Arabic Dialect Identification with a Few Labeled Examples Using Generative Adversarial Networks

Mahmoud Yusuf **Marwan Torki** **Nagwa El-Makky**
Computer and Systems Engineering Department
Alexandria University
Alexandria, Egypt
{es-mahmoud.yusuf1217, mtorki, nagwamakky}@alexu.edu.eg

Abstract

Given the challenges and complexities introduced while dealing with Dialect Arabic (DA) variations, Transformer based models, e.g., BERT, outperformed other models in dealing with the DA identification task. However, to fine-tune these models, a large corpus is required. Getting a large number high quality labeled examples for some Dialect Arabic classes is challenging and time-consuming. In this paper, we address the Dialect Arabic Identification task. We extend the transformer-based models, ARBERT and MARBERT, with unlabeled data in a generative adversarial setting using Semi-Supervised Generative Adversarial Networks (SS-GAN). Our model enabled producing high-quality embeddings for the Dialect Arabic examples and aided the model to better generalize for the downstream classification task given few labeled examples. Experimental results showed that our model reached better performance and faster convergence when only a few labeled examples are available.

1 Introduction

While Arabic is the first language of most of the Middle East and North Africa (MENA) region, different countries have different dialects of Arabic. These Dialect Arabic (DA) forms are all different from the Modern Standard Arabic (MSA). MSA is used in formal writing and speaking situations, like academia and media. In contrast, DA is the language of the street. DA is spoken by people informally in their daily conversations and on social media platforms.

The task of automatically identifying the dialect of Arabic is beneficial since it contributes to many downstream tasks and applications, such as Speech Recognition and Machine translation.

Some Arabic Dialects are very close to each other (e.g. Levantine region dialects such as Lebanese and Syrian). On the other hand, other dialects are significantly different (e.g. Egyptian and Moroccan dialects) like in Table 1. This similarity is affected by the geographic locations of the countries and their respective dialects.

Class	Example
English	Excuse me, can you take a picture of me?
MSA	معذرةً، هل يمكنك أن تلتقط صورةً لي؟
Egyptian	لا مؤاخذة، ممكن تصورني؟
Lebanese	عن اذنك، فيك تاخدلي صورة؟
Moroccan	سمح ليا، واخا تصورني عافاك؟
Qatarian	لو سمحت، ممكن تصورني؟

Table 1: Comparison between MSA and DA variations for the same sentence

Similar dialects are one of the main challenges in the Dialect Identification task. In addition, further challenges are introduced due to the lack of balanced datasets for DA.

Some datasets are imbalanced with few classes dominating the whole dataset. Figure 1 illustrates the classes distribution in the NADI (Abdul-Mageed et al., 2021b) 2021 dialect dataset. Some other datasets suffer from a limited number of dialects. Another problem is mislabeled DA examples due to noise in the labeling procedure, e.g., depending only on the geographic location.

Given these challenges, getting a large corpus of labeled DA examples for all Arab countries is challenging and time-consuming. These complexities represent a major challenge in the Arabic Dialect Identification task. We aim to improve the transformer-based models, i.e., BERT (Devlin et al., 2019), that handle the task given the lack of large enough datasets.

In this paper, we extend BERT-based models, ARBERT and MARBERT (Abdul-Mageed et al., 2021a), with a generative adversarial setting using

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 196–204
November 20–23, 2022. ©2022 Association for Computational Linguistics

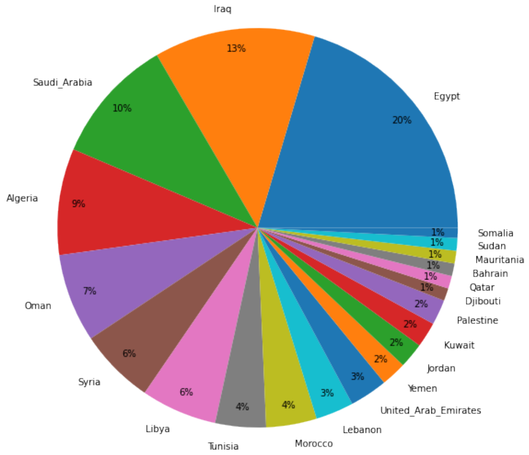


Figure 1: NADI 2021 DA training set label distribution. Only 4 classes represents more than 50% of the dataset

Semi-Supervised Generative Adversarial Networks (SS-GAN) (Salimans et al., 2016). This setting makes use of a set of unlabeled data, which can easily be obtained, to better generalize for the Arabic Dialect Identification task given a few labeled examples. Semi-supervised learning with adversarial nets was previously used for some tasks and languages, but to the best of our knowledge, it has not been used for Arabic Dialect Identification before.

