicly available Arabic PLMs on ORCA using a unified *ORCA score*. **(3)** To facilitate future work, we design a public leaderboard for scoring PLMs on ORCA. Our leaderboard is *interactive* and offers *rich meta-data* about the various datasets involved as well as the language models we evaluate. **(4)** We distribute ORCA with a new *modular toolkit* for pretraining and transfer learning for NLU. The toolkit is built around standard tools including PyTorch (Paszke et al., 2019) and HuggingFace datasets hub (Lhoest et al., 2021).

The rest of the paper is organized as follows: In Section 2, we provide an overview of related work. Section 3 introduces ORCA, our Arabic NLU benchmark. In Section 4, we describe multilingual and Arabic pretrained language models we evaluate on ORCA, providing results of our evaluation in Section 5. Section 6 is an analysis of model computational cost as measured on ORCA. We conclude in Section 7.

2 Related Work

Most recent benchmarks propose a representative set of standard NLU tasks for evaluation. These can be categorized into English-centric, multilingual, Arabic-specific, and X-specific (X being a language other than English or Arabic such as Chinese, Korean, or French). We briefly describe each of these categories next. We also provide a comparison of benchmarks in the literature in terms of task clusters covered and number of datasets in Table 1.

2.1 English-Centric Benchmarks

GLUE. The General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018) is one of the early English benchmarks. It is a collection of nine publicly available datasets from different genres. GLUE is arranged into three task clusters: sentence classification, similarity and paraphrase, and NLI.

SuperGLUE. Wang et al. (2019) propose SuperGLUE, a benchmark styled after GLUE with a new set of more challenging tasks. SuperGLUE is built around eight tasks and arranged into four task clusters: QA, NLI, WSD, and coreference resolution. The benchmark is accompanied by a leaderboard with a single-number performance metric (i.e., the *SuperGLUE score*).

2.2 Multilingual Benchmarks

bAbI. Early attempts to create multilingual benchmarks are limited in their language coverage. An example is bAbI (Weston et al., 2015), which covers only English and Hindi. It consists of a set of 20 tasks for testing text reasoning and understanding using different question-answering and coreference resolution strategies.

XGLUE. XGLUE is a cross-lingual benchmark proposed by Liang et al. (2020) to evaluate the performance of PLMs. It provides 11 tasks in both NLU and NLG scenarios that cover 19 languages. The XGLUE tasks are arranged into four *understanding tasks* (structured predictions, text classifications, QA, NLI, semantic search), and two *generation tasks* (question and title generation).

XTREME. The Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) (Hu et al., 2020) is a benchmark for evaluating the

Task Cluster	Bench.	English-Centric GLUE	SGLUE	X-Specific FLUE	IndoNLU	CLUE	JGLUE	KorNLU	Multilingual bAbI	XGLUE	XTREM	Arabic ALUE	ARLUE	ORCA
	Lang.	En	En	Fr	Id	Zh	Jp	Ko	En, Hi	19	40	Ar	Ar	Ar
Sentence Classification		✓		✓	✓	✓	✓				✓	✓	✓	✓
Structured Prediction			✓	✓	✓					✓	✓		✓	✓
STS and Paraphrase		✓		✓	✓	✓		✓		✓		✓		✓
Text Classification		✓				✓	✓						✓	✓
Natural Language Inference			✓	✓	✓	✓		✓		✓		✓		✓
Word Sense Disambiguation			✓	✓										✓
Coreference Resolution		✓	✓			✓			✓					
Question-Answering		✓	✓		✓	✓	✓		✓		✓		✓	✓
# Datasets		11	10	7	12	9	6	4	20	11	9	9	42	**60**
# Task Clusters Covered		5	5	5	5	6	3	2	2	3	4	3	4	**7**

Table 1: Comparison of NLU benchmarks proposed in the literature across the different covered task clusters. **STS**: Semantic Textual Similarity. **GLUE**: (Wang et al., 2018). **SGLUE**: SuperGLUE (Wang et al., 2019). **XGLUE**: (Liang et al., 2020). **FULE**: (Le et al., 2020). **FULE**: (Le et al., 2020). **IndoNLU**: (Wilie et al., 2020). **CLUE**: (Xu et al., 2020). **KorNLI**: KorNLI and korSTS (Ham et al., 2020). **bAbI**: (Weston et al., 2015). **XTREM**: (Hu et al., 2020). **ALUE**: (Seelawi et al., 2021). **ARLUE**: (Abdul-Mageed et al., 2021). **ORCA:** Our proposed Arabic NLU benchmark.

cross-lingual generalization capabilities of multilingual models. It covers 40 languages and includes nine datasets across four task clusters: classification (i.e., NLI and paraphrase), structured prediction (i.e., POS and NER), question answering, and sentence retrieval. Ruder et al. (2021) extend XTREME to XTREME-R (for XTREME Revisited). This new benchmark has an improved set of ten NLU tasks (including language-agnostic retrieval tasks). XTREME-R covers 50 languages. Authors also provide a multilingual diagnostic suite and evaluation capabilities through an interactive public leaderboard.

Big-bench. The Beyond the Imitation Game Benchmark or shortly BIG-bench (Srivastava et al., 2022) is a collaborative[2] NLP benchmark aimed to explore and evaluate the capabilities of large language models. It currently consists of 204 advanced NLP tasks, from diverse topics such as common-sense reasoning, linguistics, childhood development, math, biology, physics, social bias, and software development.[3]

2.3 Arabic-Specific Benchmarks

ALUE. To the best of our knowledge, two benchmarks for Arabic currently exist, ALUE (Seelawi et al., 2021) and ARLUE (Abdul-Mageed et al., 2021). ALUE is focused on NLU and comes with eight datasets arranged into three task clusters: sentence classification, NLI, and similarity and paraphrase. The sentence classification cluster involves five datasets for offensive and hate speech detection, irony prediction, sentiment analysis, and dialect identification. The NLI cluster involves two datasets, both for NLI aiming at predicting whether a premise sentence contradicts, entails, or is neutral toward a hypothesis sentence. ALUE has one dataset for semantic similarity comprising a collection of questions pair labelled with "1" (semantically similar) or "0" otherwise. The task is to predict these similarity scores. While datasets in ALUE are publicly available and the benchmark is accompanied by a leaderboard, its size and diversity (geographical and linguistic) are modest.

ARLUE. (Abdul-Mageed et al., 2021) also targets Arabic NLU tasks and is composed of 42 datasets arranged into four task clusters: sentence classification, text classification, structured prediction, and QA. Many of the datasets in ARLUE, however, are not publicly available which presents a barrier to widespread adoption. Nor is ARLUE accompanied by a leaderboard. *ORCA ameliorates these challenges.*

2.4 X-Specific Benchmarks

Other X-specific benchmarks include **CLUE.** (Xu et al., 2020), **FLUE.** (Le et al., 2020), **IndoNLU.** (Wilie et al., 2020), **JGLUE.** (Kurihara et al., 2022), and **KorNLI and KorSTS.** (Ham et al., 2020). We review these benchmarks in Ap-

[2]Contributed by 444 authors across 132 institutions.

[3]We exclude the Big-Bench benchmark from Table 1 because it has a very large number of tasks that we cannot fit into the table. It also involves task clusters that are unrelated to the ORCA benchmark.

pendix B.

3 ORCA Benchmark

We present ORCA, a benchmark for Arabic NLU that is challenging and diverse. ORCA involves 60 datasets arranged into 29 tasks and seven task clusters. In the following, we will first introduce our design principles for developing ORCA then introduce the different task clusters covered.

Cluster	Task	Level	#Data	Train	Dev	Test
SC	SA	Sent	19	50K	5K	5K
	SM	Sent	11	50K	5K	5K
	Dia-b	Sent	2	50K	5K	5K
	Dia-r	Sent	3	38.5K	4.5K	5K
	Dia-c	Sent	4	50K	5K	5K
	CL	Sent	1	3.2K	0.9K	0.4K
	MG	Sent	1	50K	5K	5K
SP	NER	Word	2	5.2K	1.1K	1.2K
	POS	Word	2	5.2K	1.1K	1.2K
TC	Topic	Doc	5	47.5K	5K	5K
QA	QA	Parag	4	101.6K	517	7.4K
STS	STS-reg	Sent	1	0.8K	0.2K	0.2K
	STS-cls	Sent	1	9.6K	1.2K	1.2K
NLI	XNLI	Sent	1	4.5K	0.5K	2.5K
	FC	Doc	2	5K	1K	1K
WSD	WSD	Word	1	21K	5K	5K
Total			**60**	487.1K	46.0K	55.1K

Table 2: The different task clusters, tasks, and data splits in ORCA. **SC:** Sentence Classification. **SP:** Structured Prediction. **TC:** Topic Classification. **STS:** Textual Semantic Similarity. **NLI:** Natural Language Inference. **QA:** Question-Answering. **SM**: Social Meaning. For abbreviations of task names, refer to Section 3.2.

3.1 Design Principles

Our goal is to offer a *challenging* and *diverse* NLU benchmark that allows evaluation of language models and measurement of progress on Arabic NLU. To this end, we develop ORCA with a number of design principles in mind. We explain these here.

Large number of public tasks. We strive to include as many tasks as possible so long as data for these tasks are public. This makes it possible for researchers to train and evaluate on these tasks without having to pay for private data. ORCA involves 60 different datasets that are *all* publicly available.

Challenging benchmark. We design ORCA to require knowledge at various linguistic levels, making it challenging. This includes knowledge at the level of tokens in context as well as at the levels of complete sentences, inter-sentence relations, whole paragraphs, and entire documents.

Coherent task clusters and tasks. Rather than listing each group of datasets representing a given task together, we group the various tasks into *task clusters*. This makes it easy for us to present the various downstream tasks. It also makes it possible to derive meaningful insights during evaluation. For example, one can compare performance at a lower-level task cluster such as structured prediction to that of performance at a higher-level cluster such as natural language inference. Within the clusters themselves, we also maintain coherent subgroupings. For example, since sentiment analysis has been one of the most popular tasks in Arabic NLP, we assign it its own sub-cluster within the sentence classification cluster. Similarly, we keep tasks such as hate speech and emotion detection that exploit social media data into a single *social meaning* cluster.

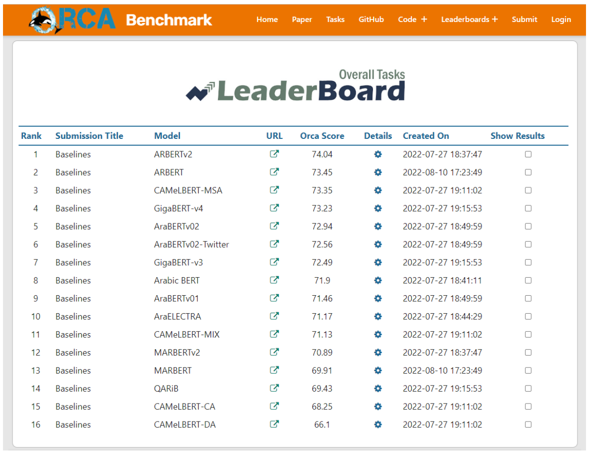
Wide linguistic variability and geographical coverage. We strive to include tasks in various Arabic varieties. This involves Modern Standard Arabic (MSA) and dialectal Arabic (DA). We include datasets collected from wide regions within the Arab world. This not only pertains our DA datasets, many of which come from various Arab countries, but also our MSA datasets as these are extracted from several news outlets from across the Arab world. This also ensures variability in topics within these datasets. To illustrate linguistic diversity within ORCA, we run an in-house binary MSA-dialect classifier on all ORCA data splits (i.e., Train, Dev, and Test).[4] For a deeper understanding of ORCA data, we also calculate several measures including the average, median, mode, and type-token ratio (TTR) of the sentences in each task. Table 3 shows the MSA vs. DA data distribution and the statistical description of ORCA datasets.

