LTG-ST at NADI Shared Task 1: Arabic Dialect Identification using a Stacking Classifier

Samia Touileb
Language Technology Group
University of Oslo
Norway
samiat@ifi.uio.no

Abstract

This paper presents our results for the Nuanced Arabic Dialect Identification (NADI) shared task of the Fifth Workshop for Arabic Natural Language Processing (WANLP 2020). We participated in the first sub-task for country-level Arabic dialect identification covering 21 Arab countries. Our contribution is based on a stacking classifier using Multinomial Naive Bayes, Linear SVC, and Logistic Regression classifiers as estimators; followed by a Logistic Regression as final estimator. Despite the fact that the results on the test set were low, with a macro F1 of 17.71, we were able to show that a simple approach can achieve comparable results to more sophisticated solutions. Moreover, the insights of our error analysis, and of the corpus content in general, can be used to develop and improve future systems.

1 Introduction

Most resources for Arabic have been developed for Modern Standard Arabic (MSA) since it is the official language in most Arabic speaking countries. MSA is used in media coverage, politics, books, and even online. However, in each individual Arabic country, the predominant language used for everyday conversation (in real life and online) is a specific dialect for that region or country (Versteegh, 2014).

Arabic dialects are not standardized. There are no formal grammar rules nor formalism to guide the speakers (Zaidan and Callison-Burch, 2014). There are some efforts on creating standards for automatically processing such dialects (Habash et al., 2018), but the language use remains non-standardized. Within one same city, people can pronounce and write the same word differently. This aspect accentuates the difficulty of automatically processing such languages, and automatically distinguishing them from each other is as challenging, since no clear structure exists. Nevertheless, many attempts have been made for Arabic dialect identification. Despite the difficulty of the task, there are still syntactic and morphological aspects of the languages that can be exploited to differentiate them from each other.

Most works rely on machine learning approaches, and span various levels of accuracy depending on the dataset used and dialects being processed. The nearer (geographically) the countries are to each other, the more similar the spoken language, and therefore the more difficult it is to automatically distinguish the dialects (Bouamor et al., 2019).

The Multi Arabic Dialect Applications and Resources (MADAR) corpus (Bouamor et al., 2018) is an important resource for Arabic dialect identification. The corpus covers parallel sentences written in 25 Arabic city dialects from the travel domain. This corpus has been used in a shared task (Bouamor et al., 2019) for both fine-grained (26 dialects) and coarse-grained (6 dialects) Arabic dialect identification where various machine learning approaches have been used. A simple Mulitnomial Naive Bayes (MNB) has shown to be very powerful in the identification of the exact city of 26 dialects from the MADAR corpus with an accuracy of 67.9%, using as features character and word 5-grams language models and the output of the coarse-grained classifier (Salameh et al., 2018). Other approaches also focused on the use of machine learning and ensemble methods, using as features word counts, language models, and embeddings (Abu Kwaik and Saad, 2019; Meftouh et al., 2019; Ragab et al., 2019; Fares et al., 2019).

Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 313–319
Barcelona, Spain (Online), December 12, 2020

	# Tweets	# Tokens
train	21,000	270,574
dev	4,957	60,700
test	5,000	64,458

Table 1: Number of tweets and total number of tokens excluding punctuation in the three splits.

In this paper, we present our participation in the Nuanced Arabic Dialect Identification (NADI) shared task (Abdul-Mageed et al., 2020). We participated in the first sub-task aiming at country-level Arabic dialect identification from tweets covering 21 Arab countries. Our contribution is based on a stacking classifier using Multinomial Naive Bayes, Linear SVC, and Logistic Regression classifiers as estimators; and Logistic Regression as final estimator.

We have experimented with various architectures, from traditional machine learning approaches as the approach presented in this paper and clustering, but also with more recent approaches as Bi-LSTMS and CNNs. We have also experimented using unlabeled tweets (provided by the shared task organizers) to train embeddings and language models. However, none of these architectures gave satisfying results.

In Section 2 we describe the NADI tweet corpus. We present our proposed model in Section 3, and describe our results and give an overall discussion of the results in Section 4. Finally, we conclude in Section 5 and discuss possible future work.

2 Data

We used the NADI corpus provided by the shared task organizers (Abdul-Mageed et al., 2020). The corpus comprises 21,000 tweets covering 21 Arab countries for the first sub-task, and 100 provinces from the same Arab countries (as sub-task 1) in the second sub-task. In the following, we will only focus on the data set for the country-level classification. Each tweet in the corpus is annotated with its associated country (i.e. dialect) label. Table 1 gives an overview of the size of the corpus in the three splits (train, dev, and test) in terms of number of tweets (size of data), and the total number of tokens excluding punctuation.