The contributions of this work are:

- Adopting the semi-supervised setting using GAN (Goodfellow et al., 2014) over ARBERT and MARBERT models. This drastically reduces large dataset requirements for the DA identification tasks. Our models outperformed BERT-based models using very small training datasets.

- We study the classification of Dialect Arabic against very small training datasets using our extended GAN models. The training sets were sampled from 4 different Arabic datasets: QADI (Abdelali et al., 2021), NADI 2021 (Abdul-Mageed et al., 2021b), ArSarcasm (Bashmal and AlZeer, 2021) and AOC (Zaidan and Callison-Burch, 2011). The sample sizes varied from 0.01% to 10% of the full training dataset.

- We applied a 2-stage setup, training the GAN extended model for some epochs and then, having a second stage of BERT-based model training. These early GAN epochs boosted BERT-based model convergence speed and performance results. The 2-stages experiment outperformed the BERT-based models for the same number of epochs.

The rest of the paper is organized as follows: in section 2, we discuss the related work in the Dialect Arabic Identification task and variations of BERT-based models. In section 3, we illustrate the system components and model architectures. We show the conducted experiments and their results, in section 4. Finally, we give a brief conclusion based on our work and the obtained results.

2 Related Work

2.1 Evolution of DA Datasets

The main challenge in Arabic Dialect Identification is the rarity of high-quality labeled datasets that represent all Arabic dialects. Recently, some datasets were introduced. However, most of them have limitations as will be shown in the next paragraphs.

The Arabic Online Commentary AOC (Zaidan and Callison-Burch, 2011) introduced rich dialectal content based on online commentary by readers of online famous Arabic newspapers. The dataset is labeled with MSA and three regional dialects: Egyptian, Gulf, and Levantine. Despite the relatively large corpus, country-level dialects are not represented in this dataset, causing the lack of many DA variations. In addition, social media data, e.g., Twitter became a richer source of DA with almost all variations available.

Dialect Identification shared tasks impassioned the Arabic DA work. The Multi Arabic Dialects Application and Resources (MADAR) (Bouamor et al., 2019) project introduced a parallel corpus that was used in MADAR shared task kin 2019. However, the examples were a translation of the Basic Traveling Expression Corpus (BTEC)(Takezawa et al., 2007). Hence, the data examples were short, and unnatural, and do not realistically represent the target dialects.

ArSarcasm (Bashmal and AlZeer, 2021) is a dataset built relying on popular Arabic Sentiment Analysis datasets, SEMEVAL 2017's (Rosenthal et al., 2017) and ASTD (Nabil et al., 2015). ArSarcasm was also annotated for dialects due to the challenges urged by dialectal variations. ArSarcasm adapted a manual annotation process with strict guidelines to guarantee the quality of the annotations. However, most of the data is either in

MSA or Egyptian dialect, and hence, the dataset suffers the rare presentation of other dialects.

The First Nuanced Arabic Dialect Identification Shared Task (NADI 2020) (Abdul-Mageed et al., 2020) included sub-tasks for the country-level and province-level DA identification. The NADI 2020 dataset covers 21 Arab countries, collected from the Twitter domain. While this data was naturally extracted from tweets, it was unbalanced with few classes dominating the dataset. In addition, the labeling criterion depends only on the user's geographic location which introduced wrong labels that prevented deep learning models from better generalization. The Second Nuanced Arabic Dialect Identification Shared Task (NADI 2021) (Abdul-Mageed et al., 2021b) dataset was based on similar collecting and labeling methods and hence has the same limitation. NADI 2021 introduced 2 new subtasks: country and province level MSA identification.

QADI (Abdelali et al., 2021) is a recent tweet dataset with a variety of country-level Arabic Dialects, with highly accurate labels and mostly evenly distributed classes. QADI represented 18 different Arab countries. QADI conducted the Dialect Identification experiments using different machine learning and deep models.

