Weighted combination of BERT and N-GRAM features for Nuanced Arabic Dialect Identification

Abdellah El Mekki[1]*, **Ahmed Alami**[2], **Hamza Alami**[2]*, **Ahmed Khoumsi**[3], **Ismail Berrada**[1]
[1]School of Computer and Communication Sciences,
Mohammed VI Polytechnic University, Ben Guerir, Morocco
[2]Faculty of Sciences Dhar EL Mehraz,
Sidi Mohamed Ben Abdellah University, Fez, Morocco
[3]Department of Electrical and Computer Engineering, University of Sherbrooke, Canada
{abdellah.elmekki,ismail.berrada}@um6p.ma
ahmed.alami@usmba.ac.ma
hamza0alami@gmail.com ahmed.khoumsi@usherbrooke.ca

Abstract

Around the Arab world, different Arabic dialects are spoken by more than 300M persons, and are increasingly popular in social media texts. However, Arabic dialects are considered to be low-resource languages, limiting the development of machine-learning based systems for these dialects. In this paper, we investigate the Arabic dialect identification task, from two perspectives: country-level dialect identification from 21 Arab countries, and province-level dialect identification from 100 provinces. We introduce an unified pipeline of state-of-the-art models, that can handle the two subtasks. Our experimental studies applied to the NADI shared task under the team name BERT-NGRAMS, show promising results both at the country-level (F1-score of 25.99%) and the province-level (F1-score of 6.39%), and thus allow us to be ranked 2nd for the country-level subtask, and 1st in the province-level subtask.

1 Introduction

Language identification is considered an important task in Natural Language Processing (NLP), as it helps personalizing applications, automatically detecting the source variety of a given text or speech segment, and collecting/tagging the data (Anshul and Arpit, 2012). In the case of Arabic language, the official language of over 20 countries, and with more than 360 million native speakers, this task becomes very challenging due to the different language variations (Dialectal Arabic), and the complex taxonomy of Arabic language (Zaidan and Callison-Burch, 2014).

In this paper, we consider the Arabic dialect identification task from two perspectives: country-level dialect identification and province-level dialect identification. In the case of the previous Multi Arabic Dialect Applications and Resources (MADAR) Shared Task (Bouamor et al., 2019) , several approaches have been proposed (Abbas et al., 2019). Zhang and Abdul-Mageed (2019) developed country-level identification models based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018), and Gated Recurrent Units (Chung et al., 2014). They ranked 1st in the subtask aiming at classifying tweets with 71.70% macro F1-score and 77.40% accuracy. Talafha et al. (2019) investigated various feature extraction methods, such as term frequency–inverse document frequency (TF-IDF) and word embedding, in order to improve the model performances. The best results, 69.86% F1-score and 76.20% accuracy, were obtained using a simple Linear Support Vector Classification (LinearSVC) classifier with a user voting mechanism.

The main contribution of this paper is the introduction of a novel approach based on a pipeline of state-of-the-art models for both sub-tasks of NADI Shared Task (Abdul-Mageed et al., 2020). For country-level identification, we build a system based on raw tweets. The core of this system is an ensemble model that applies a weighted voting technique on two classifiers, namely *M_NGRAM* and *M_BERT* (will be used as labels for our models on the rest of this paper, the M stands for model). *M_NGRAM* uses TF-IDF

*equal contribution

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 268–274
Barcelona, Spain (Online), December 12, 2020

with word and character n-grams to represent tweets. A stochastic gradient descent (SGD) classifier is then trained to optimize the Huber loss (Zhang, 2004). *M_BERT* fine-tunes AraBERT weights (Antoun et al., 2020) with a softmax classifier trained to optimize the multi-class entropy loss. For province-level identification, we build a hierarchical classification, by first performing the country-level identification, and then fine-tuning for each identified country a *M_BERT* model to predict its provinces. Figure 1 illustrates the overall architecture of the proposed solution. The proposed system generates F1-scores of 25.99% and 6.39% for the country-level identification and province-level identification, respectively.

The rest of this paper is organized as follows. Section 2 describes the Nadi Shared Task dataset. Section 3 describes the proposed approaches, models and our data preparation. Section 4 presents experimental results, error analysis and discussion. Finally, the conclusion is given in Section 5.

