BFCI at AraFinNLP2024: Support Vector Machines for Arabic Financial Text Classification

Nsrin Ashraf[1] **Hamada Nayel**[1] **Mohammed Aldawsari**[3] **Shashirkha H. L.**[2]

Tarek Elshishtawy [1]

[1]Faculty of Computers and Artificial Intelligence, Benha University, Egypt
[2]Department of Computer Science, Mangalore University, India
[3]Faculty of Engineering, Prince Sattam bin Abdulaziz University, Saudi Arabia
Correspondence: nisrien.ashraf19@fci.bu.edu.eg

Abstract

In this paper, a description of the system submitted by BFCAI team to the AraFinNLP 2024 shared task has been introduced. Our team participated in the first subtask, which aims at detecting the customer intents of cross-dialectal Arabic queries in the banking domain. Our system follows the common pipeline of text classification models using primary classification algorithms integrated with basic vectorization approach for feature extraction. Multi-layer Perceptron, Stochastic Gradient Descent and Support Vector Machines algorithms have been implemented and support vector machines outperformed all other algorithms with an f-score of 49%. Our submission's result is appropriate compared to the simplicity of the proposed model's structure.

1 Introduction

Natural Language Processing (NLP) is massively growing area of research. It has been implemented in general domain as well as specific domains such as biomedical (Shashirekha and Nayel, 2016), scientific, and financial (Patil et al., 2023) domains. A system that summarizes financial documents automatically using NLP approaches has been developed by Patil et. al (Patil et al., 2023).

Processing Arabic texts is extensively growing due to the increasing amount of texts written in Arabic. The complex morphological structure is the major challenge that faces researchers in Arabic Natural Language Processing (NLP) (AbuElAtta et al., 2023). Financial Arabic NLP is introduced as a result of the exponential growth of Middle Eastern stock markets with abroad range of diverse sectors. This rise, which is occurring in multiple regions, is indicative of the dynamic financial landscape in the region and is drawing interest and investment from around the world. NLP technologies are becoming essential as these markets develop in order to handle regional linguistic quirks and serve the international financial community that is involved in these exchanges. The efficient analysis and interpretation of financial data depends on the development of Arabic NLP tools in the banking industry (Zmandar et al., 2023, 2021). The Arabic Financial NLP (AraFinNLP) shared task aims at analysing financial texts written in Arabic (Malaysha et al., 2024).

AraFinNLP shared task aims at improving Financial Arabic NLP. It consists of two subtasks namely; Multi-dialect Intent Detection, and Cross-dialect Translation and Intent Preservation, in the banking domain. These subtasks are crucial for interpreting and managing the diverse and complex banking data prevalent in Arabic-speaking regions. Detecting intent in financial communications, particularly in bots, across various Arabic dialects, can enhance customer service, and automate query handling. This ensures inclusivity and efficiency in catering to a linguistically diverse customer base. The Dialectical Translation aspect is particularly significant given the linguistic diversity in the Arab world. It ensures that NLP models are not only accurate when dealing with Modern Standard Arabic (MSA) but also effective across diverse Arabic dialects. This advancement will open up new applications in areas like automated customer support, real-time financial news analysis, and enhanced accessibility for diverse Arabic-speaking populations, making financial services more inclusive and efficient.

ArBanking77 (Jarrar et al., 2023), a translated version of English Banking77 dataset (Casanueva et al., 2020) into MSA and Palestinian Arabic has been used for AraFinNLP 2024. Task organizers shared the dataset across all participants, providing training, development and blind test sets. Participants have to develop their models using train and development sets, and the performance of their models will be evaluated using the blind test set.

Our team participated on the first subtask: Multi-

Proceedings of The Second Arabic Natural Language Processing Conference, pages 446–449
August 16, 2024 ©2024 Association for Computational Linguistics

dialect Intent Detection in the banking domain. The subtask automates the classification process of customer intents given the query in different Arabic dialects. We proposed a machine learning-based model for the first subtask, by integrating a basic vectorization approach and support vector machines algorithm. Such approach has been applied for various arabic NLP tasks (Ashraf et al., 2022).

