AlcLaM: Arabic Dialectal Language Model

Murtadha Ahmed [*]**, Saghir Alfasly** [†]**, Wenbo** [*]**, Jamaal Qasem** [‡]**, Mohammed Ahmed** [§]**, Yunfeng Liu** [*]
[*] Zhuiyi AI Lab, China, [†] Mayo Clinic, USA, [‡] Dongbei University of Finance and Economics, China, [§] Northwestern Polytechnical University, China
[*] {a.murtadha,brucewen,glenliu}@wezhuiyi.com, [†] alfasly.saghir@mayo.edu
[‡] ja.qasem@dufe.edu.cn, [§] majeedi@mail.nwpu.edu.cn

Abstract

Pre-trained Language Models (PLMs) are integral to many modern natural language processing (NLP) systems. Although multilingual models cover a wide range of languages, they often grapple with challenges like high inference costs and a lack of diverse non-English training data. Arabic-specific PLMs are trained predominantly on modern standard Arabic, which compromises their performance on regional dialects. To tackle this, we construct an Arabic dialectal corpus comprising 3.4M sentences gathered from social media platforms. We utilize this corpus to expand the vocabulary and retrain a BERT-based model from scratch. Named AlcLaM, our model was trained using only 13 GB of text, which represents a fraction of the data used by existing models such as CAMeL, MARBERT, and ArBERT, compared to 7.8%, 10.2%, and 21.3%, respectively. Remarkably, AlcLaM demonstrates superior performance on a variety of Arabic NLP tasks despite the limited training data. AlcLaM is available at: https://github.com/amurtadha/Alclam.

1 Introduction

Pre-trained Language Models (PLMs) utilizing self-supervised learning techniques, such as BERT (Devlin et al., 2018a) and RoBERTa (Liu et al., 2019), have become pivotal in advancing the field of natural language processing (NLP) through transfer learning. These models have significantly enhanced performance across a variety of NLP tasks by leveraging vast amounts of textual data and extensive computational resources. However, the necessity for large corpora and the substantial computational demand, often requiring weeks of training time (Conneau et al., 2020; Raffel et al., 2020; Adiwardana et al., 2020), has primarily confined the development of such models to the English language and a few other major languages.

This limitation has sparked an increased interest in creating multilingual models capable of understanding and processing multiple languages simultaneously. Innovations such as mBERT (Devlin et al., 2018a), XLM-RoBERTa (Conneau et al., 2020) and LaBSE (Feng et al., 2022) aim to address this gap. Despite these efforts, the performance of these multilingual models typically lags behind their monolingual counterparts. This discrepancy is largely due to smaller, language-specific vocabularies and less comprehensive language-specific datasets (Virtanen et al., 2019; Antoun et al., 2020; Dadas et al., 2020; de Vries et al., 2019; Malmsten et al., 2020; Nguyen and Nguyen, 2020).

Furthermore, while languages with similar structures and vocabularies may benefit from shared representations (Conneau et al., 2020), this advantage does not extend to languages such as Arabic. Arabic's unique morphological and syntactic structures share little in common with the morphosyntactic frameworks of more abundantly represented Latin-based languages. To address this, various Arabic-specific PLMs have been developed, including AraBERT (Antoun et al., 2020), ArBERT (Abdul-Mageed et al., 2021), and CAMeL (Inoue et al., 2021). These models significantly enhance Arabic NLP tasks over multilingual models. However, they are predominantly trained on Modern Standard Arabic (MSA) datasets. This focus on MSA introduces two primary limitations: first, there is reduced recognition of dialectal tokens, which vary widely across different Arabic-speaking regions; second, there is a biased weighting towards MSA tokens in the models, which may not accurately reflect the linguistic nuances present in everyday Arabic usage.

In this paper, we first introduce a new corpus of 3,372,744 Arabic dialectal texts, meticulously sourced from social media platforms such as YouTube and Facebook. Second, we outline the procedure for pretraining the transformer model

Proceedings of The Second Arabic Natural Language Processing Conference, pages 153–159
August 16, 2024

(Devlin et al., 2018a) specifically for the Arabic language, which we dub AlcLaM. Note that we only train AlcLaM on 13GB text due to computational resources limitation. Finally, we assess AlcLaM's performance on three Arabic NLU downstream tasks, each distinct in nature: (i) Arabic Dialect Identification (DID), (ii) Sentiment Analysis (SA), and (iii) Hate Speech and Offensive Language Detection. Despite the limited training data, our experimental results demonstrate that AlcLaM attains state-of-the-art results on most datasets, surpassing several baseline models, including previous multilingual and single-language approaches.