2.2 Transformer based models for DA Identification

BERT model variants showed impressive results on text classification and other NLP tasks. (Mansour et al., 2020) fine-tuned Multilingual BERT (mBERT) (Devlin et al., 2019) for the NADI 2020 (Abdul-Mageed et al., 2020) shared task on DA Identification. AraBERT (Antoun et al., 2020) pretrained BERT for Arabic. AraBERT outperformed multilingual BERT model in Arabic NLP tasks and became the state-of-the-art model for these tasks in 2020.

(Abdul-Mageed et al., 2021a) introduced ARBERT and MARBERT, which are very powerful transformer-based models trained on large and massive Arabic datasets from different domains. MARBERT was pre-trained on dialectal Arabic which helped for better generalization and more powerful results on diverse tasks. ARBERT and MARBERT models achieved state-of-the-art results in different Arabic downstream NLP tasks. In Dialect Identification, both models outperformed AraBERT and other previous models in all popular DA datasets.

In (AlKhamissi et al., 2021), the authors targeted the NADI 2021 shared task using a MARBERT model and their submission was ranked the first for this shared task. However, the model still did not overcome being biased toward the dominating classes in the training dataset.

2.3 Semi-Supervised Models

Adversarial settings were also introduced on top of BERT-based models to generate different examples, which help in various text classification tasks. BAE(Garg and Ramakrishnan, 2020) presented a model for adversarially generating examples through perturbations based on the BERT Masked Language Model. GAN-BERT (Croce et al., 2020) extended fine-tuning BERT-based models with unlabeled examples using a Generative Adversarial Network (GAN)(Goodfellow et al., 2014) that helped train models with few labeled examples and generally enhance BERT-based model classification capabilities.

3 Adopted Model

3.1 Motivation

One of the key challenges in Arabic Dialect Identification research is insufficient labeled datasets. Many datasets don't fairly represent all classes, i.e., imbalanced datasets. Other datasets suffer from labeling noise.

Although having a sufficient amount of unlabeled data is extremely easy, e.g. crawling tweets, the process of labeling these examples with correct labels is expensive, impractical, and time-consuming. Some easier methods are adopted while labeling such data, e.g., depending on Twitter users' geographic location or account metadata. Unfortunately, these methods are not accurate to representing correct classes and lead to many miss-labeled examples.

Arabic is a highly inflected and derivational language. The inflection and derivation rules may change from one Arabic Dialect to another. Moreover, the same word might have totally different meaning in different Arabic Dialects. For instance, the word مهضوم (Mahdoum) meaning in MSA and Egyptian dialect is digested, which is used to describe food. While in Levantine Arabic (dialects spoken in Syria, Lebanon, Jordan and Palestine), its meaning is joyful or delightful, and used to describe persons. These specific characteristics of

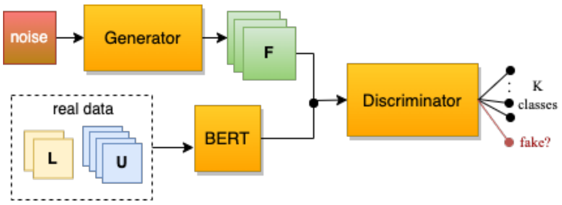


Figure 2: GAN-BERT model architecture. The discriminator D input is: labeled L and unlabeled U examples vector representations computed by BERT, in addition to the fake examples F generated by the generator G given noise input. (Adapted from (Croce et al., 2020))

Arabic Dialects make it challenging to generate human-like examples.

Traditional methods like Data Augmentation are usually used to generate more examples to solve for the rarity of available training examples. However, these methods aren't able to generate human-like real examples in our case. Traditional data augmentation like word swapping fail to generate meaningful examples. Augmenting examples by changing words to their synonyms is also inappropriate due to rarity of synonyms resources for Arabic dialects. Similarly, Back Translation always translate examples back to Modern Standard Arabic (MSA) which leads to losing the dialectal nature of the examples.

In contrast, Semi-Supervised Generative Adversarial Networks (SS-GAN) (Salimans et al., 2016) can act as an additional source of information in a semi-supervised setting. SS-GAN can capture the characteristics of the training examples and generate similar examples that are nearly indistinguishable from the real training examples.

