Fired_from_NLP at AraFinNLP 2024: Dual-Phase-BERT - A Fine-Tuned Transformer-Based Model for Multi-Dialect Intent Detection in The Financial Domain for The Arabic Language

Md. Sajid Alam Chowdhury, Mostak Mahmud Chowdhury, Anik Mahmud Shanto, Hasan Murad, Udoy Das
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
u1904{064, 055, 049}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd,
u1804109@student.cuet.ac.bd

Abstract

In the financial industry, identifying user intent from text inputs is crucial for various tasks such as automated trading, sentiment analysis, and customer support. One important component of natural language processing (NLP) is intent detection, which is significant to the finance sector. Limited studies have been conducted in the field of finance using languages with limited resources like Arabic, despite notable works being done in high-resource languages like English. To advance Arabic NLP in the financial domain, the organizer of AraFinNLP 2024 has arranged a shared task for detecting banking intents from the queries in various Arabic dialects, introducing a novel dataset named ArBanking77 which includes a collection of banking queries categorized into 77 distinct intents classes. To accomplish this task, we have presented a hierarchical approach called Dual-Phase-BERT in which the detection of dialects is carried out first, followed by the detection of banking intents. Using the provided ArBanking77 dataset, we have trained and evaluated several conventional machine learning, and deep learning models along with some cutting-edge transformer-based models. Among these models, our proposed Dual-Phase-BERT model has ranked 7^{th} out of all competitors, scoring 0.801 on the scale of F1-score on the test set.

1 Introduction

In the Arabic-speaking world, where a multitude of dialects coexist alongside Modern Standard Arabic (MSA), the need for advanced NLP models capable of handling multi-dialectal variations is increasingly evident. One critical application of such models lies in the domain of intent detection, particularly in financial sectors.

A significant amount of intent detection work has been done in high-resource languages, like English (Liu and Lane, 2016), (Chen et al., 2019), (Wang et al., 2018). However, there is a noticeable gap in significant studies addressing intent detection in the financial sector for languages with limited resources such as Arabic. In response to this challenge, AraFinNLP 2024 has arranged a shared task named Task 1: Multi-Dialect Intent Detection (Malaysha et al., 2024), introducing a novel dataset called ArBanking77 (Jarrar et al., 2023) to classify the banking intents from queries in different Arabic dialects.

The key objective of the paper is to accurately understand user queries expressed in different dialects and categorize user intents accordingly. To achieve this objective, we have utilized multiple models, including three distinct deep learning architectures (LSTM, BiLSTM), two transformer-based approaches (AraBERT, Dual-Phase BERT), and three different machine learning algorithms (XGBoost, Random Forest, SVM). We have trained and assessed every model using the ArBanking77 dataset (Jarrar et al., 2023) provided for the AraFinNLP 2024 Task 1 (Malaysha et al., 2024), in order to do a comparative analysis. Ultimately, our proposed Dual-Phase-BERT model achieved the highest Micro-F1 score of 0.801 on the test dataset among all evaluated models.

The following are our research's main contributions:

- We have developed a fine-tuned Dual-Phase-BERT model that significantly helps in predicting Arabic banking intents.
- To identify banking intents in different Arabic dialects, we have compared and contrasted several traditional machine learning, and deep learning models along with some state-of-the-art transformer-based models.

The implementation details are available in this GitHub repository[1].

[1] https://github.com/Fired-from-NLP/ArabicNLP2024-subtask1

Proceedings of The Second Arabic Natural Language Processing Conference, pages 410–414

2 Related Work

Prior works on Cross-dialect Intent Detection fall into three primary categories: machine learning, deep learning, and transformer-based methods.

Traditional machine learning approaches (Al-Tuama and Nasrawi, 2022), (Bokolo et al., 2023) have been applied for detecting intentions on augmented and without augmented data. Machine learning algorithms such as Multinomial Naïve Bayes, SVM, Logistic regression, and Random Forest have been used where random forest provided the best performance (Al-Tuama and Nasrawi, 2022). In another previous work (Bokolo et al., 2023), class-specific (crime) intent detection in social media posts has been explored leveraging machine learning algorithms where logistic regression has shown superior performance in terms of accuracy.