In summary, our contributions are twofold.

- We constructed a massive corpus of Arabic dialects, derived from the content and comments on Arabic pages on Facebook and videos from Arabic-speaking YouTubers. This corpus represents a rich variety of regional dialects and everyday language usage that has been underrepresented in previous models.

- We developed an Arabic pre-trained language model, namely AlcLaM, specifically optimized to handle the diversity and complexity of Arabic dialects based on the newly created corpus, enhancing its applicability across a wider range of NLP tasks involving Arabic text.

2 Related Work

Pre-trained language models (PLMs) using a self-supervised masking objective, such as BERT (Devlin et al., 2018a) and RoBERTa (Liu et al., 2019), have significantly advanced NLP. These models have multilingual versions, including mBERT (Devlin et al., 2018a), XLM-RoBERTa (Conneau et al., 2020) and LaBSE (Feng et al., 2022). Additionally, models featuring different objectives or architectures, such as ALBERT (Lan et al., 2020), T5 (Raffel et al., 2020), its multilingual variant mT5 (Xue et al., 2021), and GPT-3 (Brown et al., 2020), LLaMA (Touvron et al., 2023), PaLM (Chowdhery et al., 2023), GPT-4 (OpenAI, 2023), and RoFormer (Su et al., 2024) have been introduced.

Non-English PLMs have also been developed. These include Bertje for Dutch (de Vries et al., 2019), CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020) for French, PhoBERT for Vietnamese (Nguyen and Nguyen, 2020), as well as models for Finnish by Virtanen et al. (2019), for Polish by Dadas et al. (2020), and for Swedish by Malmsten et al. (2020). Pyysalo et al. (2021) have created monolingual LMs using Wikipedia data for 42 languages. For Arabic, MSA-based PLMs includes AraBERT (Antoun et al., 2020) ArabicBERT (Safaya et al., 2020), ArBERT (Abdul-Mageed et al., 2021). Another line of research involves pre-training models on a combination of MSA and dialectal data, such as MDBERT (Abdul-Mageed et al., 2021) and CAMeL (Mubarak et al., 2021). Our contributions to this field include a comprehensive Arabic dialectal corpus spanning various dialects and the development of an Arabic PLM. Our model, named AlcLaM, enhances the representation of linguistic diversity in Arabic NLP.

3 Methodology

In this paper, we develop AlcLaM, an Arabic dialect language representation model that enhances the performance on several Arabic NLP tasks. This model builds upon the BERT architecture, a stacked Bidirectional Transformer Encoder (Devlin et al., 2018a). Recognized as the foundation for many state-of-the-art results in various NLP tasks across multiple languages, BERT's architecture has proven highly effective. Below, we detail the dialectal corpus used for AlcLaM's pretraining, the pretraining setup, and the fine-tuning process.

3.1 Arabic Dialectal Corpus.

The original BERT model was trained on a corpus comprising 3.3 billion words extracted from English Wikipedia and the Book Corpus (Zhu et al., 2015). Due to the comparatively smaller size of Arabic Wikipedia dumps in comparison to English ones, we opted to utilize Arabic text from the English-Arabic bilingual corpora of opensubtitles [1] (Itamar and Itai, 2008).

It is noteworthy that publicly available Arabic corpora are heavily dominated by MSA, while social media and online reviews predominantly feature Arabic dialects. This creates a bias towards MSA tokens in Arabic PLMs, potentially leading to tokenizers failing to recognize a significant portion of dialectal vocabulary. To address this, we manually scraped Arabic texts from social media platforms. Initially, we scrape posts and comments from popular Arabic YouTube channels and Facebook Pages. However, we observed that many of

[1] opensubtitles

these comments consisted of verses from the Holy Quran and Hadith, typically written in MSA. Since our focus was on dialectal texts, we trained a binary classifier (MSA-Dialect) to filter out MSA texts. Specifically, we treated all dialectal instances of the MADAR corpus as one class, labeled "Dialect" (Bouamor et al., 2019; Murtadha et al., 2022), and utilized it to fine-tune the CAMeL model (Inoue et al., 2021), which achieved a remarkable 98% accuracy. Our final corpus comprises 3,372,744 dialectal sentences with 54,557,408 tokens. To the best of our knowledge, this marks the first attempt to assemble such a comprehensive Arabic dialectal corpus.