In addition, we acquire a country-level dialect distribution analysis over the data using AraT5 model (Nagoudi et al., 2022) fine-tuned on the ORCA dialect country-level dataset (DIA-C). We run this country-level classifier only on the dialectal portion of ORCA (i.e., datasets of tweets predicted as *dialectal* with our in-house MSA-dialect classifier). Figure F.1 shows that ORCA datasets are *truly diverse* from a geographical perspective.[5]

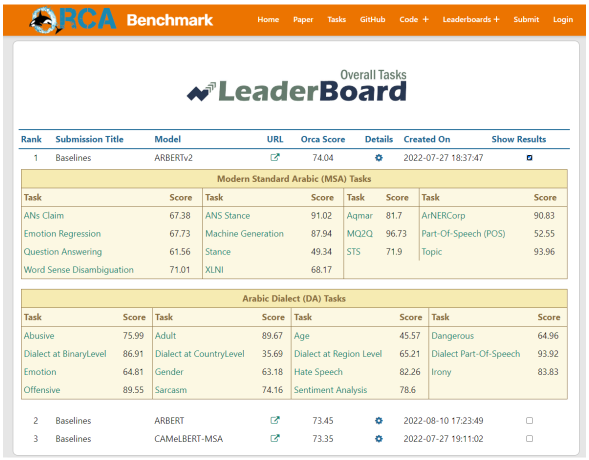
Accessible evaluation. To facilitate evaluation in

[4] As our classifier is trained using ORCA_{DA}, we exclude the ORCA dialect component from this analysis.

[5] Again, the country-level classifier is also trained using ORCA_{DIA}, so we exclude the dialect tasks from this analysis.



(a) Models ranked by our ORCA score.



(b) Detailed scores for a given model across all tasks.

Figure 2: ORCA main leaderboard.

a reasonable time frame in a GPU-friendly setting, we cap data sizes across our Train, Dev, and Test splits to ***50k, 5k, 5k*** samples respectively. This allows us to avoid larger data sizes in tasks such as *Age* and *Gender* that have *1.3m, 160k, 160k* samples for the Train, Dev, and Test splits each and both the *sentiment* and *dialect country-level* tasks that have *190k, 6.5k, 44.2k* and *711.9k, 31.5k, 52.1k* for the Train, Dev, and Test data (respectively). Table 2, shows a summary of the data splits across tasks and task clusters in ORCA.

Simple evaluation metric. We adopt a simple evaluation approach in the form of an *ORCA score*. The ORCA score is simply a macro-average of the different scores across all tasks and task clusters, where each task is weighted equally.

Modularity. We design ORCA to allow users to score models on the whole benchmark but also on individual task clusters. In both cases, the leaderboard returns results averaged across the datasets within either the whole benchmark or the individual tasks (sub-leaderboards). This allows us to invite submissions of dedicated models that take as its target subsets of ORCA datasets. Figure 2 shows ORCA's main screen with models sorted by ORCA score. We provide more screenshots illustrating ORCA's modularity in Appendix D.

Public leaderboard. We allow scoring models against ORCA through an intuitive, easy-to-use leaderboard. To facilitate adoption, we also provide a Google Colab notebook with instructions for finetunining any model on ORCA tasks.

Faithful evaluation. For each submission, we require users to provide meta-data such as the number of parameters, amount of pretraining data, and number of finetuning epochs. This facilitates comparisons across the different models. We make this meta-data available via our interactive leaderboard.

Proper credit for individual dataset authors. One issue with evaluation benchmarks is that once a benchmark is created there is a concern of not giving credit to original datasets. To overcome this limitation, we distribute a simple text file with bibliographic entries for all papers describing the 60 datasets in ORCA and strongly encourage all future use to cite them.

3.2 Tasks and Task Clusters

As explained, we arrange ORCA into 7 task clusters. These are (1) sentence classification, (2) structured prediction (3) semantic textual similarity and paraphrase, (4) text classification, (5) natural language inference, (6) word sense disambiguation, and (7) question answering.

Sentence Classification. This cluster involves the following sentence-level classification tasks: *(1) Sentiment Analysis:* 19 publicly available sentiment datasets have been used to construct this task. We merge 17 datasets benchmarked by Abdul-Mageed et al. (2021) with two new datasets: Arabizi sentiment analysis dataset (Fourati et al., 2020) and AraCust (Almuqren and Cristea, 2021), a Saudi Telecom tweets corpus for sentiment analysis. *(2) Social Meaning:* Refers to eight social meaning datasets covering prediction of hate speech and offensive language (Mubarak et al., 2020), dangerous speech (Alshehri et al., 2020), sarcasm (Farha and Magdy, 2020), adult content Mubarak et al. (2021), irony (Ghanem et al., 2019), emotion, age and gender (Mohammad et al.,

Task cluster with more likely dialectal data								
Cluster	**Task**	**Avg-char**	**Avg-word**	**Median**	**Mode**	**TTR**	**MSA**	**DIA**
	abusive	43.71	12.45	11	8	27.69	29.55	70.45
	adult	86.15	15.44	14	3	18.65	65.8	34.2
	age*	60.73	11.82	15	19	42.44	41.52	58.48
	claim	48.23	8.16	8	7	37.96	99.78	0.22
	dangerous	38.27	8.17	8	7	35.69	10.16	89.84
	dialect-B	89.92	17.17	10	4	37.84	60.27	39.73
	dialect-C	82.66	15.39	17	22	37.62	27.44	72.56
	dialect-R	80.40	15.66	8	8	36.32	12.57	87.43
SC	emotion-cls	73.58	14.60	16	10	14.25	25.72	74.28
	emotion-reg	63.57	12.60	14	9	14.25	60.3	39.70
	gender*	60.73	11.82	9	4	42.44	40.9	59.10
	hate†	99.20	19.76	16	9	24.97	25.79	74.21
	offensive†	99.20	19.76	16	9	24.97	25.79	74.21
	irony	106.67	19.70	18	17	31.15	45.32	54.68
	machine G	218.11	39.92	33	31	14.44	99.44	0.56
	sarcasm	88.80	15.69	16	18	28.29	71.49	28.51
	sentiment	127.27	22.9	16	10	79.31	64.06	35.94
Avg		86.31	16.53	14.41	11.47	32.25	47.41	52.59
Task clusters with more likely MSA data								
TC	topic	$2.7k$	474.78	286	152	5.2	99.71	0.29
QA	arlue-qa	101.6	517	7.4	4	50	100	0.0
One input avg		1.4k	495.89	146.7	78	27.6	99.86	0.15
	ans-st	50.53/45.35	8.48/7.70	44780	44749	36.46/38.70	99.85	0.15
NLI	baly-st	$7.2k$/147.65	$1.3k$/25.40	25/807	8.12/5.24	21/251	100	0.0
	xlni	90.12/44.15	16.23/7.97	15/7	9/7	13.74/28.76	99.16	0.84
STSP	sts-reg	78.72/96.44	14.19/17.26	14/13	12/8	60.07/58.13	98.86	1.14
	sts-cls	80.38/77.25	14.25/13.13	11/10	7/7	10.03/10.33	99.31	0.69
Two inputs avg		1.5k/82.17	270.63/14.29	171/12.4	8.62/6.65	28.26/77.38	99.44	0.56

Table 3: Descriptive statistics of ORCA across the different data splits. * and †: Same data with multiple labels. **SC:** Sentence Classification. **TC:** Topic Classification. **STS:** Textual Semantic Similarity. **NLI:** Natural Language Inference. **QA:** Question-Answering. For the NLI and STPS tasks we compute the statistics in both inputs (e.g., sentence 1 and sentence 2 in ASTS task). We don't include the word-level datasets in this table (i.e., SP tasks.)

2018; Abdul-Mageed et al., 2020b). *(3) Dialect Identification:* Six datasets are used for dialect classification. These are ArSarcasm$_{\text{Dia}}$ (Farha and Magdy, 2020), the Arabic Online Commentary (AOC) dataset (Zaidan and Callison-Burch, 2014), NADI-2020 (Abdul-Mageed et al., 2020a), MADAR (Bouamor et al., 2019), QADI (Abdelali et al., 2020), and Habibi (El-Haj, 2020). The dialect identification task involves three dialect classification levels. These are the binary-level (i.e., MSA vs. DA), region-level (four regions), and country-level (21 countries). *(4) Claim Prediction*: we use ANS-claim (Khouja, 2020), which is a factuality prediction of claims corpus created using the credibility of the editors as a proxy for veracity (true/false). *(5) Machine Generation:* for machine generated text detection (i.e, machine vs. human), we use the machine manipulated version of AraNews dataset (Nagoudi et al., 2020). To create this dataset, a list of words are selected (based on their POS tags) and substituted by a chosen word from the k-most similar words in an Arabic word embedding model.

Structured Prediction. This task cluster includes two tasks: *(1) Arabic NER*: we consider two publicly available Arabic NER datasets, ANERcorp (Benajiba and Rosso, 2007) and AQMAR (Schneider et al., 2012). *(2) Arabic POS Tagging*: we use two POS Tagging datasets, the multi-dialect Arabic POS dataset Darwish et al. (2018) and the Arabic POS tagging part of XGLUE (Liang et al., 2020).

Text Classification. In this task cluster, we explore *topic classification* employing three document-level classification datasets: Khaleej (Abbas et al., 2011), Arabic News Text (ANT) (Chouigui et al., 2017), and OSAC (Saad and Ashour, 2010).

Semantic Textual Similarity. This cluster aims to measure the semantic relationship between a pair of sentences. For this, we use the *(1) STS re-*

gression: data from Ar-SemEval-2017 (Cer et al., 2017) (which is a set of Arabic sentence pairs each labeled with a numerical score from the interval $[0..1]$ indicating the degree of semantic similarity). We also use *(2) STS classification* where a pair of questions is assumed to be semantically similar if they have the same exact meaning *and* answer. We use the semantic question similarity in Arabic dataset (Q2Q) proposed by Seelawi et al. (2019) where each pair is tagged with "1" (question has the same meaning and answer) or "0" (not similar).

Natural Language Inference. This cluster covers the following two tasks: *(1) Arabic NLI:* we use the Arabic part of the cross-lingual natural language inference (XNLI) corpus (Conneau et al., 2018). The goal is determining whether a text (hypothesis) is false (contradiction), undetermined (neutral), or true (entailment), given a another text (premise). *(2) Fact-checking*: in order to build a fact-checking benchmark component, we use Unified-FC (Baly et al., 2018) and ANS (Khouja, 2020). Both of these datasets target stance and factuality prediction of claims from news and social media. The two datasets are manually created by annotating the stance between a claim-document pair with labels from the set *{agree, disagree, discuss, unrelated}*.

Word Sense Disambiguation. We use the Arabic WSD benchmark (El-Razzaz et al., 2021), a context-gloss pair dataset extracted from an MSA dictionary. It consists of 15k senses for 5k unique words with an average of three senses for each word.

Question Answering. We concatenate four Arabic and multilingual QA datasets. These are ARCD (Mozannar et al., 2019), MLQA (Lewis et al., 2020b), TyDi QA (Artetxe et al., 2020), and XQuAD (Artetxe et al., 2020).

4 Language Models

In this section, we list multilingual PLMs that include Arabic in its coverage only by name, for space but provide a description of each of them in Appendix A.

Multilingual LMs. These are mBERT (Devlin et al., 2019) , XLM-R (Conneau et al., 2020b), GigaBERT (Lan et al., 2020), and mT5 (Xue et al., 2021).

Arabic LMs. These are AraBERT (Antoun et al., 2020), ArabicBERT (Safaya et al., 2020), Arabic-ALBERT (Safaya, 2020), QARiB Chowdhury et al. (2020), ARBERT & MARBERT (Abdul-Mageed et al., 2021), CamelBERT (Inoue et al., 2021), JABER and SABER (Ghaddar et al., 2021), and AraT5 (Nagoudi et al., 2022).

