Domain-Adapted BERT-based Models for Nuanced Arabic Dialect Identification and Tweet Sentiment Analysis

Gıyaseddin Bayrak
Marmara University / Istanbul - Turkey
giyaseddinalfarkh@marun.edu.tr

Abdul Majeed Issifu
Marmara University / Istanbul - Turkey
abdul.majeed@marun.edu.tr

Abstract

This paper summarizes the solution of the Nuanced Arabic Dialect Identification (NADI) 2022 shared task. It consists of two subtasks: a country-level Arabic Dialect Identification (ADID) and an Arabic Sentiment Analysis (ASA). Our work shows the importance of using domain-adapted models and language-specific pre-processing in NLP task solutions. We implement a simple but strong baseline technique to increase the stability of fine-tuning settings to obtain a good generalization of models. Our best model for the Dialect Identification subtask achieves a Macro F-1 score of 25.54% as an average of both Test-A (33.89%) and Test-B (19.19%) F-1 scores. We also obtained a Macro F-1 score of 74.29% of positive and negative sentiments only, in the Sentiment Analysis task[1].

1 Introduction

The Arabic language is one of the rich languages in the world, spoken in large geographical regions. It is officially spoken by people from the Middle East and North Africa (MENA) countries, covering a population of approximately 400 million people. It's a culturally and grammatically rich language, with a complex morphological structure. Arabic is one of the Semitic languages and has a widely varying collection of more than 30 different dialects (according to the Summer Institute of Linguistics a.k.a. SIL International). These dialects are affected by geopolitical and religious influence. The question of how to classify the different varieties of spoken Arabic is a long-standing problem in the fields of Arabic and Semitic linguistics. Researchers still develop tools and systems to keep the language in the race of Natural Language Processing (NLP) tasks on both Modern Standard Arabic (MSA) and its Dialects (DA).

[1]The code of the implementation is available at https://github.com/giyaseddin/NADI

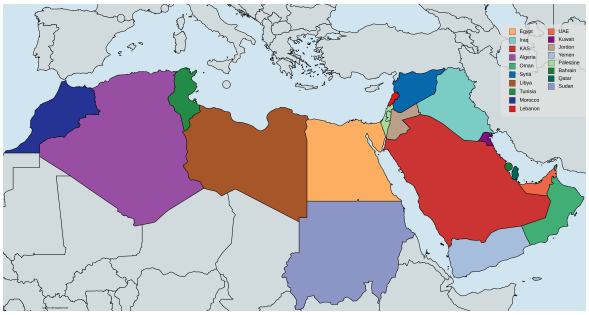


Figure 1: Geographical distribution of the countries listed in Subtask 1.

A dialect of the Arabic language can have a different meaning of a word or a vernacular dialect can differ syntactically, morphological, and orthographically in the choice of vocabulary and pronunciation. Each of these variations of dialects is distinct enough to make users resort to formal Arabic to understand each other. This prompts the need to develop a system that can automatically detect the source, region, and/or specific dialect of a given sequence of tokens or text segments. The NADI shared task series .... (Abdul-Mageed et al., 2020b) (Abdul-Mageed et al., 2021) (Abdul-Mageed et al., 2022) is one of the prominent competitions that provides datasets and modeling opportunities for researchers to improve NLP work in Arabic. Social media provides an environment for the use of both formal and informal language. This makes it more difficult when Arabic is used on social media since both dialects and the formality of the language will be taken into consideration when processing text data from social media like Twitter. This variety of dialects can be classified and used for more semantic and linguistic findings and work using machine learning and deep learning models.

