BabelBot at AraFinNLP2024: Fine-tuning T5 for Multi-dialect Intent Detection with Synthetic Data and Model Ensembling

Murhaf Fares
Independent researcher
murhaf@proton.me

Samia Touileb
University of Bergen
samia.touileb@uib.no

Abstract

This paper presents our results for the Arabic Financial NLP (AraFinNLP) shared task at the Second Arabic Natural Language Processing Conference (ArabicNLP 2024). We participated in the first sub-task, Multi-dialect Intent Detection, which focused on cross-dialect intent detection in the banking domain. Our approach involved fine-tuning an encoder-only T5 model, generating synthetic data, and model ensembling. Additionally, we conducted an in-depth analysis of the dataset, addressing annotation errors and problematic translations. Our model was ranked third in the shared task with an F1-score of 0.871.

1 Introduction

Financial Natural Language Processing (NLP) currently plays a crucial role in the Arab world. This is particularly true in the light of the considerable growth observed in the stock markets in the Middle East, and the diverse sectors contributing to this expansion (Zmandar et al., 2021). Given this evolution, it is paramount to develop Arabic NLP tools that are able to address local linguistic nuances.

Despite its widespread use, Arabic faces a scarcity of labelled data especially when it comes to dialects and domain-specific tasks (Darwish et al., 2021). This scarcity poses a challenge to most tasks focusing on Arabic, but more particularly for largely unexplored tasks such as intent detection (Jarrar et al., 2023).

The Arabic Financial NLP (AraFinNLP) (Malaysha et al., 2024) shared task is dedicated to enhancing Arabic NLP capabilities within the financial domain while also addressing the linguistic diversity of the Arab world. To this end, AraFinNLP 2024 introduces two subtasks:

1. *Multi-dialect Intent Detection*: in this subtask, the focus is to develop models that can handle cross-dialect intent detection in the banking domain; *i.e.*, classify customer intents from queries expressed in various Arabic dialects.
2. *Cross-dialect Translation and Intent Preservation*: this subtask aims to ensure precise intent preservation during translation across various Arabic dialects.[1]

In this paper, we present our participation in sub-task 1. Our contribution involves fine-tuning pre-trained language models (namely, a T5 model) for the task at hand, along with generating synthetic data using (much) large(r) language models.

In Section 2 we describe the ArBanking77 (Jarrar et al., 2023) dataset, conduct an analysis of its annotation quality and introduce the synthetic data we created to improve the performance of our model. In Section 3 we present our proposed model, and describe our experiments and results in Section 4. Finally, we summarise our main findings, and discuss possible future work in Section 5.

2 Data

2.1 ArBanking77

ArBanking77 was derived from the English Banking77 dataset by Casanueva et al. (2020). The queries in ArBanking77 were automatically translated to Modern Standard Arabic (MSA) and Palestinian Arabic (PAL) using Google Translate. The translated queries were then manually corrected by native speakers (Jarrar et al., 2023). The training and development splits in ArBanking77 contain queries in both MSA and PAL; Table 1 shows the number of queries per split and dialect.

While the distribution of intent queries is quite similar between MSA and PAL, the distribution of intent classes reveals a clear class imbalance. Additionally, there are differences in the distribution of intent classes within each language. For example,

[1] See the shared task's website for further details https://sina.birzeit.edu/arbanking77/arafinnlp

Proceedings of The Second Arabic Natural Language Processing Conference, pages 433–440

MSA train	MSA dev	PAL train	PAL dev
10,733	1,230	10,821	1,234

Table 1: Distribution of total number of intent queries in the train and dev splits of MSA and PAL data.

in MSA the most frequent class is '*transfer not received by recipient*', whereas in the Palestinian data, the most frequent class is '*beneficiary not allowed*'. The least frequent class in both dialects is the '*contactless not working*' intent class. Figure 1 in Appendix C shows the class distributions in the train splits of MSA and Palestinian dialect.

2.1.1 Data Quality

As previously mentioned, the dataset was derived using machine translation (MT) and manually reviewed afterwards. In addition, Ying and Thomas (2022) report that 14% of the training utterances in the English dataset "may have been incorrectly labelled". Consequently, we conducted an investigation into the quality of the ArBanking77 dataset and were able to identify several types of errors. We list those errors in the following and provide more examples in Table 6 in Appendix C.

