On The Arabic Dialects' Identification: Overcoming Challenges of Geographical Similarities Between Arabic dialects and Imbalanced Datasets

Salma Jamal
School of Information Technology
and Computer Science, Nile University
Giza, Egypt
sagamal@nu.edu.eg

Aly M. Kassem
School of Computer Science
University of Windsor, Canada
kassem6@uwindsor.ca

Omar Mohamed and **Ali Ashraf**
Faculty of Computers and Artificial Intelligence
Helwan University
Helwan, Egypt
{omar_20170353, aliashraf}@fci.helwan.edu.eg

Abstract

Arabic is one of the world's richest languages, with a diverse range of dialects based on geographical origin. In this paper, we present a solution to tackle subtask 1 (Country-level dialect identification) of the Nuanced Arabic Dialect Identification (NADI) shared task 2022 achieving third place with an average macro F1 score between the two test sets of 26.44%. In the preprocessing stage, we removed the most common frequent terms from all sentences across all dialects, and in the modeling step, we employed a hybrid loss function approach that includes Weighted cross entropy loss and Vector Scaling(VS) Loss. On test sets A and B, our model achieved 35.68% and 17.192% Macro F1 scores, respectively.

1 Introduction

The Arabic language is spoken in many regions of the world, including North Africa, Asia, and the Middle East. It is the official language of over 25 nations and one of the most widely used languages on the internet, with 164 million and 121 million internet users from the Middle East and North Africa, respectively. The expansion of the Arabic language over the centuries formed widely dispersed groups, which in turn transformed the language through time and separated it into different dialects, which are a specialized form of the Arabic language that is specific to a given region or social group, such as Egyptian, Jordanian, Lebanese, and Palestinian, etc. The closer the countries are geographically, the less variance between their dialects. Furthermore, in formal situations such as the media and education, all Arab nations use Modern Standard Arabic (MSA), but Arabic dialects are used in informal everyday life communication. Due to the intricacy of the language morphology and the scarcity of relevant datasets as the majority of the available datasets are data imbalanced, Arabic research received little attention in its early phases, particularly in the Arabic dialect identification task, because of the many challenges posed by the high similarity of dialects, especially in short phrases, as the same words are all commonly used in all dialects, in fact, the same word can have different meanings in the same dialect. However, it is a significant problem in many applications since being able to recognize the dialect effectively helps enhance specific applications and services, such as Automatic Speech Recognition, remote access, e-health, and e-learning. The majority of the research is focused on classifying the language into four regions: Gulf, Egyptian, Maghrebi, and Levantine, because it's less challenging than country-level dialect identification. Recent studies, however, have concentrated on classifying the language into finer-grained variants such as country-level dialects. In this paper, we present our approach to solving Nuanced Arabic Dialect Identification (NADI) shared task 2022 subtask-1 (Country-level dialect identification) (Abdul-Mageed et al., 2022). Our approach is divided into two main phases. The first step is in the prepossessing phase, where we removed the most frequent terms from all sentences across dialects to decrease the model confusion as the

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 458 - 463

same word can appear in different dialects. The second step is in the modeling phase, where we employed a hybrid loss function technique combining Weighted Cross Entropy loss and VS loss (Kini et al., 2021) to overcome the imbalanced data problem. The rest of the paper is organized as follows: section 2 provides a review of previous Arabic Dialect Identification literature, section 3 describes the proposed dataset, section 4 proposes the model of Arabic Dialect Identification, in section 5 and section 6 we show the results of the proposed model and discuss the experiments of the different parameter settings and various loss functions. Finally, we conclude in section 7.