Table A.1 (Appendix A) shows a comparison between the multilingual as well as the Arabic PLMs in terms of (1) training data size, (2) vocabulary size, (3) language varieties, and (4) model configuration and architecture.

5 Model Evaluation on ORCA

This section shows experimental settings and performance of 18 multilingual and Arabic language models on ORCA downstream tasks.[6]

Baselines. For comparison, we finetune the multilingual language models mBERT and XLM-R_{Base} on all training data of ORCA benchmark.

Evaluation. For all models and baselines, across all tasks, we identify the best model on the respective development data split (Dev) and blind-test on the testing split (Test). We methodically evaluate each task cluster, ultimately reporting a single *ORCA score* following Wang et al. (2018); Abdul-Mageed et al. (2021). ORCA score is simply the macro-average of the different scores across all tasks and task clusters, where each task is weighted equally. We compute the ORCA score for all 18 language models.

Results. We present results of all language models and the baselines on each task cluster of ORCA independently using the relevant metric, for both Dev (see Table C.1 in Appendix C) and Test (see Table 4). As Table 4 shows, $\text{ARBERT}_{\text{v2}}$ (M3) achieves the highest ORCA score across all the tasks and also for MSA tasks only[7] (*with ORCA score*=74.04 and *Avg. MSA score*=75.13), followed by $\text{CamelBERT}_{\text{msa}}$ (M11) in both cases with 73.35 and 73.64, respectively. Regarding the dialect tasks, we note that $\text{MARBERT}_{\text{v2}}$ (M4) achieves the best dialect ORCA score (*Avg. DA score*=74.62) followed by QARIB (M9) with 74.47. We also note that AraELECTRA (M8) achieves the best results in six individual tasks out of 26, followed by $\text{MARBERT}_{\text{v2}}$ (M4) which excels in five individual tasks.

Analysis. As our experiments imply, ORCA allow us to derive unique insights. Example insights that

[6] We exclude JABER and SABER (Ghaddar et al., 2021) from the evaluation as these are not supported by the Transformers library.

[7] We consider a task an MSA task if it has more than 98% of samples predicted as MSA using the MSA vs. DA classifier (see Table 3).

Cluster	Task	B1	B2	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16
SC	abusive†	72.68	71.31	76.53	78.36	75.99	78.03	75.92	78.06	76.22	76.87	**79.66**	75.49	76.98	73.57	74.43	77.34	72.28	67.68
	adult†	89.52	88.49	89.7	90.76	89.67	**90.97**	88.97	89.9	89.65	90.18	90.89	90.33	90.09	90.76	88.68	89.35	89.74	88.88
	age†	42.68	44.14	44.76	47.11	45.57	46.24	44.10	44.33	42.02	**47.26**	46.35	45.89	45.97	45.29	43.29	43.83	45.23	43.61
	claim*	65.72	66.66	70.25	67.91	67.38	67.83	69.74	69.34	70.35	**71.53**	69.2	68.96	70.32	65.66	63.06	68.81	66.29	65.88
	dangerous†	64.94	66.31	**67.32**	66.2	64.96	67.11	64.72	62.6	67.13	65.66	66.25	64.03	65.31	66.92	61.97	62.83	64.56	63.41
	dialect-B†	84.29	84.78	86.48	86.78	86.92	86.91	86.64	87.01	87.76	87.21	**87.85**	86.79	87.40	86.64	84.58	86.57	86.13	85.94
	dialect-R†	63.12	63.51	67.71	66.08	65.21	66.32	64.63	67.5	64.46	66.34	66.71	65.59	65.05	68.55	63.36	**69.22**	63.98	62.87
	dialect-C†	25.52	30.34	35.26	35.83	35.69	36.06	31.49	36.33	27.00	**36.50**	34.36	33.90	35.18	30.83	27.05	33.96	32.99	28.25
	emotion-cls†	56.79	60.05	63.6	68.85	64.81	**70.82**	60.6	64.89	60.98	66.70	68.03	65.25	63.85	64.8	59.66	61.92	62.2	55.22
	emotion-reg*	37.96	52.37	65.37	73.96	67.73	**74.27**	62.02	67.64	61.51	70.31	71.91	66.73	65.75	64.34	48.46	66.57	62.77	45.72
	gender†	61.78	64.16	64.38	66.65	63.18	**67.64**	62.41	64.37	64.24	65.65	66.64	66.38	65.19	64.25	63.37	63.97	64.35	63.50
	hate†	72.19	67.88	82.41	81.33	82.26	83.54	82.21	82.39	81.79	**85.30**	83.88	81.99	79.68	83.38	74.1	82.25	79.77	74.26
	irony†	82.31	83.13	83.53	83.27	83.83	83.09	83.63	84.51	81.56	84.62	**85.16**	84.01	83.07	81.91	79.68	80.91	83.03	79.05
	offensive†	84.62	87.18	89.28	91.84	89.55	**92.23**	87.5	90.73	89.4	91.89	91.17	90.05	89.32	90.44	86.52	88.76	87.52	85.26
	machine G.*	81.4	84.61	88.35	85.14	87.94	86.69	87.45	89.82	**90.66**	87.96	86.35	86.73	88.62	83.17	83.35	87.43	86.28	83.91
	sarcasm†	69.32	68.42	73.11	74.74	74.16	76.19	73.46	74.06	74.81	**76.83**	75.82	74.17	75.18	72.57	69.94	72.02	73.11	71.92
	sentiment†	78.99	77.21	77.78	79.08	78.60	80.83	78.45	80.50	79.56	**80.86**	80.33	79.51	79.75	77.8	76.76	78.68	78.46	76.46
SP	ner-anerc.*	85.92	86.76	90.27	86.59	90.83	87.86	90.68	**90.85**	90.17	90.03	87.5	89.27	90.71	83.61	82.94	89.52	88.77	86.54
	ner-aqmar*	75.95	76.16	80.72	74.57	**81.70**	74.22	77.34	79.2	73.43	77.66	73.72	76.84	78.54	73.77	70.71	74.97	79.5	73.15
	pos-dia†	92.04	92.78	92.92	94.14	93.92	93.38	91.65	93.79	94.70	93.554	**94.70**	93.95	94.37	93.95	92.05	92.05	93.24	92.57
	pos-xglue*	57.68	69.37	51.39	55.02	52.55	55.45	34.65	37.84	41.28	26.61	41.36	27.70	32.89	10.37	17.04	62.58	**63.89**	42.40
NLI	ans-st*	84.49	81.00	91.77	73.82	91.02	80.57	87.59	**93.23**	92.33	90.17	50.2	82.49	89.21	46.01	71.81	85.31	82.86	80.02
	baly-st*	34.48	38.27	45.63	29.07	49.34	36.52	**51.19**	46.63	41.32	37.12	31.58	48.94	49.67	30.58	48.85	49.21	49.22	47.19
	xlni*	61.88	65.06	67.22	60.50	68.17	62.22	64.69	67.93	**70.20**	66.67	55.67	63.82	66.02	54.29	61.18	66.53	62.15	61.62
STS	sts-r$^{\star\,\ddagger}$	63.91	62.24	73.00	63.48	71.90	66.12	71.27	75.4	**76.01**	70.50	41.15	70.61	74.42	71.23	70.13	73.68	73.56	66.75
	sts-c*	62.34	63.35	85.95	74.43	96.73	63.47	96.81	64.11	64.24	63.52	84.11	63.28	**97.10**	59.57	96.41	85.87	96.69	62.91
TC	topic*	92.55	93.53	94.17	93.53	93.96	93.9	94.31	**94.58**	94.11	94.02	93.32	93.72	94.38	93.18	93.41	94.05	93.86	93.27
QA	arlue-qa*	56.39	56.51	57.65	49.35	61.5	57.9	56.79	**61.56**	60.70	57.65	45.27	53.98	57.46	30.91	52.11	58.71	55.94	53.89
WSD	ar-wsd*	69.82	52.90	33.29	72.94	71.01	33.28	51.72	**76.68**	73.54	72.92	70.13	74.12	75.86	65.18	75.68	74.31	69.76	47.19
Avg. DIA†		69.39	69.98	72.98	74.07	72.95	**74.62**	71.76	73.40	<u>74.47</u>	71.97	<u>74.47</u>	73.18	73.06	72.79	69.70	72.25	71.65	69.10
Avg. MSA*		66.75	67.77	73.91	65.76	**75.13**	67.17	71.16	72.48	70.65	70.37	64.39	69.09	73.64	<u>59.42</u>	66.80	74.20	72.15	47.19
ORCA$_{score}$		68.07	68.88	73.45	69.91	**74.04**	72.95	71.46	72.94	72.56	71.17	69.43	71.13	<u>73.35</u>	66.10	69.15	73.23	71.90	58.14

Table 4: Performance of Arabic Bert-based models on ORCA Test splits (F_1) ‡ Metric for STSP taks is spearman correlation. **B1, B2**: Two baselines mBERT (Devlin et al., 2019) and XLM-R (Liu et al., 2019a). **M1, M2**: ARBERT, MARBERT (Abdul-Mageed et al., 2021). **M3, M4**: ARBERT$_{V2}$ and MARBERT$_{V2}$. **M5, M6, M7, and M8**: AraBERT$_{v1[v2,\ tw]}$, and AraElectra (Antoun et al., 2020, 2021). **M9**: QARiB (Chowdhury et al., 2020) **M10, M11, M12, and M13**: CamelBERT$_{mix[msa,\ da,\ ca]}$ (Inoue et al., 2021). **M14**: GigaBERT$_{v4}$ (Chowdhury et al., 2020). **M15**: Arabic BERT (Chowdhury et al., 2020). **M16** : Arabic Albert (Lan et al., 2020). **Avg. DIA and Avg. MSA:** The average of dialect and MSA tasks. **ORCA**$_{score}$: Average overall Dia and MSA tasks. *MSA tasks. ‡DIA tasks. A task is considered as an MSA if it has more than 98% samples predicted as MSA using an MSA vs. DIA classifier (see Table 3).

can be derived from Table 4 are: (a) a model such as M6 (i.e., AraBERTv2) that is pretrained with historical data (AlSafeer newspaper) would excel on old datasets (e.g., TC, QA, and WSD); while M4 (i.e., MARBERTv2) excels especially on datasets from social media since it is pretrained with a large Twitter collection. In addition, since ORCA arranges the metrics into one dedicated to dialect, another to MSA, and a third to both (ORCA score), it is much easier to compare model performance across the DA-MSA dimensions.

6 Analysis of Model Computational Cost

We also compare the Arabic language models in terms of computational cost using the average time needed for convergence (in minutes) and average number of epochs to convergence as identified on Dev sets. For this, we finetune all models for a maximum of 25 epochs on all ORCA tasks. *We report results in terms of average of three runs.* Figure E.2 (Appendix E) shows for each model the *total time needed for convergence* (out of 25 epochs), and Figure E.1 (Appendix E) shows *average convergence time* and *average number of epochs till convergence*. As E.2 (Appendix E) shows, Arabic Albert is the fastest model (52.26 min) to finetune for 25 epochs, but it achieves the lowest ORCA score. Excluding Arabic Albert, we observe a near constant time (between 60.32-63.69 mins) for all other models. Among the top five models, as Figure E.1 (Appendix E) shows, we also observe that ARBERT$_{v1}$ is the fastest (in terms of average convergence time and number of epochs needed to converge) and is followed by QARiB.

7 Conclusion

We presented ORCA, a large and diverse benchmark for Arabic natural language understating tasks composed of 60 datasets that are arranged in seven task clusters. To facilitate future research and adoption of our benchmark, we offer a publicly-available interactive leaderboard with a useful suite of tools and extensive meta-data. In addition, we provide a comprehensive and methodical evaluation as well as meaningful comparisons between 18 multilingual and Arabic language models on ORCA. We also compare the models in terms of computing needs. As our results show, ORCA is challenging and we hope it will help standardize comparisons and accelerate progress both for Arabic and multilingual NLP.

