SemanticCuetSync at AraFinNLP2024: Classification of Cross-Dialect Intent in the Banking Domain using Transformers

Ashraful Islam Paran, Symom Hossain Shohan, Md. Sajjad Hossain, Jawad Hossain, Shawly Ahsan and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904029,u1904048,u1904031,u1704039,u1704057}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

Intention detection is a crucial aspect of natural language understanding (NLU), focusing on identifying the primary objective underlying user input. In this work, we present a transformer-based method that excels in determining the intent of Arabic text within the banking domain. We explored several machine learning (ML), deep learning (DL), and transformer-based models on an Arabic banking dataset for intent detection. Our findings underscore the challenges that traditional ML and DL models face in understanding the nuances of various Arabic dialects, leading to subpar performance in intent detection. However, the transformer-based methods, designed to tackle such complexities, significantly outperformed the other models in classifying intent across different Arabic dialects. Notably, the AraBERTv2 model achieved the highest micro F1 score of 82.08% in ArBanking77 dataset, a testament to its effectiveness in this context. This achievement, which contributed to our work being ranked 5^{th} in the shared task, AraFinNLP2024, highlights the importance of developing models that can effectively handle the intricacies of Arabic language processing and intent detection.

1 Introduction

Textual intent detection is a natural language processing (NLP) process that involves identifying and understanding a text input's underlying purpose or objective. This task is essential for enabling machines to comprehend what a user wants to achieve with their message. The goal is to classify the text into predefined intent categories, such as asking a question, requesting, providing feedback, or initiating a transaction. Intent detection plays a pivotal role in enhancing user experiences, automating processes, ensuring accuracy, scaling operations, providing business insights, and supporting multilingual environments. It is a cornerstone technology for creating intelligent and responsive systems in various domains. In the banking sector, intent detection is vital for improving customer service, enhancing operational efficiency, ensuring accuracy and security, gaining data-driven insights, supporting a multilingual customer base, and maintaining regulatory compliance. It helps banks deliver superior service while managing risks and optimizing operations.

This research approach considered intent detection a classification problem, aiming to assign a predetermined intent label to the user's input. While intent detection in English has been extensively studied (Kumar et al.; Wang et al., 2024; Zhang et al., 2021), the task is challenging for low-resource languages such as Arabic and its dialects. Very little research has been conducted on intent detection in Arabic banking, making creating precise and efficient systems challenging. Especially in particular fields like banking, the lack of research has led to fewer datasets, tools, and models customized in Arabic. As a result, this gap restricts the development of efficient NLU systems in the financial sector for Arabic-speaking users. The AraFinNLP shared task aims to bridge this gap. Under this shared task, we developed a transformer-based system to detect intent in various Arabic dialects within the banking domain. The system was evaluated on the ArBanking77 dataset, which consists of 77 classes that cover different Arabic dialects, including Palestinian Arabic and Modern Standard Arabic (MSA). The main contributions of this work are:

- Proposed a fine-tuned transformer model for intent detection in numerous Arabic dialects.
- Investigated the performance of nine classification models (Random Forest (RF), Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), CNN, CNN+LSTM, CNN+BiLSTM, AraBERTv2,

Proceedings of The Second Arabic Natural Language Processing Conference, pages 422–427

BERT-Banking77, CamelBert) to discover a reasonable model for intent detection in Arabic.

2 Related Work

Many studies have been accomplished on textual intent detection in recent years. Yolchuyeva et al. (2020) proposed self-attention networks, including SAN and BiLSTM for intent detection. They used 300-dimensional Word2Vec and FastText embeddings pre-trained on English Wikipedia as word representations. Their model achieved an F1 score of 96.81. Zhang et al. (2020b) presented transformer-based solutions for English and multilingual datasets. This model also demonstrated vital zero and few-shot performance, reaching over 75% accuracy using only 100 training examples in all datasets. A dual channel model for intent detection was proposed by Wang et al. (2022). It creates a dual-channel feature extraction technique by combining the high-level features obtained from the pooling operation and the capsule network. The final intent detection is achieved by utilizing the feature fusion of the two channels. Their model achieved an F1 score of 88.5% and 89.76% in ATIS and SMP2019-ECDT datasets, respectively. Mezzi et al. (2022) suggested a system that uses the concepts of intent recognition to make mental health diagnoses of Arabic-speaking patients based on bidirectional encoder representations from the transformers (BERT) model. Their system achieved an F1 score of 94%. Shams et al. (2022) fine-tuned several BERT models to detect intent for Urdu. Their analysis shows that on their two datasets, the fine-tuned models of mBERT and RoBERTa-urdu-small reach 96.38% and 93.30% accuracy, respectively. The intent detection problem can be treated as a classification problem as well. Aldjanabi et al. (2021) proposed a multi-task learning technique developed on a pre-trained Arabic language model for determining offensive and hate speech. The developed model surpassed past methods for offensive and hate speech detection tasks. Jarrar et al. (2023a) proposed a solution for detecting intent in modern and dialectical Arabic using a neural model based on AraBERT. Their proposed model achieved F1 scores of 0.9209 and 0.8995 on MSA and Palestinian dialect, respectively. Considering the current constraints of the task, this work applies a transformer-based method to identify intent in various Arabic dialects used in the banking sector.

