AraT5: Text-to-Text Transformers for Arabic Language Understanding and Generation

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

Abstract
Transfer learning with a unified Transformer framework (T5) that converts all language problems into a text-to-text format has recently been proposed as a simple, yet effective, transfer learning approach. Although a multilingual version of the T5 model (mT5) has been introduced, it is not clear how well it can fare on non-English tasks involving diverse data. To investigate this question, we apply mT5 on a language with a wide variety of dialects–Arabic. For evaluation, we use an existing benchmark for Arabic language understanding and introduce a new benchmark for Arabic language generation (ARGEN). We also pre-train three powerful Arabic-specific text-to-text Transformer based models and evaluate them on the two benchmarks. Our new models perform significantly better than mT5 and exceed MARBERT, the current state-of-the-art Arabic BERT-based model, on Arabic language understanding. The models also set new SOTA on the generation benchmark. Our new models and are publicly released at https://github.com/UBC-NLP/araT5 and ARGEN will be released through the same repository.

1 Introduction
Due to their remarkable ability to transfer knowledge from unlabeled data to downstream tasks, pre-trained Transformer-based language models have emerged as important components in modern natural language processing (NLP) systems. In particular, the unified framework that converts all text-based language problems into a text-to-text format presented through the T5 model is attractive. In addition to its simplicity, this approach is also effective since it allows knowledge from high-resource tasks to transfer to low-resource tasks without the need for changes in model architecture. Unlike models such as BERT (Devlin et al., 2019), which are based on encoders only, the T5 model is also an encoder-decoder that can be naturally used not only for natural language understanding tasks, but language generation as well.

Although the T5 (Raffel et al., 2019) model originally pre-trained for the English language was recently extended to the multilingual setting as mT5 (Xue et al., 2020), it is not clear how suited it is to individual languages and varieties of these languages. In addition, systematic issues have been discovered in low-resource corpora on which language models have been trained (Caswell et al., 2021). In absence of comparisons with monolingual pre-trained language models that serve the different non-English contexts, it remains unknown if (and if so to what extent) language specific models can fare against multilingual models. In this work, we offer the first comparison of the mT5 model to models dedicated to Arabic. We choose Arabic as our context due to its large set of diverse varieties as well as its wide use on social media. Our work aims at uncovering the extent to which mT5 can serve Arabic’s different varieties. Our work also meets an existing need for pre-trained Transformer-based sequence-to-sequence models. In other words, while several BERT-based models have been pre-trained for Arabic (Antoun et al., 2020; Abdul-Mageed et al., 2021), no such attempts have been made to create sequence-to-sequence models that we know of. Our main contributions are as follows:

1. We pre-train three powerful variants of the text-to-text transformer (T5) model dedicated to Modern Standard Arabic (MSA) and Ara-
bic dialects.

2. We introduce a unified benchmark for ARabic natural language GEneration (ARGEN) composed of four tasks, namely, machine translation, summarization, news title generation and question generation. ARGEN is collected from a total of ten datasets, including two new large datasets proposed in this work.

3. We fine-tune our new models on a large benchmark of Arabic language understanding tasks, ARLUE (Abdul-Mageed et al., 2021), in both single- and multi-task settings and establish new SOTA on the majority of tasks.

The rest of the paper is organized as follows: Section 2 describes our Arabic pre-tained models. In Section 3, we introduce ARGEN, our new natural language generation benchmark. Section 4 briefly describes ARLUE, the natural language understanding benchmark. We evaluate our models on both benchmarks in Section 5. In Section 6, we provide an overview of related work. We conclude in Section 7. We now introduce our new pre-trained models.

2 Our Models

2.1 Training Data

**MSA Data.** We use 70GB of MSA text (7.1B tokens) from the following sources: AraNews (Nagoudi et al., 2020), El-Khair El-Khair (2016), Gigaword,\(^1\) OSCAR (Suárez et al., 2019), OSIAN (Zeroual et al., 2019), Wikipedia Arabic, and Hindawi Books.\(^2\)

**Twitter Data.** We randomly sample 1.5B Arabic tweets from a large in-house dataset of about 10B tweets. We use string matching to only include tweets with at least 3 Arabic words, regardless whether the tweet has non-Arabic string or not. The dataset makes up 178GB of text (21B tokens). More information about each of these datasets is in Table 1.

**Data Distribution.** In order to analyze MSA-Dialect distribution in our Twitter training data, we run an in-house binary MSA-dialect classifier on a random sample of 100M tweets. We find that the Twitter data involve 35.74% dialect tweets and 64.26% predicted as MSA.

2.2 Pre-Processing and Vocabulary

For all our models, we remove diacritics and replace URLs, user mentions, and hashtags with the tokens URL, USER, and HASHTAG respectively. The T5 (Raffel et al., 2019) model is based on a vocabulary acquired by the SentencePiece library\(^3\) using English, French, German, and Romanian web pages from “Colossal Clean Crawled Corpus” (or C4 for short). We use a similar procedure to create our languages model vocabulary. Namely, we use SentencePiece (Kudo, 2018) to encode text as WordPiece tokens (Sennrich et al., 2016) with a vocabulary size of 110K WordPieces. In order to allow further pre-training on data from additional languages, we extract our vocabulary as follows: 70M MSA sentences, 200M Arabic twitter data, 15M sentences from Wikipedia English, and 5M sentences from Wikipedia of ten other languages (Spanish, French, Italian, Portuguese, Russian, Deutsch, Turkish, Greek, Bulgaria, and Czech).\(^4\)