8 Limitations

We identify the following limitations:

1. Although we strive to include tasks in all Arabic varieties, available downstream datasets from certain countries such as Mauritania and Djibouti are almost nonexistent and so are not covered in ORCA. In addition, there is a need in the community to create more datasets for several Arabic dialects. This includes, for example, dialects such as Iraqi, Sudanese, and Yemeni. With the introduction of more datasets for such dialects, ORCA's coverage can be further extended. Regardless, as Figure F.1 (Appendix F) shows, ORCA datasets are quite diverse from a geographical perspective.

2. Although ORCA currently covers both dialectal Arabic (DA) and MSA, it does not pay as much attention to the classical variety of Arabic (CA) due to historical reasons. That is, the community did not invest as much efforts creating and releasing datasets involving CA. However, as more unlabeled datasets become available and with an undergoing positive change in the culture around data sharing, this is likely to change in the near future. Again, this will make it possible to extend ORCA to better cover CA in the future.

3. Although benchmarks in general are useful in encouraging standardize evaluations and meaningful comparisons, and can help motivate progress within the community, they also run the risk of contributing to a culture of leaderboard chasing that is not necessarily useful. That is, although scientific research advances due to competition, it also thrives through partnerships and collaborations that bring the best from diverse groups. It is in the context of this collaborative culture that we hope ORCA will be perceived and used.

9 Ethics Statement and Broad Impact

Encouraging standardized evaluations and contributing to a collaborative research culture. Similar to some other research communities, progress in the Arabic NLP community has been hampered for a long time by absence of standardized and meaningful evaluations for some tasks. This is due to several reasons, including the culture around data sharing but also as a result of insufficient funding and lack of strong training programs. This has made it challenging to measure progress. The Arabic NLP community is now expanding, and a culture of collaboration is being built as part of the larger positive developments within the overall NLP community itself. As such, it is now ripe time to introduce benchmarks that can help this ongoing progress. We hope there will be wide adoption of ORCA and that our work will trigger more efforts to create more benchmarks, including for newer tasks in what could be a virtuous cycle.

Data privacy. Regarding data involved in ORCA, we develop the benchmark using data from the public domain. For this reason, we do not have serious concerns about privacy.

Sufficient assignment of credit to individual data sources. Another important consideration in benchmarking is how credit is assigned to creators of the individual datasets. To ensure sufficient credit assignment, we refer users to the original publications, websites, GitHub repositories where a dataset originated and link all these sources in our leaderboard. We also provide bibliographic entries for all these sources that users can easily copy and paste in order to cite these original sources. By encouraging citation of original sources in any publications in the context of ORCA use, we hope to afford additional visibility to many of the individual datasets.

Acknowledgements

MAM gratefully acknowledges support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada (DRAG),[8] UBC ARC-Sockeye,[9] Advanced Micro Devices, Inc. (AMD), and Google. Any opinions, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of CRC, NSERC, SSHRC, CFI, DRAG, AMD, Google, or UBC ARC-Sockeye.

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, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

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, 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.

[8]https://alliancecan.ca
[9]https://arc.ubc.ca/ubc-arc-sockeye

Latifah Almuqren and Alexandra Cristea. 2021. AraCust: a Saudi Telecom Tweets corpus for sentiment analysis. *PeerJ Computer Science*, 7:e510.

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.

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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195.

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, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

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, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec

Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Shammur Absar Chowdhury, Ahmed Abdelali, Kareem Darwish, Jung Soon-Gyo, Joni Salminen, and Bernard J. Jansen. 2020. Improving Arabic text categorization using transformer training diversification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 226–236, Barcelona, Spain (Online). Association for Computational Linguistics.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

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

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.

Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, Tongshuang Sherry Wu, Jascha Narain Sohl-Dickstein, Jinho D. Choi, Eduard H. Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa

Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline De Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, M. Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Rishabh Gupta, Louanes Hamla, Sanghyun Han, Fabrice Harel-Canada, Antoine Honoré, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey J. Levinson, H.-J. Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxime Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Muennighoff, Timothy Sum Hon Mun, Kenton W. Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan A. Pfister, Richard Plant, Vinay Uday Prabhu, Vasile Florian Pais, Libo Qin, Shahab Raji, Pawan Rajpoot, Vikas Raunak, Roy Rinberg, Nicolas M. Roberts, Juan Diego Rodriguez, Claude Roux, S. VasconcellosP.H., Ananya B. Sai, Robin M. Schmidt, Thomas Scialom, T. Sefara, Saqib Nizam Shamsi, Xudong Shen, Haoyue Shi, Yi ning Shi, Anna V. Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyanka Soni, Taylor M Sorensen, William Soto, Aman Srivastava, K V Aditya Srivatsa, Tony Sun, T MukundVarma, Afsheen Tabassum, Fiona Anting Tan, Ryan Teehan, Monalisa Tiwari, Marie Tolkiehn, Athena Wang, Zi-Hao Wang, Gloria Wang, Zijie J. Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyi Wu, Witold Wydmański, Tianbao Xie, Usama Yaseen, M. Yee, Jing Zhang, and Yue Zhang. 2021. Nl-augmenter: A framework for task-sensitive natural language augmentation. *ArXiv*, abs/2112.02721.

Mahmoud El-Haj. 2020. Habibi: a multi Dialect multi National Arabic Song Lyrics Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1318–1326.

Ibrahim Abu El-Khair. 2016. 1.5 Billion Words Arabic Corpus. *arXiv preprint arXiv:1611.04033*.

Mohammed El-Razzaz, Mohamed Waleed Fakhr, and Fahima A Maghraby. 2021. Arabic gloss WSD using BERT. *Applied Sciences*, 11(6):2567.

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.

Chayma Fourati, Abir Messaoudi, and Hatem Haddad. 2020. TUNIZI: a Tunisian Arabizi sentiment analysis Dataset. In *AfricaNLP Workshop, Putting Africa on the NLP Map. ICLR 2020, Virtual Event*, volume arXiv:3091079.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Abbas Ghaddar, Yimeng Wu, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang, Duan Xinyu, Zhefeng Wang, Baoxing Huai, et al. 2021. JABER: Junior Arabic BERT. *arXiv preprint arXiv:2112.04329*.

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*.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

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 *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Jude Khouja. 2020. Stance prediction and claim verification: An Arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 8–17, Online. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. Jglue: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966.

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.

Trung Le, Tuan Nguyen, Nhat Ho, Hung Bui, and Dinh Phung. 2021. LAMDA: Label Matching Deep Domain Adaptation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6043–6054. PMLR.

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

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

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

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.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. 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.

Hamdy Mubarak, Sabit Hassan, and Ahmed Abdelali. 2021. Adult content detection on Arabic Twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 136–144, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-Text Transformers for Arabic Language Generation. Online. Association for Computational Linguistics.

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.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Ali Safaya. 2020. Arabic-ALBERT.

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.

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

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A Smith. 2012. Coarse lexical semantic annotation with supersenses: an arabic case study. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 253–258.

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.

Haitham Seelawi, Ahmad Mustafa, Hesham Al-Bataineh, Wael Farhan, and Hussein T Al-Natsheh. 2019. Nsurl-2019 task 8: Semantic question similarity in arabic. In *Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers*, pages 1–8.

Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. ALUE: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

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.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, volume 2012, pages 2214–2218.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

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. *arXiv preprint arXiv:1804.07461*.

Junqiu Wei, Qun Liu, Yinpeng Guo, and Xin Jiang. 2021. Training multilingual pre-trained language model with byte-level subwords. *arXiv preprint arXiv:2101.09469*.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, et al. 2020. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. *arXiv preprint arXiv:2009.05387*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

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.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics: System Demonstrations, pages 270–278, Online. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Appendices

In this appendices, we provide an addition which organized as follows:

Sections list:

Tables and Figures List:

A Language Models

In this section, we provide a description of the multilingual MLM that include Arabic in its coverage.

A.1 Multilingual LMs

mBERT is the multilingual version of BERT (Devlin et al., 2019) which is a multi-layer bidirectional encoder representations from Transformers (Vaswani et al., 2017) trained with a masked language modeling. Devlin et al. (2019) present two architectures: *Base* and *Large*. BERT models were trained on English Wikipedia[10] and BooksCorpus (Zhu et al., 2015). mBERT is trained on Wikipedia for 104 languages (including $\sim$ 153M Arabic tokens).

XLM-R (Conneau et al., 2020b) is a transformer-based multilingual masked language model pre-trained on more than 2TB of filtered CommonCrawl data in 100 languages, including Arabic (2.9B tokens). XLM-R uses a Transformer model trained a multilingual version of masked language modeling of XLM (Conneau and Lample, 2019). XLM-R comes with two sizes and architectures: *Base* and *Large*. The XLM-$\mathrm{R_{Base}}$ architecture contains 12 layers, 12 attention heads, 768 hidden units, and 270M parameters. The XLM-$\mathrm{R_{Large}}$ architecture has 124 layers, 16 attention heads, 1024 hidden units, and 550M parameters. While both XLM-R models use the same masking objective as BERT, they do not include the next sentence prediction objective used in BERT.

GigaBERT (Lan et al., 2020) is a customized bilingual BERT-based model for Arabic and English pretrained on a corpus of 10B tokens collected from different sources, including: English and Arabic Gigaword corpora,[11], OSCAR (Suárez et al., 2019), and Wikipedia. GigaBERT is designed specifically for zero-shot transfer learning from English to Arabic on information extraction tasks.

mT5 (Xue et al., 2021) is the multilingual version of **T**ext-**to**-**T**ext **T**ransfer **T**ransformer model (T5) (Raffel et al., 2020). The T5 model architecture is

[10] https://www.wikipedia.org/

[11] https://catalog.ldc.upenn.edu/LDC2011T07

	Models	Training Data			Vocabulary		Configuration
		Type	Text Size (ar)	Tokens (ar/all)	Tok.	Size	#Param.
ML LMs	mBERT	MSA	1.4GB	153M/1.5B	WP	110K	110M
	XLM-R	MSA	5.4GB	2.9B/295B	SP	250K	270M
	GigaBERT	MSA	42.4GB	4.3B/10.4B	WP	50k	125M
Arabic LMs	ARBERT	MSA	61GB	6.2B	WP	100K	163M
	ARBERT$_{v2}$	MSA, DA	243GB	27.8B	WP	100K	163M
	MARBERT	MSA, DA	128GB	15.6B	WP	100K	163M
	MARBERT$_{v2}$	MSA	198GB	21.4B	WP	100K	163M
	AraBERT	MSA	27GB	2.5B	WP	64K	135M
	AraELECTRA	MSA	77GB	8.8B	WP	64K	135M
	ArabicBERT	MSA	95GB	8.2B	WP	64K	135M
	Arabic-ALBERT	MSA	33GB	4.4B	WP	32K	110M
	QARiB	MSA, DA	97GB	14B	WP	64K	135M
	CAMeLBERT	MSA, DA, CA	167GB	8.8B	WP	30K	108M
	JABER	MSA	115GB	–	BBPE	64K	135M
	SABER	MSA	115GB	–	BBPE	64K	135M
	AraT5	MSA, DA	248GB	29B	SP	110K	220M

Table A.1: Configuration comparisons of Arabic pre-trained LMs and multilingual LMs which covered Arabic. **WP**: WordPiece (Schuster and Nakajima, 2012). **SP**: SentencePiece (Kudo and Richardson, 2018). **BBPE**: Byte-level Byte Pair Encoding (Wei et al., 2021). **ARBERT$_{v2}$:** a new model proposed in this paper.

essentially an encoder-decoder Transformer similar in configuration and size to BERT$_{\text{Base}}$. The T5 model treats every text-based language task as a "text-to-text" problem, (i.e. taking text format as input and producing new text format as output), where multi-task learning is applied with several NLP tasks: question answering, document summarization, machine translation, and sentiment classification. mT5 is trained on the "Multilingual Colossal Clean Crawled Corpus" (or mC4 for short), which is $\sim$ 26.76TB for 101 languages (including Arabic with more than $\sim$ 57B tokens) generated from 71 Common Crawl dumps.

