SENIT at AraFinNLP2024: Trust Your Model or Combine Two

Abdel Momen Ben Naser
National Engineering School of Tunisia
University of Tunis El Manar, Tunisia
abdelmomen.bennaser@etudiant-enit.utm.tn

Moez Ben Hajhmida
National Engineering School of Tunisia
University of Tunis El Manar, Tunisia
moez.benhajhmida@enit.utm.tn

Abstract

In this paper, we present our system submitted to the 2024 Shared Task on Arabic Financial NLP (AraFinNLP), specifically addressing Subtask 1: Multi-dialect Intent Detection. This task utilized a dataset derived from ArBanking77. To approach this challenge, we leveraged state-of-the-art pretrained contextualized text representation models, which we fine-tuned for the downstream task. Our methodology began with fine-tuning multilingual BERT and various Arabic-specific models, including MARBERTv1, MARBERTv2, and CAMeLBERT. To enhance our classification performance, we applied an ensembling technique, combining embeddings from MARBERT and CAMeLBERT. Our experimental results indicate that MARBERTv2 outperformed all other models. Based on the F1-score from the leaderboard, our submission for Intent Detection achieved sixth place, with a marginal difference of 0.010 from the fifth-ranked submission.

1 Introduction

The AraFinNLP (Malaysha et al., 2024) initiative underscores the critical role of Financial Natural Language Processing (NLP) within the Arab World's financial landscape. This importance is accentuated by the significant growth experienced in Middle Eastern stock markets, fueled by diverse sectors contributing to expansion. This growth, spanning multiple countries, reflects the dynamic nature of the region's financial markets, attracting global attention and investment. As these markets evolve, the demand for advanced Arabic NLP tools becomes essential to address local linguistic intricacies and serve the global financial community engaging with these markets. The progression of Arabic NLP capabilities in finance is vital for effectively analyzing and interpreting financial data.

The Arabic language includes many variants and dialects in addition to Modern Standard Arabic (MSA). While some dialects may share certain vocabulary, they still vary significantly between countries, each with its own unique characteristics, which impacts language processing tasks. For example, the sentence "فرجيني كيف اشبك البطاقة الجديدة" represents a query text in Palestinian dialect(Jarrar et al., 2017) and means "Show me how to link the new card". The equivalent sentence in MSA is: "أرني كيفية ربط البطاقة الجديدة". For a speaker not used to the Palestinian dialect this can be very confusing as "اشبك" is literally translated to "clas p" English. These sentences were extracted from ArBanking77 (Jarrar et al., 2023), a MSA and Palestinian Arabic dataset for Intent Detection in the banking domain. This dataset is being proposed in the 2024 Shared Task on The Arabic Financial NLP (AraFinNLP) (Malaysha et al., 2024). In the Intent Detection task, extracting the necessary information to identify intent is challenging when dealing with multiple dialects. Training a model on MSA and Palestinian data and expecting it to interpret a range of Arabic dialects, such as Gulf, Levantine, and North African, adds to the complexity of the task.

The contributions of this paper are summarized as follows:

- We present a detailed exploration of various state-of-the-art Arabic-specific pretrained models, including MARBERTv1, MARBERTv2, and CAMeLBERT, for multi-dialect intent detection.
- We propose an ensembling technique that combines embeddings from MARBERT and CAMeLBERT to enhance classification performance.
- We provide a comprehensive analysis of our experimental results, highlighting the effectiveness of MARBERTv2.

Proceedings of The Second Arabic Natural Language Processing Conference, pages 428–432

The structure of this paper is delineated as follows: Section 2 explores related work approaches and datasets, Section 3 furnishes a succinct exposition delineating the dataset employed in this study. Section 4 elaborates on the systems utilized and the experimental configuration undertaken to construct models for Multi-dialect Intent Detection. Section 5 elucidates the results acquired from these endeavors. Subsequently, Section 6 engenders a comprehensive discussion. The paper culminates with Section 7, offering conclusions and delineating prospective avenues for future research.