2.3 AraT5

We leverage our unlabeled MSA and Twitter data (see Section 2.1), to pre-train three models: AraT5\(_{\text{MSA}}\) on MSA data, AraT5\(_{\text{TW}}\) on twitter data, and AraT5 on both MSA and twitter data using the T5\(_{\text{Base}}\) encoder-decoder architecture (Raffel et al., 2019). Each of the encoder and decoder is similar in size and configuration to BERT\(_{\text{Base}}\) (Devlin et al., 2019), with 12 layers each with 12 attention heads, and 768 hidden units. In total, this results in a model with about 220 million parameters.\(^5\)

Raffel et al. (2019) pre-trained T5\(_{\text{Base}}\) on a mixture of unsupervised and supervised single tasks (or multi-task), where each task is converted into a text-to-text format. In fact, they propose a denoising objective that does not require labels. The main idea is feeding the model with masked (corrupted) versions of the original sentence, and training it to reconstruct the original sentence. Inspired by BERT’s objective (Devlin et al., 2019), the denoising objective (Raffel et al., 2019) works by randomly sampling and dropping out 15% of tokens in the input sequence. All consecutive spans of dropped-out tokens are then replaced by a single

---

1\(^{https://catalog.ldc.upenn.edu/LDC2009T30.}\)
2\(^{https://www.hindawi.org/books.}\)
3\(^{https://github.com/google/sentencepiece.}\)
4\(^{The MSA and twitter data are extracted from our training data presented in Section 2.1.}\)
5\(^{The output dimensionality is d_{ff} = 3,072 and inner dimensionality of d_{kv} = 64.}\)
We refer to the Arabic-to-English MT benchmark we consolidated to create ARGEN Arabic language generation. ARGEN includes ten (with QG). We describe each of these next.

In order to evaluate our pre-trained language models, we use a learning rate of 0.01, a batch size of 128 sequences, and a maximum sequence length of 512, except for AraT5_TW where the maximum sequence is 128. We use the original implementation of T5 in the TensorFlow framework. We train the models for 1M steps. Training took ∼80 days on 1 on Google Cloud TPU with 8 cores (v3.8) from TensorFlow Research Cloud (TFRC).

### 3 ARGEN: A Benchmark for Arabic Language Generation

In order to evaluate our pre-trained language models, we introduce ARGEN—a new benchmark for Arabic language generation. ARGEN includes ten datasets and covers four tasks, namely, machine translation (MT), text summarization (TS), news title generation (NGT), and question generation (QG). We describe each of these next.

#### 3.1 Machine Translation

We refer to the Arabic-to-English MT benchmark component of ARGEN as ARGEN_MT. To build this benchmark, we consider six different datasets (with 37 test splits) derived from MSA and Arabic dialects. Details about the datasets and their splits are in Table 2. We now briefly describe the datasets we consolidated to create ARGEN_MT.

| Source          | Size  | #Tokens |
|-----------------|-------|---------|
| AraNews         | 8.6GB | 847.8M  |
| Books           | 650MB | 72.5M   |
| El-Khair        | 16GB  | 1.6B    |
| Gigawords       | 10GB  | 1.1B    |
| OSIAN           | 2.8GB | 292.6M  |
| OSCAR-MSA       | 31GB  | 3.4B    |
| OSCAR-Egyptian  | 32MB  | 3.8M    |
| Wiki            | 1.4GB | 156.5M  |
| MSA-Total       | 70GB  | 7.1B    |
| Twitter (1.5B)  | 178GB | 21.9B   |
| ALL             | 248GB | 29.0B   |

Table 1: The MSA and Twitter resources used to pre-train AraT5_MSA, AraT5_TW, and AraT5.

For all our pre-training models, we use a learning rate of 0.01, a batch size of 128 sequences, and a maximum sequence length of 512, except for AraT5_TW where the maximum sequence is 128. We use the original implementation of T5 in the TensorFlow framework. We train the models for 1M steps. Training took ∼80 days on 1 on Google Cloud TPU with 8 cores (v3.8) from TensorFlow Research Cloud (TFRC).

### 3.2 Text Summarization

In order to build our text summarization benchmark component (ARGEN_TS), we use the two following Arabic text summarization datasets:

#### Essex Arabic Summaries Corpus (EASC).

EASC is an Arabic natural language resources proposed by El-Haj et al. (2010). It contains 153 Arabic articles, each with 5 human-generated extrac-
### ARGENMT

For MADAR and Bible datasets we use the same splits used by Sajjad et al. (2020).

**MADAR I**: corpus consists of 2k sentences (Test) of 21 city-level dialects each. **MADAR II**: 12k (5.5k sentences for Dev, and 6.5k for Test sets) sentences each of five other city-level dialects and of MSA.

**Bible I**: 600 sentences each as Dev and Test sets for Moroccan, Tunisian, and MSA. **Bible II**: Two Dev and Test splits (600 sentences each) are used for Bible MSA. **IWSLT 2016**: Three Test sets are used (TED 2015, TED 2016 and QED 2016 (Cettolo et al., 2016)) with the same Dev (TED talks 2014) (Cettolo et al., 2014).

### WikiLingua

An abstractive summarization dataset in 18 languages (including Arabic) proposed by Faisal Ladhak and McKeown (2020). It contains 770K articles and their summaries from WikiHow—a human written resource of how-to guides on a diverse set of domains. Each step in a WikiHow article, contains a one sentence summary followed by a paragraph detailing the instruction. The paragraphs and summaries are combined from all the steps in order to create cross-lingual article-summary pairs. The Arabic part includes summaries for 29,229 articles. We split this data into 80% Train (23.4K), 10% DEV (2.9K), and 10% TEST (2.9K).