Language Models (LM) have evolved over the years from the birth of the NLP domain, starting with simple n-gram LMs, with many computational and performance limitations. After the introduc-

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 425 - 430

Subset	Training			Dev	Test-A	Test-B	Test
	Total	Train	Validation				
Subask 1	20398	18358	2040	4758	4758	500	-
Subask 2	1500	1425	75	500	-	-	3000

Table 1: Data subset sizes for Task 1 and Task 2

tion of Deep Learning (DL), language modeling switched to language modeling using Recurrent Neural Networks (RNN), Gated Recurrent Unit (GRU), and Long-Short Term Memory (LSTM) with chronologically better deployability than the earlier methods. The drastic improvement was after introducing the Transformer architecture for language modeling (Vaswani et al., 2017) using the self-attention mechanism.

Transformer-based LM like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), are currently widely used in the NLP field to achieve state-of-the-art (SOTA) results in various tasks. BERT and its variations (RoBERTa, DistilBERT, ALBERT, etc.) are outstanding models, and they are close to becoming a de facto baseline for almost all NLP tasks, especially for Natural Language Understanding (NLU) downstream tasks. This is because of the capability of these general models to be fine-tuned on narrower tasks in different domains with high accuracy and low cost.

In this paper, we develop a system for the classification of Arabic dialects at the country level. Arabic Dialect Identification (ADID) problem is challenging because adjacent countries influence each other, with the present intermediate dialects (Abdul-Mageed et al., 2021). Our system also provides Arabic Sentiment Analysis (ASA) of given tweet texts. We improve both ADID and ASA tasks using AraBERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2020a), Arabic language-specific pre-trained BERT models.

The rest of this paper is organized as follows. In section 2, we provide a detailed explanation of the problem and datasets provided by NADI-2022(Abdul-Mageed et al., 2022). Section 3 talks about the methodology and general system development. We provide the results in section 4 and discuss the results and model limitations in section 5. The paper is concluded in section 6.

2 Data

The NADI-2022 shared task provides two problem definitions of country-level dialect identification

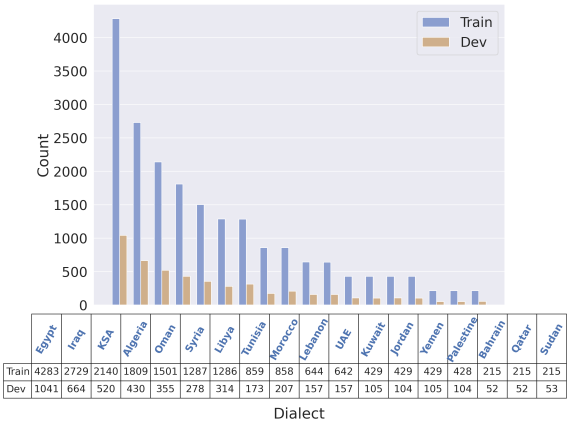


Figure 2: Country-level dialect distribution for the TRAIN and DEV data subsets of Subtask 1.

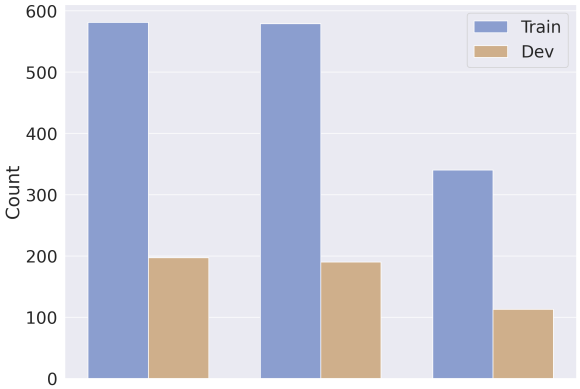
and sentiment analysis in Arabic, posted as Subtask 1 and 2 respectively. The geographical distribution of the countries covered in the dataset of Subtask 1 is shown in the map in Fig 1. Dialect distributions in the training and development sets vary based on the countries. In the datasets, for each country, we present the count of tweets included in both training and development sets, as seen in Fig 2. In the general collections of the tweets, there was no MSA taken into consideration in both datasets provided, rather just spoken dialects in the various countries as used in NADI-2021 (Abdul-Mageed et al., 2021). In Subtask 1, the test set is divided into TEST-A which includes 18 dialects on the country level, and TEST-B which covers k country-level dialects, where k is kept unknown. In Subtask 2 there's only one set for the test as shown in Table 1.