3.2 Model Architecture

Our work is mainly based on GAN-BERT model (Croce et al., 2020) that enriches the BERT fine-tuning process with an SS-GAN perspective. Semi-Supervised GAN (SS-GAN) (Salimans et al., 2016) is a Generative Adversarial Network (Goodfellow et al., 2014) with a multi-class classifier as its Discriminator. Rather than learning to discriminate between only two classes (actual and fake), it learns to distinguish between K + 1 classes, where K is the number of classes in the training dataset, plus one for the Generator's fake generated examples. The Generator input is a vector of random noise, The Generator's objective is to generate fake examples that are indistinguishable from the real dataset examples.

The Discriminator has 3 inputs: fake examples generated by the Generator (x*), real unlabeled examples (x), and real labeled training examples (x, y), with y denoting the label for the given example x.

In this work, we extend BERT-based models using SS-GAN. We use BERT-based models pretrained on Arabic datasets, namely ARBERT and MARBERT (Abdul-Mageed et al., 2021a), and adapt the fine-tuning by adding task-specific layer in addition to the SS-GAN layers to enable semi-supervised learning.

Given an input example, $e = (t_1, t_2, , .., t_n)$, BERT model's output is an $n + 2$ vector representation in R^d, i.e., $(h_{CLS}, h_1, h_2, .., h_{SEP})$. As advised in (Devlin et al., 2019), h_{CLS} is used a the example sentence embedding for the identification task.

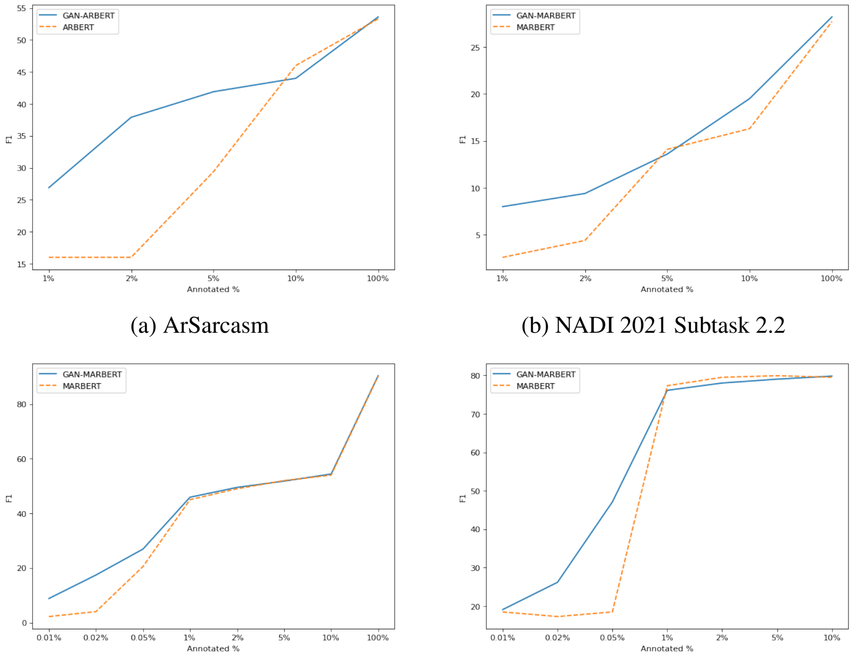
The generator G is a Multi-Layer Perceptron (MLP) that takes an input of a 100-dimensional random noise vector drawn from Normal Distribution $N(\mu, \sigma^2)$ and outputs a vector $h_{fake} \in R^d$. As shown in Figure 2, the discriminator D receives input $h_* \in R^d$ which can be the fake generator output h_{fake} or examples from the real distribution $hCLS$ (labeled or unlabeled). The Discriminator D is another Multi-Layer Perceptron (MLP) where its last layer is a softmax layer that outputs a $k + 1$ vector of logits. True examples from the real distribution are classified into the (1, ..., k) classes, while generated fake samples are classified into the additional $k + 1$ class.

When updating the discriminator, BERT-based model weights are also changed in order to consider both labeled and unlabeled examples to better fine-tune their inner representations. At evaluation the generator is discarded while keeping rest of the model, which means no additional cost at inference time compared to standard BERT-based models.

