MA at AraFinNLP2024: BERT-based Ensemble for Cross-dialectal Arabic Intent Detection

Asmaa Ramadan*, **Manar Amr***, **Marwan Torki**, **Nagwa El-Makky**
Alexandria University, Alexandria, Egypt
{es-AsmaaRamadan2023, es-ManarAmr2023, mtorki, nagwamakky}@alexu.edu.eg

Abstract

Intent detection, also called intent classification or recognition, is an NLP technique to comprehend the purpose behind user utterances. This paper focuses on Multi-dialect Arabic intent detection in banking, utilizing the ArBanking77 dataset. Our method employs an ensemble of fine-tuned BERT-based models, integrating contrastive loss for training. To enhance generalization to diverse Arabic dialects, we augment the ArBanking77 dataset, originally in Modern Standard Arabic (MSA) and Palestinian, with additional dialects such as Egyptian, Moroccan, and Saudi, among others. Our approach achieved an F1-score of **0.8771**, ranking first in subtask-1 of the AraFinNLP shared task 2024. The code is available at https://github.com/asmaaramadan99/AraFinNLP-SharedTask.git

1 Introduction

Intent detection aims at parsing the semantics of the user input to generate the best response. It is typically considered a classification task, where each utterance is associated with one, and sometimes multiple, intents. It is critical to task-oriented conversational systems in various domains, including the banking sector. While the intent detection task, has been widely investigated in the customer assistance domain (Xu et al., 2017), it is under-explored in the banking sector due to the limited availability of datasets, especially Arabic datasets.

In this paper, we present our submission to subtask-1 of shared task 1: AraFinNLP2024 in The Second Arabic Natural Language Processing Conference (Malaysha et al., 2024). The task focuses on cross-dialectal Arabic intent detection in the banking domain. The banking domain's first Arabic intent detection dataset is ArBanking77 (Jarrar et al., 2023), which is arabized from the Banking77 dataset (Casanueva et al., 2020). However, ArBanking77 (Jarrar et al., 2023) is limited to Modern Standard Arabic (MSA) and Palestinian dialects. To help overcome this limitation, we first augment the ArBanking77 dataset with additional dialects such as Egyptian, Moroccan, and Saudi to improve the model's ability to generalize to various Arabic dialects.

In addition, for enhanced performance in cross-dialectal intent detection, we propose an ensemble architecture of fine-tuned BERT-based classification models. We utilize BERT models pre-trained on dialectal Arabic data, then we fine-tune these models on the augmented ArBanking77, integrating contrastive loss in the training process.

2 Related Work

Various previous works existed in intent detection in the banking domain in English. (Casanueva et al., 2020) introduced the Banking77 dataset and established baseline accuracy scores of 93.6% for full data by fine-tuning BERT and using Universal Sentence Encoders (Casanueva et al., 2020). (Ying and Thomas, 2022) later improved these results by correcting label errors in the dataset with confident learning and cosine similarity, achieving a 92.0% F1 score on a trimmed dataset. (Li et al., 2022) showed that pre-training intent representations could enhance performance in financial intent classification, attaining an 87.3% Macro-F1 score on the Banking77 dataset with prefix-tuning and fine-tuning of the last LLM layer.

While many previous works focused on English datasets, Arabic has a limited number of labeled datasets, especially for dialectal and domain-specific tasks. One work that addressed the Arabic intent detection problem is (Mezzi et al., 2022). They presented a model for intent detection in the mental health domain in Tunisian Arabic. Their model classifies patient utterances into five differ-

*These authors contributed equally to this work.

Proceedings of The Second Arabic Natural Language Processing Conference, pages 441–445

ent categories. They used BERT as the encoder with five binary classifiers, one for each intent, and achieved an F1 score of 0.94. (Algotiml et al., 2019) introduced the ArSAS dataset, comprising approximately 21,000 tweets that were manually labeled with speech-act and sentiment categories. They trained two models on the ArSAS dataset: a Bidirectional Long Short-Term Memory (BiLSTM) network and a Support Vector Machine (SVM), achieving a macro F1 score of 0.615. (Jarrar et al., 2023) introduced the ArBanking77 dataset, significantly contributing to research in Arabic intent detection in the banking domain. ArBanking77 was used to fine-tune the BERT-based model, achieving an F1-score of 0.9209 and 0.8995 on MSA and Palestinian dialects, respectively.

3 Data

ArBanking77 Dataset (Jarrar et al., 2023) has a total of 31,404 samples in MSA and Palestinian dialect, with 21,559 samples for training, and 2,464 samples for validation. Each sample is a pair of the query text and its corresponding label, which is one of 77 intent classes. Table 1 shows a sample of intents. The number of queries per intent class ranges between 75 to 227, with an average of 170 queries per intent.

English Intent Label	Arabic Intent Label
Card Arrival	وصول البطاقة
Exchange rate	سعر الصرف
Transaction charged twice	تم التحصيل مرتين
Pin blocked	رمز التعريف الشخصي محظور
Edit personal details	تحرير التفاصيل الشخصية

Table 1: Examples for Intent Classes.