Deep learning-based approaches perform better on cross-dialect intent detection due to their layered structure and less dependency on explicitly defining features. A deep learning neural network, Bi-model based recurrent neural network (Wang et al., 2018), has been capitalized to detect intention and slot filling. In another work (Niu et al., 2019), for simultaneous intent detection and slot-filling, a unique bi-directional interconnected model has been presented.

A fine-tuned transformer named joint-BERT (Chen et al., 2019) has been utilized for joint intent classification and slot filling. Another approach (Qin et al., 2021) uses a co-interactive transformer to detect intention and joint filling. (Chen et al., 2022) highlights the sensitivity of the state-of-the-art multi-intention model to threshold setting. They have taken advantage of several layers within a transformer-based encoder to create multi-grain representations. In (Büyük et al., 2021), researchers enhance intent detection performance by utilizing datasets with different intent categories.

3 Dataset

We have utilized the Multi-Dialect Intent Detection dataset ArBanking77 (Jarrar et al., 2023) provided for the shared task 1 of the AraFinNLP 2024 (Malaysha et al., 2024). The dataset consists of Arabic banking queries in the MSA and the PAL dialect, labeled one of the 77 categories (banking intents).

Table 1 shows that the training and validation dataset contains samples only in MSA and PAL dialects whereas the testing dataset contains samples not only in MSA and PAL dialects but also in a variety of Arabic dialects, such as Levantine, Gulf, and North African. This inclusion of multiple dialect samples in the test dataset has made this task more challenging. We have merged the MSA-train dataset with the PAL-train dataset for training purposes and merged the MSA-dev dataset with the PAL-dev dataset to validate the training. The dataset that has been provided is highly unbalanced; samples under some categories, such as beneficiary disapproval and transaction adjustments twice, are significantly more than samples under other categories, such as virtual cards not working and contactless not working.

Dialect	Train	Dev	Test
MSA	10733	1230	-
PAL	10821	1234	-
Multi-Dialect	-	-	11721
Total	21554	2464	11721

Table 1: Dataset distribution.

4 Methodology

In this section, we have presented an explanation of our methodology to develop models for the detection of multi-dialect banking intents. Figure 1 presents a summary of our methodology.

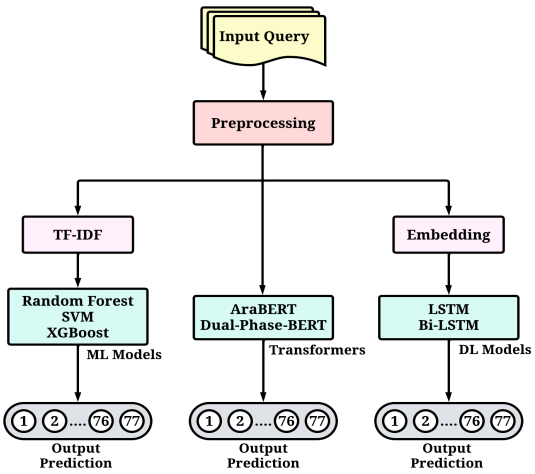

Figure 1: An outline of our approach.

4.1 Machine Learning Based Approaches

We have employed traditional machine learning techniques Random Forest, Support Vector Machine (SVM), and XGBoost to identify intents. To determine the significance of each word within a text, we have employed the TF-IDF vectorizer. Using the modified training data, we have employed

a Random Forest model with 100 estimators and a random state of 42, an SVM model with a linear kernel, and up to 1000 iterations as the maximum. Additionally, we have developed an XGBoost model with 100 estimators and a max depth of 20, that builds an ensemble of decision trees repeatedly using gradient boosting.

4.2 Deep Learning Based Approaches

We have also utilized deep learning-based models such as LSTM and BiLSTM. At first, we have defined an embedding layer to convert words into vectors, followed by a spatial dropout for regularization. Then we have defined an LSTM layer for the LSTM model and a bidirectional LSTM layer for the BiLSTM model to capture sequential dependencies where a dense output layer for multi-class classification that uses a softmax activation function comes next. Both the LSTM model and the BiLSTM model are compiled with the Sparse Categorical Cross-Entropy loss function and the Adam optimizer.

