# Country-level Arabic dialect identification using RNNs with and without linguistic features

**Elsayed Issa**[1*] **Mohammed AlShakhori**[1†] **Reda Al-Bahrani**[2**] **Gus Hahn-Powell**[1§]

[1]University of Arizona

[2]Independent Researcher

{*elsayedissa,†mkalshakhori,§hahnpowell}@email.arizona.edu

**reda.bahrani@gmail.com

## Abstract

This work investigates the value of augmenting recurrent neural networks with feature engineering for the Second Nuanced Arabic Dialect Identification (NADI) Subtask 1.2: Country-level DA identification. We compare the performance of a simple word-level LSTM using pretrained embeddings with one enhanced using feature embeddings for engineered linguistic features. Our results show that the addition of explicit features to the LSTM is detrimental to performance. We attribute this performance loss to the bivalency of some linguistic items in some text, ubiquity of topics, and participant mobility.

## 1 Introduction

Arabic exhibits *diglossia*—the existence of two spoken varieties of a language side by side in a community (Ferguson, 1959); while there are a multitude of informal regional varieties, Modern Standard Arabic (MSA) serves as the chief formal variety. Not only the existence of the two spoken varieties is a complex situation for linguists to investigate (Bassiouney, 2009), but it is more complex for data scientists to classify text data of such a language. While phonological differences are apparent in speech, the distinction is lost in writing, as all varieties use the same orthographic system.

Additionally, short vowels in the orthographic system are represented by a diacritic above each phoneme as "أَحَبَّ" meaning "he loved"; recently, however, these vocalic diacritics are dropped from any word as in "احب." The omission is common in news articles, institutional texts, and most obviously on social media platforms. This issue causes what we term *bivalent linguistic unit*, which means that a written text without any vocalic diacritic can belong to any dialect depending on its readers' dialects even if it is written in a local context, a concept that we adopt from (Woolard, 1998). With so few orthographic contrasts, classifying written varieties (MSA, Arabic regional dialects) poses a challenge.

Over the years, there have been several attempts at classifying Arabic dialects, starting from classical natural language processing methods to deep learning whether throughout individual work or shared tasks such as MADAR series (Bouamor et al., 2019), which continued to enhance Arabic dialect identification followed by the NADI series starting in 2019 (Abdul-Mageed et al., 2020). In 2013, Elfardy and Diab (2013) implemented a supervised system for identifying MSA and Egyptian Arabic at the sentence level, by predicting the level of formality of a sentence harvested from the web. Observing the lack of the other Arabic dialects' representation in previous work, Zaidan and Callison-Burch (2014) constructed a corpus focused on including other Arabic varieties. Using $n$-gram and word character models, they were able to evaluate annotators' biases towards labeling text written in their own dialects.

Deep Learning (DL) methods have revolutionized tasks such as large-scale language modeling (Bengio et al., 2003; Dauphin et al., 2017; Jozefowicz et al., 2016), language identification (Joulin et al., 2017), and sentiment analysis (Dong et al., 2014; Severyn and Moschitti, 2015; Araque et al., 2017). The orthographic overlap between MSA and regional dialects has posed a serious challenge to past work on fine-grained dialect identification. Elaraby and Abdul-Mageed (2018) demonstrated that both recurrent and convolutional neural networks can surpass linear models such as logistic regression, multinomial Naive Bayes, and linear kernel support vector machines (SVM) classifiers. Other methods such as word vector modeling are able to identify some linguistic features of Arabic



*Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 276–281
Kyiv, Ukraine (Virtual), April 19, 2021.

tweet corpus (Abdul-Mageed et al., 2018).

## 2 Data

In our experiments, we restricted ourselves to using only the official Twitter corpus provided by the Second NADI Shared Task (Abdul-Mageed et al., 2021). As a preprocessing step, we normalized all partitions of the data by removing non-Arabic words, emojis, links, and excess white space. After normalization, we tokenized the tweets using Keras (Chollet et al., 2015). 10% of the training partition was set aside for monitoring validation loss in an effort to avoid overfitting through early stopping.[1]

## 3 Experiments

In this work, we explored two approaches[2] to fine-grained dialect classification. The first one involved using pretrained word embeddings as the input to an LSTM (Hochreiter and Schmidhuber, 1997) used to encode each tweet. In Experiment 2, we combined the LSTM from Experiment 1 with a feed-forward neural network that encodes a concatenation of low-dimensional dense embeddings representing explicit linguistic features. These linguistic features were used side-by-side with the word-level RNN from Experiment 1 with the aim of supplementing our input with features deemed salient to dialect classification.