2 Related Work

This section discusses previous research addressing Arabic Dialect Identification challenges in the Arabic language, the methodologies, strengths, and drawbacks. (Zaidan and Callison-Burch, 2011) labeled the Arabic Online Commentary Dataset (AOC) through crowd-sourcing by collecting reader's comments from three Arabic newspapers: Al-Ghad, Al-Youm Al-Sabe, and Al-Riyadh each of which represents one of the three dialects Egyptian, Gulf, and Levantine. The final dataset is composed of 108,173K comments. They built a model to classify even more crawled data and achieved an accuracy of 82.45%. (Abdelali et al., 2021) gathered dialectal Arabic tweets and labeled them based on account descriptions. The resulting dataset comprises 540k tweets from 2,525 users spread over 18 Arab nations. (Talafha et al., 2020) present a solution that won the 2020 NADI shared task (Subtask 1.2) (Abdul-Mageed et al., 2020) by adapting to the task's unlabeled data (task-adaptive pretraining), then fine-tuning the dialect identification task using AraBERT on 10M unlabeled tweets. (El Mekki et al., 2020) the solution that won Subtask 2.2 employed a hierarchical classifier with a combination of TF-IDF and AraBERT features to classify the country at the first level then at the second level to classify the province. (AlKhamissi et al., 2021) the solution that won the 2021 NADI shared (Abdul-Mageed et al., 2021b) employed an ensemble learning model by fine-tuning the MARBERT model with adapters and Vertical Attention (VAtt). Embedding two additional layers at each transformer block at the MARBERT model allows for preserving the pre-trained embedded knowledge in the MARBERT layers. At NADI-2021, an

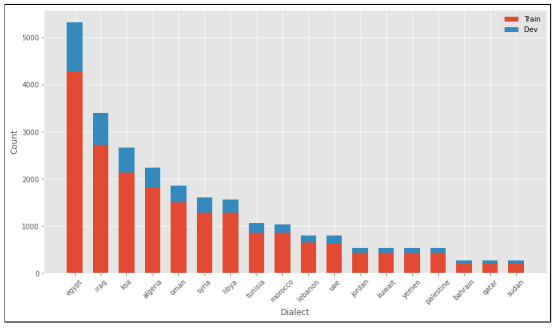

Figure 1: Train/Dev Sizes per Country

end-to-end deep Multi-Task Learning (MTL) approach (El Mekki et al., 2021) was used to address both country-level and province-level identification. The MTL model combines the contextualized word embedding of MARBERT with two task-specific attention layers that extract task-discriminative characteristics. The results of this study show that most studies relied on the robustness of the model to solve the issue of high word similarity in different dialects rather than conducting additional research to address the issue. In order to reduce the likelihood of model confusion between the different dialects, this study aims to overcome past limitations by adopting a strategy to exclude the most common phrases from the tweets in the dataset. In the modeling phase, we also employed a hybrid loss function combining VS loss and Weighted Cross Entropy loss to solve the issue of data imbalance, which is a common issue in most Arabic datasets.

3 Dataset

The proposed NADI-2022 dataset - Country-level dialect identification (sub-task 1) (Abdul-Mageed et al., 2022) has 18 distinct dialects with a total of 20k tweets for training, 4871 tweets for development, and two test sets for testing, test-A with 18 dialects and 4758 tweets, and test-b with an unknown number of dialects and 1474 tweets. However, the dataset's distribution is significantly uneven and skewed (see Figure 1), with Egypt being the most common dialect with a total of 4283 tweets and Sudan, Qatar, and Bahrain being the least common classes with a total of 215 tweets for each dialect. Arabic Dialect Identification has two major challenges. First, there is a significant degree of similarity between dialects in short words; numerous short phrases are utilized in all dialects. Second, there is an imbalance in data distribution. To overcome these issues, we eliminated the most frequently occurring phrases from the corpus and

used a hybrid loss function composed of Weighted cross-entropy and VS loss.

4 System Description

This section will outline the methods we followed in developing our approach to solving subtask 1, starting with data pre-processing to the model's experiments with different loss functions.

