dzNLP at NADI 2024 Shared Task: Multi-Classifier Ensemble with Weighted Voting and TF-IDF Features

Mohamed Lichouri
LCPTS-FGE. USTHB
Algiers-ALGERIA
mlichouri@usthb.dz

Khaled Lounnas
CRSTDLA
Algiers-ALGERIA
k.lounnas@crstdla.dz

Zahaf Boualem Nadjib and Rabiai mehdi ayoub
University of Algiers 01
Algiers-ALGERIA
{zahafndjib,Rabiaimehdiayoub}@gmail.com

Abstract

This paper presents the contribution of our dzNLP team to the NADI 2024 shared task, specifically in Subtask 1 - Multi-label Country-level Dialect Identification (MLDID) (Closed Track). We explored various configurations to address the challenge: in Experiment 1, we utilized a union of n-gram analyzers (word, character, character with word boundaries) with different n-gram values; in Experiment 2, we combined a weighted union of Term Frequency-Inverse Document Frequency (TF-IDF) features with various weights; and in Experiment 3, we implemented a weighted major voting scheme using three classifiers: Linear Support Vector Classifier (LSVC), Random Forest (RF), and K-Nearest Neighbors (KNN).

Our approach, despite its simplicity and reliance on traditional machine learning techniques, demonstrated competitive performance in terms of F1-score and precision. Notably, we achieved the highest precision score of 63.22% among the participating teams. However, our overall F1 score was approximately 21%, significantly impacted by a low recall rate of 12.87%. This indicates that while our models were highly precise, they struggled to recall a broad range of dialect labels, highlighting a critical area for improvement in handling diverse dialectal variations.

1 Introduction

Arabic, with its rich linguistic diversity encompassing numerous dialects alongside Modern Standard Arabic (MSA), presents a significant challenge for natural language processing (NLP) tasks. Despite its importance, many Arabic dialects remain understudied due to resource constraints such as limited research funding and datasets. Addressing this gap, the Nuanced Arabic Dialect Identification (NADI) shared task series (Abdul-Mageed et al., 2020, 2021, 2022, 2023) serves as a pivotal initiative to overcome these hurdles. By providing extensive datasets and modeling opportunities, NADI aims to facilitate dialect identification and processing tasks, thereby advancing the understanding and utilization of Arabic dialects.

In the context of feature extraction for Arabic text analysis, standard approaches frequently employ Term Frequency-Inverse Document Frequency (TF-IDF). Our previous work (Abbas et al., 2019) in the MADAR'2019 shared task (Bouamor et al., 2019) demonstrated the effectiveness of using TF-IDF for efficient feature extraction in Arabic text classification. Drawing inspiration from this prior research, we adopted a similar strategy in our current study. Specifically, in our first experiment, we employed a union of TF-IDF features with various n-gram analyzers (word, character, and character with word boundaries) to enhance the representation of Arabic text.

Moreover, the utilization of weighted feature fusion has emerged as a promising technique for enhancing classification accuracy. (Lichouri et al., 2021) demonstrated the efficacy of weighted feature fusion in multilingual text classification, as showcased in our earlier study in NADI'2023 for Arabic dialect identification (Lichouri et al., 2023), emphasizing its adaptability across various linguistic datasets. This methodology closely corresponds to our Experiment 2, which focuses on the weighted concatenation of TF-IDF features.

These earlier studies, along with recent advancements in TF-IDF-based feature extraction and weighted feature fusion, have collectively inspired innovative ideas. In our third experiment, we implemented a weighted hard major voting approach.

This paper, instead of introducing innovative solutions or groundbreaking insights for NADI 2024 (Abdul-Mageed et al., 2024), serves as a concise

Proceedings of The Second Arabic Natural Language Processing Conference, pages 754–757
August 16, 2024 ©2024 Association for Computational Linguistics

consolidation of existing knowledge. The rest of the paper is structured as follows: Section 2 provides an overview of the dataset used in our study. In Section 3, we present our proposed system, followed by discussing the findings and their significance in Section 4. Finally, our paper concludes in Section 5, summarizing the key takeaways and contributions.