### 3.1 Experiment 1: CBOW and LSTM

The neural architecture used for Experiment 1, shown in Figure 1, consists of pretrained word embeddings and an LSTM to model sequential information. We compared two different sets of available word embeddings: Aravec (Soliman et al., 2017) and Mazajak (Abu Farha and Magdy, 2019). Both sets of pretrained word embeddings were developed using Twitter data with different vocabulary, vector, and corpus sizes. Although both Aravec and Mazajak achieved similar results, the Mazajak word embeddings trained using Continuous Bag of Words (CBOW) ($n$ = 100M tweets) achieved the optimal results in these experiments. As a result, an embedding matrix of the shape [maximum features, embedding size] was created to serve as the weights in the embedding layer in our neural network model.

[1]The early stopping patience was set to 2.

[2]Code: github.com/clu-ling/wanlp-2021

Our neural architecture for Experiment 1 consists of three layers.[3] Input to the first layer is restricted to a maximum 80 tokens. The second layer is an embedding layer initialized using the pretrained Mazajak word embeddings. The third layer is an LSTM layer with 300 units, a dropout rate of 0.3, and a recurrent dropout rate of 0.2.

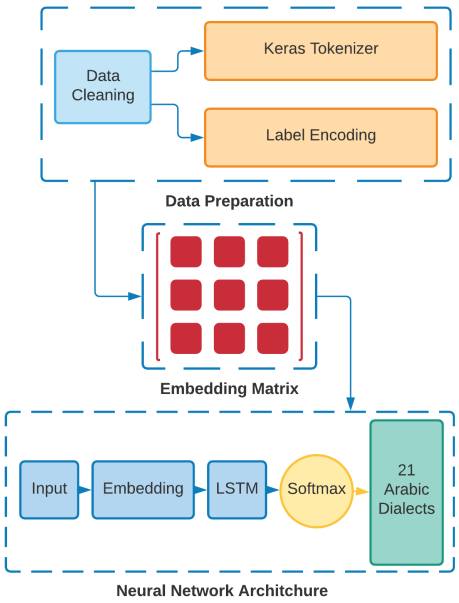


Figure 1: Experiment 1 - The Pipeline for the Arabic Identification System involves 1) data preparation, 2) the extraction of embedding matrix from the CBOW, and 3) a neural network with three layers.

### 3.2 Experiment 2: Engineered features

Experiment 2 extends the architecture of Experiment 1 by injecting linguistic information using engineered features to learn low-dimensional dense embeddings. These linguistic units are unique distinctive features that signify each dialect from each other. These features vary in terms of their linguistic types starting from demonstrative markers to degree markers. Figure 2 shows the architecture of this two-component network.

The first component works the same way as the model in the Experiment 1 in which the embedding layer receives its weights from the embedding matrix of the pretrained word embeddings. The second component takes a binary vector representing features present in a document (tweet). We use 56 linguistic to represent all 21 Arabic dialects. The input vector of the 56 binary values is

[3]The hyperparameters used are as follows: embedding size of 300, vocabulary size of 50000, batch size of 64, and maximum sequence length of 80.

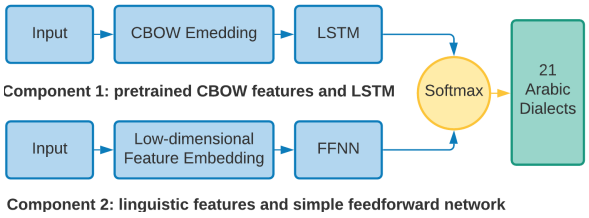


Figure 2: Experiment 2 - the two-component architecture combining word embeddings with embeddings learned for explicit linguistic features. One input consists of pretrained word embeddings fed into an LSTM. The second input is a concatenation of learned embeddings for linguistic features. Finally, the two inputs are combined through concatenation prior to classification.

used to select low-dimensional feature embeddings which are concatenated and fed through a simple feed-forward network consisting of two 100-unit hidden layers with ReLU activation followed by an element-wise multiplication before being concatenated to the output of the LSTM described in Experiment 1.

Table 1 shows a sample of engineered linguistic features. These simple features represent expressions and terms commonly used in each dialect. If one of these features is present, it is assigned 1 otherwise 0. Though here we only report results for the model using positive features, we also explored learning representations for the absence of features (NOT_X).

| Dialect | Sample features | Gloss |
|---|---|---|
| Iraqi | خوش | ok / good |
| Saudi | كذا | like this |
| Moroccan | ديال / ديالي | of-genitive |

Table 1: A small sample of the engineered linguistic features for Egyptian, Iraqi, Saudi, Moroccan dialects from DA_Subtask 1.2.Each binary feature was used to learn a dense low-dimensional embedding.