4.1 Data Pre-Processing

4.1.1 Text cleaning

In this step, we focused on text cleaning by removing certain irrelevant letters and symbols from the tweets:

- We eliminated any non-Arabic characters, numbers, or Arabic diacritics.
- Each word in the tweet was normalized to its base form.
- Since the dataset is made up of Arabic dialect tweets, certain users have a tendency to repeat certain characters within words (text elongation). These extra characters were removed from each word.
- We eliminated the emojis from the tweets because they don't provide any additional context for classifying the tweets into their dialects.

4.1.2 Common Terms Removal

The removal of the most prevalent terms from each tweet is one of the key components of the proposed method. All Arabic dialects are derived from Modern Standard Arabic (MSA), as we previously stated. Additionally, the closer geographically located countries are, the less variance there is between their dialects. For these reasons, we noticed that many terms overlap between Arabic dialects, which could potentially confuse the model during the learning phase. Therefore, we decided to remove the most frequently occurring words in the dataset from all tweets.

Counting the number of times each term appeared in the whole dataset, the results varied from "1" occurrences, which we regarded as distinct terms from a particular dialect, to "4529" occurrences, which we regarded as words that confuse the model. As anticipated, words that can be used in multiple dialects are the most frequently occurring words outside of stop words. For example, the word "قلبي" was repeated "347" times in the dataset and can have various meanings depending on the dialect it is used in, such as "my love" in the Egyptian dialect, "heart (the organ)" in UAE, etc. Another illustration is the term "طيب" which appeared "286" times and may signify either "ok" in the Egyptian dialect or "delicious" in Lebanon. Setting a hyperparameter that is the removal threshold to regard the term as common or distinct; if the count of the term exceeds the threshold, we remove it from the whole corpus.

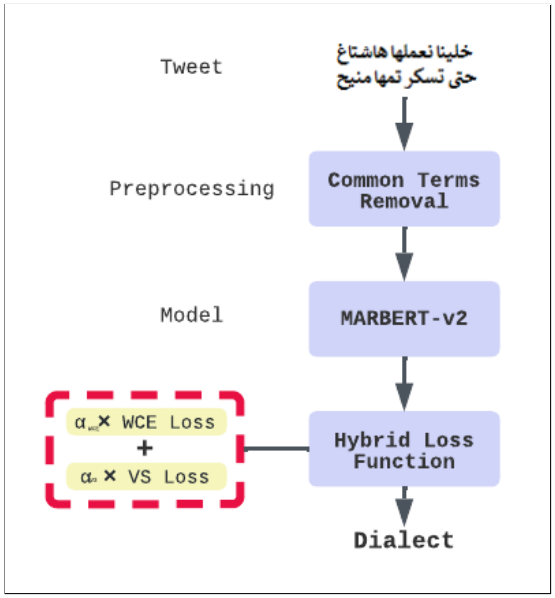

Figure 2: Pipeline of the proposed method

4.2 Loss Functions

In this subsection, we discuss the two main loss functions and the hybrid approach between them.

4.2.1 Weighted Cross-Entropy Loss

Weighted Cross-Entropy loss is a variant of regular Cross-Entropy loss that differs by assigning sample weights inversely proportional to class frequency rather than treating all classes equally.

$$\mathrm{CE} = -\frac{1}{N}\sum_{i}\sum_{j\in\{0,1\}} y_{ij}\log p_{ij} \qquad (1)$$

Equation 1, demonstrates that each x_i contributes equally to the overall objective When we don't want all x_i to be treated equally, the standard approach is to assign different weighting factors to different classes. Adding α_i as a weighting factor modifies the standard cross-entropy(Equation 1) as follows:

Data Pre-Processing	Loss Function	Macro-F1(%)
Cleaning only	Weighted CE	28.315
	VS loss	34.207
	Hybrid Loss	34.274
Frequent Removal	VS loss	34.461
	Hybrid Loss	35.8

Table 1: Dev-set result of our method on subtask 1

$$\text{Weighted CE} = -\frac{1}{N}\sum_{i}\alpha_i \sum_{j\in\{0,1\}} y_{ij}\log p_{ij} \tag{2}$$

where $\alpha_i \in [0,1]$ is set by assigning sample weights inversely proportionately to the class frequency.