3.2 Model Training

For AlcLaM, we adhere to the original BERT (Devlin et al., 2018a). Each training input sequence is generated using whole word masking, where 15% of the N input tokens are chosen for replacement. These selected tokens undergo replacement as follows: 80% are substituted with the [MASK] token, 10% with a random token, and 10% remain unchanged. Following Liu et al. (2019), we exclude the next sentence prediction (NSP) loss from our training process. This decision is based on the observation that removing the NSP loss either matches or slightly improves downstream task performance. We employ the same network configuration as BERT-base: consisting of 12 layers, 768 hidden units, and 12 attention heads, resulting in approximately 125 million parameters. During training, we utilize a batch size of 64 sequences and set a maximum sequence length of 128 tokens and 5 training epochs. Throughout training, we set the learning rate to $5e-5$.

3.3 Fine-tuning

To fine-tune AlcLaM for sequence classification, we utilize the final hidden state of the first token, corresponding to the embedding of the special "[CLS]" token that is prepended to the beginning of each sentence (Murtadha et al., 2024). A simple feed-forward layer with a Softmax activation function is added to compute the probability distribution over the predicted output classes. During fine-tuning, both the classifier and the pre-trained model weights are jointly trained to maximize the log-probability of the correct class (Ahmed et al., 2023).

4 Empirical Evaluation

4.1 Datasets

We evaluated AlcLaM on the following datasets that cover various NLP tasks in Arabic. Sentiment analysis (SemEval 2017 task 4 (Kiritchenko et al., 2016), ASAD (Alharbi et al., 2020), ASTD (Nabil et al., 2015), ArSAS (Elmadany et al., 2018), LABR (Aly and Atiya, 2013)), offensive language detection (Adult (Mubarak et al., 2021), Offensive and HateSpeech (Mubarak et al., 2020)), dialect identification (MADAR-6, MADAR-26(Bouamor et al., 2019) and NADI (Abdul-Mageed et al., 2020)). For experiments, MADAR-2 and MADAR-9 are derived from MADAR-26. MADAR-2 is binary (MSA-dialect), while MADAR-9 categorizes dialects into 9 regions: Yemen, MSA, Maghreb, Nile Egypt, Libya, Gulf, Nile Sudan, Iraq, and Levant.

4.2 Baselines

We compare our AlcLaM model with:

1. Multilingual PLMs like mBERT (Devlin et al., 2018b) and LaBSE (Feng et al., 2022);
2. MSA-based Arabic PLMs such as AraBERT (Antoun et al., 2020) and ArBERT (Abdul-Mageed et al., 2021);
3. MSA-Dialect-based PLMs, including MdBERT (Talafha et al., 2020), and MARBERT (Abdul-Mageed et al., 2021) and CAMeL (Inoue et al., 2021).

4.3 Results

For each dataset, we report the average results of five runs, each with different random seeds, to ensure statistical significance. The results for various Arabic NLP tasks are presented in Table 1 and Table 2 in terms of F1 and accuracy metrics, respectively. From these results, we make the following observations:

1. Multilingual models such as mBERT and LaBSE are outperformed by Arabic-specific models that are pre-trained with larger vocabularies and more extensive language-specific datasets. This observation aligns with the findings of Abdul-Mageed et al. (2021).
2. Models that incorporate dialectal data during pre-training, such as MdBERT, CAMeL and MARBERT, not only excel in DID task, but