To enhance the model's performance across various Arabic dialects, MSA train queries are translated using the 600M-distilled version of the No Language Left Behind translation language model (NLLB), which provides various Arabic dialects.(Costa-jussà et al., 2022). Table 2 shows an example of a translated MSA query from the training dataset. The translation process occasionally introduced diacritics and unnecessary symbols, necessitating preprocessing the translated queries to eliminate them. Table 3 provides an overview of the augmented dataset and included dialects.

The test set used for model evaluation in subtask-1 of the AraFinNLP shared task has 11,721 queries of multiple Arabic dialects.

Dialect	Query
MSA	لماذا يختلف سعر الصرف الخاص بك في أيام مختلفة؟
EG	ليه سعر الصرف الخاص بك مختلف في أيام مختلفة؟
NL	ليش سعر الصرف مختلف في أيام مختلفة؟
TUN	علاش سعر الصرف متباين في أيام مختلفة؟
MOR	علاش كايختلف سعر الصرف ديكم في أيام مختلفة؟

Table 2: Example of translation of a query from MSA to Egyptian, North Levantine, Tunisian, and Moroccan.

Dialect	Train
MSA	10,733
Palestinian (PAL)	10,826
South Levantine (SL)	10,733
North Levantine (NL)	10,733
Egyptian (EG)	10,733
Tunisian (TUN)	10,733
Moroccan (MOR)	10,733
Saudi (SD)	10,733
Total	85,957

Table 3: Number of samples of each dialect in the training set.

4 System

This section presents the architecture and key components of our intent detection architecture for multi-dialect Arabic intents.

4.1 Architecture Overview

Our proposed architecture is an ensemble of 6 BERT-based models. Each model is a pre-trained BERT model that produces text embeddings of the input query text. The BERT model is followed by a single classification layer that takes text embeddings as input and outputs a single label for the input query.

During training, the BERT model is finetuned for the intent classification task using the augmented multi-dialect ArBanking77 dataset.

During inference, we use average ensemble to combine the class probabilities' predicted by each model, producing a vector of probabilities of length 77. Then, the intent with the maximum probability is selected. All models contribute equally to the ensemble process. Table 4 shows the configurations of each model.

Model	Base Model	Preprocessed Data	Train Data	Contrastive Loss
m1	AraBERTv0.2	✓	Augmented data	✗
m2	AraBERTv0.2	✗	Augmented data	✗
m3	AraBERTv0.2	✓	Augmented data	✓
m4	AraBERTv0.2	✗	MSA and PAL	✗
m5	MARBERTv2	✗	Augmented data	✗
m6	MARBERTv2	✗	Augmented data	✓

Table 4: Configurations of each model.

4.2 Pre-trained BERT Models

We used two pre-trained BERT models as baselines for our models.

4.2.1 AraBERTv0.2-Twitter-base

AraBERTv0.2-Twitter (Antoun et al., 2021) is a transformer model pre-trained following the architecture of Google's BERT (Devlin et al., 2019). Like BERT, it is trained on a Masked Language Model (MLM) task, utilizing a 77 GB MSA text dataset plus 60 million Arabic multi-dialect tweets. We employ the AraBERTv0.2-base variant, which consists of approximately 136 million parameters.

4.2.2 MARBERTv2

Like AraBERTv0.2-Twitter (Antoun et al., 2021), MARBERTV2 (Abdul-Mageed et al., 2021) is a transformer model based on Google's BERT (Devlin et al., 2019). MARBERTV2 is pre-trained on 1B Arabic multi-dialect tweets, the same as MARBERTV1, in addition to the same MSA as ARBERT and the AraNews dataset. MARBERTv2 consists of approximately 160 million parameters.

4.3 Contrastive Loss

Contrastive loss directs the learning process to bring similar instances closer together within the embedding space while encouraging dissimilar instances to move farther apart. We use the normalized temperature-scaled cross-entropy loss (NT-Xent), following (Chen et al., 2020). We combine NT-Xent loss with cross-entropy loss for training each model. The overall loss function is defined as follows

$$Loss = Loss_{\text{Cross-Entropy}} + \alpha Loss_{\text{Contrastive}} \quad (1)$$

We set α to 10 in our experiments. We also implement a custom batch sampler which limits the number of unique classes in each batch to ensure the existence of both positive and negative examples for each sample in the batch.

4.4 Ensemble

Our ensemble technique employs a straightforward method where we use average ensemble. This simple yet effective approach leverages the strengths of each model to improve overall performance.

To ensure diversity within our ensemble, we used two different pre-trained BERT-based models, AraBERTv0.2-Twitter and MARBERTV2, to learn diverse hidden representations for each class. Additionally, we applied a mixture of loss functions, with some models using only cross-entropy loss, while others used both contrastive loss and cross-entropy loss.