2 Literature Review

Research in Arabic NLP has seen significant advancements in recent years, with various studies exploring techniques and models tailored for different dialects and applications. Pretrained language models, such as BERT (Jacob Devlin and Toutanova., 2019), have been adapted for Arabic, leading to the development of models like MARBERT (Muhammad Abdul-Mageed and Nagoudi, 2021) and CAMeLBERT (Inoue et al., 2021), which focus on Modern Standard Arabic and various dialects. These models have shown improvements in tasks like sentiment analysis, named entity recognition, and intent detection. Other notable works include the use of ensembling techniques and transfer learning to enhance model performance across different NLP tasks. The AraFinNLP shared task builds on these advancements, providing a platform to further explore the potential of these models in a financial context.

3 Data

The dataset employed in the context of the Multi-dialect Intent Detection competition's subtask 1 originates from ArBanking77. ArBanking77 was obtained by translating the English Banking77 dataset (Jarrar et al., 2023) into MSA and Palestinian Arabic. This dataset is being expanded in this shared task to include a set of Arabic dialects in addition to Palestinian. ArBanking77 includes a substantial collection of 31,404 queries categorized into 77 distinct intent classes, encompassing a broad spectrum of banking-related inquiries and requests. The dataset provided for Subtask-1 of the competition consists of 24,018 sentences in MSA and Palestinian dialect, labeled with 77 intent classes. The initial released training data included the validation dataset as well. The data distributions for Subtask-1 are shown in Figure 1.

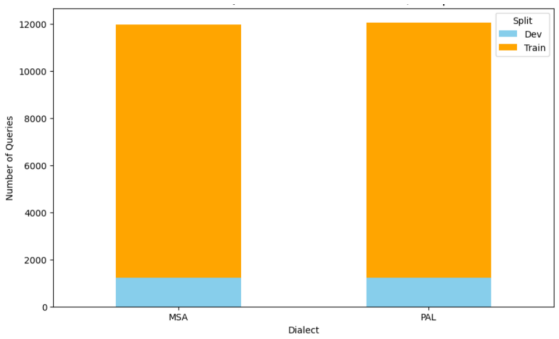

Figure 1: Distribution Of Queries Across Dialects

	Train	Validation	Total
MSA	10,733	1,230	11,963
PAL	10,821	1,234	12,055
Total	21,554	2,464	24,018

Table 1: The dataset description for Subtask 1 - Multi-dialect Intent Detection.

Table 1 provides a detailed description of the data distribution across dialects, as well as the train and validation split within the dataset.

4 System Description

Pretrained contextualized text representation models are highly effective for natural language comprehension. Among these, BERT (Jacob Devlin and Toutanova., 2019) is the state-of-the-art model, surpassing its predecessors in NLP. Arabic-specific studies have also been conducted. In our approach, we explored BERT variations for Arabic: CAMeLBERT, MARBERTv1, and MARBERTv2. Additionally, we combined CAMeLBERT and MARBERTv2 embeddings with a fully connected layer for fine-tuning.

4.1 CAMeLBERT

CAMeLBERT, introduced by (Inoue et al., 2021), is a pretrained Language Model (LLM) based on the BERT architecture, specifically designed for Modern Standard Arabic (MSA), dialectal Arabic (DA), and classical Arabic (CA). It uses a WordPiece tokenizer with a 30,000-word vocabulary, trained on 167 GB of text with Hugging Face's tokenizers. CAMeLBERT retains original letter casing and accents, incorporating whole word masking and a duplicate factor of 10. The model can make up to 20 predictions per sequence for datasets with

128 tokens and up to 80 predictions for datasets with 512 tokens.

4.2 MARBERTv1 and MARBERTv2

MARBERT, (Muhammad Abdul-Mageed and Nagoudi, 2021), is a BERT-based pretrained language model specifically targeting diverse Arabic dialects. It was trained on 128 GB of Arabic tweets, totaling 15.6 billion tokens, including both Arabic and non-Arabic words. MARBERT aims to enhance language representation for underrepresented Arabic variations. MARBERTv2, an updated version, includes additional training on Modern Standard Arabic (MSA) and the AraNews dataset, featuring an increased sequence length of 512 tokens over 40 epochs.