A.2 Arabic LMs

Several Arabic LMs have been developed. We describe the most notable among these here.

AraBERT (Antoun et al., 2020) is the first pre-trained language model proposed for Arabic. It is based on the two BERT$_{\text{Base}}$ and BERT$_{\text{Large}}$ architectures. AraBERT$_{\text{Base}}$ (Antoun et al., 2020) is trained on 24GB of Arabic text (70M sentences and 3B tokens) collected from Arabic Wikipedia, Arabic news, Open Source International dataset (OSIAN) (Zeroual et al., 2019), and 1.5B words corpus from (El-Khair, 2016). In order to train BERT$_{\text{Large}}$ Antoun et al. (2021) use the same AraBERT$_{Base}$ data augmented with the unshuffled Arabic OSCAR dataset (Suárez et al., 2019) and news articles provided by As-Safir newspaper[12] (77GB or 8.8B tokens) . The augmented data is also used to train AraELECTRA$_{Large}$–an Arabic language model that employs an ELECTRA objective (Clark et al., 2020).

ArabicBERT is an Arabic BERT-based model proposed by Safaya et al. (2020) Authors pretrain fourr variants: ArabicBERT$_{\text{Mini}}$, ArabicBERT$_{\text{Medium}}$, ArabicBERT$_{\text{Base}}$, ArabicBERT$_{\text{Large}}$.[13] The models are pretrained on unshuffled Arabic OSCAR (Suárez et al., 2019), Arabic Wikipedia, and other Arabic resources which sum up to 95GB of text ($\sim$ 8.2B tokens).

Arabic-ALBERT (Safaya, 2020) is an Arabic language representation model based on **A** **Lite** **Bert** (ALBERT) (Lan et al., 2019). ALBERT is a Transformer-based neural network architecture (similar to BERT and XLM-R) with two parameter reduction techniques proposed to increase the training speed and lower memory consumption of the BERT model. Arabic-ALBERT is pretrained on $\sim$ 4.4B tokens extracted from Arabic OSCAR (Suárez et al., 2019) and Arabic Wikipedia. Arabic-ALBERT comes with three dif-

[12] https://www.assafir.com/

[13] https://github.com/alisafaya/Arabic-BERT

ferent architectures: Arabic-ALBERT$_{\text{Base}}$, Arabic-ALBERT$_{\text{Large}}$, Arabic-ALBERT$_{\text{XLarge}}$.

QARiB. Chowdhury et al. (2020) propose the **Q**CRI **AR**abic and D**i**alectal **B**ERT (QARiB) model. QARiB is trained on a collection of 97GB of Arabic Text (14B tokens) on both MSA (180 Million sentences) and Twitter data (420 Million tweets). Authors use the Twitter API to collect Arabic tweets, keeping only tweets identified as Arabic by Twitter language filter. For MSA data in QARiB is a combination of Arabic Gigaword,[14], Abulkhair Arabic Corpus (El-Khair, 2016), and OPUS (Tiedemann, 2012).

ARBERT (Abdul-Mageed et al., 2021) is a pre-trained language model focused on MSA. ARBERT is trained using the same architecture as BERT$_{\text{Base}}$ with a vocabulary of 100K WordPieces, making $\sim$ 163M parameters. ARBERT exploits a collection of Arabic datasets comprising 61GB of text (6.2B tokens) from the following sources: El-Khair El-Khair (2016), Arabic Gigaword,[15], OSCAR (Suárez et al., 2019), OSIAN (Zeroual et al., 2019), Arabic Wikipedia, and Hindawi Books.[16]

ARBERT$_{\text{v2}}$. We provide a new Arabic version of **ARBERT**, by further pretraining ARBERT on 243GB MSA dataset (70GB MSA data from various sources and 173GB extracted and cleaned from the Arabic part of the multilingual Colossal Clean Crawled Corpus (mC4) (Xue et al., 2021).

MARBERT (Abdul-Mageed et al., 2021) is a pre-trained language model focused on both dialectal Arabic and MSA. This model is trained on a sample of 1B Arabic tweets (128GB of text, 15.6B tokens). In this dataset, authors keep only tweets with at least 3 Arabic words (based on character string matching) regardless of whether the tweet has non-Arabic string or not. MARBERT uses the same vocabulary size (100K WordPieces) and network architecture as ARBERT (BERT$_{\text{Base}}$), but without the next sentence prediction objective since tweets are short. **MARBERT$_{\text{v2}}$.** Abdul-Mageed et al. (2021) further pretrain MARBERT with additional data using a larger sequence length of 512 tokens for 40 epochs.

CamelBERT (Inoue et al., 2021) is pre-trained using BERT$_{\text{Base}}$ architecture on four types of Arabic datasets: MSA (107GB), dialectal Arabic (54GB), classical Arabic (6GB), and a mixture of the last three datasets (167GB). CamelBERT is trained using a small vocabulary of 30K tokens (in WordPieces).

JABER and SABER (Ghaddar et al., 2021) are BERT-based models (Base and Large) pretraind on 115GB of text data collected from Common Crawl (CC), OSCAR (Suárez et al., 2019), OSIAN (Zeroual et al., 2019), El-Khair El-Khair (2016), and Arabic Wikipedia. In order to overcome the out-of-vocabulary problem and improve the representations of rare words, JABER is trained using a Byte-level Byte Pair Encoding (BBPE) (Wei et al., 2021) tokenizer with a vocabulary size of 64K.

AraT5 (Nagoudi et al., 2022) is an Arabic text-to-text Transformer model dedicated to MSA and dialects. It is essentially an encoder-decoder Transformer similar in configuration and size to T5 (Raffel et al., 2020). AraT5 is trained on more than 248GB of Arabic text (70GB MSA and 178GB tweets), where the data is from the following sources: AraNews (Nagoudi et al., 2020), El-Khair El-Khair (2016), Gigaword,[17], OSCAR (Suárez et al., 2019), OSIAN (Zeroual et al., 2019), Wikipedia Arabic, and Hindawi Books.[18]

Table A.1 shows a comparison between the multilingual as well as the Arabic language models in terms of (1) training data size, (2) vocabulary size, (3) language varieties, and (4) model configuration and architecture.

B X-Specific Benchmarks

CLUE. Xu et al. (2020) introduce CLUE, a benchmark for Chinese NLU. It covers nine tasks spanning single-sentence/sentence-pair classification, text classification, coreference resolution, semantic similarity, and question answering.

FLUE. Le et al. (2020) offer FLUE, a French NLU benchmark involving six datasets with different levels of difficulty, degrees of formality, and domains. FLUE is arranged into three tasks: text classification, paraphrasing, and NLI.

IndoNLU. Wilie et al. (2020) present IndoNLU, a benchmark for Bahasa Indonesian NLU with 12 downstream tasks organized into five task clusters: sentence classification, structure protection, text classification, semantic similarity, and question answering.

JGLUE. Kurihara et al. (2022) propose JGLUE, a Japanese NLU benchmark consisting of six

[14] https://catalog.ldc.upenn.edu/LDC2011T11
[15] https://catalog.ldc.upenn.edu/LDC2009T30
[16] https://www.hindawi.org/books
[17] https://catalog.ldc.upenn.edu/LDC2009T30
[18] https://www.hindawi.org/books

datasets arranged into three task clusters: sentence classification, text classification, and question answering.

KorNLI and KorSTS. Ham et al. (2020) release KorNLI and KorSTS, two benchmark datasets for NLI and STS in the Korean language.

C ORCA Evaluation

In this section, we provide additional information about the evaluation as follows:

- Performance of Arabic BERT-based models on ORCA Dev splits are shown in Table C.1.
- Figure C.1 shows ORCA scores from the different PLMs across all 29 tasks in ORCA benchmark.
- Figure C.2 shows models' ORCA scores across all clusters in ORCA benchmark.
- Figure C.3 shows models' F_1 scores across all tasks in sentence classification cluster.

D Public leaderboard.

In this section, we provide additional screenshots for ORCAleaderboard, as follows:

- Figure D.1 shows an example of tasks sorted alphabetically.
- Figure D.3 shows detailed scores by all models for a given task..
- Figure D.3 shows detailed information about each task cluster and associated tasks, with each task assigned an identifier, language variety, evaluation metric, a link to the dataset website/GitHub/paper and bibliographic information.

E Analysis of Model Computational Cost

In this section, we provide additional information about the models' computational cost, as follows:

- Figure E.1 shows the average number of epochs (in orange), and time needed to converge (mins, in blue) for all the studied pretrained language models across all ORCA tasks.
- Figure E.2 shows the time needed in minutes to fine-tune (25 epochs). We compute the average time of three runs across all ORCA tasks.

F ORCA Data

In this section, we provide additional information about ORCAData, as follows:

- Table F.1 shows a randomly picked examples from the dialectal portion of ORCA Train datasets.
- Figure F.1 shows the predicted country-level distribution, in percentage, in the dialectal portion of ORCA.