Misspelled English words that were directly translated to Arabic There are misspelled words in the English data that were directly translated into Arabic. For instance, the sentence "*I tired but an unable to activate my card*" was translated to لقد تعبت ولكني غير قادر على تفعيل بطاقتي. The English sentence was clearly conveying that the person was trying to activate their card, rather than that they were tired, as translated in the Arabic version.

Translation errors These are straightforward translation errors, which may have arisen due to the MT system and escaped detection during the human review process. For example the sentence "*Where do you have locations at?*" was translated as أين لديك مواقع في. Certain erroneous translations occurred repeatedly in the dataset. One such example is the word "charge" that has been translated as تهمة (accusation) on 17 occasions in the dataset; e.g. *"I've never been to that store. That's a fraudulent charge"* was translated as لم أذهب إلى هذا المتجر أبدا. هذه تهمة احتيالية. The word "support" has also been inaccurately translated on many occasions to convey a sense of endorsement or agreement, as for example in the sentence "*Do you support all countries?*" translated as هل تؤيد كل الدول؟.

Mislabelling This type of errors might have originated from the mislabelling of the English data as pointed out by Ying and Thomas (2022). For example the sentence "*How can I get a new card?*" has been labelled as "Contactless not working".

Unusual translations These instances could be categorised as translation errors, although they do not strictly fall into the traditional definition of such errors. Interestingly, many of those sentences often contain uncommon words or relatively new concepts (such as contactless payment), resulting in a somewhat peculiar sentence in Arabic. For example, the sentence "*I can't get a contactless payment to work*" was translated in a rather awkward-sounding way as لا يمكنني الحصول على مدفوعات لاتلامسية للعمل.

Given the presence of such annotation errors and the relatively small size of the dataset, we decided to merge the MSA and PAL splits in both the training and development sets. However, while this approach may yield a model proficient in handling these specific Arabic variations, the overarching challenge of developing a system that accommodates other dialects persists. Consequently, we opted to create a synthetic dataset, with the hope of achieving improved results across various Arabic dialects.

2.2 Synthetic Data

We augmented the dataset provided by the shared task organisers by translating a large subset of the ArBanking77 data into Moroccan, Tunisian, and Saudi dialects using Cohere's multilingual model `Command R+`, which supports Arabic among other languages.[2]

We prompted `Command R+` to translate 8,694 queries from MSA into Moroccan, 8,067 into Tunisian, and 7,885 into Saudi, ensuring that the label distribution in each dialect closely mirrors that of the original (MSA) data. Appendix A shows the prompts we used for this process. We applied a few simple post-processing steps to: (1) ensure that the generated queries are different from the MSA source (viz. surface form comparison between the

[2]We used `Command R+` via Cohere's (free tier) API, but the model is also available on https://huggingface.co/CohereForAI/c4ai-command-r-plus

Dialect	Train	Dev
Moroccan	8,694	833
Tunisian	8,067	949
Saudi	7,885	1,022
MSA Cohere	4,014	
Total	28,660	2,804

Table 2: Number of queries in the synthetic dataset.

source and generated strings) and (2) exclude the occasional gibberish output by `Command R+` (using a length-based heuristic).[3]

We also translated a subset of the development data to the aforementioned dialects using `Command R+` because the ArBanking77 development set only contains queries in PAL and MSA. We used the synthetic development set to guide our experiments, but the ultimate utility of the dialect-augmented data could only be determined on the test set, which includes queries in all target dialects.

Finally, to further enhance the variability in the training dataset, we (re-)translated an additional 4,014 queries from English into MSA Arabic using `Command R+`, as the ArBanking77 dataset includes the original queries in English. The rationale behind this step is to introduce a broader range of expressions and potentially improve the model's robustness on the test set. Table 2 summarises the number of synthetic data we generated per dialect.[4]

3 System Description

The main idea behind our system aligns with earlier methods as it involves fine-tuning a pre-trained language model (Jarrar et al., 2023). Nevertheless, we opted for a different pre-trained model and augmented the training data.