The dialects of the following countries are included in the NADI corpus: Algeria, Bahrain, Djibouti, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tunisia, United Arab Emirates, and Yemen. Despite the relatively large amount of tweets used for training, the corpus is extremely unbalanced, which we believe has made this task even more difficult. As can be seen in Figure 1, the Egyptian dialect on it's own represents more than 21% of the training data, while the dialects from Djibouti, Bahrain, Sudan, Mauritania, and Somalia only represent 1% of the training data.

A further analysis of the corpus also showed that for all dialects except Egypt and Iraq, most of the word types were not representative of the language. Most vocabulary items are shared between dialects, which might confuse any classification system. In Figure 2 we show that the lightest (at the bottom) color represents the normalized number of unique words for each dialect. The color in the middle represents the normalized number of words shared between each dialect and up to four other dialects. The darkest color (at the top) represents the normalized number of words shared between each dialect with more than four other dialects.

From Figure 2 it is evident that 58% of the word types in the Egyptian and Iraqi tweets are unique to these dialects. However, for the remaining 19 dialects, over 50% of the word types were shared with other dialects. Our preliminary investigations have shown that many of these shared vocabulary items were actually MSA words. It is important to note that many dialectal words are shared with MSA.

For example the words شعور، أعتقد، مدريد، أجمل، النادي، عمرك، الفرح، يدخل respectively *enters*, *joy*, *your age*, *the club*, *more beautiful*, *Madrid*, *I believe*, *feelings* in English, were shared by more than 10 dialects. Another example is the word هُناك (*there*) which in the corpus is uniquely present in tweets labeled as Libyan, actually exists also in other forms in the corpus. When written هناكَ, it is only present

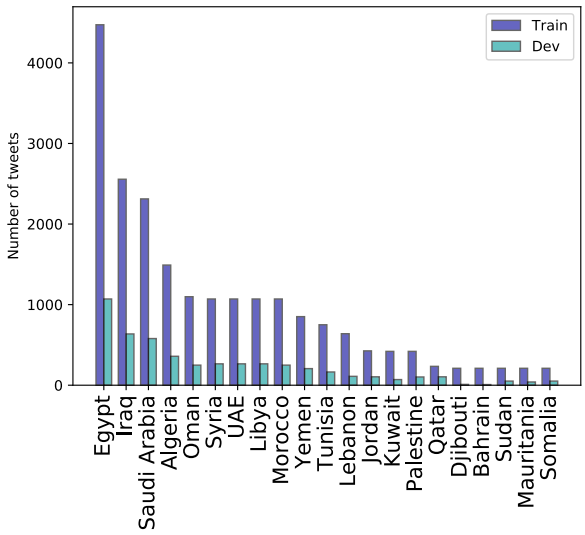

Figure 1: Number of tweets per dialect in both train and dev splits. This shows a very skewed distribution and the predominance of the Egyptian dialect in the data set.

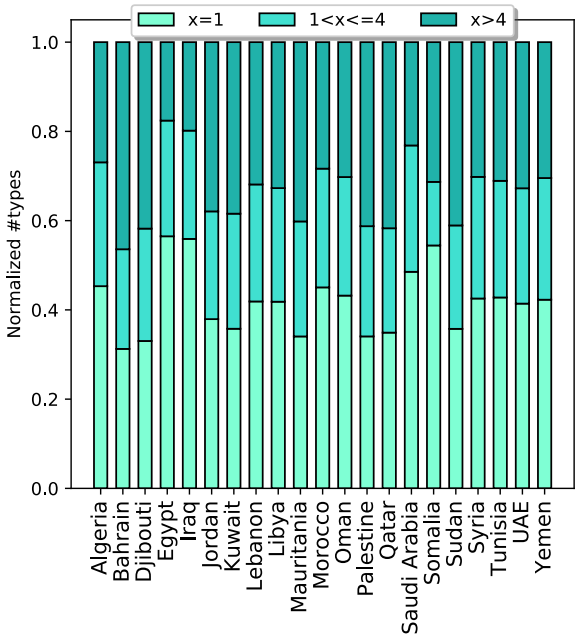

Figure 2: Normalized number of word types in each dialect. The lightest color represents unique words to each vocabulary. The color in the middle represents words shared with one to four other dialects, and the darkest color represents words shared with more than four other dialects.

in tweets from Iraq and Morocco, while when written in its most basic form without vowels, هناك, it is present in tweets labeled with the 17 dialects from Algeria, Djibouti, Egypt, Iraq, Jordan, Lebanon, Libya, Morocco, Oman, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tunisia, United Arab Emirates, and Yemen. This also plays a role in the difficulty of the task. As most dialects share their vocabulary, the words distinguishing them from each others might actually not be that different from common shared words. Therefore distinguishing the dialects from each other gets more challenging.