## 4 Results & Discussion

We evaluated the architectures from both experiments on development data provided for the task. Based on the performance of the two systems, our submission for the shared task uses the architecture from Experiment 1 which does not incorporate any engineered linguistic features.

Our results, shown in Table 2, emphasize the main findings of this article: linguistic features (at least of the forms explored) do not provide sufficient information for fine-grained Arabic dialect identification. Rather, we believe that pretrained word embeddings and models such as BERT are amongst the optimal solutions for feature extraction for Arabic dialectal classification.

| Metric (macro) | Model 1 | Model 2 |
|---|---|---|
| Accuracy | 41.36 | 37.82 |
| Precision | 30.12 | 21.65 |
| Recall | 21.56 | 18.72 |
| F1 | 22.10 | 18.60 |

Table 2: Results of the development data for Experiment 1 & 2. The F1 score for Experiment 1 (our simpler model consisting of pretrained word embeddings and an LSTM) outperforms the Experiment 2 architecture which incorporated engineered linguistic features.

There are several observations that underscore the decline in the performance of our model in Experiment 2. Though we treated Experiment 1 as a baseline system to dialect/language identification, it achieved a better macro F1 score than our proposed hybrid method in Experiment 2 which incorporates simple engineered features. This suggests that pretrained word embeddings already provide richer information than what was encoded in our engineered features.

From a linguistic perspective, we believe that explicitly modeling salient features is a promising direction for improving our model; however, there are are a number of reasons this approach was unsuccessful here.

**Sparse features** Our system has few features relative to the number of classes, and the frequency feature in the corpus (and thus their coverage) is low. That is, these features are insufficient to cover the set of documents available for each dialect.

**Genre** Much of the data is characterized by what we call *global genre*—meaning that the content of the text contains global shared topics such as sports and popular culture. For instance, examples 3, 4, and 5 in Table 3 indicate that the content of the tweets is governed by a global genre topic which imposes less presence of the local linguistic features of the participants' dialects. In order to improve performance through feature engineering, the content of the data has to be characterized by more *local genre*. Sociologists have shown that participants of different linguistic communities in

| No | Label | Tweet | Gloss | Type |
|---|---|---|---|---|
| 1. | Oman | هههههه بايخه هههههه | hahaha stale joke hahaha | bivalent linguistic units |
| 2. | Syria | هههههه لاتصدق بس تستاهل | haha don't believe but you deserve | bivalent linguistic units |
| 3. | Syria | ميسي وضعية الميت | Messi ... the state of being dead | global genre |
| 4. | Iraq | اعطي المزيد من الوقت ليسي ونيمار راح يكون البرسا ديناميت | more time to Messi & Nymear, Barca will be dynamite | global genre |
| 5. | Algeria | ميسي ونيمار طلب يا حبيبي | oh my dear, Messi & Nymar are demands | global genre |
| 6. | Algeria | شلون تفوز بالابطال و انت مدربك فالفيردي | How will you win the Europe Champion while the coach is Valverde? | participant mobility |
| 7. | Djibouti | بدنا ترند في كل مكان اليسا في أعياد بيروت | We want a trend everywhere Alissa_in_Beriut_festival | participant mobility |
| 8. | Iraq | لا عاد الحساويين دبغين | Oh the Hassawi people are not likable | shared Gulf lexical items |
| 9. | Oman | عيل اجمل وجه راح اتامله | a guy with good-looking face that I like to stare at | cross-dialect linguistic units |

Table 3: A sample of the tweets from the labeled data subtask 1.2. The above tweets are either linguistically constrained or governed by participant mobility. Tweets that are linguistically *bivalent* as 1 & 2 can be classified in any dialect. *Globally genre tweets* as in 3-5 are governed by a global topic rather a local one in the region of the dialect. The *shared Gulf lexical item* indicates that some linguistic segments are shared by other dialects in the same region. No. 9 presents a sample of how one lexical item can be semantically different from another language.

social events exhibit global identities by projecting shared linguistic features; however, the local identities of the same participants emerge by projecting more local linguistic features as long as the topic touches their personality or their communities (De Fina, 2000; Schilling-Estes, 2004; Bucholtz and Hall, 2005; Bucholtz, 2010; Benor, 2010; Becker, 2009, *inter alia*). Modeling these kinds of linguistic features seems a promising direction for future research.