2 Description of the Dataset

The Nuanced Arabic Dialect Identification (NADI) 2024 shared task provided a valuable dataset for Subtask 1: Multi-label Country-Level Dialect Identification (MLDID). The dataset encompassed dialects from various Arabic-speaking countries, including Egyptian, Saudi Arabian, Algerian, Syrian, Palestinian, and Lebanese.

The dataset provided for MLDID consisted of two parts: a development set, comprising 100 samples, served as a training ground for system development and evaluation, and a test set, containing 1,000 samples, was used for the final evaluation of participant models.

Additionally, participants were granted access to NADI datasets from 2020, 2021, and 2023 for training purposes. These datasets provided a larger pool of Arabic text for system development and refinement.

3 Proposed system

Our proposed system is based on two concatenation or union process of features vs classifiers where we opted for three experiments:

- **Experiment 1 (Lichouri et al., 2018; Abbas et al., 2019):** In this experiment, we employ the TF-IDFVectorizer, which employs three analyzers (Word, Char, and Char_wb), each with varying n-gram ranges. In the default configuration, we combine these three features, assigning equal weights of 1 to each. During feature extraction, we varied the n-gram values (ranging from $n = 1$ to 5). Finally, the SVC classifier was trained.

- **Experiment 2 (Lichouri et al., 2021, 2023):** In this instance, we combine the three TF-IDF features using a weight vector comprising three distinct values (w1, w2, w3) corresponding to the Word, Char, and Char_wb analyzer, respectively. The LSVC classifier was then trained.

- **Experiment 3:** In this instance, we applied a weighted hard major voting for three classifier (LSVC, RF and KNN).

Table 1: The various parameters used in our work

Settings	Values
N-GRAM	(1,n) with n=1 to 5
TFIDF	transformer_weights: 0.1 - 1 max_features: 300-1000

LSVC	C= 1 - 5 and
	class_weight='balanced'

RF	Default
KNN	n_neighbors=3
Major Voting	weight = 0.1 - 0.6

The various parameters used in these three experiments are reported in Table 1 where we specify the settings and corresponding values for each parameter. For N-GRAM, we explored n-gram ranges from unigrams to 5-grams. In TFIDF, we varied the transformer weights from 0.1 to 1 and the maximum number of features from 300 to 1000. Additionally, for the LSVC classifier, we adjusted the regularization parameter C from 1 to 5 and set the class weight to 'balanced'. RF classifier was used with default settings, while for KNN, we set the number of neighbors to 3. Lastly, in the Major Voting ensemble technique, we experimented with weights ranging from 0.1 to 0.6.

4 Obtained results

During our participation in the MLDID (Closed Track) project, we iteratively explored various approaches to enhance our multi-label dialect identification models (see Table 1). This exploration encompassed feature engineering with TF-IDF, incorporating ensemble methods, and experimenting with different voting combination strategies.

As a baseline experiment, we utilized a basic TF-IDF representation with 1-gram features and a linear kernel SVC classifier. Despite its simplicity, it achieved a low F1-score of 19.43%.

In the first experiment, we expanded the feature space by incorporating a combination of word and character-level n-grams (word, char and char_wb) in TF-IDF representation. We employed the Linear Support Vector Classifier (LSVC) classifier with balanced class weights, resulting in notable improvements in F1-score of 20.64%.

In the second experiment, we experimented with a different combination of n-grams while maintain-