Both ARBERT and MARBERT showed suboptimal performance in question answering (QA) tasks due to their initial training with a restricted sequence length of 128 tokens. To address this, MARBERTv2 was enhanced with additional pretraining on the same MSA data and the AraNews dataset, using a longer sequence length, resulting in a refined model with 29 billion tokens.

4.3 CAMeLBERT + MARBERTv2

The effort to combine the capabilities of multiple models has been a long-standing pursuit. For instance, ensemble methods aggregate the outputs of different models to improve prediction performance and robustness (Jiang et al., 2023). Motivated by ensemble methods, in this section, we describe how we combined both BERT variations to try to get a good performance. First, we defined a neural network model using PyTorch for the downstream task at hand. Combining the outputs of two pre-trained BERT models and processing them through a linear layer for classification. We used CAMeLBERT and MARBERT as pre-trained BERT models. These models take input sequences and produce contextualized representations of the tokens in those sequences. Then a forward function is applied to take the input sequences for both BERT models (Marbert's input ids, Marbert's attention mask, CAMeLBERT's input ids and CAMeLBERT's attention mask) and pass them through the respective BERT models.

The last hidden state of each BERT model is extracted, and the output corresponding to the [CLS] token (a special token used for classification tasks) is selected. These outputs are concatenated and processed through a linear layer to obtain the final logits for classification.

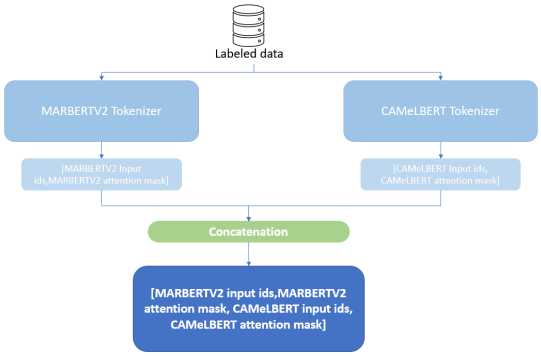

Figure 2: Combined tokenizer outputs from MARBERTV2 and CAMeLBERT.

Figure 2 provides a visual representation of the tokenization process and the concatenation of the outputs, supporting the text description effectively.

Our new model architecture involves techniques like concatenation, pooling, and linear transformation to combine information from multiple Large Language Models and make predictions based on their combined representations. Inspired by transfer learning scenarios, where pre-trained models are fine-tuned for specific downstream tasks.

4.4 System Submission

We leverage pre-trained language models as the foundation for our final models. In addition to surpassing previous methodologies in performance, substantial volumes of unlabeled text have been employed to train these versatile models. Fine-tuning these models on significantly smaller annotated datasets yields favorable outcomes due to the insights acquired during the initial pre-training phase, which typically demands considerable computational resources. Consequently, given the relatively limited size of our dataset, we opted to fine-tune BERT pre-trained models. Fine-tuning entails the addition of an untrained layer of neurons atop the pre-trained model, with only the weights of the final layers adjusted to align with the characteristics of the new labeled dataset. For hardware, We have used the Tesla P100 GPU available on Kaggle Notebooks.

5 Results

Our models were validated using balanced validation sets provided by the competition. Among the models, those built on MARBERT showed the best performance, likely due to MARBERT's focus on

MSA, which is 50% of our dataset. MARBERT's specialization in MSA allows it to excel compared to models trained on multiple dialects. We also experimented with different training epochs, using a batch size of 16 for training and testing.

Model	#epochs	F1 Dev	F1 Test
CAMeLBERT	7	0.8346	0.7743
CAMeLBERT	5	0.8317	0.7744
CAMeLBERT	3	0.8177	0.7123

Table 2: Performance of CAMeLBERT with a batch size of 16.

Table 2 shows the F1-score results for CAMeLBERT with different numbers of epochs, indicating that 5 epochs yielded the best result.