3 Dataset and Task Description

This work used ArBanking77 (Jarrar et al., 2023b), which originated from the Banking77 dataset (Casanueva et al., 2020). It includes 77 different classes (intents) and features queries in MSA and Palestinian (PAL) dialect. Table 1 shows the distribution of train, development, and test sets for MSA and PAL. The dataset comprises 21,559 texts in the training set, 2,464 in the development set, and 11,721 in the test set.

Language	Train set	Dev set	Test set
MSA	10733	1230	11721
PAL	10826	1234	
Total	21559	2464	11721

Table 1: Dataset statistics for Subtask-1.

Subtask-1 (Malaysha et al., 2024) concerns cross-dialect intent detection in the banking domain. This subtask mainly focused on creating systems that accurately classify consumer intents from queries in multiple Arabic dialects. Table 2 illustrates a few examples of the query and intent of the dataset.

Query	Intent
ما زلت أنتظر بطاقتي؟ (I am still waiting on my card?)	card arrival
أرني كيفية ربط البطاقة الجديدة ، (Show me how to link the new card,)	card linking
كيف تحسب أسعار الصرف؟ (How do you calculate exchange rates?)	exchange rate

Table 2: Subtask-1 task sample with Query and Intent for MSA and PAL.

4 System Overview

This work exploited a variety of ML, DL, and transformer-based models to conduct multi-dialect intent detection in the Arabic banking industry. Figure 1 illustrates the schematic configuration of the proposed technique.

Textual data was converted into high-dimensional vectors using BoW and TF-IDF (Term Frequency-Inverse Document Frequency)

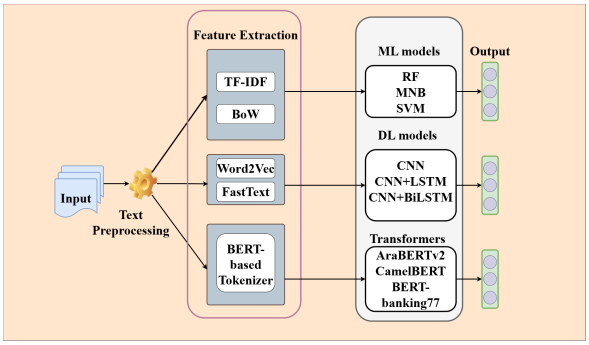

Figure 1: Schematic process for intent detection.

(Takenobu, 1994) in ML-based techniques. For DL models, Word2Vec and FastText word embeddings were utilized (Mikolov et al., 2013), while BERT-based tokenizers were employed for transformer models.

4.1 Classifiers

This work explored nine classification models (3 ML, 3 DL, and 3 transformers).

- **ML Models**: This work tested RF (Breiman, 2001), SVM (Boser et al., 1992), and MNB with TF-IDF and BoW features. Table 3 shows several ML model implementation parameters.

Classifier	Parameters	Value
RF	max-depth	20
	n-estimator	500
MNB	alpha	1.0
	fit-prior	False
SVM	kernel	linear
	gamma	auto

Table 3: Parameters of the employed ML models.

- **CNN:** The CNN model used in this work includes an embedding layer with an output dimension of 256. It comprises one convolution layer with 512 filters and a global maximum pooling layer for downsampling. The dense layer with L2 regularization introduces non-linearity to prevent overfitting. The output layer allows for multi-class classification with its 77 units and softmax activation. The loss function used here was '*SparseCategorical-Crossentropy*' with the '*Adam*' optimizer.

- **CNN+LSTM:** This approach used one convolution layer, one maximum pooling layer, three LSTM layers, and three dense layers with a dropout layer. The number of units for the convolution layer was 512; for the LSTM layers, they were 512, 256, and 128, respectively. For the first dense layer, the unit size was 256 with '*relu*' activation; the second dense layer had 128 units with '*tanh*' activation, and the third dense layer was for classification with 77 units.

