Mavericks at NADI 2023 Shared Task: Unravelling Regional Nuances through Dialect Identification using Transformer-based Approach

Vedant Deshpande [*], **Yash Patwardhan**[*], **Kshitij Deshpande**[*],
Sudeep Mangalvedhekar[*] and **Ravindra Murumkar**[*]
Pune Institute of Computer Technology, Pune
{vedantd41, yash23pat, kshitij.deshpande7, sudeepm117}@gmail.com,
rbmurumkar@pict.edu

Abstract

In this paper, we present our approach for the "Nuanced Arabic Dialect Identification (NADI) Shared Task 2023". We highlight our methodology for subtask 1 which deals with country-level dialect identification. Recognizing dialects plays an instrumental role in enhancing the performance of various downstream NLP tasks such as speech recognition and translation. The task uses the Twitter dataset (TWT-2023) that encompasses 18 dialects for the multi-class classification problem. Numerous transformer-based models, pre-trained on Arabic language, are employed for identifying country-level dialects. We fine-tune these state-of-the-art models on the provided dataset. The ensembling method is leveraged to yield improved performance of the system. We achieved an F_1-score of 76.65 (11th rank on the leaderboard) on the test dataset.

1 Introduction

Dialects, which are variations of a language, often differ in their vocabulary, grammar, pronunciation, and occasionally even cultural quirks. The practice of identifying the particular dialect or regional variety of a language that is used in a text or speech sample is known as dialect identification. The goal of dialect identification is to categorize a text or speech into one of the many dialects or regional adaptations that may exist. For many NLP applications, including language modeling, speech recognition, and data retrieval, this task may be vital.

Arabic, with its plethora of dialects, is a rich language. However, many of these dialects are not studied in depth because of a dearth of monetary backing and available datasets. Arabic dialect identification can assist in perpetuating linguistic diversity by acknowledging and valuing various dialects. It contributes to addressing the gap between existing NLP techniques and the rich fabric of regional dialectal differences in a globalized setting. Rule-based strategies for Arabic dialect identification have given way to data-driven techniques, with a focus on machine learning, deep learning, and the creation of corpora of languages and datasets. The accuracy of dialect detection has risen significantly with the use of multilingual pre-trained models such as BERT and its derivatives.

This paper presents our approach for subtask 1: Country-level Dialect Identification, which poses a multiclass classification problem (Abdul-Mageed et al., 2023). Multiclass classification is a form of statistical modeling or machine learning problem where the objective is to classify data into more than two unique classes or labels. We aim to classify the tweets and map them into their respective dialect labels. We have demonstrated the use of various transformer-based models on the given Arabic data. The ensembling method has been leveraged to enhance the performance of the proposed system.

2 Related Work

Dialect detection in Arabic is an arduous task due to several factors, including the lack of a consistent spelling system, the medium's characteristics, and the scarcity of data. Surveys on deep learning and Natural Language Processing methods for processing Arabic data were presented in 2015 (Shoufan and Alameri, 2015) and 2017 (Al-Ayyoub et al., 2018), focusing on the identification of Arabic dialects. However, only 6 Arabic dialect classes had been examined until that time. The MADAR project was launched in 2018 to provide a large corpus of 25 Arabic city dialects (Bouamor et al., 2018). A study on the classification of dialects in 25 Arab cities used multi-label classification methods and examined a wide range of features, yielding promising results (Salameh et al., 2018). Employing supervised machine learning methods on Arabic NLP tasks was found to be a difficult