4.3 Transformer-Based Approaches

Transformer-based methods are currently being widely applied in numerous domains. We have utilized AraBERT (Antoun et al., 2020) to address this task. As Palestinian Arabic (PAL) and Modern Standard Arabic (MSA) represent two distinct varieties of Arabic, the differences between these two dialects are significant. That is why we have introduced a dual-phase method, as seen in Figure 2. In this method, we have classified the query into MSA or PAL dialect at first, then further classified the query into 77 banking intents. We have used three AraBERT models for these classification tasks.

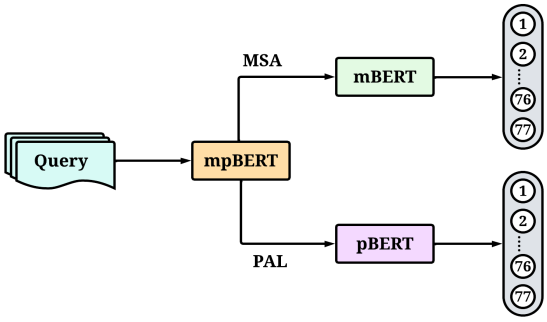

Figure 2: Dual-Phase-BERT: Transformer-based model architecture for Multi-Dialect Intent Detection.

We have fine-tuned one AraBERT model named mpBERT to distinguish between MSA and PAL queries. Then we have fine-tuned two more AraBERT models, one for the MSA dialect named mBERT and another for the PAL dialect named pBERT. the mBERT model is used to classify the MSA queries into one of the 77 banking intents and the pBERT model is used to classify the PSA queries into one of the 77 banking intents.

The classification of banking intents is done in two phases. At first, the mpBERT model recognizes the dialect of the queries. Then, based on the decision of mpBERT, the queries are passed to either the mBERT model or the pBERT model. Then the mBERT model will detect intents for queries in the MSA dialect and the pBERT model detects intents for the queries in the PAL dialect.

In our approach, we have preprocessed the datasets by applying Arabic text preprocessing using the ArabertPreprocessor. We have used the BertTokenizer to configure tokenization and initialized a BERT model (TFBertModel) with pretrained weights. We have also defined model configuration parameters like maximum sequence length and model architecture. We have built a classification model using the Keras API[2] of TensorFlow[3]. The output of the BERT model is passed through a dropout layer, then a tanh activation function-containing dense layer, followed by another dropout layer, and ultimately a dense layer with the softmax activation function for output. We have constructed the model utilizing the Adam optimizer and sparse categorical cross-entropy loss function. The trained model is then assessed using the validation data after being trained on the training data for a predetermined amount of epochs.

5 Experimental Setup

5.1 Parameter Settings

The parameter setups for each model are displayed in Table 2.

Model	lr	optim	bs	epoch
LSTM	$1e^{-3}$	Adam	64	15
BiLSTM	$1e^{-3}$	Adam	64	15
AraBERT	$5e^{-6}$	Adam	-	30
mpBERT	$5e^{-6}$	Adam	-	20
mBERT	$5e^{-6}$	Adam	-	30
pBERT	$5e^{-6}$	Adam	-	30

Table 2: Configurations of parameters for different models.

[2] https://keras.io/api/
[3] https://www.tensorflow.org/

In Table 2, learning rate, optimizer, batch size, and the number of epochs are represented by the variables lr, optim, bs, and epoch respectively.

5.2 Evaluation Metrics

According to the instruction provided by the organizer of Shared Task 1, we have evaluated our models by calculating the F1 score on the test dataset. Equation 1 gives the mathematical description of the F1 score.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

where

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

Equations 2 and 3 denote True Positive, False Positive, and False Negative, respectively, with the symbols TP, FP, and FN.

6 Experimental Findings

In this section, we have provided the experimental results of different conventional models alongside our proposed model on the test dataset.

Table 3 shows a comparative analysis of different types of machine learning, deep learning, and transformer-based models by evaluating the Precision (Macro) score, Recall (Macro) score, and Micro-F1 score on the test dataset.