4 Experimental Results

4.1 Semi-Supervised Setting: GAN-MARBERT and GAN-ARBERT

In this section, we evaluate the impact of GAN-BERT-Based models, namely GAN-MARBERT and GAN-ARBERT over the Arabic Dialect Identification task under different training environments, i.e., number of dialectal classes and number of labeled training examples. We compare our proposed method with MARBERT / ARBERT which are the existing methods that achieve state-of-the-art results in the Arabic Dialect Identification task. With



(a) ArSarcasm

(b) NADI 2021 Subtask 2.2

(c) QADI

(d) AOC

Figure 3: Learning curves for the Dialect Identification task against the 4 datasets. We run all the models for 10 epochs with the same learning rate 2e-5. The same sequence length of 40 was used in all experiments.

very few training examples, we assess our model in the DI task against the following datasets: QADI (Abdelali et al., 2021) that has 18 classes, NADI 2021 Subtask 2.2 (Abdul-Mageed et al., 2021b) that has 21 classes, ArSarcasm (Bashmal and AlZeer, 2021) that has 5 classes, and AOC (Zaidan and Callison-Burch, 2011) that has 4 classes.

We use the macro-F1 score as the evaluation metric for our models. The macro-F1 score is the standard evaluation metric in the dialect identification task.

As discussed in section 3, we extend BERT-based models with a generative adversarial setting. The generator G is an MLP with a single hidden layer activated by a leaky relu function. The generator G input is a random noise vector drawn from the Normal distribution $N(0, 1)$. The generator G output is a 768-dimensional vector that represents the fake generated examples. The discriminator D is another similar MLP with a final softmax layer for the final dialect classification. We use a dropout rate of 0.2 after the hidden layer in both G and D.

We chose the best performing BERT-based pre-trained model as the base model for each dataset, as reported in (Abdul-Mageed et al., 2021a). For QADI, NADI, and AOC, the chosen base model is MARBERT. While for ArSarcasm, the base model is ARBERT.

We start training the models by sampling only 0.01% or 1% of the full training dataset, depending on the size of the dataset, in order to have a very small training set. The process is repeated with incremental larger training samples.

For the unlabeled examples, we use a set of $10K$ randomly sampled tweets from the unlabeled set provided in the NADI 2021 (Abdul-Mageed et al., 2021b) dataset.

The ArSarcassm (Bashmal and AlZeer, 2021) Dialect Identification task results are shown in figure 3a. The training dataset consists of 8438 examples, and the test dataset consists of 2111 examples, labeled with 5 dialect classes. The plot shows the macro-F1 scores of the GAN-ARBERT and ARBERT models. When 1% of the training data is used (around 85 examples), ARBERT almost diverges, while GAN-ARBERT achieves F1 of more than 25%. With 2% of the training data, GAN-ARBERT achieved F1 of 38%, obviously outperforming ARBERT. The same trend continued until 10% of the training data is used.

For NADI 2021 (Abdul-Mageed et al., 2021b) sub-task 2.2 dataset, similar outcomes were observed as shown in figure 3b. The dataset consists of 21000 training examples and 5000 test examples labeled with 21 dialect classes. NADI has a large number of classes with unbalanced training exam-

Sample Size	GAN-ARBERT	ARBERT
1%	**32.4**	20.5
2%	**37.9**	28.9
5%	43.7	**47**
10%	45.3	**48.5**

(a) ArSarcasm

Sample Size	GAN-MARBERT	MARBERT
1%	**11.2**	7.2
2%	13.3	**14.8**
5%	19.9	**20**
10%	20.8	**21.9**

(b) NADI

Sample Size	GAN-MARBERT	MARBERT
0.01%	**8.8**	2.2
0.02%	**17.4**	4
0.05%	**26.9**	20.5
1%	**45.9**	45
2%	**49.5**	49
5%	51.7	**52**
10%	**54.4**	54

(c) QADI

Sample Size	GAN-MARBERT	MARBERT
0.01%	**19.1**	18.5
0.02%	**26.2**	17.3
0.05%	**47.1**	18.5
1%	76.2	**78.7**
2%	78	**79.5**
5%	79	**79.9**
10%	**79.8**	79.5

(d) AOC

Table 2: Experimental results for the Semi-Supervised setting. The evaluation metric is Marco F1 score.