[*]Equal contribution

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 678–682

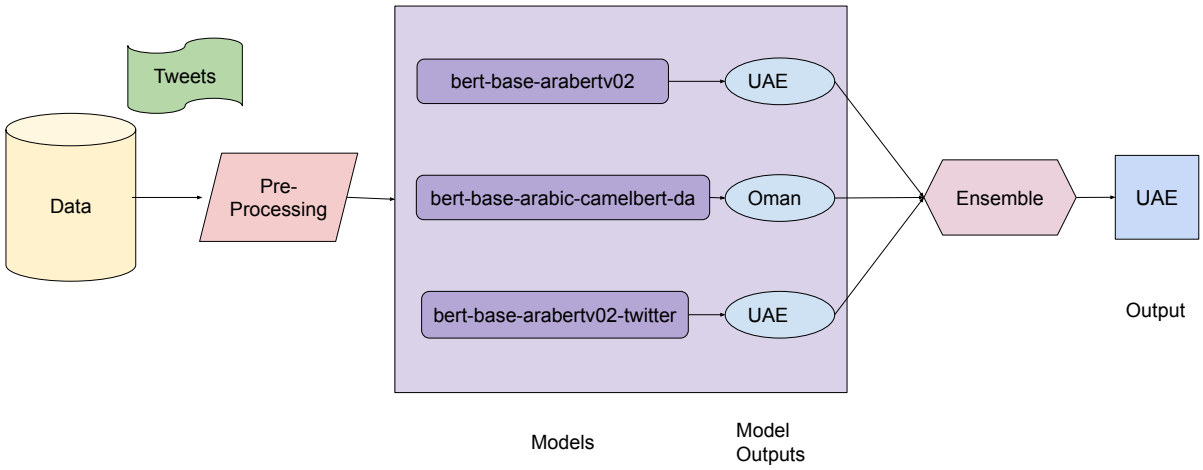

Figure 1: System architecture

feat because of the lack of resources in the Arabic language (El Mekki et al., 2020). As a result, scholars and researchers have introduced plenty of initiatives to make new datasets available and encourage more people to work in the field of Arabic NLP. One of the initiatives, Nuanced Arabic Dialect Identification (NADI) shared tasks, was started in 2020 which comprised country-level and province-level dialect detection (Abdul-Mageed et al., 2020). BERT (Bidirectional Encoder Representation from Transformers) (Devlin et al., 2019) models have been commonly used for these dialect detection tasks. A multilingual BERT model was pretrained on unlabeled tweets and fine-tuned for the classification task by Mansour et al. (2020). AraBERT was finetuned on an additional dataset produced by reverse translating the NADI dataset and employed for the dialect detection task by Tahssin et al. (2020). Furthermore, Gaanoun and Benelallam (2020) utilized ensembling methods and semi-supervised methods along with Arabic-BERT. A system comprising an ensembling of multiple models was created using MARBERT as the base model, which yielded promising results (AlKhamissi et al., 2021). In the NADI 2022 shared task, AlShenaifi and Azmi (2022) pretrained AraBERT model and BiLSTM model for dialect detection. Various models were combined and performance was enhanced using a combination of TF-IDF and n-grams. An ensembling of transformer-based models, predominantly using variations of MARBERT was employed for dialect detection as well as sentiment analysis, in the NADI 2022 shared tasks (Bayrak and Issifu, 2022), (Khered et al., 2022). (Oumar and Mrini, 2022) addressed the issue regarding an imbalance in the classes of the NADI dataset by using focal loss and employed various Arabic BERT-based models.

This paper proposes a system that employs an ensemble of transformer-based models, specifically variations of BERT for the classification task.

Dataset	Number of Samples
Training	18000
Development	1800
Testing	3600

Table 1: Dataset's training, development, and test split

3 Data

The dataset provided for subtask 1: Country-level dialect identification contains tweets. The given Twitter dataset comprises 18 dialects and a corpus of a total of 23400 tweets. The entire dataset is split into training (76.92%), development (7.69%), and test (15.38%). Additionally, datasets of previous years (Abdul-Mageed et al., 2020, 2021; Bouamor et al., 2018) are also provided for the training purpose. As shown in table 1, the training data has 18000 tweet samples, development data has 1800 samples and testing data has 3600 samples. The dataset contains features such as id, content, and label. Every sample's tweet content in the training dataset is labeled with its dialect. This subtask falls under the category of multi-class classification.

The provided dataset needed to be pre-processed before passing it to the model. We make use of regular expressions to remove "noisy" elements from the input texts. Texts like "USER", "NUM" and "URL" are removed from the input because they don't contribute additional information to the model's understanding.

4 System

The given subtask tackles the problem of country-level dialect identification. This comes under the umbrella of multi-class classification problems for which Language Models have been extensively used and have achieved impressive results. The models are trained for 10 epochs with a learning rate of 1e-5, a batch size of 32, and the AdamW optimizer. We experiment and use several language models and ensembling methods in our research, as shown in Figure 1.

4.1 AraBERT

Antoun et al. (2020) addresses how BERT models that have been pre-trained on a sizable corpus of a particular language, such as Arabic, do well on language comprehension tasks. They point out several such models, which are used in our study to help deliver cutting-edge outcomes for the Arabic language.

The 70 million phrases that make up the pre-training dataset, which is around 24 GB in size, are used to train the models. The news in the data covers a wide range of topics that is valuable for many downstream applications. The pre-training tasks that aid in the models' contextual knowledge of the input sequence include the Next Sentence Prediction Task and Masked Language Modelling Tasks. To demonstrate AraBERT's efficacy across diverse tasks and domains, it was tested on three NLP tasks: entity recognition, sentiment analysis, and question-answering.

Small adjustments have been made to the pre-training phases and parameters for the selected AraBERT model versions. AraBERT v1 or v0.1 are the original models, and v2 or v0.2 are the more recent versions with improved pre-processing and vocabulary. In addition to the dataset used for the other v0.2 models, AraBERTv0.2-Twitter-base is pre-trained with 60 million multi-dialect tweets. It possesses 136 parameters. Pre-trained examples for AraBERTv2-base include 207M instances with a sequence length of 512 and 420M examples with a sequence length of 128.

4.2 CAMeLBERT

Inoue et al. (2021) introduced the CAMelBERT model collection, which consists of more than eight pre-trained models for NLP tasks in Arabic. The parameters taken into consideration for the experiment were the task type, language variant, and size. Language models were provided in several variants, including classical Arabic (CA), dialectal Arabic (DA), and Modern Standard Arabic (MSA), with the DA variant being chosen for this study. The models were pretrained on variations of the MADAR dataset and NADI datasets for the Dialect Identification task. CAMelBERT was trained with the Adam optimizer and a learning rate of 1e-4. The pre-trained models are evaluated on five major tasks in NLP: Sentiment Analysis, Dialect Identification, POS tagging, Named Entity Recognition, and Poetry Classification.