2 Dataset

The dataset given for the shared task was obtained from English Banking77 Dataset (Casanueva et al., 2020) consists of 13,083 costumer service queries divided into 77 classes according to the customer intents. ArBanking77 dataset (Jarrar et al., 2023) contains MSA and Palestinian Arabic, For the proposed Model Our team participated in Subtask-1 using the MSA dataset, as the statistics of dataset is shown in Table 1.

Train	Validation	Test
10733	1230	11721

Table 1: ArBanking77 Dataset Statistics.

3 Proposed Model

In this section, a detailed description of the proposed model including the main structure, text representation, algorithms, experimental setup and evaluation metrics has been presented. As shown in Figure 1, the first phase of the proposed model is data distribution for MSA data presented in Table 1. The second phase is feature engineering, which aims at converting the text into numerical values efficiently. The third phase is training the model, or implementation of the classification algorithm. The final phase is evaluation, which is performed in the organizers side. The following subsections give more details for each phase separately.

3.1 Feature Engineering

Feature engineering phase proposes a conversion of the row text data into numerical values, as most algorithms accept numerical input of a fixed size rather than text data of different sizes. A collective technique employs a document-term vector, in which each document is encoded as a discrete vector that adds occurrences for each word in the vocabulary it includes (Rameshbhai and Paulose, 2019). Term Frequency/Term Document Frequency, *TF/IDF*, is one of simple and efficient text vectorization approaches where it takes into account the importance of each term in the document. It can handle high-dimensional data by reducing the feature space to the most important terms. In this work, we employed *TF-IDF* with n-gram model using *uni*-gram ($n = 1$) and *bi*-gram ($n = 2$). N-gram model captures more complex relationships between words, by considering both single words and pair of words, other parameters set CountVectorizer CV=5, MaxDF=1.0. This method can lead to more accurate text analysis (Nayel, 2020).

3.2 Training the Model

In this phase, the resulted numerical features are used as input to the classification algorithm for training. Three classification algorithms have been implemented to broaden the experiments namely; Multi-Layer Perceptron (MLP), Support Vector Machine (SVM) and Stochastic Gradient Descent (SGD). These algorithms have been successfully implemented for different Arabic NLP tasks such as dialect identification (Sobhy et al., 2022) and offensive language identification (Nayel, 2020)

- **SVM** is one of the most effective ML text classifiers in multi-label text classification which has the ability to handle high-dimensional data with a relatively minimal number of training samples.

- **MLP** is an artificial neural network that is widely utilized in classification and regression problems. In the context of multi-label text categorization, MLPs can be utilized to learn the complicated relationships between input text and output labels.

- **SGD** is more efficient in Multi-label text classification compared to other optimization algorithms which has several parameters that can be tuned to improve performance (Diab, 2019).

3.3 Evaluation

F1-score, also known as F-measure, is a harmonic mean of Precision (P) and Recall (R). It has been used to evaluate the model performance. Equation 1, shows that calculation of F1-score using P and R.