4.2.2 Vector Scaling Loss

(Kini et al., 2021) proposed an extension of the VS-loss to handle imbalanced datasets, which is an improved version of cross-entropy loss but with the addition of three parameters that combine additive and multiplicative logit modifications. The VS-loss formula for multiclass datasets is as follows:

$$\ell_{\text{VS}}(y, f_w(x)) = -\omega_y \log\left(e^{\Delta_y f_y(x)+\iota_y} / \sum_{c\in[C]} e^{\Delta_c f_c(x)+\iota_c}\right) \tag{3}$$

weight parameters $\omega_{\pm} > 0$, additive logit parameters $\iota_{\pm} \in R$, and multiplicative logit parameters $\Delta_{\pm} > 0$:

4.2.3 Hybrid Loss Function

In data-imbalanced scenarios, using Focal loss, Dice loss, Tversky loss, and VS loss functions instead of standard weighted cross-entropy enhances model performance, as stated at (Mostafa et al., 2022). We tested various loss functions to see how well they performed in overcoming the imbalance problem in the provided dataset. The VS loss and Weighted cross-entropy(WCE) were the top performers, so we attempted to combine them. Because each loss function does not produce the same mistakes as the other, combining the two loss functions results in two predictions instead of simply one. Each of these predictions has its own loss. As a dynamic ensemble learning approach, gradients from all of these losses are propagated back through the model. The balancing weights α_{VS}, α_{WCE} are 0.7 and 0.3 respectively. The proposed hybrid loss function is defined as follows:

$$Hybrid.loss = \alpha_{VS} VS + \alpha_{WCE} WCE \tag{4}$$

4.3 Pre-Trained Model

Because the proposed dataset is a collection of tweets, we have to choose a pre-trained model that was trained on social media data (Twitter data) with dialect diversity, as the dataset contains 18 dialectics. According to (Abdul-Mageed et al., 2021a), fine-tuning phase performance increases if the model was pre-trained on the same dataset domain.

MARBERT-v2: Our model is based on the publicly available transformer model MARBERT-V2, which was trained on 1B multidialectal Arabic tweets (Abdul-Mageed et al., 2021a). It is based on the BERT-BASE architecture (Devlin et al., 2018) and has 163M parameters, including 12 encoder layers, 12 attention heads, and 768 hidden sizes, but without the next sentence prediction (NSP) component. MARBERT-V2 is an extension to MARBERT that has been trained on more data such as Books (Hindawi), El-Khair (El-Khair, 2016), Gigaword, OSCAR (Suárez et al., 2019), OSIAN (Zeroual et al., 2019), and AraNews dataset (Nagoudi et al., 2020), as well as a longer sequence length of 512 tokens totaling 29B. Figure 2 illustrates the proposed pipeline of our method.

5 Results

In the two main steps of the suggested technique, we tested with various settings. The best macro F1-score is obtained by eliminating the most common terms during the pre-processing step and then feeding the pre-processed data into MARBERT-V2 using the suggested hybrid loss function. The proposed methodology achieved an F1 score of 39%

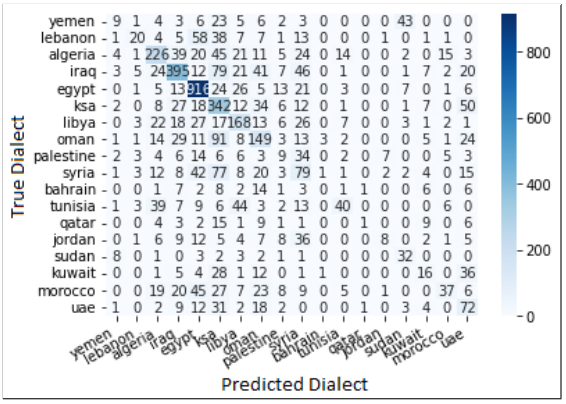

Figure 3: The confusion matrix of the proposed results

on the development set, 35.68% on test set A, and 17.1924% on test set B, and the average macro F1 score between the two test sets is 26.44%.