	Dataset	Multilingual PLMs		MSA-based PLMs		MSA-Dialect-based PLMs			
		mBERT	LaBSE	AraBERT	ArBERT	MdBERT	CAMeL	MARBERT	AlcLaM
DID	MADAR-2	72.9 ± 16.9	86.6 ± 0.5	87.1 ± 0.2	87.1 ± 0.2	86.0 ± 0.6	87.5 ± 1.0	85.3 ± 3.8	**98.2 ± 0.1**
	MADAR-6	91.3 ± 0.1	91.1 ± 0.2	91.6 ± 0.1	91.6 ± 0.2	91.6 ± 0.0	92.0 ± 0.1	92.2 ± 0.2	**93.2 ± 0.1***
	MADAR-9	75.5 ± 0.5	75.7 ± 0.2	76.8 ± 0.3	74.5 ± 4.3	75.9 ± 0.5	77.5 ± 0.4	78.2 ± 0.3	**81.9 ± 0.3**
	MADAR-26	60.5 ± 0.2	62.0 ± 0.2	62.0 ± 0.1	61.7 ± 0.1	60.2 ± 0.4	62.9 ± 0.1	61.5 ± 0.4	**66.3 ± 0.1***
	NADI	17.6 ± 0.5	17.6 ± 0.5	22.6 ± 0.5	22.6 ± 0.5	24.9 ± 0.6	25.9 ± 0.5	**28.6 ± 0.8***	25.6 ± 0.6
SA	SemEval	51.3 ± 1.3	64.2 ± 0.7	65.4 ± 0.5	64.4 ± 0.9	65.6 ± 0.3	67.1 ± 0.7	66.4 ± 0.3	**69.2 ± 0.4***
	ASAD	59.8 ± 0.0	62.4 ± 0.0	41.3 ± 0.0	66.9 ± 0.0	**67.5 ± 0.0**	65.8 ± 0.0	66.8 ± 0.0	66.7 ± 0.0
	AJGT	86.4 ± 0.3	92.4 ± 0.7	92.7 ± 0.3	92.6 ± 0.4	93.6 ± 0.0	93.6 ± 0.3	93.7 ± 0.1	**95.0 ± 0.3***
	ASTD	46.3 ± 1.4	55.7 ± 0.4	57.5 ± 2.3	59.7 ± 0.1	61.9 ± 0.4	60.2 ± 0.2	61.0 ± 0.5	**64.6 ± 0.1***
	LABR	81.1 ± 0.0	85.4 ± 0.0	85.9 ± 0.0	85.9 ± 0.0	84.7 ± 0.0	**86.3 ± 0.0**	85.0 ± 0.0	84.9 ± 0.0
	ARSAS	73.2 ± 0.7	76.2 ± 0.6	76.8 ± 0.3	76.1 ± 0.2	76.3 ± 0.2	77.1 ± 0.3	76.2 ± 0.2	**77.9 ± 0.3***
HSOD	HateSpeech	67.9 ± 1.4	73.7 ± 1.1	76.4 ± 1.2	76.8 ± 1.4	80.0 ± 0.1	78.8 ± 0.6	80.0 ± 0.8	**81.4 ± 0.5***
	Offense	85.3 ± 0.5	87.2 ± 0.5	90.5 ± 0.4	90.5 ± 0.4	90.8 ± 0.2	89.2 ± 0.5	90.8 ± 0.3	**91.3 ± 0.3***
	Adult	87.9 ± 0.1	87.2 ± 0.3	88.6 ± 0.1	88.4 ± 0.6	88.1 ± 0.0	88.6 ± 0.3	88.3 ± 0.1	**89.3 ± 0.3***

Table 1: F1 Score Evaluation of Various Arabic NLP Models. Best scores are highlighted in bold. An asterisk (*) denotes statistical significance, determined by a t-test with a p-value (< 0.05). Our AlcLaM not only excels in DID task but also shows improvements in most other tasks. This performance is expected as most Arabic NLP datasets are collected from social media, which is dominated by dialectal expressions.