- **CNN+BiLSTM:** This model employs a comparable structure to the CNN+LSTM model but it incorporates Bidirectional LSTM.

Models	LR	WD	WS	EP
AraBERTv2	$4e^{-5}$	0	50	15
BERT- Banking77	$4e^{-5}$	0.3	0	8
CamelBERT	$4e^{-5}$	0	0	5

Table 4: Hyperparameters for the transformers.

- **AraBERTv2:** AraBERTv2 (Antoun et al., 2020) is the successor of AraBERT, which leverages the BERT architecture. This model is evaluated on different downstream tasks like sentiment analysis, named entity recognition, and Arabic question answering. This work fine-tuned AraBERTv2 on the Arabic banking dataset and obtained a decent outcome.

- **BERT-Banking77:** BERT-Banking77[1] detects intent in English and is fine-tuned using BANKING77 (Casanueva et al., 2020) dataset. In this approach, we used '*Helsinki-NLP/opus-mt-ar-en*' (Zhang et al., 2020a) and '*llama-3-8b-instruct*' (Touvron et al., 2023) for zero-shot translation before feeding the input text into the model.

- **Arabic CamelBERT:** CamelBERT (Inoue et al., 2021) was built by fine-tuning BERT on different dialects of the Arabic language for domain-specific tasks. It used the ASTD, ArSAS, and SemEval datasets for fine-tuning. This model understands the Arabic dialects well and was further fine-tuned in this work on the Arabic banking dataset.

Table 4 illustrates different hyperparameters such as learning rate (LR), weight decay (WD), warmup steps (WS), and number of epochs (EP) used in the transformer-based models.

[1]https://huggingface.co/philschmid/BERT-Banking77

5 Results and Analysis

Table 5 demonstrates the evaluation results of ML, DL, and transformer-based models on the test set. The superiority of the models is determined based on the F1-score).

ML Models				
Classifier	**P(%)**	**R(%)**	**F1(%)**	**A(%)**
SVM (TF-IDF)	**69.06**	**61.65**	**61.69**	**61.69**
RF (TF-IDF)	62.95	41.94	42.07	42.07
MNB (TF-IDF)	63.73	61.71	61.05	61.05
SVM (BoW)	63.68	56.54	56.37	56.37
RF (BoW)	59.46	53.57	53.39	53.39
MNB (BoW)	64.48	55.89	55.30	55.30
DL Models				
Classifier	**P(%)**	**R(%)**	**F1(%)**	**A(%)**
CNN (Word2Vec)	**49.12**	**43.55**	**43.33**	**43.33**
CNN (FastText)	46.12	40.98	40.52	40.52
CNN+LSTM (Word2Vec)	44.84	38.53	38.50	38.50
CNN+LSTM (FastText)	45.79	40.11	39.72	39.72
CNN+BiLSTM (Word2Vec)	48.11	41.91	42.04	42.04
CNN+BiLSTM (FastText)	44.56	38.10	37.82	37.82
Transformers				
Classifier	**P(%)**	**R(%)**	**F1(%)**	**A(%)**
AraBERTv2	**82.65**	**82.44**	**82.08**	**82.08**
BERT-Banking77	70.22	68.63	68.56	68.56
CamelBERT	74.50	73.47	73.00	73.00

Table 5: Performance of the models on the test set.

Results revealed that SVM with TF-IDF features earned the most elevated F1-score (61.69%) among the ML approaches. On the other hand, CNN outperformed the CNN+LSTM (38.50%) and CNN+BiLSTM (42.04%) by obtaining the highest F1-score of 43.33%, which is approximately 18.36% lower than the best ML approach (SVM). DL models utilizing FastText word embeddings also performed relatively poorly, with CNN, CNN+LSTM, and CNN+BiLSTM achieving F1-scores of 40.52%, 39.72%, and 37.82%, respectively.

However, the transformer model outperforms the top-scoring ML and DL models. AraBERTv2 achieved the highest F1 score of 82.08%, whereas Bert-Banking77 and CamelBert scored 68.56% and 73.00%, respectively.

While ML and DL models performed well on the development set, their performance declined on the test sets. DL models significantly underperformed compared to ML models. AraBERTv2 is pre-trained with Wikipedia dataset, encompassing a range of Arabic dialects and demonstrating superior performance over all exploited models.

5.1 Error Analysis

A comprehensive quantitative and qualitative error analysis was conducted to provide detailed insights into the proposed model's performance.