2.1 Subtask 1: Arabic Dialect Identification Dataset

The country-level dialect identification task is a multi-class classification problem that aims to identify and categorize which country, province, or dialect an Arabic tweet comes from. This task has a training dataset covering about 18 dialects of Arabic tweets summing up to 20K tweets. Subsets of both Subtasks data are in Table 1.

2.2 Subtask 2: Sentiment Analysis Dataset

The second task (subtask2) is a sentiment analysis problem aimed at determining whether an Arabic tweet is either positive, negative, or neutral. NADI-2022 provided a total of 5,000 tweets covering 10 Arab countries involving both MSA and DA. These tweets are manually labeled with tags from the set positive, negative, neutral.



	Positive	Negative	Neutral
Train	581	579	340
Dev	197	190	113

Sentiment

Figure 3: Sentiment distribution for the TRAIN and DEV data subsets of Subtask 2.

3 System Development

In recent advancements of NLP, models with state-of-the-art (SOTA) results like SMART-RoBERTa Large (Jiang et al., 2019) have shown that using transformer models, it is reasonable to expect SOTA performance in tasks such as sentiment analysis (Aghajanyan et al., 2021) and question answering (Yamada et al., 2020). The SOTA leaderboard of SST-2 dataset (Socher et al., 2013) shows clearly that transformer models are currently the best for text classification with almost the top 50 models using transformer architecture[2]. We use the same approach in solving both Subtask 1 and Subtask 2 of the shared task. We used pre-trained transformer models in all experiments.

Domain-specific transfer learning and fine-tuning of transformer models is proven to be more robust by Issifu et al. (Çelkmasat et al., 2022) and Bayrak et al. (Akça et al., 2022), (Bayrak et al., 2022). They fine-tuned transformer models on Biomedical and Turkish law datasets respectively to achieve results better than their original general transformer models. Better performance obtained in these works are accredited to

- General domain pre-training: when the transformer model is being trained on a huge corpus collected from various sources.
- Domain-specific LM fine-tuning: a continuation of the pre-training but with a relevant domain corpus instead of the general one for getting more accurate token representations.
- Task-specific fine-turning: done using a supervised training dataset.

For a more robust performance of the system, we adopt Arabic language domain-specific pre-trained transformer models, AraBERT and MARBERT. These pre-trained models gained SOTA in SA on AJGT,HARD,LABR (2-class, unbalanced) datasets.

3.1 Pre-processing

Since social media is a platform where everyone can showcase their opinions, text data from Twitter (especially in Arabic) comes in raw, unclean, and with variations. Noise in tweets commonly comes from the use of slang words, non-ASCII characters like emoji, spelling mistakes, URLs, etc. (Wadhawan, 2021).

The measures we took to clean and prepossessed the data are adopted from AraBERT [3] as follows; 1) Removing HTML markup tags, eliminating non-text and out-of-context tokens. 2) Replacing URLs, Emails and user mentions in Twitter with the tokens: [رابط], [بريد], and [مستخدم] respectively [4]. 3) Stripping Tashkeel (diacritics) and Tatweel (elongation). Tashkeel is the use of short vowel/consonant marks that manifest a word's pronunciation. E.g. the word الْعَرَبِيّة becomes العربية. Tatweel is adding horizontal stroke between two Arabic letters to elongate its visual appearance. For example, the word كـــــلمة becomes كلمة after striping tatweel. We stripped these two (Tashkeel and Tatweel) to reduce the lexical sparsity of the words. They do not constitute the actual word's body and are not usually used in tweets. 4) For the same reasons mentioned, we insert white space before and after all non-Arabic digits. 5) Mapping all the Hindi numbers (٠ ١ ٢ ٣ ...) to Arabic numbers (0 1 2 3 ...). 6) Similarly, we reduced the repetition of characters to 2 characters by replacing the repeated characters with 2 of its kind. For example, the word مرررررررة becomes مرة. This helps normalize the words used in the tweets. 7) Replacing the slash / with a dash – since it is absent in the vocabulary of AraBERT. 8) We do not cancel out