We also trained some models on noisy data and others on preprocessed data, providing evidence that this approach can enhance model robustness. By training on different data qualities, the models learn to handle various data imperfections, which contributes to better generalization. Furthermore, some models were trained on subsets of the dialects. This encouraged each model to learn unique hidden representations for each class, increasing the diversity among ensemble members.

Ensemble learning is particularly effective when diverse models with different representations are combined (Dietterich, 2000). This diversity, resulting from different architectures, loss functions, and data preprocessing methods, facilitated improved generalization and robustness, ultimately enhancing the overall performance of our intent detection system.

5 Experiments and Results

5.1 Experiments Setup

For our experimental settings, all experiments were conducted using a single RTX 3070 GPU. The batch size was set to 32, and the maximum input sequence length was fixed at 128 tokens.

Model	Test F1	Test Precision	Test Recall	Test Accuracy
m1	0.8547	0.8558	0.8578	0.8547
m2	0.8434	0.8492	0.8479	0.8434
m3	0.8578	0.8603	0.8619	0.8578
m4	0.7913	0.8083	0.7971	0.7913
m5	0.8198	0.8294	0.8237	0.8198
m6	0.8302	0.8350	0.8348	0.8302
ensemble	0.8773	0.8799	0.8803	0.8773

Table 5: Test scores for each classification model and their ensemble on the test set used for evaluation in the shared task.

5.2 Results

In this section, we show the results of each of the models and the final ensemble architecture. Table 5 summarizes these results. As shown in the results, the ensemble model outperformed all individual models, achieving an F1 score of 0.8773. This shows how the ensemble technique highly contributed to enhancing the final predictions.

Among all individual models, the worst-performing model is m4, which is trained on only MSA and Palestinian dialects, achieving an F1 score of 0.7913. Meanwhile, other models, trained on more dialects, achieved higher results. This can indicate that using a BERT model which is fine-tuned on only MSA, and Palestinian dialects as a base model is not enough on its own to achieve generalization during testing.

AraBERTv0.2-Twitter-based models fine-tuned on dialectical data generally achieve better test scores than MARBERTv2. This can be shown by comparing m2 with m5, having F1 scores of 0.8434 and 0.8198 respectively. This explains why the majority of the models in our ensemble are AraBERTv0.2-Twitter-based models.

The effect of applying data preprocessing is highlighted via the superior performance of m1 compared to m2, having F1 scores of 0.8547 and 0.8434 respectively.

6 Discussion

6.1 The Ensemble Technique

As shown in the results section, using the ensemble technique achieved the highest test scores among all other experiments. This superior performance can be attributed to the ensemble technique's ability to reduce variance, noise, and errors in individual models, thereby making more robust and stable predictions.

6.2 Dataset Augmentation

Augmenting the ArBanking77 dataset (Jarrar et al., 2023) with additional Arabic dialects has also been shown to improve the performance of individual models, and hence the overall performance of our architecture. However, the additional dialects are obtained with a translation language model, which can produce errors and inaccurate translations. Refining the generated translations manually could significantly improve the accuracy of the predictions.

6.3 Future Work

In our architecture, we used AraBERTv0.2-Twitter (Antoun et al., 2021) and MARBERTV2 (Abdul-Mageed et al., 2021) as baselines in our models, having approximately 136M and 160M parameters respectively. These models are pre-trained on a large corpus of dialectal Arabic data, which highly contributed to our model's enhanced performance in cross-dialectal intent detection. As a future investigation, using larger language models can potentially add significant improvement to individual models, and hence, their ensemble. Higher quality translation would contribute to the improvement of the classification task.

7 Conclusion

In this paper, we presented an ensemble architecture for intent classification in the banking domain, ranking first in subtask-1 of the AraFinNLP shared task 2024. We also augmented the ArBanking77 dataset with additional dialects, enhancing the model's generalizability to various dialects. Additionally, we incorporated contrastive loss for enhanced text representations. Future work can explore techniques to further improve generalization and classification performance.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. 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.

Bushra Algotiml, AbdelRahim Elmadany, and Walid Magdy. 2019. Arabic tweet-act: Speech act recognition for Arabic asynchronous conversations. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 183–191, Florence, Italy. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. Arabert: Transformer-based model for arabic language understanding. *Preprint*, arXiv:2003.00104.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. *Preprint*, arXiv:2003.04807.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *Preprint*, arXiv:2002.05709.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, and Sana Ghanem. 2023. Arabanking77: 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.

Xianzhi Li, Will Aitken, Xiaodan Zhu, and Stephen W. Thomas. 2022. Learning better intent representations for financial open intent classification. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 68–77, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Ridha Mezzi, Aymen Yahyaoui, Mohamed Wassim Krir, Wadii Boulila, and Anis Koubaa. 2022. Mental health intent recognition for arabic-speaking patients using the mini international neuropsychiatric interview (mini) and bert model. *Sensors*, 22(3).

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media.

Cecilia Ying and Stephen Thomas. 2022. Label errors in BANKING77. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 139–143, Dublin, Ireland. Association for Computational Linguistics.