Arabic dialect identification using machine learning and transformer-based models: Submission to the NADI 2022 Shared Task

Nouf AlShenaifi, Aqil Azmi
Department of Computer Science, King Saud University, Riyadh, Saudi Arabia
{noalshenaifi, aqil}@ksu.edu.sa

Abstract

Dialect is the language variation of a specific community. In this paper, we show the models we created to participate in the third Nuanced Arabic Dialect Identification (NADI) shared task (Subtask 1) that involves developing a system to classify a tweet into a country-level dialect. We utilized several machine learning techniques as well as deep learning transformer-based models. For the machine learning approach, we build an ensemble classifier of various machine learning models. In our deep learning approach, we consider bidirectional LSTM model and AraBERT pretrained model. The results demonstrate that the deep learning approach performs noticeably better than the other machine learning approaches with 68.7% accuracy on the development set.

1 Introduction

Dialect identification is the task of automatically identifying the dialect of a particular part of the text (Zaidan and Callison-Burch 2011). Arabic dialects differ by region, and there are no available dictionaries for their vocabulary or written rules for the words that are specific to those dialects. Developing an Arabic dialect identification system experimenting with different corpora and working at different levels of representation has attracted increasing attention in recent years (Elnagar et al. 2021). In this paper, we present our work to tackle the third Nuanced Arabic Dialect Identification (NADI) shared task that targets country-level dialects. We built multiple classifiers based on machine learning and deep learning techniques. We experimented with an approach of combining different Machine Learning models using a combination of n-grams and TF-IDF as features to enhance the performance. Another method applied in this study is a deep learning approach including Bidirectional Long-Short Term Memory (BiLSTM) model and pre-trained AraBert model. This paper is organized as follows: Section 2 details the used dataset. Section 3 presents the applied preprocessing steps and the proposed approach for Arabic Dialect Identification. In Section 4, we discuss the obtained results. Section 5 contains the conclusion.

2 Datasets

The dataset used in NADI 2022 shared task (Subtask 1) is the same as the prior NADI shared task (Abdul-Mageed et al. 2020) (Abdul-Mageed et al. 2021). It consists of 20k labeled tweets for training, 4,871 for development that covers 18 Arabic dialects. For testing, two test sets were provided TEST-A and TEST-B. TEST-A includes 18 country-level dialects. In the second test (TEST-B), K country-level dialects are covered where k is kept unknown. The training data which consists of 20K tweets is unbalanced as you can see in Figure 1. Figure 1 displays how tweets are distributed among Arab countries. Most of the tweets belong to Egypt (4283 tweets) and only 215 belong to Bahrain, Qatar, and Sudan. The provided data is normalized in which all URLs are replaced with the word 'URL' and mentions replaced with the word 'USER'. Around 10M unlabeled tweets were also provided to participating teams by the NADI shared task organizers.

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 464 - 467

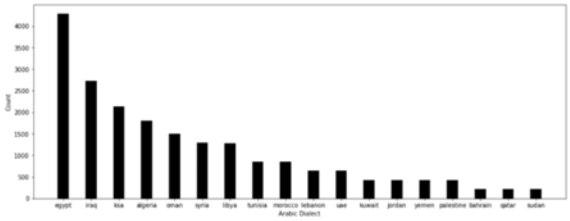

Figure 1: NADI 2022 shared task Training Dataset statistics.

3 Methodology

This section shows the models we used in our experiments starting with machine learning methods and moving on to deep learning and transformer-based approaches.

3.1 Data Preprocessing

Even though NADI training data set is normalized by replacing mentioned user with the token "user" and all links with the token "URL", further cleaning and preprocessing was required. Hence before training our proposed models, we used pre-processing steps including tokenizing, removal of punctuation marks, emojis, Arabic stop words and diacritics, and repeated chars such as "هههههههههههه". We also performed several experiments to test the effects of different preprocessing tasks such as stemming, and we found that stemming has a negative impact on the results. To deal with data imbalances, we applied Synthetic Minority Oversampling Technique (SMOTE) (Fernández et al. 2018) as an imbalance correction technique for oversampling imbalanced classification datasets.