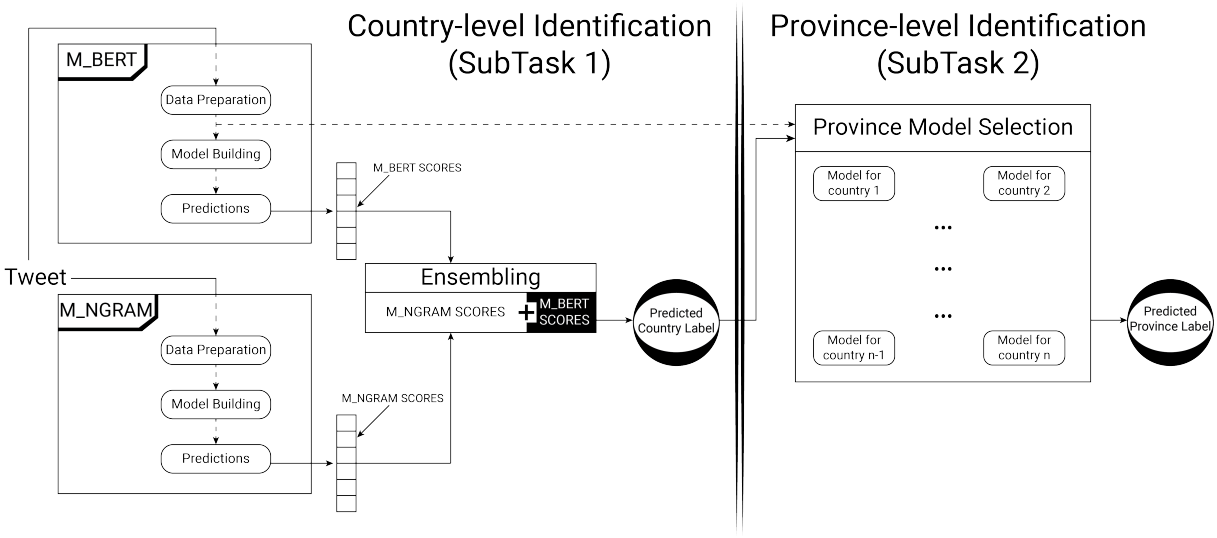

Figure 1: Global overview of the proposed system

2 Dataset presentation

The Nuanced Arabic Dialect Identification (NADI) shared task is the first task focusing on naturally-occurring fine-grained dialect. It has been divided into two subtasks: 1) the country-level identification subtask and 2) the province-level identification subtask. The organizers of the shared task provide four sets of collected tweets: the train set (21K), the development set (4,957), the test set (5,000), and the unlabeled tweets set (10M). As we can see in Figure 2, the NADI task is quite challenging due to the unbalanced distribution of tweets (as example a very low frequency of tweets for Djibouti (DJ) and Bahrain (BH) while the number of Egypt (EG) tweets is very high) and the nuance between Arabic dialects.

3 Methods

In this section, we review the data preparation pipeline and the proposed models, namely *M_NGRAM* and *M_BERT*, used to build our ensemble model.

3.1 Data preparation

As the final system is based on models that rely on different pre-processing steps, below we describe the data preparation pipeline for each model.

3.1.1 *M_BERT* data preparation

For tweet preprocessing, we applied the approach of Hamza et al. (2020) to handle the pretrain-fine-tune discrepancy challenge. The latter can be explained by the fact that the special tokens such as *[MASK]*

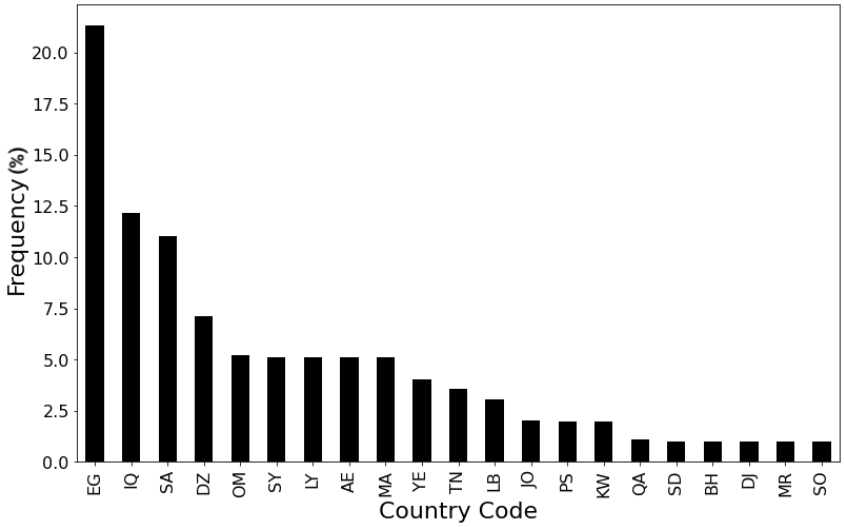

Figure 2: Label distribution of the training set. Country code following the ISO 3166-1 alpha-2 (Wikipedia, 2020)

used by *AraBERT* during pretraining are absent from specific datasets at fine-tuning step. As a tweet may contain words and emojis, the preprocessing pipeline is composed of the following steps:

1. Detecting emojis: the position and the meaning of each emoji is extracted within a tweet.

2. Substituting emojis: each emoji is replaced with the special token *[MASK]* and its meaning is translated from English to Arabic. This allows our model to overcome the pretrain-fine-tune discrepancy.