### News Title Generation

The purpose of the News Title Generation (NGT) task is to generate proper titles for given news article (Liang et al., 2020). We introduce the NGT task as a new task for Arabic language generation. Given an article, a title generation model needs to output a short grammatical sequence of words suited to the article content. For NGT, we create a novel dataset from an existing news dataset. We refer to this new dataset as **ARGENNTG**. We extract 120K articles along with their titles from AraNews (Nagoudi et al., 2020). We only include titles with at least three words in this dataset. We split ARGENNTG data into 80% (93.3K), 10% (11.7K) and 10% (11.7K) for training, development, and test respectively. Details about ARGENNTG are in Table B.1 (Appendix). A sample of a news article from our TEST split and example titles generated by our models are provided in Table B.6 (Appendix).

### Question Generation

In the Question Generation (QG) task, a question is produced for a passage (Gehrmann et al., 2021). Given the absence of an Arabic QG dataset, we follow Kriangchaivech and Wangperawong (2019) in creating a new Arabic QG dataset (**ARGENQG**) using a publicly-available Arabic question answering (QA) resource. Kriangchaivech and Wangperawong (2019) train a model to generate simple questions relevant to passages and answers extracted from SQuAD (Rajpurkar et al., 2016). Similarly, we build ARGENQG by extracting 96K (passage, question) pairs from the Arabic QA dataset ARLUEQA16 (see Section 4). More particularly, we construct ARLUEQA from the follow-
ing QA datasets: ARCD (Mozannar et al., 2019) and the three human translated Arabic test sets of the XTREME benchmark (Hu et al., 2020): MLQA (Lewis et al., 2019), XQuAD (Artetxe et al., 2020), and TyDi QA (Artetxe et al., 2020).

4 ARLUE: A Benchmark for Arabic Language Understanding

Recently, Abdul-Mageed et al. (2021) introduced ARLUE, a natural language understanding benchmark for Arabic. ARLUE is composed of 42 publicly available datasets, making it the largest and most diverse Arabic NLP benchmark. ARLUE is arranged into the six cluster tasks of sentiment analysis (SA), social meaning (SM), topic classification (TC), dialect identification (DI), named entity recognition (NER), and question answering (QA). We methodically evaluate each cluster task, ultimately reporting a single ARLUE score following Abdul-Mageed et al. (2021). Table A.1 shows a summary of the ARLUE benchmark (Appendix).

We briefly describe ARLUE tasks next. ARLUESenti. To construct this task cluster Abdul-Mageed et al. (2021) merged 17 MSA and DA publicly available datasets. ARLUESM. ARLUESM refers to eight social meaning datasets covering prediction of age, dangerous speech, emotion, gender, hate speech, irony, offensive language, and sarcasm. used in this benchmark. We will follow Abdul-Mageed et al. (2021) in not merging the social meaning datasets, but rather report performance on each individual dataset as well as average performance across all tasks as part of an overall ARLUE score. ARLUETopic. This benchmark component is a concatenation of three topic classification datasets: Arabic News Text (ANT) (Chouigui et al., 2017), Khaleej (Abbas et al., 2011), and OSAC (Saad and Ashour, 2010). ARLUEDia. Five datasets are used for dialect classification. These are AOC Zaidan and Callison-Burch (2014), ArSarcasmDia (Farha and Magdy, 2020), MADAR (sub-task 2) (Bouamor et al., 2019), NADI-2020 (Abdul-Mageed et al., 2020a), and QADI (Abdelali et al., 2020). ARLUEDia involve three categories, namely, ARLUEDia-B for MSA-dialect classification (binary), ARLUEDia-R, and ARLUEDia-C for the region and country level classification into four classes (region), and 21 classes (country) respectively.

ARLUQA. Four Arabic and multilingual QA datasets are concatenated to build ARLUQA: ARCD (Mozannar et al., 2019) MLQA (Lewis et al., 2019), XQuAD (Artetxe et al., 2020), and TyDi QA (Artetxe et al., 2020).

5 Evaluation

5.1 Arabic LU Evaluation

Baselines. For comparison, we fine-tune a number of models on the same training data as our new models. These include the multilingual sequence-to-sequence model mT5 (Xue et al., 2020), and the powerful Arabic-specific BERT-based model MARBERT (Abdul-Mageed et al., 2021). We note that MARBERT achieves the SOTA across the majority of 6 cluster tasks of ARLUE, with the highest ARLUE score.

Settings and Evaluation. We evaluate our models on the language understanding benchmark, ARLUE, under two settings: (i) single task learning and (ii) multi-task learning. We present results on all the task clusters included in ARLUE except for NER which is a token-level task that is not straightforward with the text-to-text set up we adopt. Table 3 shows our evaluation results using the relevant metric for each task.

Abdul-Mageed et al. (2021) introduced ARLUE score, a metric used to score pre-trained language model performance on multiple datasets. ARLUE score is a simply macro-average of the different scores across all task clusters, where each task is weighted equally following (Wang et al., 2018). We compute the ARLUE score (i.e., overall macro-average) for each of our three models (i.e., AraT5\textsubscript{MSA}, AraT5\textsubscript{TW}, and AraT5) and the baseline (mT5).

Single Task. We fine-tune our three models and mT5 individually on each of the six tasks of ARLUE. We typically (i.e., in all our experiments) identify the best checkpoint for each model on the development set, and report its performance on both development and test data. As Table 3

17Detailed information about ARLUE tasks and datasets are provided in A.1 (Appendix).
18We note that the classes were straightforwardly merged without modifying any class labels.
19All corresponding splits from the different QA datasets are merged.
20MARBERT outperform both multilingual encoder-only Transformers mBERT, XLM-R\textsubscript{base}, XLM-R\textsubscript{large}, and Arabic-specific BERT-based AraBERT (Antoun et al., 2020), ARBERT (Abdul-Mageed et al., 2021).
shows, our AraT5 model achieves the highest ARLUE score (77.52), followed by AraT5_{MSA} (77.50) and AraT5_{TW} (75.33). We note that all our models outperform mT5 and the MARBERT (SOTA) by \( \sim +2.74 \) and \( \sim +1 \) ARLUE score points, respectively.