3.2 Machine Learning-Based Models

3.2.1 Logistic Regression

Logistic regression is used to assess the statistical significance of each independent word with regard to probability (Shah et al. 2020). We have applied a logistic regression classifier on the concatenation of word n-grams (n=1 to 3) and char n-grams (n=1 to 4) TF-IDF features using one-vs-the-rest scheme for multi-class training.

3.2.2 Support Vector Machine

Support Vector Machine (SVM) which is based on structural risk minimization is recommended to use for handling large textual features (Fanny, Muliono, and Tanzil 2018). We build a Support Vector Machine (SVM) classifier for country-level dialect identification task based on CountVectorizer and TF-IDF word n-gram features. CountVectorizer to transform each tweet into a vector on the basis of the frequency of each word that appears in the whole dataset. For the extracted features, we used TF-IDF vectors with word n-grams where (n=1 to 3).

3.2.3 Ensemble Classifier

This classifier is a soft voting classifier of three individual machine learning models Stochastic Gradient Descent (SGD) Classifier, Multinomial Naive Bayes, and Bernoulli Naive Bayes as shown in Figure 2. Naive Bayes classifier is still used for text classification as a fast and easy to implement machine learning classifier (Kowsari et al. 2019). Stochastic Gradient Descent (SGD) is an efficient approach to fitting linear classifiers and regressors under convex loss functions such as Support Vector Machines and Logistic Regression. We used TF-IDF with character (2-5)-grams, and word (1-4) grams as a feature for training our ensemble classifier.

3.3 Deep Learning models

3.3.1 Embedding Layer with bidirectional LSTM model

Bidirectional Long-Short Term Memory (BiLSTM) network was used with pretrained word embeddings as an input. In word embedding, we obtain values for word vectors or embeddings by training a neural language model to capture semantic and syntactic relationships between words in the corpus (Soliman, Eissa, and El-Beltagy 2017) (Mikolov et al. 2017). In our model, we used Aravec (Soliman, Eissa, and El-Beltagy 2017)a pretrained word embeddings developed using Twitter data based on the continuous bag-of-words and another on the Skip-gram mode. We also built a word vectors model (word2vec model) using 300K tweets from the NADI unlabeled dataset (the 10M tweets) (Srinivasa-Desikan 2018). Fast text skigram model is trained on the corpus, to create an embedding matrix that contains embedding words each one represents a word in the corpus. Our BiLSTM model consists of an embedding layer, 128 hidden units, and a dense layer with 18 hidden units and softmax activation function to identify dialects. For the network configuration, we used 300 as input sequence length 0.1 for dropout rate, and 10 for epochs, because more than that the model overfits.

Moreover, Adam was the optimization technique we used, and Categorical cross-entropy was used as the loss function.

3.3.2 Fine-tuning Arabert Transformer

AraBERT is pretrained transformer model based on BERT transformer model (Devlin et al. 2018) specifically for the Arabic language (Antoun, Baly, and Hajj 2020). We used the pre-trained AraBERT model and fine-tuning hyperparameters for Arabic dialect identification tasks on NADI Dataset. We utilize the Hugging Face Trainer utility (McMillan-Major et al. 2021), which allows us to fine-tune AraBERT by changing parameter options. The final configuration of the model we used is Adam optimizer with 1e-8 for adam epsilon, Learning Rate of 1e-5, Maximum Sequence Length is 128, Batch Size is 40, and number of Epochs is 6.

4 Results & Discussion

In our experiments, we have reported the result of multiple models starting with machine learning approaches and moving on to transformer-based methods. The evaluation measures include F-score, Accuracy, Precision, and Recall. However, the Macro Averaged F-score is the official metric of evaluation. Table 1 shows the performance in terms of F1-score and accuracy of various Machine Learning and deep learning models evaluated on dev and test sets. According to the results shown in Table 1, the three best classifiers are Ensemble Classifier, Bidirectional LSTM, and Fine-tuning Arabert Transformer on both dev and test set for the first sub-task of NADI shared task. The best results on the development set are obtained by Embedding Layer with Bidirectional LSTM classifier with an F1-score of 50.5%. The obtained results show that deep learning approach significantly outperforms the other machine learning approaches.