3 System

Our model, as shown in Figure 3, is based on a stacking classifier. We first train three different models on the NADI train split, and use their respective predictions to train a final estimator to get our final predictions. Stacking allows us to use the strengths of each individual classifier by using their output as input of a final estimator. All of the experiments are implemented using Python and the `scikitlearn` library (Pedregosa et al., 2011).

Our model classifies tweets into the NADI 21 dialects. It uses a combination of word and character n-grams and skipgrams concatenated using Feature Union estimator in sklearn. We give different weights to each feature vector, which were selected after a thorough analysis by experimenting with various possible weights (and combination of weights) using grid search. We used the following features:

- TF-IDF vectors of word bi-grams, with a vector weight of 0.5.
- TF-IDF vectors of character n-grams in the range (2,5) when using word boundaries, with a vector weight of 0.5.
- TF-IDF vectors of character n-grams in the range (3,5), with a vector weight of 0.5.
- TF-IDF vectors of skip grams with both one word or one character skipping. The weight for the vector transformation for both these features was set to 0.3.

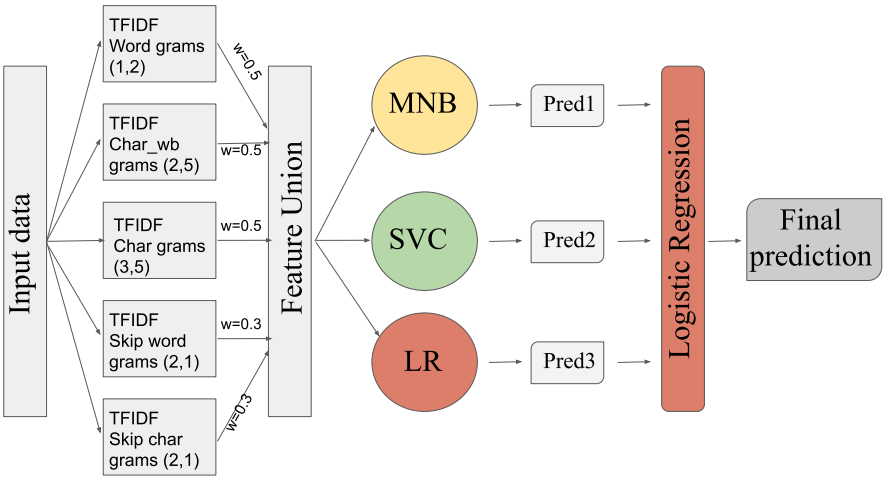

Figure 3: Model architecture. Our model is a stacking classifier. First, three different classifiers are trained on the NADI train split. Their respective predictions are thereafter used with a final estimator to get final predictions.

	dev	**test**
Macro average F1	17.87	17.71
Overall accuracy	37.26	36.22

Table 2: Accuracy and macro average F1 for both dev and test split.

Thereafter, we use these feature vectors to train three different classifiers:

- a Multinomial Naive Bayes (MNB) with alpha set to 0.004.
- a Support Vector Classification (SVC) using `hinge` loss with a tolerance for stopping criterion of 0.9.
- a Logistic Regressor (LR) with a `liblinear` solver.

The predictions of these classifiers are thereafter stacked and fed to a final classifier, in our case, a Logistic Regression using a tolerance for stopping criterion equal to 0.01 and `hinge` loss. As shown in Figure 3, the classifiers MNB, SVC, and LR are fitted on the full feature vector sets, while our final LR estimator is trained using cross-validated predictions of the three base estimators MNB, SVC, and LR.

4 Results and Discussion

We report in Table 2 our model's scores on both dev and test splits using the macro F1-score and accuracy metrics. Both metrics give very low scores, which reflect the difficulty of the task.