We fine-tuned (the encoder of) a T5-based model (Raffel et al., 2020) on the combination of the ArBanking77 dataset and the synthetic data introduced in Section 2. The T5 architecture is an encoder-decoder transformer model (*i.e.* sequence-to-sequence model), but it has been previously used for regression, classification, ranking, and sentence embedding tasks (Ni et al., 2022; Zhuang et al., 2023; Do et al., 2024). To use T5 for classification, we added a classification head on top of its encoder and fine-tuned the model to classify intents.[5]

Concretely, we fine-tuned the model by Fares (2024), which in itself is a fine-tuned version of AraT5$_{v2}$ by Nagoudi et al. (2022).[6] We chose the model by Fares (2024) because it was trained to translate five regional Arabic dialects into MSA.[7] Those regional dialects include: the Gulf, Egyptian, Levantine, Iraqi, and Maghrebi dialects.

As we explain in Section 4.1, for our final submission we trained three separate T5-based classifiers and ensembled their output. Our ensembling approach is rather straightforward as it boils down to collecting the labels from three models and finding the majority label. In case of a tie, we select the label with the highest prediction score.

Implementation Details

We trained our model using the Transformers library (Wolf et al., 2020), but we had to implement a custom class to enable training an encoder-only T5 model for intent classification.[8] Otherwise, all of our models and experiments use the same hyperparameters and configuration in Table 5 in Appendix B, except for the number of epochs and batch size (which had to be adjusted due to memory constraints).

4 Experiments and Results

In this section, we describe the series of experiments we conducted to develop our system. These experiments aimed to assess the usefulness of synthetic data and determine the extent to which a model can generalise to unseen dialects.

We fine-tuned the T5 model, by Fares (2024), using five different combinations of the gold and synthetic datasets and evaluated its performance on four distinct development sets.[9] The results of these experiments are presented in Table 3, where we can clearly see that using the combination of the ArBanking77 dataset and the synthetic dataset (Synthetic+Joint) leads to the best results across

[3]The length-based heuristic is only to exclude self-repetition in the output of `Command R+`.

[4]We prompted the `Command R+` model to translate the same number of queries into all dialects but the final numbers reported vary across dialects due to post-processing.

[5]The classification head is simply a dense layer, a dropout layer and an output layer.

[6]The model is available on https://huggingface.co/Murhaf/AraT5-MSAizer

[7]We experimented with masked language models such as `ARBERTv2` and `MARBERTv2` (Abdul-Mageed et al., 2021) and achieved similar results to the T5 model we chose.

[8]The custom class ensures that only the encoder part is used and updated during training. It also adds a classification head on top of the encoder.

[9]Due to time constraints, we trained each of the models for 10 epochs only.

Eval \ Train	MSA	PAL	Joint	Synthetic	Synthetic+Joint
MSA-Dev	0.8894	0.8146	0.9268	0.9089	**0.9406**
PAL-Dev	0.8128	0.8890	0.9230	0.8995	**0.9441**
Joint-Dev	0.8510	0.8518	0.9249	0.9042	**0.9424**
Synthetic-Dev	0.7222	0.7496	0.8099	0.8726	**0.8973**

Table 3: Train indicates the dataset used for training and Eval indicates the development dataset used for evaluation. The reported results are given in terms of micro-F1 scores. Joint refers to the concatenation of MSA and PAL.

all development splits. We also observe that using the MSA-PAL joint dataset improves performance compared to training with either dataset alone. To understand the reasons behind this improvement, we closely inspected the ArBanking77 dataset and found that translation of the same English queries can appear in both the MSA training set and the PAL development set (or vice versa). This overlap can lead to data leakage between the training and development sets when the dataset is combined at the MSA-dialect level In other words, training on the joint MSA-PAL datasets effectively trains the model on examples from the development set, albeit in a different dialect. This issue is arguably an oversight in the dataset design.

Further, it is evident that the model's performance drops on unseen dialects; for example, the model trained on MSA only exhibits a decrease in micro-F1 score by approximately 7 points on Palestinian Arabic (PAL-Dev). Lastly, one might be tempted to conclude that using only synthetic data is sufficient. However, it is important to remember that the synthetic data is simply derived from the gold data; that is, the two datasets are highly similar.

4.1 Test Results

Precision	Recall	F1
0.8723	0.8728	0.8709

Table 4: Precision, recall, and F1 values of our proposed model, ranked 3rd at AraFinNLP2024.