[2]Papers with code SOTA models https://paperswithcode.com/sota/sentiment-analysis-on-sst-2-binary. 2014 -2022 results

[3]https://github.com/aub-mind/arabert

[4]Steps 1 and 2 of the pre-processing are redundant in our setting, they're already replaced in Subtask 1 and 2 data.

all the emojis; instead, we apply a normalization used in AraBERT, this helps eliminate the sparsity of the emojis.

3.2 Arabic BERT-based Model

Arabic transformer language models MARBERT and AraBERT are based on the original BERT architecture (Devlin et al., 2018). AraBERT is trained on 23GB of Arabic text, making ∼70M sentences and 3B words, from Arabic Wikipedia, the Open Source International dataset (OSIAN) (Zeroual et al., 2019). MARBERT, however, is trained on 1B Arabic tweets, each tweet with at least 3 words. In our work we use AraBERT v0.2 Twitter-base [5] which is a further pre-training of AraBERT v02 on additional 60M Multi-Dialect tweets. We refer to this model in the result tables as *AraBERTtw*. We trained our models to classify Arabic language tweets into their various dialects on the country level using very selective hyper-parameters. To avoid local minima, overfitting, and related training issues, we adopt the setup and the hyper-parameters from the work of (Mosbach et al., 2020). We trained the model for 4 *epochs* with *batch size* of 16, and using ADAMW optimizer (Loshchilov and Hutter, 2017) with *learning rate* of $2e-5$, and *weight decay* $\lambda = 0.01$. The bias correction terms are set as $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e-6$ with the use of gradient clipping and a *warmup ratio* of 10% of the total training data.

We use the same setting for training the model for Subtask 2. The difference in ASA model is the number of neurons in the last classification layers is changed to 3, the number of classes in the problem.

4 Results

To evaluate the results, we use the official metrics defined by the shared task: Macro-Averaged F-score for Dialect ID (Subtask 1) and Macro-F1-PN score -neglecting the neutral class- over the positive and negative for the sentiment classification (Subtask 2).

In Table 2 we report the baseline model in the first row that is trained using similar hyper-parameters except the arbitrarily chosen: *5 epochs*, *batch size* of 32 *warmup steps=500*, *learning rate=5e-5* and optimizer's $\epsilon = 1e-8$. The results are also reported in the shared task's leader-board as *giyaseddin* team. In the same table, we show the macro F-1 score with the accuracy for the experimented models against each of the DEV, TEST-A, and TEST-B set provided by the shared-task for Subtask 1. Similarly, the test results of Subtask 2 are presented in Table 3. Our best-performing model (MARBERTv2) achieved 33.89% F-1 score in Subtask 1 TEST-A for Dialect ID with pre-processing. This is also the best-performing model in the average scores of both test sets with 25.54%. The model with the best generalization on TEST-B with a *k* number of countries, is MARBERT with pre-processing with F-1 score of 17.51%. For ASA in Subtask 2, MARBERT trained without the use of pre-processing performed better on the test set than other models with the best Macro-F1-PN score of 74.29%. We see from the confusion matrix of the best model on DEV subset of Subtask 1 in Fig 2 that dialects with a high number of examples are classified better than dialects with a lower number.

Model	PP	Dev		Test-A		Test-B	
		F-1	Acc.	F-1	Acc.	F-1	Acc.
AraBERTtw$_{init}$	Yes	30.47	48.49	30.55	47.65	14.30	29.92
AraBERTtw	No	30.16	49.13	30.71	48.17	14.98	30.46
AraBERTtw	Yes	30.80	49.56	31.30	48.57	15.35	30.19
MARBERT	No	32.56	50.30	32.20	49.41	16.04	32.56
MARBERT	Yes	32.86	50.03	31.66	49.18	**17.51**	**35.14**
MARBERTv2	No	**33.18**	**52.27**	33.40	51.24	17.08	34.33
MARBERTv2	Yes	32.19	51.22	**33.89**	**51.66**	17.19	34.87

Table 2: F-1 Macro and Accuracy results of different models on Subtask 1. PP column indicates using pre-processing before training.