**Multitask.** We also investigate multitask learning (Caruana, 1997; Ruder, 2017) with our AraT5 models. This approach consists of training the model on multiple tasks simultaneously (i.e., the model and its parameters are shared across all tasks) in order to eventually improve performance on each individual task. In our case, we fine-tune our models on many tasks at the same time using: (i) The three dialect datasets: ARLUE_{Dia-B}, ARLUE_{Dia-R}, and ARLUE_{Dia-C} and (ii) the social meaning datasets of ARLUE_{SM}. Table 4 and Table 5 show the results of multi-task experiments for dialect settings and social meaning, respectively. Our results show that multi-task training outperforms single task models in the majority of the dialects experiments (n=7 out of 9 experiments, 77.78% of the tasks) and half of the social meaning tasks (n=18 out of 36 experiments, 50% of the tasks). These results are promising, and hence we plan to further investigate multi-task learning with our new models in the future.

### 5.2 Arabic LG Evaluation

**Baselines.** For all tasks of ARGEN, we compare our models to models fine-tuned with mT5 using the same data. In addition, for the machine translation task we compare to a vanilla sequence-to-sequence (S2S) Transformer (Vaswani et al., 2017) trained from scratch as implemented in Fairseq (Ott et al., 2019).

**Machine Translation.** We fine-tune our three models, mT5 (baseline I), on 2M MSA-English parallel sentences extracted from the Open Parallel Corpus (OPUS) (Tiedemann, 2012). We also train two S2S Transformers models (baseline II) on 2M (S2S2M) and 10M (S2S10M) MSA-English\(^{21}\) parallel sentences extracted from the same source (Tiedemann, 2012).

For all models and baselines, we identify the best model on TED14 (Cettolo et al., 2014) (DEV). Then, we test on TED15, TED16 and QED16 test splits (Cettolo et al., 2016). Results of ARGEN_{MT} are reported in Table 6. As Table 6 shows, our models achieve best BLEU score in 32 out of the 37 tests splits. AraT5_{MSA} archives best results in 25 of these test splits, outperforming all the baselines (S2S2M), (S2S10M), and mT5 with +5.67, +4.99, and +0.44 BLEU points. These results are striking since our language models are pre-trained on Arabic data only (although they include English vocabulary). In other words, even under this current zero-shot setting, the models perform very well. In addition, our AraT5 model outperforms even the S2M model trained with 5X more data.

For completeness, we also provide the current SOTA on each of our datasets. We do not compare our results to SOTA since these are acquired by models fine-tuned on much larger datasets than ours. For example, Sajjad et al. (2020) exploit \( \sim 42M \) parralel sentences to train their models. To limit GPU needs during our experiments, especially given the long fine-tuning process typical of T5 models, we do not fine-tune the models on the full amounts of available parallel data. However, in the future we plan to compare our models under the full data setting.

**Text Summarization.** For both ARGEN_{ST} datasets, we fine-tune on the Train split of WikiLingua (Faisal Ladhak and McKeown, 2020). Then, we identify the best checkpoint for each model on Dev of the same data, and report its performance on the Test split of WikiLingua (Faisal Ladhak and McKeown, 2020) and all EASC data El-Haj et al. (2010) (i.e., we consider all EASC as test set). We report results of all our models under different settings in ROUGE scores (Lin, 2004). As Table 7, AraT5\_{Tweet} acquires best results on WikiLingua data, while mT5 outperforms us on EASC (we hypothesize since EASC is older data that is likely part of the mC4 on which mT5 was pre-trained). On both datasets, we outperform previous SOTA.

**News Title and Question Generation.** For both tasks, we fine-tune all our models on the Train split of our new datasets ARGEN_{NTG} and ARGEN_{QG} respectively. We report results of all our models and mT5 in BLEU scores (Papineni et al., 2002). Table 7 shows that all our models outperform mT5 on both tasks. We also observe that AraT5_{MSA} and AraT5 achieve the best BLEU score on ARGEN_{NTG} (20.61%) and ARGEN_{QG} (16.99%) respectively.

\(^{21}\)We use the same 2M MSA-English used to fine-tune our models and mT5.
Wikipedia XLM-R (Conneau et al., 2019) is also a (including Arabic, with ∼

Table 4: Performance of our models on ARLUE Di-

Table 3: Performance of our models on ARLUE TEST datasets (Acc / F

6 Related Work
6.1 Multilingual LMs
mBERT mBERT is the multilingual version of BERT (Devlin et al., 2019) which is a multi-layer bidirectional encoder representations from Transformers trained with a masked language modeling objective. BERT models were trained on English Wikipedia 22 and BooksCorpus (Zhu et al., 2015). mBERT is trained on Wikipedia for 104 languages (including Arabic, with ∼153M tokens) with 12 layers, 12 attention heads, 768 hidden units each and 110M parameters.