Sample Size	2-Stage	ARBERT
1%	**32**	20.5
2%	**38.1**	28.9
5%	45.7	**47**

(a) ArSarcasm

Sample Size	2-Stage	MARBERT
1%	**10.9**	7.2
2%	**16.5**	14.8
5%	**20.3**	20

(b) NADI

Sample Size	2-Stage	MARBERT
0.01%	**7.8**	2.2
0.02%	**8.9**	4
0.05%	**23**	20.5

(c) QADI

Sample Size	2-Stage	MARBERT
0.01%	**20.2**	18.5
0.02%	**20.9**	17.3
0.05%	**43.9**	18.5

(d) AOC

Table 3: Experimental results for the 2-stages setup. The evaluation metric is Marco F1 score.

ples distribution. GAN-MARBERT outperforms the MARBERT model in most settings. When 1% of the training set is used (210 examples), GAN-MARBERT achieves more than 3 times the F1 score obtained by MARBERT, GAN-MARBERT achieves F1 of 8% while MARBERT achieves F1 of 2.8%. The same trend continues with different sample sizes. The semi-supervised setting shows performance improvement over MARBERT for most of the sample sizes.

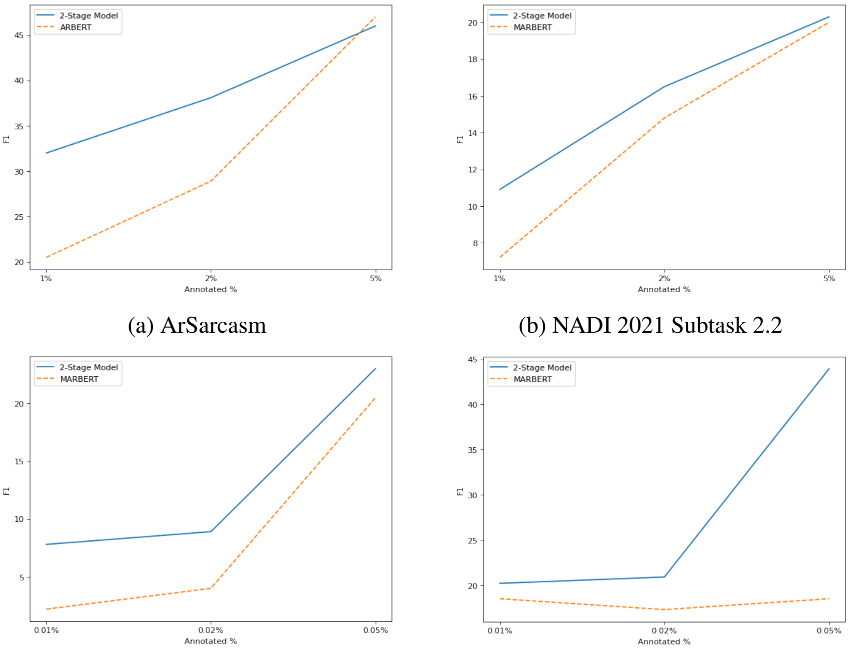
The observations were confirmed against QADI (Abdelali et al., 2021) dataset in figure 3c. QADI is the largest dataset used in these experiments with 367,353 training examples and 3304 test examples labeled with 18 dialects classes. QADI fairly represents most of the dialect classes and guarantees clean and correct labels. However, the same trend was shown in small training sample sizes. Using 0.01% (37 examples) and 0.02% (74 examples) of the training dataset, GAN-MARBERT achieves more than 4 times the macro-F1 score obtained by MARBERT model for the corresponding number of examples. Noticeable improvements in the F1 score continued until 2% of the training set is used.

Finally, we evaluate the models against AOC (Zaidan and Callison-Burch, 2011) dataset, which consists of 86,542 training examples and 10,812 test examples, labeled with 4 classes. For 0.02% of the training set (only 17 examples), GAN-MARBERT obtains F1 of more than 26% while MARBERT got 17% F1. When using a 0.05% of the training set (184 examples), GAN-MARBERT achieves F1 of 47% while MARBERT only got F1 of 18%, i.e, more than 2.5X F1 improvement. For larger training sample sizes, both models performed similarly.