	Dataset	Multilingual PLMs		MSA-based PLMs		MSA-Dialect-based PLMs			
		mBERT	LaBSE	AraBERT	ArBERT	MdBERT	CAMeL	MARBERT	AlcLaM
DID	MADAR-2	97.3 ± 0.8	98.0 ± 0.1	98.1 ± 0.0	98.1 ± 0.0	98.0 ± 0.1	98.1 ± 0.1	97.2 ± 0.7	**99.7 ± 0.0**
	MADAR-6	91.3 ± 0.1	91.1 ± 0.2	91.6 ± 0.1	91.6 ± 0.2	91.6 ± 0.0	92.0 ± 0.1	92.2 ± 0.2	**93.2 ± 0.1***
	MADAR-9	78.5 ± 0.5	79.1 ± 0.1	80.4 ± 0.2	77.7 ± 3.6	79.1 ± 0.5	80.5 ± 0.2	81.1 ± 0.3	**83.4 ± 0.4**
	MADAR-26	60.6 ± 0.2	61.9 ± 0.2	61.9 ± 0.1	61.7 ± 0.2	60.1 ± 0.3	62.9 ± 0.2	61.3 ± 0.3	**66.1 ± 0.2***
	NADI	33.4 ± 0.6	33.4 ± 0.6	38.9 ± 1.7	38.9 ± 1.7	41.9 ± 1.9	42.7 ± 1.6	**47.3 ± 0.1***	46.6 ± 1.0
SA	SemEval	53.4 ± 1.5	65.0 ± 0.6	66.1 ± 0.5	65.1 ± 0.8	66.1 ± 0.3	68.0 ± 0.3	66.9 ± 0.3	**69.5 ± 0.3***
	ASAD	74.6 ± 0.0	75.2 ± 0.0	70.6 ± 0.0	78.4 ± 0.0	77.6 ± 0.0	77.0 ± 0.0	77.6 ± 0.0	**79.5 ± 0.0**
	AJGT	86.4 ± 0.3	92.4 ± 0.7	92.8 ± 0.3	92.6 ± 0.4	93.6 ± 0.0	93.6 ± 0.3	93.8 ± 0.1	**95.0 ± 0.3***
	ASTD	46.7 ± 1.7	55.6 ± 0.6	57.7 ± 2.4	59.7 ± 0.3	62.0 ± 0.3	60.1 ± 0.2	61.0 ± 0.3	**64.9 ± 0.1***
	LABR	90.4 ± 0.0	92.3 ± 0.0	92.8 ± 0.0	92.8 ± 0.0	91.9 ± 0.0	**93.0 ± 0.0**	92.6 ± 0.0	92.6 ± 0.0
	ARSAS	74.5 ± 0.8	77.2 ± 0.7	77.6 ± 0.3	77.0 ± 0.3	77.5 ± 0.3	78.0 ± 0.3	77.4 ± 0.4	**78.6 ± 0.5***
HSOD	HateSpeech	75.2 ± 2.2	80.0 ± 0.7	80.5 ± 1.4	80.8 ± 1.9	84.3 ± 0.3	83.3 ± 0.6	84.4 ± 0.4	**84.6 ± 0.7***
	Offense	91.7 ± 0.1	92.8 ± 0.4	94.5 ± 0.2	94.6 ± 0.4	94.6 ± 0.2	93.6 ± 0.2	94.8 ± 0.0	**94.9 ± 0.1***
	Adult	95.0 ± 0.0	94.4 ± 0.2	95.2 ± 0.1	94.9 ± 0.4	95.1 ± 0.1	95.2 ± 0.2	95.1 ± 0.0	**95.6 ± 0.0**

Table 2: Accuracy Evaluation of Various Arabic NLP Models

also perform significantly across a broader range of Arabic NLP tasks. This suggests that the similarities among Arabic dialects may not always have positive effects on other tasks beyond ADI. The experimental results underscore the value of integrating more dialectal information during training, as the tokenizers in these models are likely to recognize more dialect-specific tokens, which are often unidentified in other models.

3. Despite being trained on less MSA text and fewer training steps, due to computational resource constraints, our model outperforms its alternatives in most tasks and achieves competitive performance in others. Although the improvements in tasks other than ADI are modest, they are significant given the inherent complexities of the Arabic language.

In tasks beyond DID, AlcLaM may show modest improvements, but it introduces vital empirical factors like stability and statistical significance, supported by a t-test ($p < 0.05$). MSA-Dialect PLMs consistently demonstrate superior performance across a range of Arabic NLP tasks. These empirical findings clearly support our claim regarding the critical importance of incorporating Arabic dialectal data in the pre-training process.

5 Conclusion

In this paper, we present AlcLaM, a novel BERT-based model trained specifically to address the challenge of Arabic dialectal variation. Leveraging a carefully curated corpus sourced from social media platforms, AlcLaM its alternatives across various Arabic NLP tasks, despite being trained on significantly less data. For future work, expanding the dialectal vocabulary without increasing inference costs, inspired by Chinese character modeling.