3. Concatenating emoji-free tweets with their respective emojis Arabic meanings: the special token *[CLS]* is added to the head of each sentence, while the special token *[SEP]* was added to delimit the tweet and the emojis Arabic meanings.

4. Tokenizing the output sentence: all words except special tokens are segmented by Farasa (Abdelali et al., 2016) and then tokenized with *AraBERT* tokenizer. The latter is based on WordPiece (Schuster and Nakajima, 2012) algorithm that is an unsupervised model and follows a sub-words units approach.

3.1.2 *M_NGRAM* data preparation

The data preparation for *M_NGRAM* model can be summarized as follows:

- **Data augmentation:** for subtask 1, we construct for each country a list of keywords used to extract tweets from the unlabeled 10 million tweets. The list contains flag emoji, country name, city names and jargon.

- **Data cleaning:** as Arabic dialects are not considered as official languages, it is hard to get rules and standards for each of them. This makes the pre-processing of the provided data for this task hard and limited. For the *M_NGRAM* model, the pre-processing of tweets is done by removing special characters, nomalizing some Arabic characters and words, nomalizing specific links using regular expressions, and removing Tatweel (characters elongation) and non-Arabic characters.

- **Feature extraction:** TF–IDF features are extracted from the pre-processed data in two levels:

  – *Word-level n-grams:* N-gram words are extracted, then vectorized using TF-IDF scores. Unigrams have been found to give the best performances.

– *Character-level n-grams:* as the task is nuanced Arabic dialect identification, dialects of many countries cannot be differentiated based on words. Moreover, Arabic dialects have no standard writing. This raises the problem of Out-of-vocabulary (OOV) words in the validation phase. To tackle this problem, we use character-level n-grams that treat subwords as features. TD-IDF vectorization is then performed on character-level n-grams. After several experiments, (3,5) range shows the best performance on the character-level n-grams.

3.2 Country-level Identification

3.2.1 Building the *M_BERT* model

The *M_BERT* model aims to classify a tweet in a predefined country. Thus, the following steps are taken in order to build the model:

1. The tokens obtained from *M_BERT* data preparation step (section 3.1.1) are grouped into two segments. The first one contains tweet's tokens, while the second segment contains the tokens of the Arabic meanings of detected emojis.
2. The token embeddings or representations are computed by feeding their indices and segments to *AraBERT* model (Antoun et al., 2020).
3. The tweet representation is then the output of a global max pooling function applied on *AraBERT* token representation.
4. The probability that a tweet belongs to a country is computed by a softmax function that takes the tweet representation as input.
5. The model is trained to minimize the multi-class cross entropy loss.

We should mention that *AraBERT* model has the same configuration as *BERT-base* model (Devlin et al., 2019). It is composed of 12 encoder blocks, 768 hidden dimensions, 12 attention heads, 512 maximum sequence lengths, and a total of about 110M parameters. The model is trained on two objectives:

- Masked Language Model where the model is trained to predict a masked token.
- Next Sentence Prediction in which the model is optimized to predict if the second sentence follows the first one.

AraBERT is pre-trained on 70 million Arabic sentences, corresponding to $\sim$24GB of text. The authors consider a vocabulary that contains 64k tokens. Another key point to mention here is that during training, *AraBERT* parameters are fine-tuned on this specific task: Arabic Country-level Dialect Identification.

3.2.2 Building the *M_NGRAM* model

Since the training data is noisy and Arabic dialects are etymologically close with each other (Habash et al., 2012) , dialect identification gets harder for many tweets. Moreover, the followed data augmentation pipeline is not accurate since the augmentation criterion is chosen based on heuristics. Therefore, we decided to train our *M_NGRAM* model using stochastic gradient descent (SGD) with the following points:

- **Weighting of samples:** the size of the augmented data is 10 times larger than the original data. This makes the classification model more biased towards the augmented samples rather than the original ones. To address this issue, we weight respectively the original samples and the augmented samples with 1 and 0.25, respectively.
- **A loss function sensitive to outliers:** we use the Modified Huber Loss (Zhang, 2004) as a loss function for the SGD classifier. This loss showed to be less sensitive to outliers.