XLM-R (Conneau et al., 2019) is also a Transformer-based multilingual masked language model pre-trained on more than 2 TB of filtered CommonCrawl data in 100 languages, including Arabic (2.9B tokens). XLM-R model uses the same masking objective as BERT, but not the next sentence prediction.

mT5 (Xue et al., 2020) is the multilingual version of Text-to-Text Transfer Transformer model (T5) (Raffel et al., 2019). The T5 model architecture is essentially an encoder-decoder Transformer (Vaswani et al., 2017) similar in configuration and size to a BERTBase (Devlin et al., 2019). As explained earlier, the basic idea behind T5 is to treat every text-based language task as a “text-to-text” problem. This greatly facilitates especially multi-task learning where all tasks can be seamlessly learned without architectural model changes. The mT5 model is trained on the “Multilingual Colossal Clean Crawl Corpus” (or mC4 for short), which is ∼26,767B for 101 languages generated from 71 Common Crawl dumps.

Table 5: Performance of our models on ARLUE social meaning (SM) Test datasets on single- and multi-tasks setting (Acc / F1). S: Single Task. M: Multi-task.
Table 6: ARGENMT datasets. S2S: Sequence-to-Sequence Transformer models trained from scratch on 2M and 10M parallel sentences. SOTA: ¹ Sajjad et al. (2020), ² Durrani et al. (2017), † Junczys-Dowmunt et al. (2016).

Table 7: Performance of our summarization models on Test. We consider mT5 as SOTA for WikiLin, and Alami et al. (2021) (ROUGE1=59.17) for EASC.

6.2 Arabic LMs

AraBERT (Antoun et al., 2020) is an Arabic pre-trained language model based on the BERTBase architecture (Devlin et al., 2019) configuration. AraBERT (Antoun et al., 2020).

ARBERT (Abdul-Mageed et al., 2021) is a large scale pre-training masked language model focused on Modern Standard Arabic (MSA). ARBERT is trained on a collection of Arabic datasets of 61GB of text (6.2B tokens). The same architecture as
BERT\textsubscript{Base} was used, but with a larger vocabulary of 100K WordPieces, making ~163M parameters. **MARBERT** (Abdul-Mageed et al., 2021) is similar to ARBERT but was trained with dialectal Arabic and MSA data using 1B tweets (128GB of text, 15.6B tokens). For training, Abdul-Mageed et al. (2021) use only tweets with at least 3 Arabic words and exclude the next sentence prediction objective.

## 7 Conclusion

We introduced three powerful Arabic-specific text-to-text Transformer models trained on large MSA and/or Arabic dialectal data. We also introduced ARGEN, a unified benchmark for Arabic Natural Language Generation composed of four tasks collected from a total of ten datasets. In addition to ARGEN, we evaluate our models on the recently introduced Arabic language understanding benchmark ARLUE. Our models establish new SOTA on ARLUE and on several ARGEN datasets, outperforming mT5 even in the machine translation task where they have not explicitly seen English data during pre-training (zero-shot). Our models involve vocabulary from 11 languages other than Arabic and hence can easily be further pre-trained on these languages. We will make our models and new generation benchmark publicly available upon acceptance.

## References

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

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The amara corpus: Building parallel language resources for the educational domain. In *LREC*, volume 14, pages 1044–1054.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2020. Arabic Dialect Identification in the Wild. *arXiv preprint arXiv:2005.06557*.

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

Muhammad Abdul-Mageed, Abdellahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In *Proceedings of the ACL-IJCNLP 2021 Main Conference*. 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 Fourth Arabic Natural Language Processing Workshop*.

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

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

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

Nabil Alami, Mohammed Meknassi, Noureddine Ennahmah, Yassine El Adlouni, and Ouafae Ammor. 2021. Unsupervised neural networks for automatic arabic text summarization using document clustering and topic modeling. *Expert Systems with Applications*, 172:114652.

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

Ali Alshehri, Muhammad Abdul-Mageed, et al. 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.

Mohamed Aly and Amir Atiya. 2013. Labr: A large scale arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 494–498.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabort: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.
Amina Chouigui, Oussama Ben Khiroun, and Bilel Mauro Cettolo, Niehues Jan, Stüker Sebastian, Luisa Machine

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa

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

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Osanna Obeid, Salam Khalfi, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).

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 (WANLP19), Florence, Italy.

Rich Caruana. 1997. Multitask learning. Machine learning, 28(1):41–75.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Alahasera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. arXiv preprint arXiv:2103.12028.

Mauro Cettolo, Niehues Jan, Stüker Sebastian, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2016. The iwslt 2016 evaluation campaign. In International Workshop on Spoken Language Translation.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th iwslt evaluation campaign, iwslt 2014. In Proceedings of the International Workshop on Spoken Language Translation, Hanoi, Vietnam, vol. 57.

Amina Chouigui, Ouussama Ben Khiroun, and Bilel Elayeb. 2017. Ant corpus: an arabic news text collection for textual classification. In 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), pages 135–142. IEEE.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

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.

Nadir Durranii, Fahim Dalvi, Hassan Sajjad, and Stephan Vogel. 2017. Qc machine translation systems for iwslt 16. arXiv preprint arXiv:1701.03924.

Mahmoud El-Hajj, Udo Kruschwitz, and Chris Fox. 2010. Using mechanical turk to create a corpus of arabic summaries.

Ibrahim Abu El-Khair. 2016. 1.5 billion words arabic corpus. arXiv preprint arXiv:1611.04033.

AbdelRahim Elmadany, Hamdy Muharak, and Walid Magdy. 2018. ArSAS: An Arabic speech-act and sentiment corpus of tweets. OSACT; 3:20.

Mohamed Elmahdy, Mark Hasegawa-Johnson, and Eiman Mustafawi. 2014. Development of a tv broadcasts speech recognition system for qatari arabic. In LREC, pages 3057–3061.