Cluster	Task	B1	B2	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16
SC	abusive†	72.68	71.31	76.53	78.36	75.99	78.03	75.92	78.06	76.22	76.87	**79.66**	75.49	76.98	73.57	74.43	77.34	72.28	67.68
	adult†	89.52	88.49	89.7	90.76	89.67	**90.97**	88.97	89.9	89.65	90.18	90.89	90.33	90.09	90.76	88.68	89.35	89.74	88.88
	age†	42.68	44.14	44.76	47.11	45.57	46.24	44.10	44.33	42.02	**47.26**	46.35	45.89	45.97	45.29	43.29	43.83	45.23	43.61
	claim*	65.72	66.66	70.25	67.91	67.38	67.83	69.74	69.34	70.35	**71.53**	69.2	68.96	70.32	65.66	63.06	68.81	66.29	65.88
	dangerous†	64.94	66.31	**67.32**	66.2	64.96	67.11	64.72	62.6	67.13	65.66	66.25	64.03	65.31	66.92	61.97	62.83	64.56	63.41
	dialect-b†	84.29	84.78	86.48	86.78	86.92	86.91	86.64	87.01	87.76	87.21	**87.85**	86.79	87.40	86.64	84.58	86.57	86.13	85.94
	dialect-r†	63.12	63.51	67.71	66.08	65.21	66.32	64.63	67.5	64.46	66.34	66.71	65.59	65.05	68.55	63.36	**69.22**	63.98	62.87
	dialect-c†	25.52	30.34	35.26	35.83	35.69	36.06	31.49	36.33	27.00	**36.50**	34.36	33.90	35.18	30.83	27.05	33.96	32.99	28.25
	emotion†	56.79	60.05	63.6	68.85	64.81	**70.82**	60.6	64.89	60.98	66.70	68.03	65.25	63.85	64.8	59.66	61.92	62.2	55.22
	emotion-reg *	37.96	52.37	65.37	73.96	67.73	**74.27**	62.02	67.64	61.51	70.31	71.91	66.73	65.75	64.34	48.46	66.57	62.77	45.72
	gender†	61.78	64.16	64.38	66.65	63.18	**67.64**	62.41	64.37	64.24	65.65	66.64	66.38	65.19	64.25	63.37	63.97	64.35	63.50
	hate†	72.19	67.88	82.41	81.33	82.26	83.54	82.21	82.39	81.79	**85.30**	83.88	81.99	79.68	83.38	74.1	82.25	79.77	74.26
	irony†	82.31	83.13	83.53	83.27	83.83	83.09	83.63	84.51	81.56	84.62	**85.16**	84.01	83.07	81.91	79.68	80.91	83.03	79.05
	offensive†	84.62	87.18	89.28	91.84	89.55	**92.23**	87.5	90.73	89.4	91.89	91.17	90.05	89.32	90.44	86.52	88.76	87.52	85.26
	machine G.*	81.4	84.61	88.35	85.14	87.94	86.69	87.45	89.82	**90.66**	87.96	86.35	86.73	88.62	83.17	83.35	87.43	86.28	83.91
	sarcasm†	69.32	68.42	73.11	74.74	74.16	76.19	73.46	74.06	74.81	**76.83**	75.82	74.17	75.18	72.57	69.94	72.02	73.11	71.92
	sentiment†	78.99	77.21	77.78	79.08	78.60	80.83	78.45	80.50	79.56	**80.86**	80.33	79.51	79.75	77.8	76.76	78.68	78.46	76.46
NER	anerc.	85.92	86.76	90.27	86.59	90.83	87.86	90.68	**90.85**	90.17	90.03	87.5	89.27	90.71	83.61	82.94	89.52	88.77	86.54
	aqmar	75.95	76.16	80.72	74.57	**81.70**	74.22	77.34	79.2	73.43	77.66	73.72	76.84	78.54	73.77	70.71	74.97	79.5	73.15
	pos-dia†	92.04	92.78	92.92	94.14	93.92	93.38	91.65	93.79	94.70	93.554	**94.70**	93.95	94.37	93.95	92.05	92.05	93.24	92.57
	pos-xglue*	57.68	69.37	51.39	55.02	52.55	55.45	34.65	37.84	41.28	26.61	41.36	27.70	32.89	10.37	17.04	62.58	**63.89**	42.40
NLI	ans-st*‡	84.49	81.00	91.77	73.82	91.02	80.57	87.59	**93.23**	92.33	90.17	50.2	82.49	89.21	46.01	71.81	85.31	82.86	80.02
	baly-st*	34.48	38.27	45.63	29.07	49.34	36.52	**51.19**	46.63	41.32	37.12	31.58	48.94	49.67	30.58	48.85	49.21	49.22	47.19
	xlni*	61.88	65.06	67.22	60.50	68.17	62.22	64.69	67.93	**70.20**	66.67	55.67	63.82	66.02	54.29	61.18	66.53	62.15	61.62
STS	sts-r†	63.91	62.24	73	63.48	71.90	66.12	71.27	75.4	**76.01**	70.50	41.15	70.61	74.42	71.23	70.13	73.68	73.56	66.75
	sts-c*	62.34	63.35	85.95	74.43	96.73	63.47	96.81	64.11	64.24	63.52	84.11	63.28	**97.10**	59.57	96.41	85.87	96.69	62.91
TC	topic	92.55	93.53	94.17	93.53	93.96	93.9	94.31	**94.58**	94.11	94.02	93.32	93.72	94.38	93.18	93.41	94.05	93.86	93.27
QA	arlue-qa	56.39	56.51	57.65	49.35	61.5	57.9	56.79	**61.56**	60.70	57.65	45.27	53.98	57.46	30.91	52.11	58.71	55.94	53.89
	pos-dia†	92.04	92.78	92.92	94.14	93.92	93.38	91.65	93.79	94.70	93.554	**94.70**	93.95	94.37	93.95	92.05	92.05	93.24	92.57
Avg. Dia		67.76	68.35	71.56	72.63	71.45	**73.28**	70.33	71.94	70.47	72.99	<u>73.07</u>	71.67	71.57	71.26	68.09	70.82	70.23	67.59
Avg. MSA		66.91	68.87	75.86	69.36	**77.35**	70.90	75.82	75.02	73.75	73.09	65.83	72.11	<u>76.85</u>	63.02	70.20	75.05	74.82	68.40
ORCA$_{score}$		67.34	68.61	73.71	70.99	**74.40**	72.12	73.08	73.48	72.11	73.04	69.45	71.89	<u>74.21</u>	67.14	69.15	72.94	72.53	67.99

Table C.1: Performance of Arabic Bert-based models on ORCA Dev splits (F_1). ‡ Metric for STSP taks is spearman correlation. **B1, B2**: Two baselines mBERT (Devlin et al., 2019) and XLM-R (Liu et al., 2019a). **M1, M2**: ARBERT, MARBERT (Abdul-Mageed et al., 2021). **M3, M4**: ARBERT$_{\text{V2}}$ and MARBERT$_{\text{V2}}$. **M5, M6, M7, and M8**: AraBERT$_{\text{v1[v2, tw]}}$, and AraElectra (Antoun et al., 2020, 2021). **M9**: Qraib (Chowdhury et al., 2020) **M10, M11, M12, and M13**: CamelBERT$_{\text{mix[msa, da, ca]}}$ (Inoue et al., 2021). **M14**: GigaBERT$_{\text{v4}}$ (Chowdhury et al., 2020). **M15**: Arabic BERT (Chowdhury et al., 2020). **M16** : Arabic Albert (Lan et al., 2020). **Avg. Dia, and Avg. MSA:** The average of dialect and MSA tasks. **ORCA$_{score}$** : Average overall Dia and MSA tasks. *MSA tasks. ‡DIA tasks. A task is considered as an MSA if it has more than 98% samples predicted as MSA using an MSA Vs DIA classifier (see Table 3).

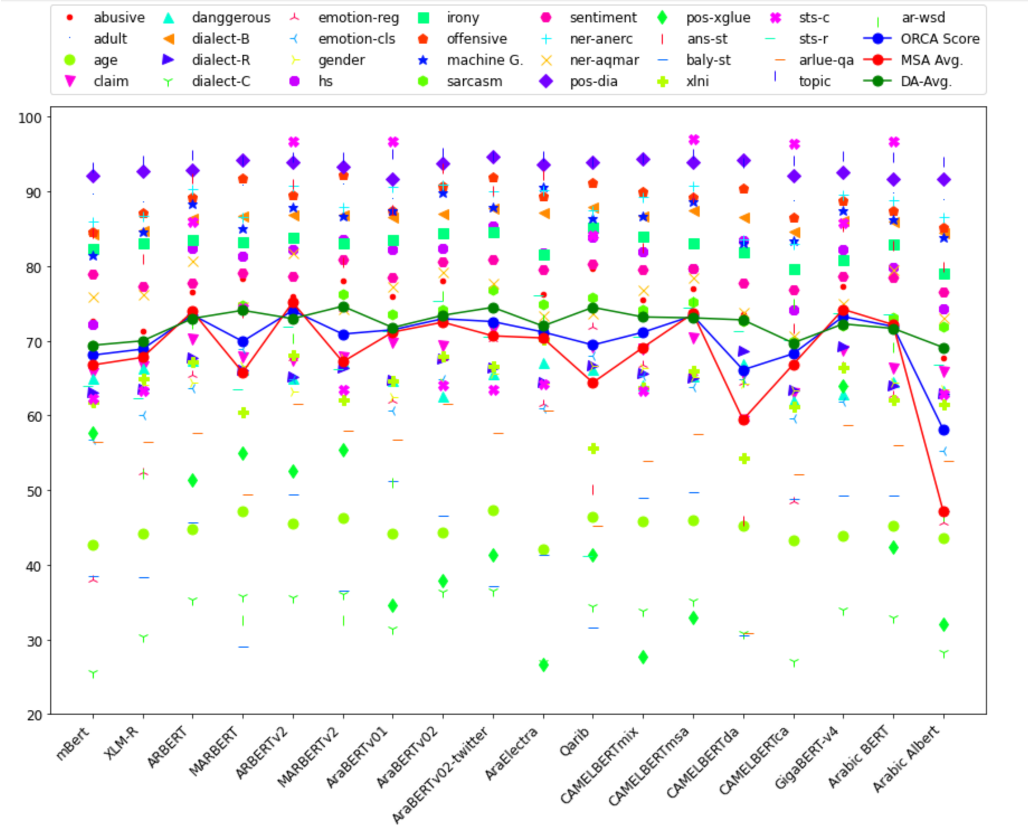


Figure C.1: Models by ORCA score across all 29 tasks in ORCA benchmark.

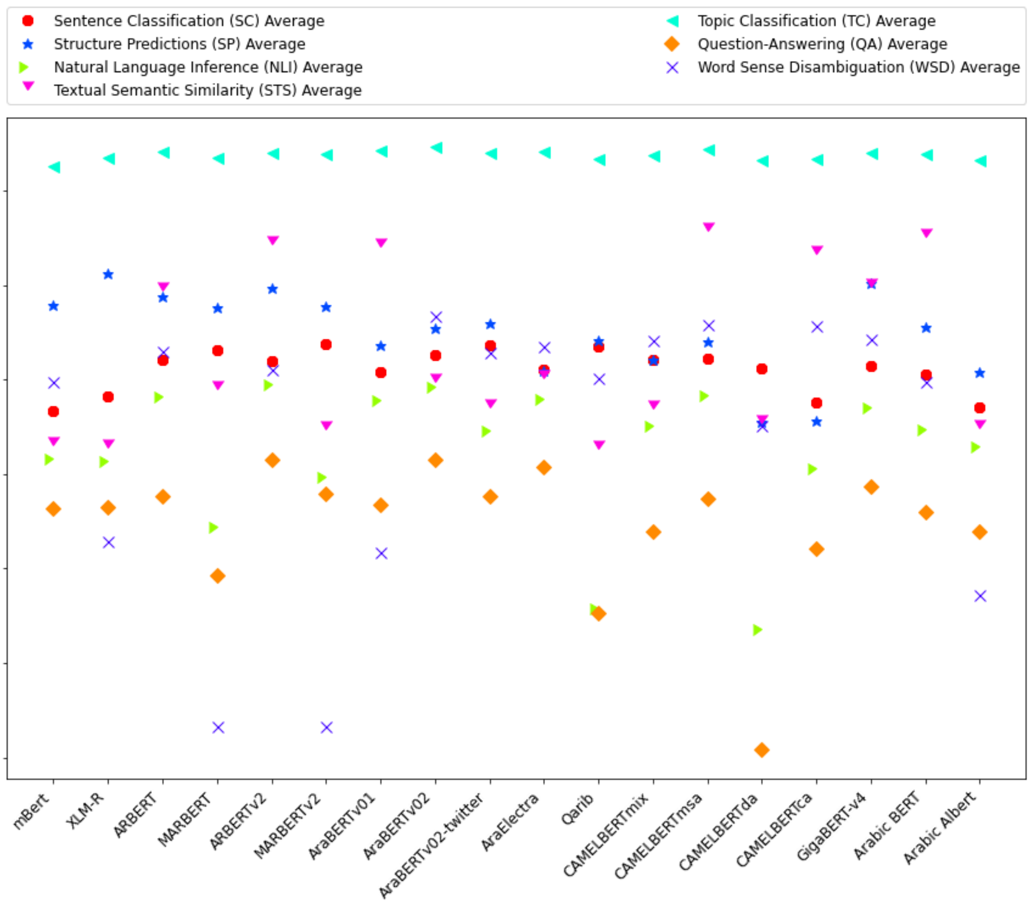


Figure C.2: Models by ORCA score across all clusters in ORCA benchmark.