Our final submission to the shared task is, in fact, the result of model ensembling—a method recently used in the somewhat related shared task on Arabic dialect identification (Elkaref et al., 2023).

Specifically, we ran the test set through three fine-tuned T5 models and performed majority voting on the predicted labels. The first model was trained for 10 epochs on a combination of synthetic and ArBanking77 joint datasets, as described in the previous section. For the second model, we included the synthetic development data in the training process and trained it for 15 epochs. The third model underwent training for 30 epochs but on 90% of the data used for the second model. Table 4 shows the result of our final submission. Our model achieved a competitive performance, securing the third position among the eleven participating teams with a micro-F1 score of 0.8709, trailing closely behind the top two teams who achieved scores of 0.8773 and 0.8762.

While such an ensembling approach can lead to higher F1 scores, we question whether the computational resources required to run multiple models in a production environment are justified by the marginal gains obtained. In fact, each of the three ensembling models individually achieves approximately 0.86 in micro-F1 score on the official test dataset, suggesting that model ensembling may not be the optimal strategy in practical applications.

5 Conclusion

In this paper we presented our participation in the AraFinNLP shared task on the Multi-dialect Intent Detection. We employed a strategy involving fine-tuning an encoder-only T5 model, generating synthetic data, and model ensembling to address the challenges of cross-dialect intent detection in the banking domain. Our analysis of the dataset revealed various challenges, including annotation errors and problematic translations.

Experimental results demonstrate the effectiveness of our approach, particularly the significant performance improvement achieved by incorporating synthetic data. Our final model secured third place in the shared task with a competitive micro-F1 score of 0.8709.

While our approach yielded promising results,

we believe further improvement of the synthetic data quality is needed. One path forward would be to use a metric like pointwise V-information (Ethayarajh et al., 2022) to filter out synthetic queries with low PVI values (*i.e.* queries with little relevant information to their intents). Lin et al. (2023) implemented a similar approach for intent classification under few-shot settings, that yielded good results.

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.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient Intent Detection with Dual Sentence Encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R El-Beltagy, Wassim El-Hajj, et al. 2021. A panoramic survey of natural language processing in the arab world. *Communications of the ACM*, 64(4):72–81.

Heejin Do, Yunsu Kim, and Gary Lee. 2024. Autoregressive Score Generation for Multi-trait Essay Scoring. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1659–1666, St. Julian's, Malta. Association for Computational Linguistics.

Mohab Elkaref, Movina Moses, Shinnosuke Tanaka, James Barry, and Geeth Mel. 2023. NLPeople at NADI 2023 shared task: Arabic Dialect Identification with Augmented Context and Multi-Stage Tuning. In *Proceedings of ArabicNLP 2023*, pages 642–646, Singapore (Hybrid). Association for Computational Linguistics.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Murhaf Fares. 2024. AraT5-MSAizer: Translating Dialectal Arabic to MSA. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 124–129, Torino, Italia. ELRA and ICCL.

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.

Yen-Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2023. Selective In-Context Data Augmentation for Intent Detection using Pointwise V-Information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1463–1476, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-T5: Scalable Sentence Encoders from Pretrained Text-to-Text Models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

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.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. RankT5: Fine-Tuning T5 for Text Ranking with Ranking Losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2308–2313, New York, NY, USA. Association for Computing Machinery.

Nadhem Zmandar, Mahmoud El-Haj, and Paul Rayson. 2021. Multilingual Financial Word Embeddings for Arabic, English and French. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4584–4589. IEEE.

A Synthetic Data Prompts

We used the following prompt to translate a selection of MSA samples into Moroccan Arabic. The same prompt was also used for Tunisian and Saudi dialects; we simply replaced the dialect name in the prompt.

> Translate the following sentences from MSA to Moroccan dialect. Make sure each translation is separated by a new line. Keep the ID in the start.
>
> ...
>
> ...

B Hyperparameters

Hyperparameter	Value(s)
Learning Rate	0.00002
Optimizer	adamw
Batch Size	4,8
Epochs	10,15,30
Seed	42
Learning Rate Scheduler Type	Linear
Weight Decay	0.0001
Dropout Rate	0.1

Table 5: Training hyperparameters

C Dataset Inspection

Figure 1 shows the class distribution in the train splits of MSA and PAL in ArBanking77.