Ashraf Elnagar, Yasmin S Khalifa, and Anas Einea. 2018. Hotel Arabic-reviews dataset construction for sentiment analysis applications. In Intelligent Natural Language Processing: Trends and Applications, pages 35–52. Springer.

Claire Cardie Faisal Ladhak, Esin Durmus and Kathleen McKeown. 2020. Wikilingua: A new benchmark dataset for multilingual abstractive summarization. In Findings of EMNLP, 2020.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset. In Proceedings of the 9th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pages 32–39.

Ibrahim Abu Farha and Walid Magdy. 2021. Benchmarking transformer-based language models for arabic sentiment and sarcasm detection. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 21–31.

Sebastian Gehrmann, Tosin Adewumi, Karmany Aggarwal, Pawan Sasanka Ammanamanchi, Arem Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. arXiv preprint arXiv:2102.01672.

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.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In Proceedings of the 37th International
Karima Meftouh, Salima Harrat, Salma Jamoussi, Hussein Mozannar, Karl El Hajal, Elie Maamary, and Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out.

Marcin Junczys-Dowmont, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. \textit{arXiv preprint arXiv:1610.01108}.

Kettip Kriangchaivech and Artit Wangperawong. 2019. Question generation by transformers. \textit{arXiv preprint arXiv:1909.05017}.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple sub-word candidates. \textit{arXiv preprint arXiv:1804.10959}.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. \textit{arXiv preprint arXiv:1910.07475}.

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 dataset for cross-lingual pre-training, understanding and generation. \textit{arXiv preprint arXiv:2004.01401}.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. \textit{Text Summarization Branches Out}.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic Dialect corpus. In \textit{Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation}, pages 26–34, Shanghai, China.

Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem Hajj. 2019. Neural arabic question answering. \textit{arXiv preprint arXiv:1906.05394}.

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

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

El Moatez Billah Nagoudi, Abdellrahim Elmadany, Muhammad Abdul-Mageed, Tariq Alhindi, and Hasan Cavusoglu. 2020. Machine generation and detection of arabic manipulated and fake news. In \textit{Proceedings of the Fifth Arabic Natural Language Processing Workshop}, pages 69–84.

Ossama Obeid, Nasser Zalmut, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. Camel tools: An open source python toolkit for arabic natural language processing. In \textit{Proceedings of The 12th Language Resources and Evaluation Conference}, pages 7022–7032.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. \textit{arXiv preprint arXiv:1904.01038}.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In \textit{Proceedings of the 40th annual meeting on association for computational linguistics}, pages 311–318. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. \textit{arXiv preprint arXiv:1910.10683}.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. \textit{arXiv preprint arXiv:1606.05250}.

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

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. \textit{arXiv preprint arXiv:1706.05098}.

Motaz K Saad and Wesam M Ashour. 2010. Osac: Open source arabic corpora. \textit{Osac: Open source arabic corpora}, 10.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. Arabench: Benchmarking dialectal arabic-english machine translation. In \textit{Proceedings of the 28th International Conference on Computational Linguistics}, pages 5094–5107.

Mohammad Salameh, Saif M Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on arabic social media posts. In \textit{Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies}, pages 767–777.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In \textit{Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)}, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. 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. 2012:2214–2218.

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

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934.

Wajdi Zaghouani and Anis Charfi. 2018. Arap-Tweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).

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

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stal- lard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies, pages 49–59.

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.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhut- dinov, 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 (ICCV).

Michał Ziemska, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pages 3530–3534.
Appendices

A ARLUE

In this section we describe briefly ARLUE benchmark Abdul-Mageed et al. (2021) and report the evaluation results in validation dataset for single and multi-tasks experiment.

A.1 datasets

As shown in Table A.1 ARLUE benchmark consist of five clusters as the following.

- **ARLUE\textsubscript{Senti}**. Consist 17 sentiment datasets (MSA and DA) collected from: AJGT (Alo-mari et al., 2017), AraNET\textsubscript{Sent} (Abdul-Mageed et al., 2020b), AraSenTi-Tweet (Ali-Twairesh et al., 2017), ArSarcasm\textsubscript{Sent} (Farha and Magdy, 2020), ArSAS (Elmadany et al., 2018), ArSenD-Lev (Baly et al., 2019), ASTD (Nabil et al., 2015), AWATIF (Abdul-Mageed and Diab, 2012), BBNS & SYTS (Salameh et al., 2015), CAMelsent (Obeid et al., 2020), HARD (Elmagar et al., 2018), LABR (Aly and Atiya, 2013), Twitter\textsubscript{Abullah} (Abdulla et al., 2013), Twitter\textsubscript{Saad}\textsuperscript{23} and SemEval-2017 (Rosenthal et al., 2017).

- **ARLUE\textsubscript{SM}**. They proposed eight tasks for social meaning collected from: Age and Gender (Zaghouani and Charfi, 2018), Dangerous Speech (Alshehri et al., 2020), Offensive Language and Hate Speech Mubarak et al. (2020), Emotion (Abdul-Mageed et al., 2020b), Irony (Ghanem et al., 2019), and Sarcasm (Farha and Magdy, 2020).

- **ARLUE\textsubscript{Topic}**. They merged three datasets Arabic News Text (ANT) (Chougui et al., 2017), Khaleej (Abbas et al., 2011), and OSAC (Saad and Ashour, 2010).

- **ARLUE\textsubscript{PA}**. Collected from: Arabic Online Commentary (AOC) (Zaidan and Callison-Burch, 2014), ArSarcasm\textsubscript{Dia} (Farha and Magdy, 2020),\textsuperscript{24} MADADAR (sub-task 2) (Bouamor et al., 2019), NADI-2020 (Abdul-Mageed et al., 2020a), and QADI (Abdelali et al., 2020).