Class	Model	Macro Average		
		P	R	F1
ML	XGBoost	0.591	0.518	0.519
	Random Forest	0.601	0.544	0.542
	SVM	0.691	0.617	0.617
DL	LSTM	0.624	0.623	0.628
	BiLSTM	0.659	0.658	0.663
TF	AraBERT	0.780	0.771	0.767
	Dual-Phase-BERT	**0.809**	**0.806**	**0.801**

Table 3: Results of different models on the test dataset. Here P, R, F1, and TF stand for Precision, Recall, F1 score, and Transformer respectively.

Among the machine learning models, we have observed that the SVM model performs slightly better than the Random Forest model and the XGBoost model by achieving an F1 score of 0.617. Between deep learning approaches, we have seen little improvement in the performance as both the LSTM model and the BiLSTM model have higher F1 scores compared to the models of machine learning. The BiLSTM model has obtained an F1 score of 0.663 which has a slight edge over the score of the LSTM model. However, the transformer-based models have surpassed both the machine learning and deep learning models convincingly. Although the AraBERT model attained a remarkable F1 score of 0.767, our proposed Dual-Phase-BERT model has excelled with an outstanding F1 score of 0.801. The Dual-Phase-BERT model has outperformed all of these models by securing the highest score.

7 Error Analysis

Although our Dual-Phase-BERT model has performed extremely well on the DEV set with an F1 score of 0.984, the score has dropped to 0.801 in the test phase. The main reason behind this is the inclusion of some other Arabic dialects such as Levantine, North African, and Gulf alongside MSA and Palestinian in the test set whereas the training set and the DEV set contained queries only in MSA and Palestinian. Besides that, the provided dataset is highly unbalanced. Therefore, the model is slightly biased toward the classes that have more samples during the training period and therefore leads to some misclassifications.

8 Conclusion

We have conducted a comparative performance analysis in this study by assessing a range of machine learning, deep learning, and transformer-based methods to detect banking intents from queries in different Arabic dialects. The ArBanking77 dataset that was given in the shared task has been used. Our results indicate that our proposed Dual-Phase-BERT model has achieved better results than all other models, achieving a remarkable F1 score of 0.801 on the test dataset. In the future, we will try different techniques to mitigate the data imbalance problem and explore more transformer-based model architectures.

9 Ethical Consideration

We commit to prioritizing privacy through informed consent, reducing biases, and transparent modeling to promote intent detection in the finance sector. To develop a transparent technology environment, our ethical approach puts a high priority on accountability, accessibility, and privacy.

References

Alaa T Al-Tuama and Dhamyaa A Nasrawi. 2022. Intent classification using machine learning algorithms and augmented data. In *ICDSIC*. IEEE.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *OSACPT*, Marseille, France. European Language Resource Association.

Biodoumoye George Bokolo, Praise Onyehanere, Ebikela Ogegbene-Ise, Itunu Olufemi, and Josiah Nii Armah Tettey. 2023. Leveraging machine learning for crime intent detection in social media posts. In *AIGC*. Springer.

Osman Büyük, Mustafa Erden, and Levent M. Arslan. 2021. Leveraging the information in in-domain datasets for transformer-based intent detection. In *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*.

Lisung Chen, Nuo Chen, Yuexian Zou, Yong Wang, and Xinzhong Sun. 2022. A transformer-based threshold-free framework for multi-intent nlu. In *ICCL*.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv*.

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.

Bing Liu and Ian Lane. 2016. Joint online spoken language understanding and language modeling with recurrent neural networks. *arXiv*.

Sanad Malaysha, Mo El-Haj, Saad Ezzini, Mohammad Khalilia, Mustafa Jarrar, Sultan Nasser, Ismail Berrada, and Houda Bouamor. 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.

Peiqing Niu, Zhongfu Chen, Meina Song, et al. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. *arXiv*.

Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu. 2021. A co-interactive transformer for joint slot filling and intent detection. In *ICASSP*. IEEE.

Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based rnn semantic frame parsing model for intent detection and slot filling. *arXiv*.