We believe that the main issue is the unbalanced nature of the dataset. We have experimented with various approaches to boost the performance. We used oversampling and under-sampling, as well as balanced sampling, but none of these gave satisfying results.

As can be seen in Figure 4 our model achieves high scores when predicting the Egyptian, Saudi Arabian, Algerian, and Iraqi dialects. From the confusion matrix it is also apparent that most dialects were miss-classified as these four dialects. We believe that these are partly due to the amount of tweets of each dialect present in the training set, as these represent the top four most frequent dialects in the train set (see Figure 1). To support this hypothesis, we trained our model only on tweets from these four dialects and achieved an F1-score of over 60%.

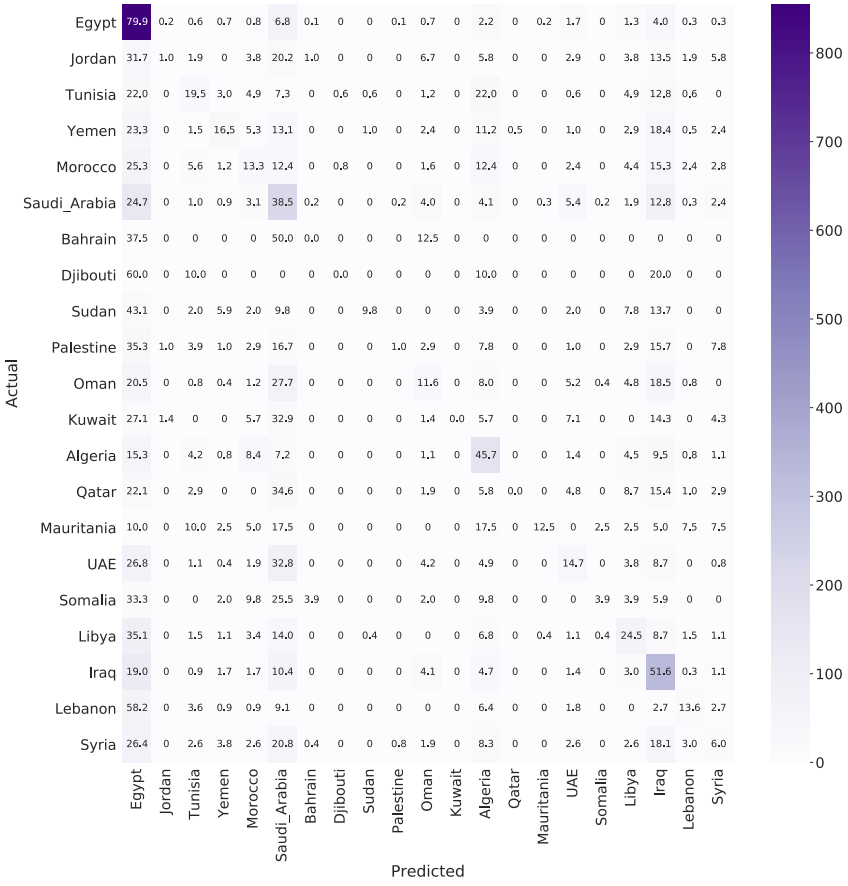

Figure 4: Confusion matrix of our model on the dev data set.

Further analysis of the output have shown that our model performs very poorly on the less frequent dialects. Our model is not able to correctly predict the dialects Qatar, Bahrain, Djibouti, and Kuwait. Once again, we believe that this is due to the skewed nature of the data. But also, the difficulty to differentiate between the dialects in general. The individual precision and recall scores (see Table 3) for each dialect also show that while our model performs poorly on these infrequent dialects, we achieve relatively high precision scores on them. This means that our model predicts very few examples of these dialects, but most of the predicted labels are correct.

A better performing system would have performed on the entire data set as our system performs on the Egyptian dialect: with high precision and high recall. A model that returns multiple predictions, with most predictions correctly labeled. We think that such a system might require a more diversified data set, with less overlap in the vocabulary.

Moreover, a close analysis of the tweets themselves revealed that many tweets labeled as dialects were actually written in (mostly) MSA, which we believe could have further skewed the classifications. As an example, consider the sentences (1) and (2). These are labelled as Algerian in the training data, despite being both written in MSA. We did not carry an exhaustive analysis on the amount of tweets that actually were written in MSA despite being labeled as Algerian, but from our preliminary analysis, it seems that many were either completely written in MSA, or mostly used MSA words. This might be partly due to which region the person tweeting is from, as people from different regions and backgrounds might prefer to write in MSA, or prefer to reach a broader international audience.