The SGD classifier is trained on the concatenation of TF-IDF vectors of the word-level n-grams and character-level n-grams extracted in section 3.1.2. We use Scikit-learn (Pedregosa et al., 2011) implementation for training our *M_NGRAM* model.

3.3 Province-level identification

We propose a hierarchical classifier to detect the province of a given tweet. We begin by grouping the dataset by countries. Then for each group (hence country), an AraBERT-based classifier is trained to predict the province label. All the province classifiers follow the same process described in section 3.2.1. To predict a tweet province, the tweet is first is prepossessed by the *M_BERT* data preparation step. Next, we identify the tweet's country with our Arabic country-level identifier. After that, the province-level classifier is chosen according to the identified country. Finally, the preprocessed tweet is fed to its province-level classifier to predict the province.

4 Experimental results, error analysis and discussion

In this section, we conduct experiments to evaluate the performance of the proposed models *M_BERT*, *M_NGRAM*, and *ensemble model* on the development set. Table 1 shows the obtained results in terms of macro precision, macro recall, macro F1-score, and accuracy. As we can see, for subtask 1 the ensemble model achieves the best results: 40.95% for the accuracy and 27.24% for the F1-score. We can notice that the *M_NGRAM* achieves 25.02% macro F1-score while the *M_BERT* scored only 22.42% macro F1-score. Thus, when applying weighted soft voting, the scores obtained by *M_NGRAM* must contribute more than M_BERT scores. Figure 3 confirms that 0.7 for *M_NGRAM* and 0.3 for *M_BERT* are the weights to reach the best F1-score. Our final models (the ensemble model and the hierarchical model) show to perform well on the test set too (Table 1). It is worth to be mentioned that our systems ranked 2nd and 1st for country-level identification and province-level identification, respectively.

Subtask	DEV/TEST	Model	Precision	Recall	F1	Accuracy
Subtask 1	DEV Set	*M_NGRAM*	31.85%	24.03%	25.02%	37.72%
		M_BERT	28.11%	21.86%	22.42%	37.32%
		Ensemble Model	**33.75%**	**25.70%**	**27.24%**	**40.95%**
	TEST Set	Ensemble Model	**30.25%**	**24.85%**	**25.99%**	**39.66%**
Subtask 2	DEV Set	Hierarchical classifier	**26.57%**	**26.02%**	**23.55%**	**26.24%**
	Test Set		**7.84%**	**6.54%**	**6.39%**	**6.50%**

Table 1: Subtask 1 and subtask 2 performances evaluation of used classifiers on dev and test sets.

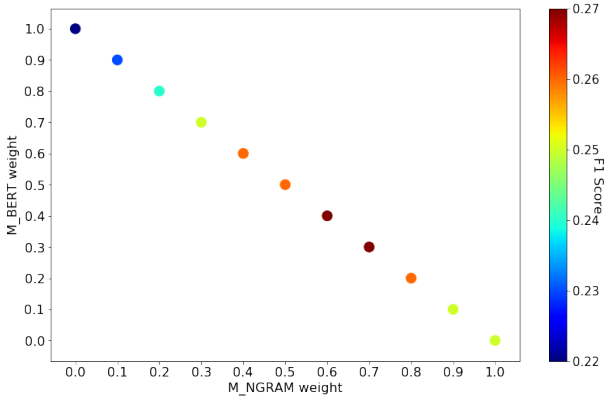

Figure 3: Performance of the weighted soft voting ensemble model with respect to the weights of M_NGRAM and M_BERT for the country-level identification.

In order to help future research in the field of automatic Arabic dialect identification, we discuss below some challenges that our ensemble model faced during the country-level identification:

1. Extremely imbalanced dataset challenge: Some countries like Egypt and Iraq present 21.30% and 12.17% of the training set (Figure 2) while other countries such as Djibouti or Bahrain present