Model	#epochs	F1 Dev	F1 Test
MARBERTv1	7	0.8363	0.8024
MARBERTv1	5	0.8378	0.8114
MARBERTv1	3	0.8260	0.7908

Table 3: Performance of MARBERTv1 with batch size of 16.

Table 3 presents the F1-score results for MARBERTv1 across various epochs, with the best results achieved using 5 epochs.

Model	#epochs	F1 Dev	F1 Test
MARBERTv2	7	0.8363	0.8134
MARBERTv2	5	0.8417	0.8204
MARBERTv2	3	0.8360	0.8060

Table 4: Performance of MARBERTv2 with batch size of 16.

Table 4 displays the F1-score results for MARBERTv2, with the best performance observed at 5 epochs.

Model	F1 Dev	F1 Test
CAM+MARv1	0.8106	0.8049
CAM+MARv2	0.8134	0.8040

Table 5: Performance of combined-model with batch size of 16.

Table 5 shows the F1-score results for the combined model using MarBert and CAMeLBERT embeddings, with 5 epochs yielding the best results.

5.1 Official Submission Results

The official results of subtask1, Multi-dialect Intent Detection, are presented in Table 6. We are ranked 6th as the SENIT team.

Team Name	Micro-F1	Rank
MA	0.87731	1
AlexuNLP24	0.8762	2
BabelBot	0.87092	3
UDEL	0.83423	4
SemanticCUETSync	0.82083	5
SENIT	0.82041	6
Fired_from_NLP	0.80138	7
MTU	0.78935	8
SMASH	0.78662	9
dzFinNlp	0.67213	10
BFCI	0.49074	11

Table 6: Official results on subtask1, Multi-dialect Intent Detection.

5.2 Discussion

Results from Tables 2, 3, 4, and 5 show that using MARBERTv2 with 5 epochs in the training phase yielded the best results. In this case, the combination of both MARBERTv2 and CAMeLBERT could potentially outperform MARBERTv2 on this downstream task if an attention mechanism layer was used to calculate attention weights, indicating the relative importance of each element within the combined representation for the target task.

It's crucial to highlight the distinction between MARBERTv2's results and those of other models. MARBERTv2, being pretrained on a large dataset containing a variety of Arabic dialects, demonstrates superior performance with dialectal data.

Another approach is to use individual Large Language Models as feature extractors, fine-tuning them on the downstream task. Their outputs are then passed to a meta-model, which can be a linear neural network, a neural network with an attention mechanism, or a Bayesian neural network. This technique may outperform the individual models.

6 Conclusion

We employed four language models to detect multi-dialect intents (CAMeLBERT, MARBERTv1, MARBERTv2, and a combination of CAMeLBERT and MARBERTv2). The best results were obtained with MARBERTv2 for subtask 1, which was then selected for the final submission.

References

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Kenton Lee Jacob Devlin, Ming-Wei Chang and Kristina Toutanova. 2019. In *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),*, page 4171–4186.

Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, and Sana Ghanem. 2023. Arbanking77: Intent detection neural model and a new dataset in modern and dialectical arabic. In *Proceedings of ArabicNLP 2023, Singapore (Hybrid), December 7, 2023*, pages 276–287. Association for Computational Linguistics.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: an annotated corpus for the palestinian arabic dialect. *Lang. Resour. Evaluation*, 51(3):745–775.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Sanad Malaysha, Mo El-Haj, Saad Ezzini, Mohammad Khalilia, Mustafa Jarrar, Sultan Nasser, and Ismail Berrada. 2024. AraFinNlp 2024: The first arabic financial nlp shared task. In *Proceedings of the 2nd Arabic Natural Language Processing Conference (Arabic-NLP), Part of the ACL 2024*. Association for Computational Linguistics.

AbdelRahim Elmadany Muhammad Abdul-Mageed and El Moatez Billah Nagoudi. 2021. Arbert marbert: Deep bidirectional transformers for arabic. *Computing Research Repository*, arXiv:2101.01785.