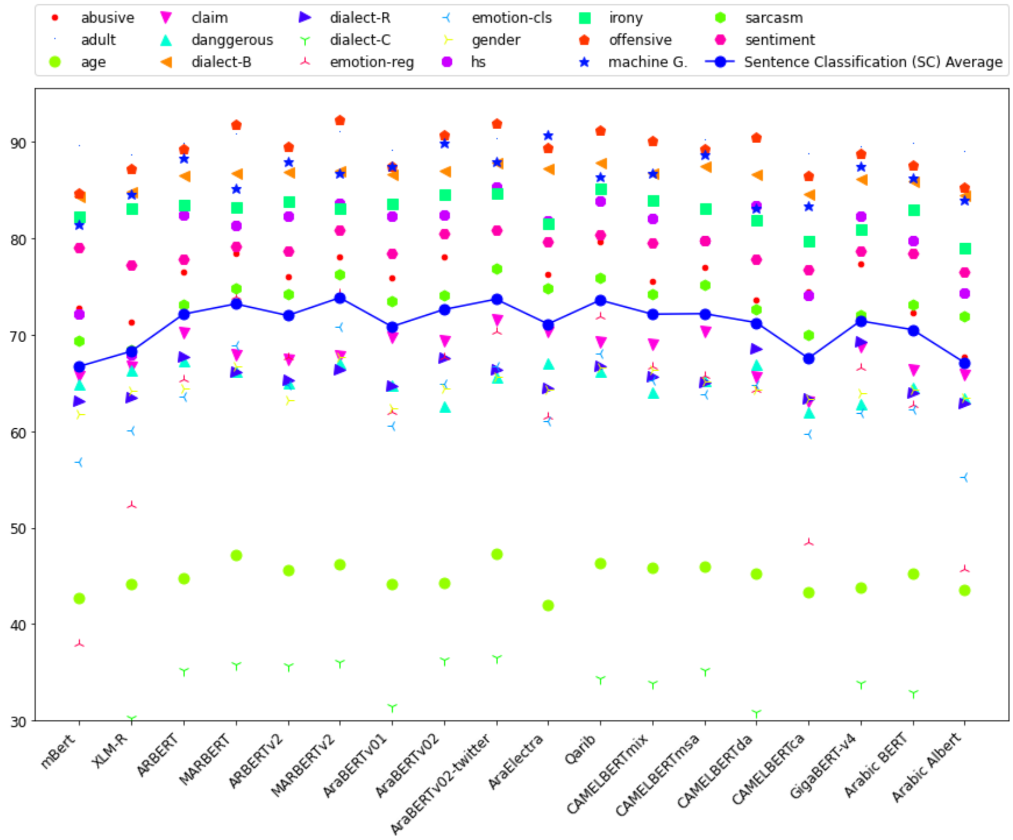
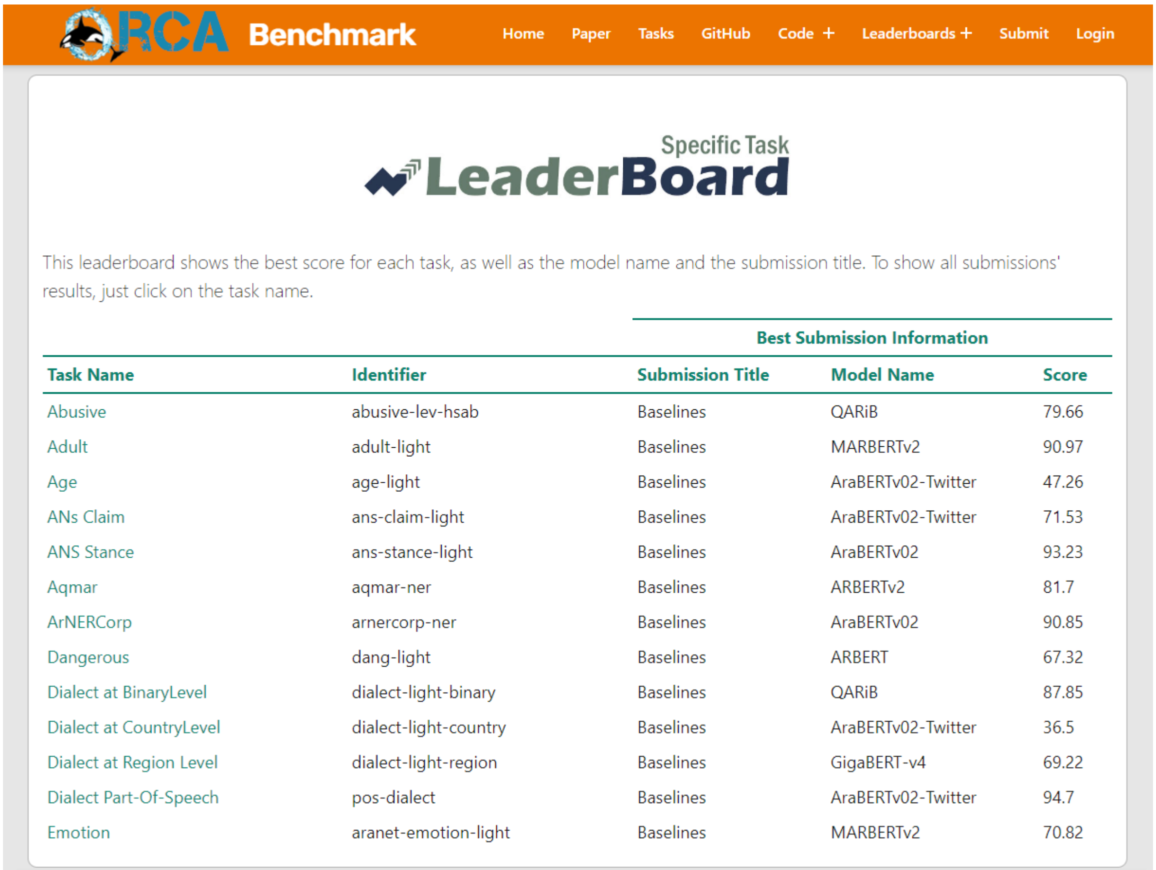


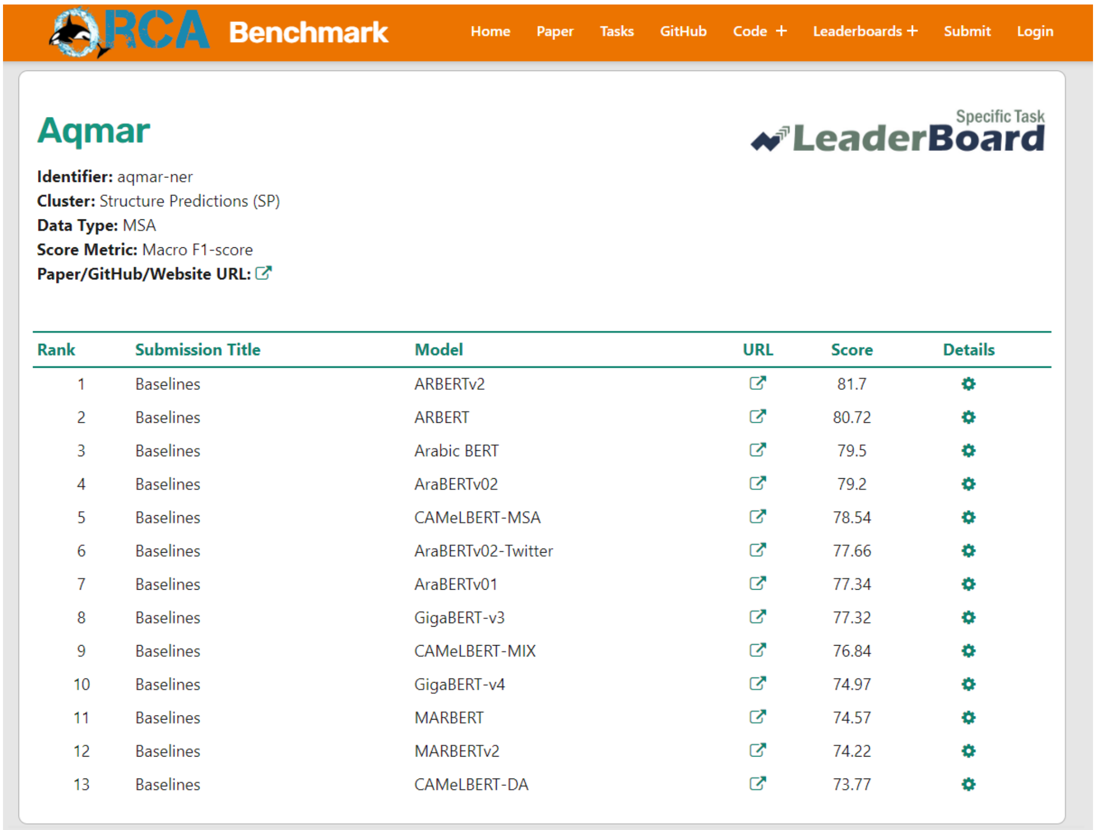
Figure C.3: Models by F_1 score across all tasks in sentence classification cluster.



This leaderboard shows the best score for each task, as well as the model name and the submission title. To show all submissions' results, just click on the task name.

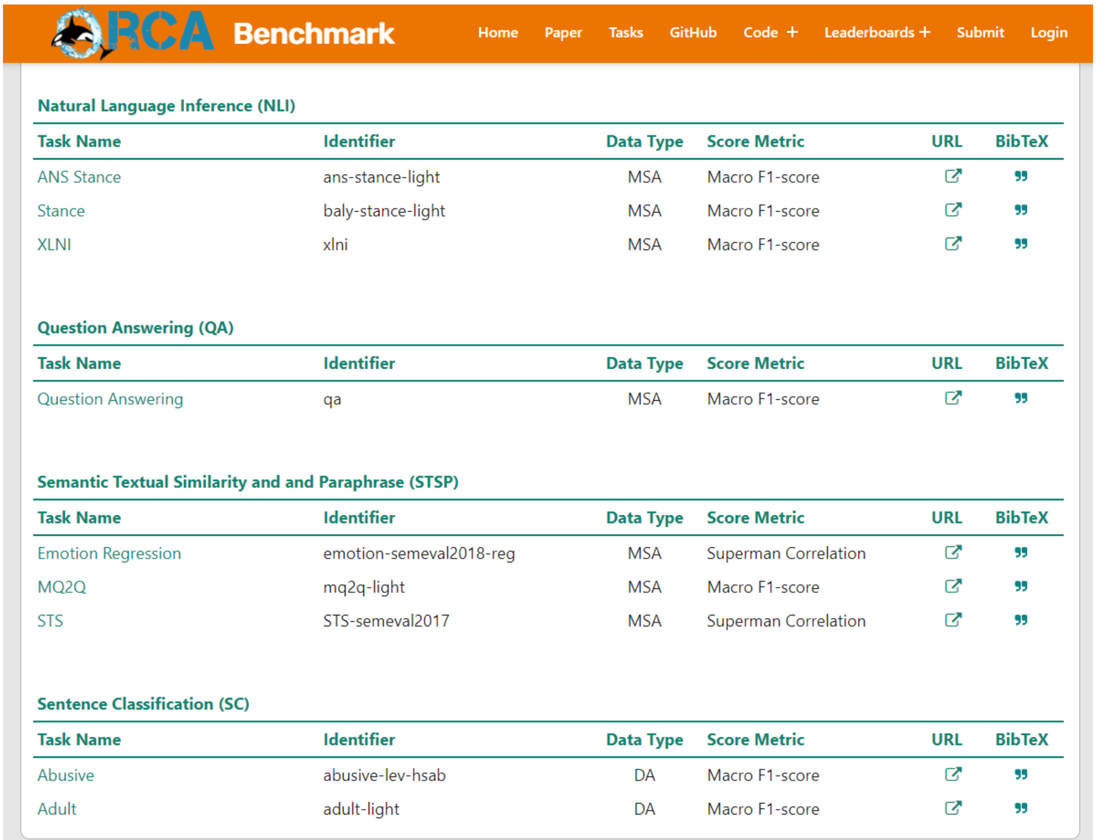
		Best Submission Information		
Task Name	**Identifier**	**Submission Title**	**Model Name**	**Score**
Abusive	abusive-lev-hsab	Baselines	QARiB	79.66
Adult	adult-light	Baselines	MARBERTv2	90.97
Age	age-light	Baselines	AraBERTv02-Twitter	47.26
ANs Claim	ans-claim-light	Baselines	AraBERTv02-Twitter	71.53
ANS Stance	ans-stance-light	Baselines	AraBERTv02	93.23
Aqmar	aqmar-ner	Baselines	ARBERTv2	81.7
ArNERCorp	arnercorp-ner	Baselines	AraBERTv02	90.85
Dangerous	dang-light	Baselines	ARBERT	67.32
Dialect at BinaryLevel	dialect-light-binary	Baselines	QARiB	87.85
Dialect at CountryLevel	dialect-light-country	Baselines	AraBERTv02-Twitter	36.5
Dialect at Region Level	dialect-light-region	Baselines	GigaBERT-v4	69.22
Dialect Part-Of-Speech	pos-dialect	Baselines	AraBERTv02-Twitter	94.7
Emotion	aranet-emotion-light	Baselines	MARBERTv2	70.82

Figure D.1: OCRA leaderboard for example tasks sorted alphabetically.