Admittedly, these sentences are generic and could in theory be present in all dialects. But it is exactly these type of sentences that can complicate the classification task. Social media platforms can be a gold mine for Arabic dialect identification, and having these type of sentences in a corpus is inevitable. We nevertheless believe that filtering these out, or even classifying them as MSA, might increase the performance of dialect identification systems.

Label	Precision	Recall	F1
Lebanon	28.30%	13.64%	18.40%
Iraq	41.57%	**51.57%**	**46.04%**
Sudan	**55.56%**	9.80%	16.67%
Yemen	37.36%	16.50%	22.90%
Qatar	0.00%	0.00%	0.00%
Saudi Arabia	26.48%	38.51%	31.39%
Morocco	20.12%	13.25%	15.98%
Kuwait	0.00%	0.00%	0.00%
Mauritania	**50.00%**	12.50%	20.00%
Tunisia	27.35%	19.51%	22.78%
Libya	32.18%	24.53%	27.84%
Palestine	20.00%	0.98%	1.87%
Syria	18.39%	6.04%	9.09%
Djibouti	0.00%	0.00%	0.00%
Somalia	33.33%	3.92%	7.02%
United Arab Emirates	25.83%	14.72%	18.75%
Bahrain	0.00%	0.00%	0.00%
Algeria	36.36%	**45.68%**	**40.49%**
Egypt	**46.64%**	**79.91%**	**58.90%**
Oman	21.97%	11.65%	15.22%
Jordan	20.00%	0.96%	1.83%

Table 3: Individual dialect precision, recall, and F1 scores. Values presented in bold are top three values for each measure.

(1) a. مساء الخير والسرور والورود اختي الكريمة

b. Good evening, with pleasures and roses, my dear sister.

(2) a. عيد ميلاد سعيد كل عام وانت بألف خير

b. Happy birthday, I hope that you will have a great and healthy year.

5 Conclusion

We proposed a model for automatically classifying 21 Arabic dialects using a corpus of tweets. Our system uses a stacking classifier relying on TF-IDF word and character features, and the three classifiers Multinomial Naive Bayes, Linear SVC, and Logistic Regression.

Our model did not achieve high scores on all dialects, but we believe that it shows that simple approaches can achieve satisfying results even when using heavily unbalanced data. We have experimented with more advanced deep learning approaches, but were not able to achieve satisfying results. We also believe that the poor results achieved reflect the data we have used. The heavy unbalance in the data has added an extra layer of difficulties to the task of dialect identification and we think that more balanced data can help us develop better models that can achieve higher accuracy.

However, the insights we have gained during the analysis of the train and dev data sets and our results, have enabled us to understand more what are the typical issues encountered during Arabic dialect identification, and we believe that these insights can be used to develop and improve future systems. We think that a more systematic approach based on contextualized word embeddings as BERT (Devlin et al., 2018) combined with sequence labeling approaches can give better results and achieve higher scores, and aim to go in this direction in future work.

References

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 (WANLP 2020)*, Barcelona, Spain.

Kathrein Abu Kwaik and Motaz K Saad. 2019. Arbdialectid at madar shared task 1: Language modelling and ensemble learning for fine grained arabic dialect identification. *ArbDialectID at MADAR Shared Task 1: Language Modelling and Ensemble Learning for Fine Grained Arabic Dialect Identification*, (Proceedings of the Fourth Arabic Natural Language Processing Workshop).

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Youssef Fares, Zeyad El-Zanaty, Kareem Abdel-Salam, Muhammed Ezzeldin, Aliaa Mohamed, Karim El-Awaad, and Marwan Torki. 2019. Arabic dialect identification with deep learning and hybrid frequency based features. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 224–228.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, et al. 2018. Unified guidelines and resources for arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Karima Meftouh, Karima Abidi, Salima Harrat, and Kamel Smaili. 2019. The SMarT Classifier for Ara-bic Fine-Grained Dialect Identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop (WANLP 2019)*, Florence, Italy.

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.

Ahmad Ragab, Haitham Seelawi, Mostafa Samir, Abdelrahman Mattar, Hesham Al-Bataineh, Mohammad Zaghloul, Ahmad Mustafa, Bashar Talafha, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2019. Mawdoo3 ai at madar shared task: Arabic fine-grained dialect identification with ensemble learning. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 244–248.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344.

Kees Versteegh. 2014. *Arabic language*. Edinburgh University Press.

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.