Quantitative Analysis

Figure 2 categorizes intents into correctly classified and misclassified. Out of 11,721 intents, 9,621 (82%) were accurately predicted, while 2,100 (18%) were misclassified. Although the test dataset includes various Arabic dialects, the proposed model was trained only on the PAL and MSA datasets. This limitation might contribute to the model's difficulty in generalizing to unseen Arabic dialects, leading to several misclassifications.

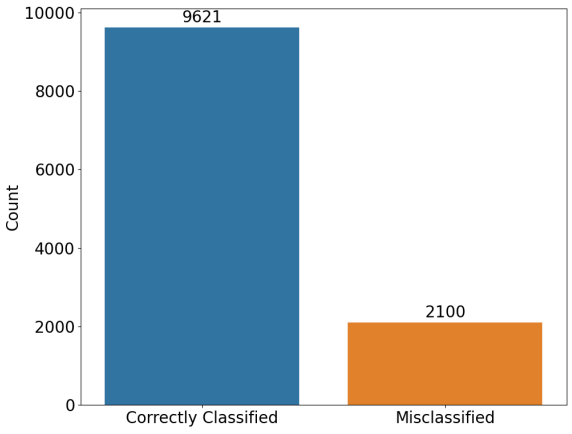

Figure 2: Bar chart showing correctly classified and misclassified intents of AraBERTv2 model.

Qualitative Analysis

Table 6 presents some predicted outputs of the proposed model. In the first and third samples, the model successfully predicted the intent of the text. On the other hand, it failed to do so in the second sample. This is challenging to predict because queries regarding top up limits and top up reverted can be very similar and difficult to differentiate.

Query	Actual	Predicted
متى بعتتلي بطاقتي الجديدة؟ (When did you send me my new card?)	card arrival	card arrival
كيف بقدر أعبي كمان مرة ؟ (How much could I top up?)	top up limits	top up reverted
هل في حد لعمر الشخص؟ (Do you have a limit for someone's age?)	age limit	age limit

Table 6: Some example predictions made by the proposed AraBERTv2 model with actual intent and predicted intent.

6 Conclusion

This study investigated the efficacy of several transformer-based, DL, and ML models on the Arabic banking dataset to detect intent. The results demonstrate that the transformer-based model (AraBERTv2) performed superiorly in subtask-1, achieving the highest F1 score of 82.08%. The study suggests that one of the main reasons for the proposed model's poor performance is its inability to generalize to unseen Arabic dialects. Future research could investigate advanced large language models (LLMs) to achieve better performance.

References

Wassen Aldjanabi, Abdelghani Dahou, Mohammed AA Al-qaness, Mohamed Abd Elaziz, Ahmed Mohamed Helmi, and Robertas Damaševičius. 2021. Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. In *Informatics*, volume 8, page 69. MDPI.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

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

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.

Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, and Sana Ghanem. 2023a. Arbanking77: Intent detection neural model and a new dataset in modern and dialectical arabic. *arXiv preprint arXiv:2310.19034*.

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

Saurabh Kumar, Suman Deep, and Pourush Kalra. Enhancing customer service in banking with ai: Intent classification using distilbert.

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):846.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Sana Shams, Bareera Sadia, and Muhammad Aslam. 2022. Intent detection in urdu queries using fine-tuned bert models. In *2022 16th International Conference on Open Source Systems and Technologies (ICOSST)*, pages 1–6. IEEE.

Tokunaga Takenobu. 1994. Text categorization based on weighted inverse document frequency. *Information Processing Society of Japan, SIGNL*, 94(100):33–40.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Lihua Wang, Huiting Yang, Feng Li, Wenzhong Yang, and Zhenwan Zou. 2022. Intent detection model based on dual-channel feature fusion. In *2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC)*, volume 6, pages 1862–1867. IEEE.

Pei Wang, Keqing He, Yejie Wang, Xiaoshuai Song, Yutao Mou, Jingang Wang, Yunsen Xian, Xunliang Cai, and Weiran Xu. 2024. Beyond the known: Investigating llms performance on out-of-domain intent detection. *arXiv preprint arXiv:2402.17256*.

Sevinj Yolchuyeva, Géza Németh, and Bálint Gyires-Tóth. 2020. Self-attention networks for intent detection. *arXiv preprint arXiv:2006.15585*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

Jianguo Zhang, Kazuma Hashimoto, Yao Wan, Zhiwei Liu, Ye Liu, Caiming Xiong, and Philip S Yu. 2021. Are pretrained transformers robust in intent classification? a missing ingredient in evaluation of out-of-scope intent detection. *arXiv preprint arXiv:2106.04564*.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020b. Intent detection with wikihow. *arXiv preprint arXiv:2009.05781*.