Rank	Submission Title	Model	URL	Score	Details
1	Baselines	ARBERTv2		81.7	
2	Baselines	ARBERT		80.72	
3	Baselines	Arabic BERT		79.5	
4	Baselines	AraBERTv02		79.2	
5	Baselines	CAMeLBERT-MSA		78.54	
6	Baselines	AraBERTv02-Twitter		77.66	
7	Baselines	AraBERTv01		77.34	
8	Baselines	GigaBERT-v3		77.32	
9	Baselines	CAMeLBERT-MIX		76.84	
10	Baselines	GigaBERT-v4		74.97	
11	Baselines	MARBERT		74.57	
12	Baselines	MARBERTv2		74.22	
13	Baselines	CAMeLBERT-DA		73.77	

Figure D.2: Modularity of OCRA leaderboard allows showing detailed scores by all models for a given task.



Task Name	Identifier	Data Type	Score Metric	URL	BibTeX
ANS Stance	ans-stance-light	MSA	Macro F1-score		
Stance	baly-stance-light	MSA	Macro F1-score		
XLNI	xlni	MSA	Macro F1-score		

Task Name	Identifier	Data Type	Score Metric	URL	BibTeX
Question Answering	qa	MSA	Macro F1-score		

Task Name	Identifier	Data Type	Score Metric	URL	BibTeX
Emotion Regression	emotion-semeval2018-reg	MSA	Superman Correlation		
MQ2Q	mq2q-light	MSA	Macro F1-score		
STS	STS-semeval2017	MSA	Superman Correlation		

Task Name	Identifier	Data Type	Score Metric	URL	BibTeX
Abusive	abusive-lev-hsab	DA	Macro F1-score		
Adult	adult-light	DA	Macro F1-score		

Figure D.3: OCRA leaderboard also provides detailed information about each task cluster and associated tasks, with each task assigned an identifier, language variety, evaluation metric, a link to the dataset website/GitHub/paper and bibliographic information.

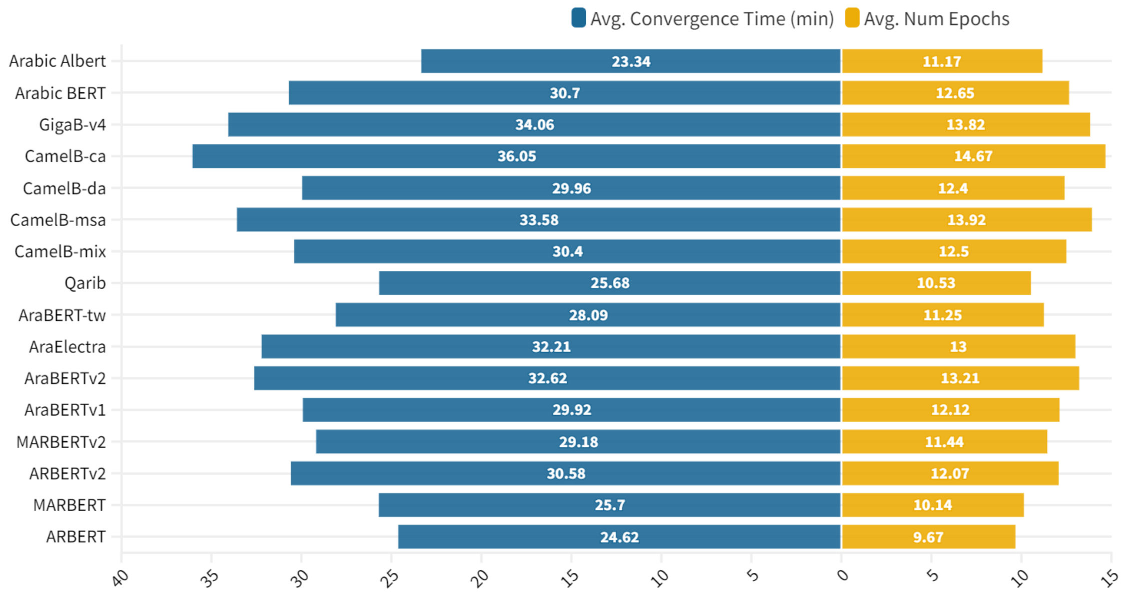


Figure E.1: The average number of epochs (in orange), and time needed to converge (mins, in blue) for all the studied pretrained language models across all ORCA tasks.

Country	Example	Dataset	Label
Egypt	ايطاليا و انجلترا خيبوا توقعاتى بس	Emotion	Happy
	لن يفهمك، فأنت تتحدث عن أمر قطعت فيه آلاف الأميال تفكيرا ولم يمش فيه خطوة	Adult	Not Adult
	الخلفة الوسخة بتجيب لاهلها التهزيق	Sarcasm	Sarcasm
Jordan	ايوون لازم انزله هاد عشان بس افوز اجمع سكور بس بعدك ضعيفه انا ٢٠ الف	Gender	Male
	ومن وراكي يا نشميه يا ام رح تتطبق علينا	Offensive	Not Offensive
	ما احد يربط هالجحش	Abusive	Abusive
KSA	اذا تسوين شى	Dangerous	Dangerous
	وش رايكم تحذفون الاغاني وتحطون ايديكم على قلوبكم !	Emotion	Happy
	اكره الي تصير مشرفة دعم متسابق وتوكل نفسها محامية وتحيك ٤٢ ساعه تراقب التايم	Age	Under 25
Kuwait	للاسف الشبكه تعيسه جدا لا بديراب ولا بالعماريه والعيينه ولا بشقرا	Sentiment	Negative
	نفسي . كل مره احط اليوز و الرقم ولمه ادشه بعد ٥ دقايق القاه طالع	Gender	Male
	هه عااد ماله امان هوو كلش اي علي اسم عيالتكم هه	Emotion	Fear

Table F.1: Randomly picked examples from the dialectal portion of ORCA Train datasets.

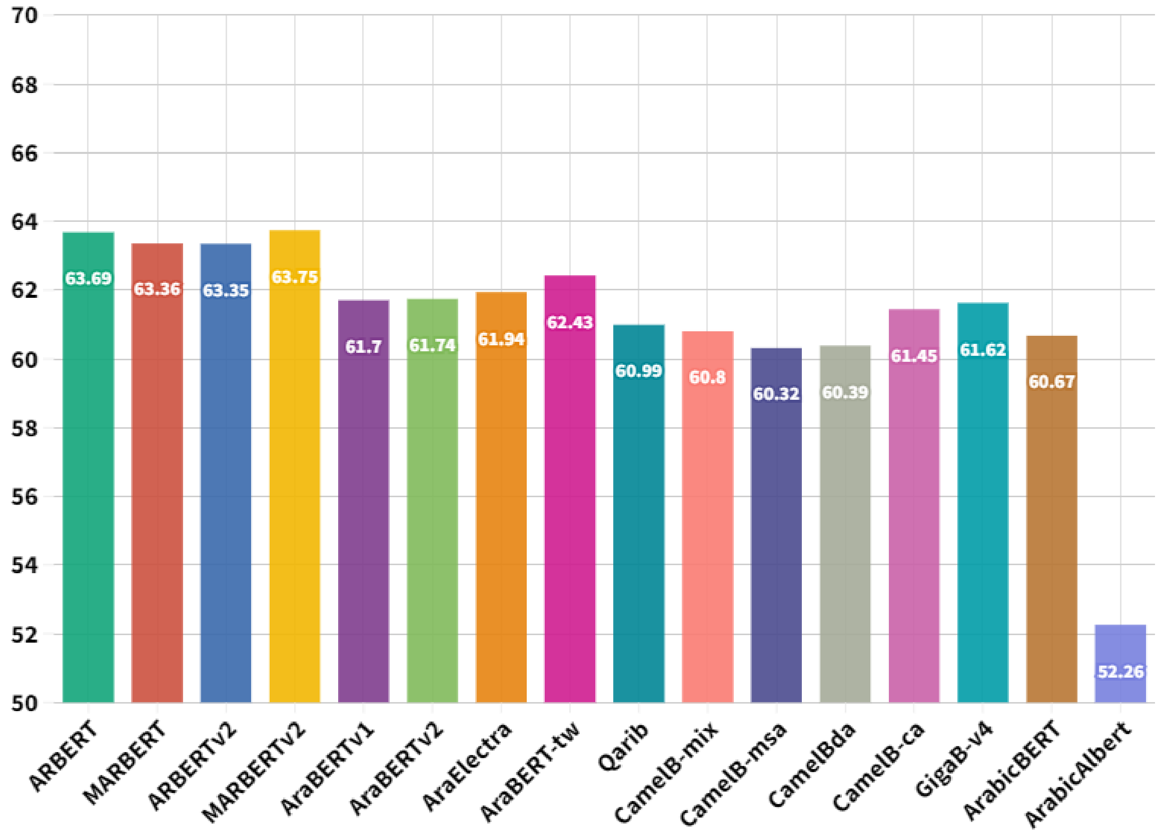


Figure E.2: The time needed in minutes to fine-tune (25 epochs). We compute the average time of three runs across all ORCA tasks.

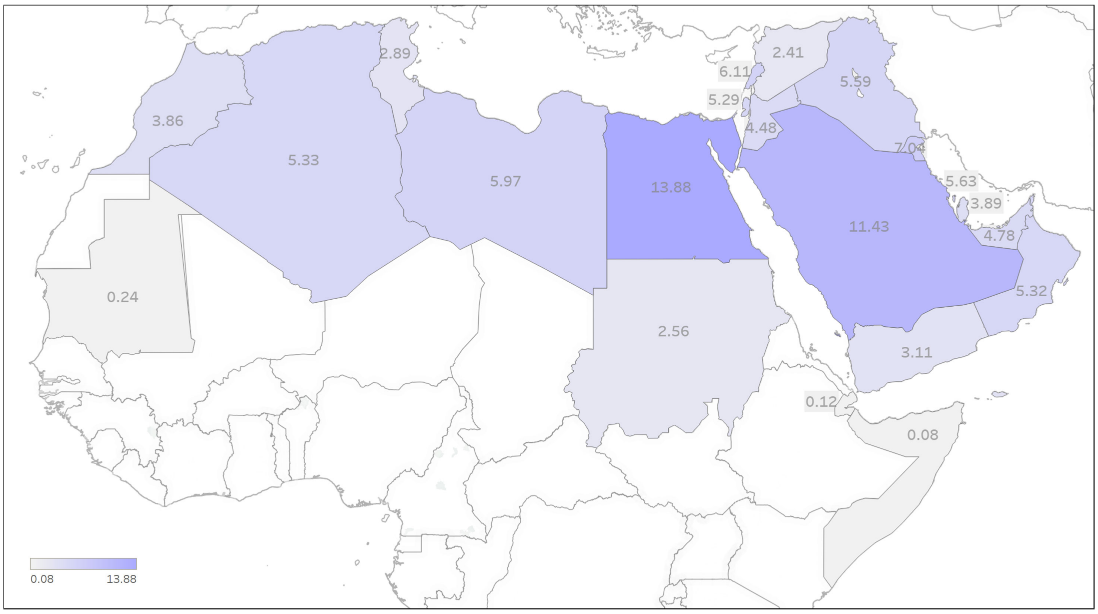


Figure F.1: Predicted country-level distribution, in percentage, in the dialectal portion of ORCA.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
abstract, section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
9

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3, 4

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Since the data belong to the public domain, we do not have serious concerns about privacy or anti-social language beyond what already exists online and is accessible to anyone.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3,4,5

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3,4

C ☑ Did you run computational experiments?

6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
6,5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5,6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.