Model	PP	Dev		Test	
		F1-PN	Acc.	F1-PN	Acc.
AraBERTtw$_{init}$	Yes	72.24	67.00	71.43	65.80
AraBERTtw	No	**72.58**	**67.60**	71.21	65.80
AraBERTtw	Yes	72.07	66.80	71.43	65.80
MARBERT	No	71.44	66.00	**74.29**	**69.00**
MARBERT	Yes	72.14	67.20	73.14	67.60
MARBERTv2	No	71.91	65.80	74.25	68.70
MARBERTv2	Yes	68.42	62.40	74.06	68.53

Table 3: Accuracy and Macro F-1 of Negatives and Positives results of different models on Subtask 2. PP column indicates using pre-processing before training.

5 Discussion and Future Work

According to our experiments, we see that the pre-processing we used has a positive impact on Dialect Identification, unlike Sentiment Analysis. Initial results say that tokens and expressions that identify

[5] Model names on HuggingFace hub are *aubmindlab/bert-base-arabertv02-twitter* and *UBC-NLP/MARBERT*.

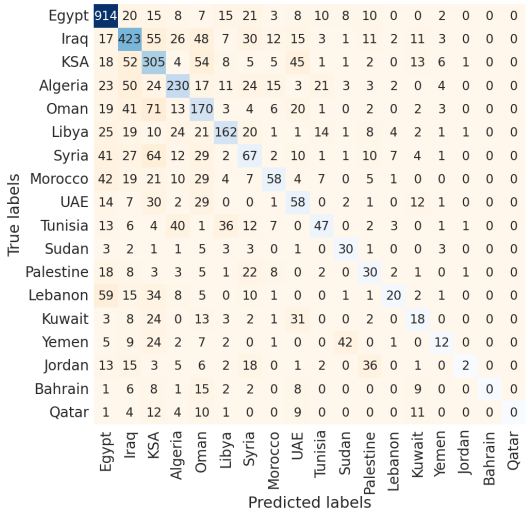


Figure 4: Confusion matrix of the predictions of the best performing MARBERT with no pre-processing against DEV subset of Subtask 1.

a dialect are not correlated with the processed (replaced or removed tokens like emojis, repetitions, etc.) so we see better results with them processed. In ASA, on the other hand, we see that they have an opposite effect on the classification. In general, MARBERT performs better on both subtasks even though its 2nd version performs better on the ARLUE benchmark (Abdul-Mageed et al., 2020a). In Fig 5 we see the Kernel Density Estimation (KDE) line of failing predictions of the model on ADID lies slightly to the left of the line of the successful predictions. This means that the probability to make correct predictions is higher when the sentence is longer.

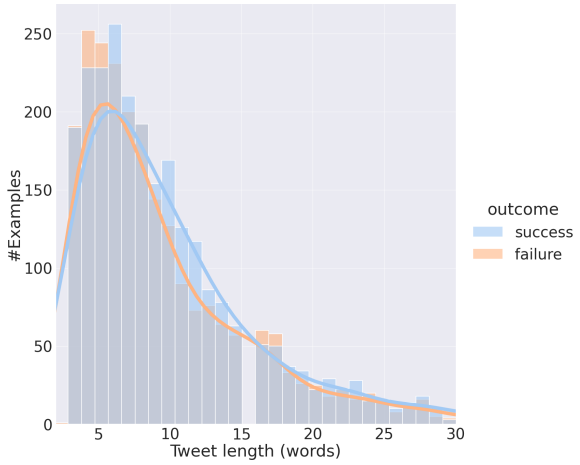


Figure 5: An overlapping histogram for both successful and failing predictions with respect to the word count (window from 1 to 30) taken from MARBERT with no pre-processing against DEV set of Subtask 1.