6 Discussion

As indicated in the Table 1, we tested the loss functions on the cleaned-only dataset first as a standalone loss and then as a hybrid loss between the two proposed loss functions in order to obtain the best results possible. As anticipated, VS loss outperformed Weighted Cross-Entropy Loss, scoring 34.207%, demonstrating that it is better able to address the issue of class imbalance. Additionally, we combined the two loss functions to create a hybrid loss, which performed better than the individual losses by achieving 34.274%. Utilizing multiple values for the removal threshold during the second phase of our method, which involves removing the most prevalent frequent terms, we found that the optimal value produced superior outcomes than utilizing the cleaned-data only, obtaining 34.461% with VS loss and 35.8% with the hybrid loss, which is the best results in our pipeline. Figure 3 illustrates the confusion matrix of the predicted results, demonstrating what we previously stated: the closer the countries are geographically, the more similar their dialects are. For instance, if we take the KSA dialect, it is most frequently confused with the UAE, Kuwaiti, and Omani dialects, and they are all GULF countries located on the same continent, thus closer to each other.

7 Conclusion

This paper outlines our method for solving Nuanced Arabic Dialect Identification (NADI) shared task 2022 subtask-1 (Country-level dialect identification). We eliminated the most frequent words from all tweets to reduce the likelihood that the model would become confused between the different dialects. Additionally, we used MARBERTv2 with a hybrid loss function approach during the modeling phase to effectively address the class imbalance issue in our dataset. The findings demonstrated that our method outperforms a standalone loss function and tweets without removing the most common terms, with an F1 score of 35.68% on test set A and 17.1924% on test set B, and the average macro F1 score between the two test sets is 26.44%. To further improve the model performance, we aim to develop better methods to handle the removal process of the standard terms. Also, collecting more data for the least common dialects may be significant in the performance.

Limitations

We employed the MARBERT-V2 pre-trained BERT-based model, which was trained on a large corpus with a reduced bias toward specific dialects. However, the proposed dataset sample size is insufficient to allow the model to generalize successfully to new data. The elimination of standard terms component is based just on frequency; introducing additional factors may improve performance.

Ethics Statement

The Arabic language is one of the world's most frequently spoken languages. Developing a system to recognize Arabic in its numerous dialects would benefit several applications in the Arabic language, such as offensive text identification on social media, because many internet users communicate using informal language (dialects).

Acknowledgements

This research is supported by the Vector Scholarship in Artificial Intelligence, provided through the Vector Institute. Also, We gratefully acknowledge support from Compute Canada.

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.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 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, Online. 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.

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, 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)*.

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*.

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*.

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

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.

Abdellah El Mekki, Abdelkader El Mahdaouy, Kabil Essefar, Nabil El Mamoun, Ismail Berrada, and Ahmed Khoumsi. 2021. Bert-based multi-task model for country and province level msa and dialectal arabic identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 271–275.

Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. 2021. Label-imbalanced and group-sensitive classification under overparameterization. In *Advances in Neural Information Processing Systems*, volume 34, pages 18970–18983.

Ali Mostafa, Omar Mohamed, and Ali Ashraf. 2022. Gof at arabic hate speech 2022: Breaking the loss function convention for data-imbalanced arabic offensive text detection. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 167–175, Marseille, France. European Language Resources Association.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Tariq Alhindi. 2020. Machine generation and detection of Arabic manipulated and fake news. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 69–84, Barcelona, Spain (Online). Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

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. *arXiv preprint arXiv:2007.05612*.

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.

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, Florence, Italy. Association for Computational Linguistics.