Table 6 shows more examples of the translation and annotation errors we found in the ArBanking77 dataset.

Issue	Example
Misspelling	Where are my funds? I topped off my car but it didn't seem to complete. أين أموالي؟ تصدرت سيارتي ولكن يبدو أنها لم تكتمل.
	I just had a look at my statement. Why have I been changed for using the ATM? اتفرجت على بياناتي، ليش تغير استخدامي للصراف؟
Translation errors	I want to top up with my Apple Watch. أريد تعبئة ساعة آبل الخاصة بي.
	If I want a friend to top off my account, can they? إذا كنت أرغب في أن يقوم صديق بإغلاق حسابي ، فهل يمكنهم ذلك؟
	I need my card topped up, بدي بطاقتي
	How long does it take for my top up to clear? الوقت الذي تطلبه لازالة التعبئة؟
	How do you top-up using cash? كيف ممكن أضيف على رصيدي ؟
Mislabelling	How will I know my PIN number? كيف اعرف الباسورد تبعي؟ *Label: Get physical card*
	can i create my own pin right away هل يمكنني إنشاء رقم التعريف الشخصي الخاص بي على الفور؟ *Label: Get physical card*
	Can you tell me what i steps i should take since my card was stolen? بتقدر اتقلي بعد ما انسرقت بطاقتي ايش بقدر أسوي؟ *Label: lost or stolen phone*
	can i go into my app to find my pin? بقدر أدخل عالتطبيق لأعرف الرقم السري *Label: get physical card*

Table 6: Selected examples of the translation and labelling errors we identified in the ArBanking77 dataset

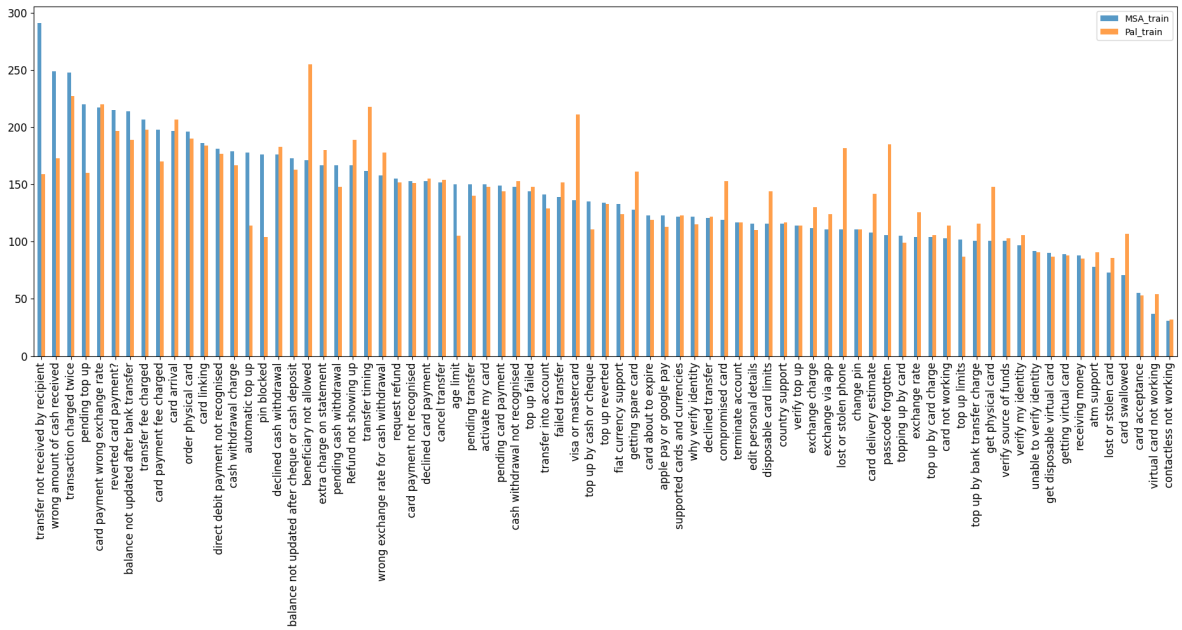

Figure 1: Distribution of intent classes in train splits of MSA and Palestinian dialect.