	Text	Label	Prediction	#Words	Comment
0	مين مدايجك بس	Egypt	Iraq	3	Mislabelled
1	الجو حلو لي درجة ودي أفل شعري	Syria	KSA	7	Mislabelled
2	احسن عدنان فالدنيا	Tunisia	Oman	3	Unclear / both
3	تعالي خاص اذا ممكن	KSA	Iraq	4	Unclear / both
4	يستاهل ابو ماجد	Syria	Oman	3	Mislabelled
5	ياارب الامتحانات تخلص ! ! زهقت : (: (	Palestine	Egypt	10	Mislabelled

Table 4: Examples of instances that are mislabelled or unclear in Subtask 1.

In our analysis, we focus more on ADID, in which we still have to face the challenge of the highly correlated dialects such as Palestinian with Jordanian, or Saudi Arabian with Emirati or Omani. Combining MSA with the dialects makes the problem harder and it is out of the scope of Subtask 1. Moreover, labeling such a dataset is hard to achieve without any confusion in the labels, even a human-level baseline might not be purely reliable. We present some of the examples that are either mislabelled or unclear in Table 4. Collecting more data can help in this problem, but focusing on increasing the quality of the data, e.g. using active learning methods. Platform bias is clear in the tweet nature, which could be considered as a limitation for the model in different use cases. The models experimented on are not bias-free, even though the used model is pre-trained on multi-source corpus keeps they're still prone to social biases (Garrido-Muñoz et al., 2021). To increase the performance of our classifier models, we intern to leverage new models from different architectures like (Nagoudi et al., 2022), since it achieved SOTA on Arabic NLU tasks. We also plan to use an ensemble model like (AlKhamissi et al., 2021), for it has a potential improvement gap in the overall performance.

6 Conclusion

This study is focused on two main tasks: Arabic Dialect Identification and Arabic Sentiment Analysis based only on the text of the tweets. We demonstrate the nuanced variations between the models before and after applying language-specific preprocessing, besides using domain-adapted models pre-trained on Arabic corpus. Understanding these variations requires knowledge of the nature of different data collections that should be considered. We conclude that it is important to choose the set of hyper-parameters of fine-tuning carefully to obtain a more stable and better generalization. Finally, we found that MARBERT outperforms other models in the generalization capability in both subtasks.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020a. Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020b. 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, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The Third Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Seven Arabic Natural Language Processing Workshop (WANLP 2022)*.

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. *arXiv preprint arXiv:2101.11038*.

Onur Akça, Gıyaseddin Bayrak, Abdul Majeed Issifu, and Murat Can Ganz. 2022. Traditional machine learning and deep learning-based text classification for turkish law documents using transformers and domain adaptation. In *2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6. IEEE.

Badr AlKhamissi, Mohamed Gabr, Muhammad El-Nokrashy, and Khaled Essam. 2021. Adapting marbert for improved arabic dialect identification: Submission to the nadi 2021 shared task. *arXiv preprint arXiv:2103.01065*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Gıyaseddin Bayrak, Muhammed Şakir Toprak, Murat Can Ganiz, Halife Kodaz, and Ural Koç. 2022. Deep learning-based brain hemorrhage detection in ct reports. In *Challenges of Trustable AI and Added-Value on Health*, pages 866–867. IOS Press.

Gökberk Çelkmasat, Muhammed Enes Aktürk, Yunus Emre Ertunç, Abdul Majeed Issifu, and Murat Can Ganz. 2022. Biomedical named entity recognition using transformers with bilstm+ crf and graph convolutional neural networks. In *2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2019. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Anshul Wadhawan. 2021. Dialect identification in nuanced arabic tweets using farasa segmentation and arabert. *arXiv preprint arXiv:2102.09749*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

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.