The experimental results scores against different training dataset sample sizes are shown in Table 2

4.2 Two-Stages Setup: Using a BERT-based model after the GAN-BERT

In this setup, we evaluate a 2-stages setup. The first stage is training the BERT-based model with the GAN extension for 5 epochs. In the second stage, the GAN module is eliminated and the BERT-based model is trained for another 5 epochs. With



(a) ArSarcasm

(b) NADI 2021 Subtask 2.2

(c) QADI

(d) AOC

Figure 4: 2-stages experiments results. We used MARBERT as the base model for NADI, QADI and AOC datasets, while using ARBERT for ArSarcasm. Each experiment consists of 10 epochs. In the 2-stage experiments, we train the base model extended with GAN component for 5 epochs, then eliminate the GAN component and train the base model alone for another 5 epochs.

smaller training set samples, the first stage gave a performance boost to the overall model result when compared to the BERT-based model alone.

Figure 4 shows the experiments results. In both setups, we use the same learning rate $2e - 5$ and sequence length 40. For QADI and AOC datasets, we used 0.01%, 0.02%, and 0.05% of the annotated samples. For NADI and ArSarcasm, we used a1%, 2%, and 5% of the training dataset.

The experiment showed that adding the first stage with the semi-supervised setting helped the base model to better generalize for a few labeled examples and to converge faster.. Overall, the 2-stages setup outperformed the base model.

For ArSarcasm (Bashmal and AlZeer, 2021) dataset, figure 4a shows how the 2-stages setup achieves higher scores and faster convergence with smaller sample sizes. For example, when using only 1% of the training set, the 2-stages setup achieves F1 of 32, while ARBERT achieves only F1 of 20.5. Similar outcomes were obtained for NADI (Abdul-Mageed et al., 2021b) dataset in figure 4b. When 1% of the training set is used, the 2-stages setup achieves F1 of 10.9, compared to 7.2 by MARBERT. For QADI (Abdelali et al., 2021) dataset, figure 4c confirms the same outcomes. When only 0.01% of the training sample is used, the 2-stages setup achieves more than 3 times the F1 score obtained by MARBERT. The 2-stages setup achieves F1 of 7.8 compared to F1 of 2.2 by the MARBERT model. The trend continues with other sample sizes, with 0.02% of the training set, the 2-stages setup achieves F1 of 8.9 compared to 4 by MARBERT. Finally, for AOC (Zaidan and Callison-Burch, 2011) dataset, the 2-stages setup converges way faster than MARBERT as shown in figure 4d. With only a 0.05% training sample, the 2-stages setup achieves more than 2 times the F1 obtained by MARBERT. It achieves F1 of 43.9 compared to 18.5 for MARBERT.

The experimental results scores against different training dataset sample sizes are shown in Table 3

5 Conclusion

One of the main challenges of the Arabic Dialect Identification task is the rarity of high-quality labeled examples. This paper addresses this problem by adopting adversarial training to allow semi-supervised learning. it applies this approach to two BERT-based models, namely, MARBERT and ARBERT. Experimental results show that the GAN extension improves the performance of the BERT-based models, given a few labeled examples. The paper also introduces a 2-stages setup, where it trains the base model extended with GAN component for 5 epochs, then eliminate the GAN component and train the base model alone for another 5 epochs. Using very small training sets, the adopted approach helps the base model for better generalization and faster convergence, with no additional cost at inference time.

Adding SS-GAN module on top of BERT-based models, empirically showed enhancements in performance and faster convergence given a few labeled examples of the datasets, which validates our hypothesis.

References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, et al. 2021a. Arbert & marbert: Deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105.

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.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021b. Nadi 2021: The second nuanced arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259.

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. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 260–264.

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.

Laila Bashmal and Daliyah AlZeer. 2021. Arsarcasm shared task: An ensemble bert model for sarcasmdetection in arabic tweets. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 323–328.

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

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. Gan-bert: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119.

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.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Moataz Mansour, Moustafa Tohamy, Zeyad Ezzat, and Marwan Torki. 2020. Arabic dialect identification using BERT fine-tuning. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 308–312, Barcelona, Spain (Online). Association for Computational Linguistics.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2515–2519.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing,* *Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324.

Omar Zaidan and Chris Callison-Burch. 2011. The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41.