\textsuperscript{23}www.kaggle.com/mksaad/arabic-sentiment-twitter.

\textsuperscript{24}ArSarcasm\textsubscript{Diial} carries regional dialect labels.

- **ARLUE\textsubscript{QA}**. They use ARCD (Mozannar et al., 2019) and the 3 human translated Arabic test sections of the XTREME benchmark (Hu et al., 2020): MLQA (Lewis et al., 2019), XQUAD (Artetxe et al., 2020), and TyDi QA (Artetxe et al., 2020).

### Table A.1: ARLUE categories across the different data splits.

| Category | Task | Train | DEVT | TEST |
|----------|------|-------|------|------|
| ARLUE\textsubscript{Sent} | 17 | SA | 190.5K | 6.5K | 44.2K |
| ARLUE\textsubscript{SM} | 8 | SM | 1.51M | 162.5K | 166.1K |
| ARLUE\textsubscript{Topic} | 5 | TC | 47.5K | 5.9K | 5.9K |
| ARLUE\textsubscript{Dia-B} | 2 | DI | 94.9K | 10.8K | 12.9K |
| ARLUE\textsubscript{Dia-D} | 2 | DI | 38.5K | 4.5K | 5.3K |
| ARLUE\textsubscript{DIA-C} | 3 | DI | 711.9K | 31.5K | 52.1K |
| ARLUE\textsubscript{QA} | 4 | QA | 101.6K | 517 | 7.45K |

\textsuperscript{2} Number of question-answer pairs (Abdul-Mageed et al., 2021).

A.2 Evaluation on DEV

The evaluation on validation dataset on all ARLUE clusters as single tasks over all models are shown in Table A.2. Moreover, Table A.3 and Table A.4 illustrate the results on dialect datasets at binary, regions, and countries levels.

### Table A.2: Performance of our models on ARLUE DEV datasets (Acc / F\textsubscript{1}). * Metric for ARLUE\textsubscript{Sent} is Acc/ F\textsubscript{1}. \textsuperscript{†} Metric for ARLUE\textsubscript{QA} is Exact Match (EM) / F\textsubscript{1}. \textsuperscript{‡} Metric for ARLUE\textsubscript{Sent} is F\textsubscript{1}. SOTA: MArBERT (Abdul-Mageed et al., 2021)

| Dataset | SOTA | ARLUE\textsubscript{Sent} | ARLUE\textsubscript{SM} | ARLUE\textsubscript{Topic} | ARLUE\textsubscript{Dia} | ARLUE\textsubscript{PA} | ARLUE\textsubscript{QA} | ARLUE\textsubscript{Diial} |
|---------|------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Arabic  | 75.99 | 74.51 | 75.16 | 77.24 | 76.96 |

| Dataset | Setting | mT5 | AraT5\textsubscript{Sent} | AraT5\textsubscript{SM} | AraT5\textsubscript{Topic} | AraT5\textsubscript{Dia} | AraT5\textsubscript{PA} | AraT5\textsubscript{QA} |
|---------|---------|------|----------------|----------------|----------------|----------------|----------------|----------------|
| Arabic  | S       | 84.7 | 89.7 | 93.7 | 80.31 | 90.97 | 90.07 | 89.74 |
|         | M       | 89.02 | 89.80 | 90.02 | 89.71 | 90.86 | 89.98 | 89.64 |
| Arabic  | S       | 88.07 | 89.90 | 90.92 | 91.42 | 90.80 | 90.14 | 89.64 |
|         | M       | 91.58 | 92.08 | 93.02 | 91.42 | 90.80 | 90.14 | 89.64 |
| Arabic  | S       | 44.2 | 50.37 | 50.87 | 91.37 | 89.19 | 53.57 | 80.25 |
|         | M       | 49.09 | 55.50 | 51.92 | 40.17 | 53.62 | 80.25 | 52.62 |

Table A.3: Performance of our models on ARLUE Diaclets DEV datasets on single and multi tasks setting (Acc / F\textsubscript{1}). We copied single tasks results from Table A.2 in this table for comparison. S: Single Task. M: Multi-task.
Table A.4: Performance of our models on ARLUE Social Meaning (SM) DEV datasets on single and multi tasks setting (Acc / F1). S: Single Task. M: Multi-task.

| Dataset | S/M | mT5 | AraT5Tweet | AraT5MSE | AraT5 |
|---------|-----|-----|------------|----------|-------|
| Age     | S   | 60.86 / 61.05 | 62.29 / 62.48 | 63.26 / 63.41 | 63.50 / 63.66 |
|         | M   | 61.44 / 61.52 | 64.22 / 64.39 | 63.96 / 64.13 | 63.97 / 64.05 |
| Dangerous | S   | 79.02 / 79.54 | 82.25 / 84.09 | 80.30 / 80.92 | 81.57 / 81.10 |
|         | M   | 80.08 / 80.54 | 82.85 / 84.69 | 80.78 / 81.35 | 81.64 / 82.10 |
| Emotion | S   | 71.82 / 70.25 | 73.30 / 71.93 | 73.87 / 73.37 | 74.18 / 73.56 |
|         | M   | 69.89 / 68.49 | 72.53 / 72.14 | 73.64 / 71.77 | 73.97 / 72.13 |
| Gender  | S   | 72.06 / 71.84 | 72.27 / 72.06 | 73.87 / 73.55 | 73.87 / 73.24 |
|         | M   | 72.06 / 72.35 | 74.34 / 74.15 | 74.45 / 74.32 | 74.44 / 74.31 |
| Hate    | S   | 56.07 / 55.61 | 56.07 / 55.22 | 56.07 / 55.41 | 56.07 / 55.46 |
|         | M   | 95.70 / 76.71 | 96.40 / 80.69 | 96.00 / 78.12 | 96.40 / 80.22 |
| Irony   | S   | 82.88 / 82.84 | 84.38 / 84.13 | 84.86 / 84.85 | 84.87 / 83.51 |
|         | M   | 81.89 / 81.18 | 83.12 / 84.12 | 84.62 / 84.54 | 85.44 / 85.58 |
| Offensive | S   | 93.07 / 85.92 | 90.20 / 91.76 | 95.20 / 91.06 | 96.60 / 92.48 |
|         | M   | 92.20 / 86.85 | 94.60 / 91.01 | 94.70 / 90.92 | 95.70 / 90.54 |
| ARLUE_SM | S   | 79.50 / 74.45 | 80.50 / 76.92 | 81.25 / 77.26 | 80.31 / 77.18 |
|         | M   | 78.97 / 74.14 | 81.19 / 77.40 | 81.33 / 77.03 | 81.23 / 77.23 |