$$F1 - score = \frac{2 * P * R}{P + R} \quad (1)$$

F1-score is commonly used for multi-class classification problems

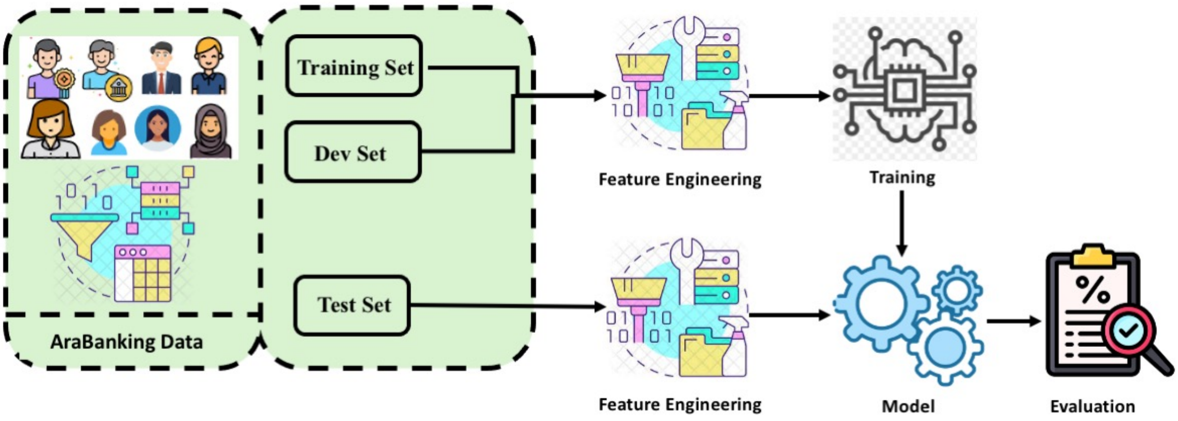

Figure 1: The general structure of the proposed model

4 Experimental Setup and Results

We used Jupyter notebook [1] for GPU support with `python` programming language and `Sklearn` (Pedregosa et al., 2011) library was utilized for machine learning classifiers. To achieve the same results, we set random state of the random seed package at a specific value for each classifier. The hyper-parameters of the implemented classifiers are as follows:

- MLP has been trained using the following back-propagation variations: `hidden layers` = 20, solver = ’`adam`’, activation function = `logistic`, maximum iteration = 500 and random state = 42
- SGD has been trained using ’`log`’ loss function, penalty = 12, maximum iteration = 1000 and random state = 10
- SVM uses ’`linear`’ kernel, regularization = ’`L1`’ and random state = 42

Table 2 shows F1-score reported on development phase for the three models. It is clear that SVM outperforms both SGD and MLP. We submitted the output of SVM to the shared task, and the result of SVM for the test set is 49.10% f1-score.

Classifier	F1-Score
MLP	84.46%
SVM	85.12%
SGD	74.65%

Table 2: ArBanking77 F1-Score for Development Phase

5 Discussion

The proposed model follows the basic pipeline of NLP-based model and no external resources have been employed. The issue of overfitting has been raised in the proposed model, where f1-scores are disparately resulted on development set and the blind test set. The SVM classifier outperform MLP and SGD classifiers in the presence of overfitting during the training phase due to its built-in regularization term that stimulate the model to find a simpler decision boundary. Using TF-IDF vectorization approach and machine learning algorithms resulted 49% Micro F1-Score achieving rank 11 in the shared task.

6 Future Work

This work can be enhanced by employing additional classification algorithms, where we implemented only three basic algorithms. In addition, deep neural networks can be implemented for the given task. On the other hand, using advanced text vectorization approach can lead to improvement of the performance. Finally, large language models (LLMs) can be customised to this task.

[1] https://jupyter.org/

References

Ahmed H. AbuElAtta, Mahmoud Sobhy, Ahmed A. El-Sawy, and Hamada Nayel. 2023. Arabic regional dialect identification (ardi) using pair of continuous bag-of-words and data augmentation. *International Journal of Advanced Computer Science and Applications*, 14(11).

Nsrin Ashraf, Hamada Nayel, and Mohamed Taha. 2022. Misinformation detection in arabic tweets: A case study about covid-19 vaccination. *Benha Journal of Applied Sciences*, 7(5):265–268.

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.

Shadi Diab. 2019. Optimizing stochastic gradient descent in text classification based on fine-tuning hyper-parameters approach. A case study on automatic classification of global terrorist attacks. *CoRR*, abs/1902.06542.

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.

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.

Hamada Nayel. 2020. NAYEL at SemEval-2020 task 12: TF/IDF-based approach for automatic offensive language detection in Arabic tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2086–2089, Barcelona (online). International Committee for Computational Linguistics.

Krutika Patil, Medha Badamikar, and Sheetal Sonawane. 2023. Nlp based text summarization of fintech rfps. In *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, pages 865–869.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Chaudhary Jashubhai Rameshbhai and Joy Paulose. 2019. Opinion mining on newspaper headlines using svm and nlp. *International journal of electrical and computer engineering (IJECE)*, 9(3):2152–2163.

H. L. Shashirekha and Hamada A. Nayel. 2016. A comparative study of segment representation for biomedical named entity recognition. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1046–1052.

Mahmoud Sobhy, Ahmed H. Abu El-Atta, Ahmed A. El-Sawy, and Hamada Nayel. 2022. Word representation models for Arabic dialect identification. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 474–478, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Nadhem Zmandar, Mo El-Haj, and Paul Rayson. 2023. FinAraT5: A text to text model for financial Arabic text understanding and generation. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 262–273, Vienna, Austria. NOVA CLUNL, Portugal.