Limitations

Despite the advancements achieved by AlcLaM, it is important to acknowledge its current limitations:

- AlcLaM is trained from scratch to build its vocabulary. However, incorporating weights of new dialectal vocabulary from existing Arabic PLMs and adjusting through continued training is a potential avenue for enhancement. Nevertheless, expanding the vocabulary size to encompass more dialectal tokens might lead to increased inference costs.
- Given that AlcLaM was trained on approximately 10% of the training data used by its alternatives, due to computational resource constraints, its performance on generative tasks may not be as significant. Nonetheless, this limitation can be mitigated by continued training on our open-source AlcLaM model.

References

Muhammad Abdul-Mageed, AbdelRahim A. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7088–7105. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. 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.

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Murtadha Ahmed, Shengfeng Pan, Wen Bo, Jianlin Su, Xinxin Cao, Wenze Zhang, and Yunfeng Liu. 2023. Rank-aware negative training for semi-supervised text classification. *Trans. Assoc. Comput. Linguistics*, 11:771–786.

Basma Alharbi, Hind Alamro, Manal Alshehri, Zuhair Khayyat, Manal Kalkatawi, Inji Ibrahim Jaber, and Xiangliang Zhang. 2020. ASAD: A twitter-based benchmark arabic sentiment analysis dataset. *CoRR*, abs/2011.00578.

Mohamed A. Aly and Amir F. Atiya. 2013. LABR: A large scale arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 494–498. The Association for Computer Linguistics.

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, Marseille, France. European Language Resource Association.

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, WANLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 199–207. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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 Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben 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 Garcia, 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 Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. 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.

Slawomir Dadas, Michal Perelkiewicz, and Rafal Poswiata. 2020. Pre-training polish transformer-based language models at scale. In *Artificial Intelligence and Soft Computing - 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II*, volume 12416 of *Lecture Notes in Computer Science*, pages 301–314. Springer.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch BERT model. *CoRR*, abs/1912.09582.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

A Elmadany, Hamdy Mubarak, and Walid Magdy. 2018. Arsas: An arabic speech-act and sentiment corpus of tweets. *OSACT*, 3:20.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. Association for Computational Linguistics.

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.

Einav Itamar and Alon Itai. 2008. Using Movie Subtitles for Creating a Large-Scale Bilingual Corpora. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

Svetlana Kiritchenko, Saif Mohammad, and Mohammad Salameh. 2016. SemEval-2016 task 7: Determining sentiment intensity of English and Arabic phrases. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 42–51, San Diego, California. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît 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, LREC 2020, Marseille, France, May 11-16, 2020*, pages 2479–2490. European Language Resources Association.

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

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden - making a swedish BERT. *CoRR*, abs/2007.01658.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7203–7219. 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.

Ahmed Murtadha, Shengfeng Pan, Bo Wen, Jianlin Su, Wenze Zhang, and Yunfeng Liu. 2022. BERT-ASC: auxiliary-sentence construction for implicit aspect learning in sentiment analysis. *CoRR*, abs/2203.11702.

Ahmed Murtadha, Bo Wen, Luo Ao, Shengfeng Pan, Jianlin Su, Xinxin Cao, and Yunfeng Liu. 2024. Towards robust learning with noisy and pseudo labels for text classification. *Inf. Sci.*, 661:120160.

Mahmoud Nabil, Mohamed A. Aly, and Amir F. Atiya. 2015. ASTD: arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2515–2519. The Association for Computational Linguistics.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1037–1042. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Sampo Pyysalo, Jenna Kanerva, Antti Virtanen, and Filip Ginter. 2021. Wikibert models: Deep transfer learning for many languages. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics, NoDaLiDa 2021, Reykjavik, Iceland (Online), May 31 - June 2, 2021*, pages 1–10. Linköping University Electronic Press, Sweden.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 2054–2059. International Committee for Computational Linguistics.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein T. Al-Natsheh. 2020. Multi-dialect arabic BERT for country-level dialect identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop, WANLP@COLING 2020, Barcelona, Spain (Online), December 12, 2020*, pages 111–118. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

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

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Richard S. 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 *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.