5 Ensembling

Ensembling is a technique that integrates the output of multiple models to get the system's eventual outcome. For this, both statistical and non-statistical methods are employed. Ensembling is beneficial since it contributes to the production of results that are superior to those provided by the individual models.

We note that the "hard voting" ensemble strategy emerges as the most effective and precise among the many strategies used for ensembling. In hard voting, the final prediction is chosen based on the majority vote or the "mode" of all the predictions. It reduces the volatility in the outcomes and aids in strengthening the system's robustness.

Model	F_1 **Score**
AraBERTv02-Twitter-base	**77.03**
CAMeLBERT-DA	72.78
AraBERTv02-base	73.07
Ensemble - Hard Voting	**77.62**

Table 2: Results for Dialect Identification Task on the Development dataset

6 Results

This section discusses the results obtained by our system and analyses its performance. Table 2 and

Model	F_1 Score
AraBERTv02-Twitter-base	**75.17**
CAMeLBERT-DA	71.99
AraBERTv02-base	72.09
Ensemble - Hard Voting	**76.65**

Table 3: Results for Dialect Identification Task on the Test dataset

Table 3 depict our scores for the individual models used and the corresponding ensembled score on the development dataset and the test dataset respectively. The F_1 score is used as the official metric for scoring the systems.

AraBERTv02-Twitter-base outperforms the other models with an F_1 score of 77.03 on the development dataset and 75.17 on the test dataset. This performance demonstrates the benefits of utilizing a model that is pre-trained on a corpus similar to the one the task demands. AraBERTv02-Twitter-base is pre-trained on 60M multi-dialect tweets besides the usual datasets used for AraBERT models, giving it an edge over other models for this particular task. We select the ensemble-based system as our final approach since it produces outcomes with minimal variation and offers more stable predictions. This is justified by the superior performance of our system in the final evaluation stage. Our final system achieved an F_1 score of 76.65 on the test dataset.

7 Conclusion

This paper compares several transformer-based models on the task of Nuanced Arabic Dialect Identification (NADI). AraBERTv02-Twitter-base is found to outperform other models for this task. It achieves an F_1 score of 76.65. We use hard voting-based ensembling as the final approach for our system as it generates predictions that are stable while also improving the overall performance. With higher computational resources at hand, the performance of the system can be improved by training it for longer and by using bigger models for the system. Models that are specifically pre-trained on data that is similar to the data used in the task at hand can help enhance understanding and in turn, give better performance. We can also experiment with other suitable ensembling methods and gauge their efficiency for our task.

Limitations

Models used for this task are computationally heavy and require significant computing resources for inference. As a result in certain real-world applications where there are compute constraints, using the system may pose a challenge. The data used for evaluation and pre-training of the models mentioned may have been biased even though the quality of the data used is high. Thus, it may not accurately represent real-world scenarios.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, ElMoatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The Fourth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

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.

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.

Mahmoud Al-Ayyoub, Aya Nuseir, Kholoud Alsmearat, Yaser Jararweh, and Brij Gupta. 2018. Deep learning for arabic nlp: A survey. *Journal of Computational Science*, 26:522–531.

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, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nouf AlShenaifi and Aqil Azmi. 2022. Arabic dialect identification using machine learning and transformer-based models: Submission to the NADI 2022 shared task. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 464–467, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic lan-

guage 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.

Giyaseddin Bayrak and Abdul Majeed Issifu. 2022. Domain-adapted BERT-based models for nuanced Arabic dialect identification and tweet sentiment analysis. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 425–430, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Abdellah El Mekki, Ahmed Alami, Hamza Alami, Ahmed Khoumsi, and Ismail Berrada. 2020. Weighted combination of BERT and n-GRAM features for nuanced Arabic dialect identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 268–274, Barcelona, Spain (Online). Association for Computational Linguistics.

Kamel Gaanoun and Imade Benelallam. 2020. Arabic dialect identification: An Arabic-BERT model with data augmentation and ensembling strategy. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 275–281, Barcelona, Spain (Online). 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*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Abdullah Salem Khered, Ingy Yasser Hassan Abdou Abdelhalim, and Riza Batista-Navarro. 2022. Building an ensemble of transformer models for Arabic dialect classification and sentiment analysis. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 479–484, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Ahmed Oumar and Khalil Mrini. 2022. Ahmed and khalil at NADI 2022: Transfer learning and addressing class imbalance for Arabic dialect identification and sentiment analysis. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 442–446, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Abdulhadi Shoufan and Sumaya Alameri. 2015. Natural language processing for dialectical Arabic: A survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 36–48, Beijing, China. Association for Computational Linguistics.

Rawan Tahssin, Youssef Kishk, and Marwan Torki. 2020. Identifying nuanced dialect for Arabic tweets with deep learning and reverse translation corpus extension system. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 288–294, Barcelona, Spain (Online). Association for Computational Linguistics.