Table B.1: Main characteristics of AraLGNTG data splits. For each split, we provide the number of article-title pairs and the average length of the articles and titles.

| Split     | Article/Title | Avg article len | Avg title len |
|-----------|---------------|-----------------|---------------|
| TRAIN     | 93.3K         | 256.46          | 10.06         |
| DEV       | 11.7K         | 253.11          | 10.03         |
| TEST      | 11.7K         | 260.92          | 10.03         |
| Total     | 11.6G.K       | 256.63          | 10.04         |

Table B.2: Performance of our models on document summarization DEV splits.

B ARGEN

In this section we describe the ARGEN_MT datasets splits and report the evaluation results in validation datasets. Details about ARGEN_NTG are in Table B.1 and ARGEN_MT datasets splits are shown in Table 2. Moreover, The evaluation on validation datasets for ARGEN_TS, ARGEN_NTG, ARGEN_QG and ARGEN_MT are shown B.2, B.3, and B.4. Finally, B.6 shows an examples from title generation test dataset (ARGEN_NTG).

| Dataset     | mT5 | AraT5Tweet | AraT5MSE | AraT5 |
|-------------|-----|------------|----------|-------|
| ARGEN_NTG   | 19.22 | 19.38 | 20.19 | 20.01 |
| ARGEN_QG    | 13.95 | 11.25 | 12.96 | 15.36 |

Table B.3: Performance of our models on title and question generation DEV split based on Bleu score.
| Varieties | Dataset | Region | Country-Level | City-Level | DEV | TEST |
|-----------|---------|--------|---------------|------------|-----|------|
| DIA       | ADPT    | Levante | -             | -          | 138K|      |
|           | Zbib et al. (2012) | Nile | Egypt | Guelma | - | 38K |
|           | Bible I  | Maghrebi | Morocco | - | 600 |      |
|           |         | Egypt | Cairo | - | 6.5k |      |
|           |         | Sudan | Khartoum | - | 2k |      |
|           |         | Qatar | Doha | - | 6.5k |      |
|           |         | Yemen | Sana’a | - | 2k |      |
|           |         | Oman | Muscat | - | 2k |      |
|           |         | KSA | Riyadh | - | 2k |      |
|           |         | Jedd | Muscat | - | 2k |      |
|           |         | Iraq | Baghdad | - | 2k |      |
|           |         | Iraq | Basra | - | 2k |      |
|           |         | Iraq | Mosul | - | 2k |      |
|           |         | Lebanon | Beirut | - | 6.5k |      |
|           |         | Palestine | Jerusalem | - | 2k |      |
|           |         | Jordan | Amman | - | 2k |      |
|           |         | Jordan | Salt | - | 2k |      |
|           |         | Syria | damascus | - | 2k |      |
|           |         | Syria | Aleppo | - | 2k |      |
|           |         | Algeria | Alger | - | 2k |      |
|           |         | Lybia | Trip | - | 2k |      |
|           |         | Lybia | Beng | - | 2k |      |
|           |         | Tunisia | Tunis | - | 6.5k |      |
|           |         | Tunisia | Safax | - | 2k |      |
|           |         | Morocco | Fes | - | 6.5k |      |
|           |         | Morocco | Rabat | - | 2k |      |
|           | MADAR I | Bouamor et al. (2018) | Lebanon | - | 600 |      |
|           |         | Palestine | Jerusalem | - | 2k |      |
|           |         | Jordan | Amman | - | 2k |      |
|           |         | Jordan | Salt | - | 2k |      |
|           |         | Syria | damascus | - | 2k |      |
|           |         | Algeria | Alger | - | 2k |      |
|           |         | Lybia | Trip | - | 2k |      |
|           |         | Lybia | Beng | - | 2k |      |
|           |         | Tunisia | Tunis | - | 6.5k |      |
|           |         | Tunisia | Safax | - | 2k |      |
|           |         | Morocco | Fes | - | 6.5k |      |
|           |         | Morocco | Rabat | - | 2k |      |
|           | MSA     | MADAR I | Bouamor et al. (2018) | - | 600 |      |
|           |         | IWSLT TED15 Cettolo et al. (2016) | - | - | 6.5k |      |
|           |         | IWSLT TED16 / Cettolo et al. (2016) | - | - | 6.5k |      |
|           |         | IWSLT QED16 (Cettolo et al., 2016) | - | - | 550 |      |
|           |         | UN Zienki et al. (2016) | - | - | 1k |      |

Table B.5: ARGENMT datasets.
Table B.6: Title generation samples from TEST set using our Models.