**Location granularity** A number of the features included in this work are regional, rather than unique to a single dialect. An example of such a situation is example 8 in Table 3 in which an Iraqi participant shares the same linguistic features "عاد" that are found in the Saudi dialect. Similarly, some word-level features exhibit different behavior across dialects. For instance, Gulf states use heavily the word "عيل" as "then," shown in Table 3 example 9, but this word semantically means "child" in the Egyptian dialect data. Such *cross-dialect linguistic units* require more sophisticated feature engineering to adequately model.

**Dataset design** This is a challenging dataset. Consider, for instance, the topic of *participant mobility*. Examples 6 and 7 show some linguistic items that belong to different dialects compared to the location of the tweets. In this sense, the word "بدنا" in example 7 is a Levantine lexicon, rather than a Djibouti one. While examining the training data, we came across tweets that, to a native speaker, could just as easily be classified into *any* dialect (see examples 1 & 2). Sociolinguists characterize these samples as *bivalent linguistic units* (Woolard, 1998). Such datapoints may provide very little signal for their assigned class.

**Impoverished features** Rather than incorporating general linguistic features, our approach focused on devising distinctive word-level features that signify each dialect. Considering the complexity of the task, our preliminary features are inadequate both in terms of their number (56) and distribution across linguistic levels (morphology, syntax, etc.). There are several Arabic morphological analysis tools that could be used to greatly enrich and expand our set of features (Habash et al., 2012; Obeid et al., 2020). In the same respect, Bouamor et al. (2019) presented a large parallel corpus of 25 Arabic city dialects in the travel domain and a lexicon of 1,045 concepts with an average of 45 words from 25 cities per concept. While the approach presented here does not make use of these resources, our architecture was designed with extensibility in mind. We hope to explore this avenue in future work.

## 5 Conclusion

In this work, we report results for two experiments: one that uses pretrained CBOW word embeddings with an LSTM, and another that integrates linguistic features as low-dimensional feature embeddings fed through a simple feed-forward network. The two experiments are used to classify Arabic tweets for the Second Nuanced Arabic Dialect Identification (NADI) Subtask 1.2.: Country-level DA identification. The results show that rare linguistic features do not enhance the performance of an LSTM with pretrained CBOW word embeddings. These results emphasize the importance of pretrained models in NLP tasks related to Arabic dialects.

## References


Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim 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 (WANLP 2021)*.

Ibrahim Abu Farha and Walid Magdy. 2019. Mazajak: An online Arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 192–198, Florence, Italy. Association for Computational Linguistics.

Oscar Araque, Ignacio Corcuera-Platas, J Fernando Sánchez-Rada, and Carlos A Iglesias. 2017. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236–246.

Reem Bassiouney. 2009. Arabic sociolinguistics: Topics in diglossia, gender. *Identity, and Politics. Georgetown University Press*.

Kara Becker. 2009. /r/and the construction of place identity on new york city's lower east side 1. *Journal of Sociolinguistics*, 13(5):634–658.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155.

Sarah Bunin Benor. 2010. Ethnolinguistic repertoire: Shifting the analytic focus in language and ethnicity 1. *Journal of Sociolinguistics*, 14(2):159–183.

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.

Mary Bucholtz. 2010. *White kids: Language, race, and styles of youth identity*. Cambridge University Press.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.

François Chollet et al. 2015. Keras. `https://keras.io`.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR.

Anna De Fina. 2000. Orientation in immigrant narratives: The role of ethnicity in the identification of characters. *Discourse Studies*, 2(2):131–157.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 49–54.

Mohamed Elaraby and Muhammad Abdul-Mageed. 2018. Deep models for arabic dialect identification on benchmarked data. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 263–274.

Heba Elfardy and Mona Diab. 2013. Sentence level dialect identification in arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461.

Charles A Ferguson. 1959. Diglossia. *WORD*, 15(2):325–340.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 2012. A morphological analyzer for egyptian arabic. In *Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology*, pages 1–9.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the 12th language resources and evaluation conference*, pages 7022–7032.

Natalie Schilling-Estes. 2004. Constructing ethnicity in interaction. *Journal of Sociolinguistics*, 8(2):163–195.

Aliaksei Severyn and Alessandro Moschitti. 2015. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 959–962.

Abu Bakr Soliman, Kareem Eissa, and Samhaa R El-Beltagy. 2017. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, 117:256–265.

Kathryn A Woolard. 1998. Simultaneity and bivalency as strategies in bilingualism. *Journal of linguistic anthropology*, 8(1):3–29.

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