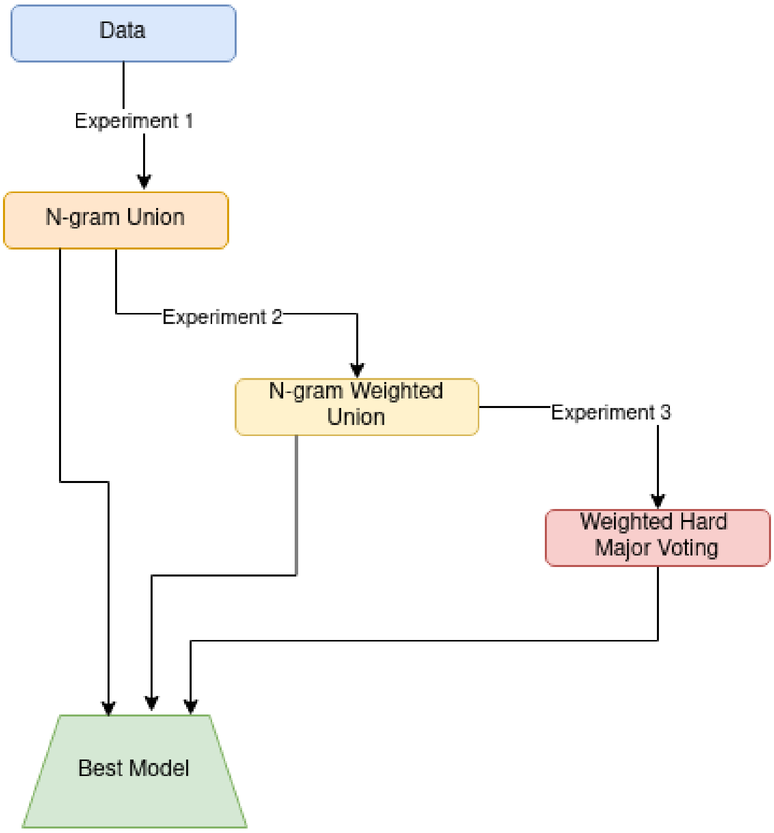

Figure 1: Proposed system for Arabic Dialect Identification.

Table 2: Obtained F1-score from merging TF-IDF representations with uniform and different weighting

Id	n-gram range	Classifier Configuration	Transformer weights	F1-score
1	(3, 5, 5)	class_weight='balanced', C=5		20.64
2	(5, 5, 5)	class_weight='balanced', C=4	tw={0.65, 0.85, 0.85}	22.51
3	(3, 4, 5)	C=4	tw={0.45, 0.5, 0.75}	21.07
4	(4, 4, 4)	C=4	tw={0.45, 0.5, 0.75}	20.78
5	(4, 4, 4)	C=4	tw={0.35, 0.45, 0.75}	20.51

ing the balanced class weights. We also explored variations in the weights assigned to different feature types within the TF-IDF representation. The obtained F1-score ranged from 20.51% to 22.51% (see Table 2).

In the third experiment, we introduced ensemble methods, namely hard voting and weighted hard voting, combining SVC, RF, and KNN classifiers. Interestingly, despite the complexity of these ensemble approaches, their F1-score were lower compared to well-tuned single LSVC configurations (16.33% and 21.44%, respectively). This suggests that the ensemble methods may not have effectively leveraged the complementary strengths of the individual classifiers or their voting strategies might not have been optimal for this task.

5 Conclusion

Our analysis of the NADI 2024 shared task highlights the critical role of feature engineering and model optimization in Arabic dialect identification. Key findings include integrating character-level features alongside word-level features consistently improved performance, balanced class weights within the LSVC classifier significantly enhanced F1-score, surpassing 20.64%, strategic assignment of transformer weights yielded the highest F1-score of 22.51%, and ensemble methods like hard voting achieved moderate F1-score but were surpassed by finely tuned single SVC configurations. In conclusion, our study emphasizes the importance of incorporating character-level information, utilizing balanced class weights, and exploring advanced feature weighting techniques to advance Arabic dialect identification systems. These insights provide valuable guidance for future research in this domain.

References

Mourad Abbas, Mohamed Lichouri, and Abed Alhakim Freihat. 2019. St madar 2019 shared task: Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 269–273.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, ElMoatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The Fourth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of The Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual).

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The third nuanced Arabic dialect identification shared task. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Mohamed Lichouri, Mourad Abbas, Abed Alhakim Freihat, and Dhiya El Hak Megtouf. 2018. Word-level vs sentence-level language identification: Application to algerian and arabic dialects. *Procedia Computer Science*, 142:246–253.

Mohamed Lichouri, Mourad Abbas, Khaled Lounnas, Besma Benaziz, and Aicha Zitouni. 2021. Arabic dialect identification based on a weighted concatenation of TF-IDF features. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 282–286, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Mohamed Lichouri, Khaled Lounnas, Aicha Zitouni, Houda Latrache, and Rachida Djeradi. 2023. Usthb at nadi 2023 shared task: Exploring preprocessing and feature engineering strategies for arabic dialect identification. *arXiv preprint arXiv:2312.10536*.