Models	Dev		Test-A		Test-B	
	Acc	F1	Acc	F1	Acc	F1
Logistic Regression	36.9	19.1	7.8	5.5	17.6	7.9
SVM	40.9	19.4	36.9	16.1	21.3	7.4
Ensemble Classifier	46.2	24.4	39.1	18.8	24.4	9.1
Bidirectional LSTM	68.7	50.5	39.9	22.4	23.7	9.3
Fine-tuning Arabert Transformer	68.7	50.5	38.2	21.9	23.5	9.1

Table 1: The obtained results of the dev & test dataset.

5 Conclusion

In this paper, we present our submitted models to the third Nuanced Arabic Dialect Identification shared task. We conducted different experiments in which we tried different preprocessing procedures and several feature combinations for model training. We combined different machine learning approach such as (Logistic Regression, Support Vector Machine, and Multinomial Naive Bayes) to build a strong Arabic dialect identification System. We further developed a transformer-based model using Embedding Layer with a bidirectional LSTM model and Fine-tuning Arabert Transformer. The obtained results have shown that our transformer-based model outperforms all machine learning model on Macro-F1 evaluation measure.

References

Abdul-Mageed, Muhammad, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. "NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task." *ArXiv Preprint ArXiv:2010.11334*.

Abdul-Mageed, Muhammad, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. "Nadi 2021: The Second Nuanced Arabic Dialect Identification Shared Task." *ArXiv Preprint ArXiv:2103.08466*.

Antoun, Wissam, Fady Baly, and Hazem Hajj. 2020. "Arabert: Transformer-Based Model for Arabic Language Understanding." *ArXiv Preprint ArXiv:2003.00104*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *ArXiv Preprint ArXiv:1810.04805*.

Elnagar, Ashraf, Sane M Yagi, Ali Bou Nassif, Ismail Shahin, and Said A Salloum. 2021. "Systematic Literature Review of Dialectal Arabic: Identification and Detection." *IEEE Access* 9: 31010–42.

Fanny, Fanny, Yohan Muliono, and Fidelson Tanzil. 2018. "A Comparison of Text

Classification Methods K-NN, Naïve Bayes, and Support Vector Machine for News Classification." *Jurnal Informatika: Jurnal Pengembangan IT* 3 (2): 157–60.

Fernández, Alberto, Salvador Garcia, Francisco Herrera, and Nitesh v Chawla. 2018. "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-Year Anniversary." *Journal of Artificial Intelligence Research* 61: 863–905.

Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. "Text Classification Algorithms: A Survey." *Information* 10 (4): 150.

McMillan-Major, Angelina, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. "Reusable Templates and Guides for Documenting Datasets and Models for Natural Language Processing and Generation: A Case Study of the HuggingFace and GEM Data and Model Cards." *ArXiv Preprint ArXiv:2108.07374*.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. "Advances in Pre-Training Distributed Word Representations." *ArXiv Preprint ArXiv:1712.09405*.

Shah, Kanish, Henil Patel, Devanshi Sanghvi, and Manan Shah. 2020. "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification." *Augmented Human Research* 5 (1): 1–16.

Soliman, Abu Bakr, 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: 256–65.

Srinivasa-Desikan, Bhargav. 2018. *Natural Language Processing and Computational Linguistics: A Practical Guide to Text Analysis with Python, Gensim, SpaCy, and Keras*. Packt Publishing Ltd.

Zaidan, Omar, and Chris Callison-Burch. 2011. "The Arabic Online Commentary Dataset: An Annotated Dataset of Informal Arabic with High Dialectal Content." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 37–41.