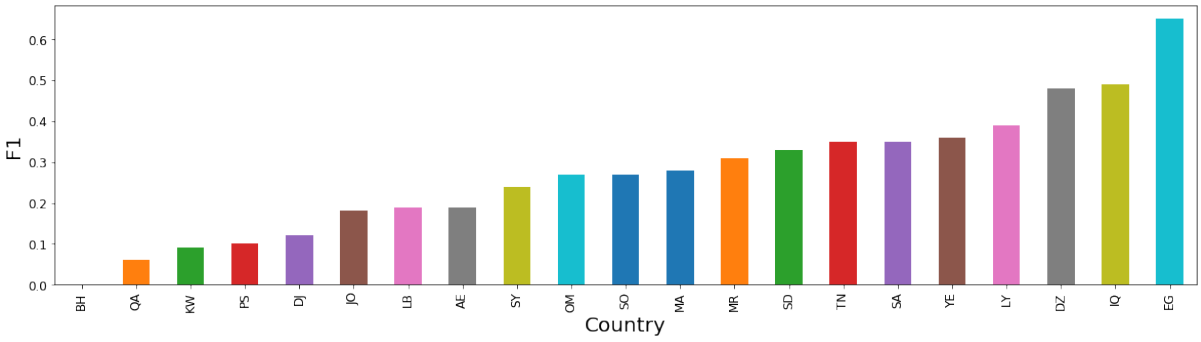

Figure 4: F1 scores of the ensemble model for the 21 countries dialects for the subtask 1. Country code following the ISO 3166-1 alpha-2 (Wikipedia, 2020)

only 1% of the training set. This led to a large gap between the F1-scores (Figure 4) of the well represented countries and the under represented countries, e.g., the F1-score of Egypt is about 68% compared to $\sim 8\%$ (respectively 0%) for Djibouti (respectively Bahrain).

2. Etymological challenge: All tweets within the dataset have Arabic as their root language. Thus, many expressions are shared between many dialects such as لا حول و لا قوة الا بالله ("There is no power but from God" /"lA Hwl w lA qwħ AlA bAllh"), إن شاء الله ("God willing" / "Ăn šA' Allh") or بسم الله الرحمن الرحيم ("In the name of Allah the Merciful" / "bsm Allh AlrHmn AlrHym"). One can notice that for each Arabic example we include its English translation and its Buckwalter-Habash-Soudi transliterations (Habash et al., 2007).

3. Unstructured and noisy nature of tweets challenge: Some model predictions are biased by the presence of some dialect-specific words or tokens within tweets. For instance, the expected label of the tweet هو مش كل حاجة بس بيعمل كل حاجة ("It is not everything, but does everything" / "hw mš kl HAjħ bs byζml kl HAjħ") is United Arab Emirates, yet the model has predicted Egypt as the country label. The words الأهلي ("AlÂhly"), حاجه ("HAjh") and مش ("mš") exist more likely in tweets labeled as Egypt, therefore, the model will attribute the highest probability to Egypt label.

4. Topic-biased challenge: The predominance of one or more topics in a set of tweets that belong to the same country. Taking the class label Djibouti as example, we notice clearly that the majority of tweets are about soccer topic. As consequence, the model predict the majority of tweets related to the soccer topic as Djibouti tweets.

5 Conclusion

In this paper, we described our submission to the NADI Shared Task. We built a system composed of two classifiers: Ensemble model for country-level identification, and a Hierarchical classifier for province-level identification. The quote "alone we are strong, together we are stronger." has been verified: our ensemble model in subtask 1 increased significantly our F1-score to 27.24% on development set and 25.99% on the test set, allowing us to rank second in the competition. In subtask 2 the hierarchical classifier achieved 6.39% F1-score and ranked 1st. This work has shown that the combination of neural network-based features (BERT) with statistical features (TF-IDF) might increase the performance in other NLP tasks.

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, Florence, Italy, August. Association for Computational Linguistics.

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for Arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16, San Diego, California, June. Association for Computational Linguistics.

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.

Mittal Anshul and Goel Arpit. 2012. Stock prediction using twitter sentiment analysis.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

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.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Nizar Habash, Mona Diab, and Owen Rambow. 2012. Conventional orthography for dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 711–718, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Alami Hamza, Ouatik El Alaoui Said, Benlahbib Abdessamad, and En-nahnahi Noureddine. 2020. Lisac fsdm-usmba team at semeval 2020 task 12: Overcoming arabert's pretrain-finetune discrepancy for arabic offensive language identification.

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.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012*, pages 5149–5152. IEEE.

Bashar Talafha, Wael Farhan, Ahmed Altakrouri, and Hussein Al-Natsheh. 2019. Mawdoo3 AI at MADAR shared task: Arabic tweet dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 239–243, Florence, Italy, August. Association for Computational Linguistics.

Wikipedia. 2020. ISO 3166-1 alpha-2 — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=ISO\%203166-1\%20alpha-2&oldid=972097545`. [Online; accessed 13-August-2020].

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

Chiyu Zhang and Muhammad Abdul-Mageed. 2019. No army, no navy: BERT semi-supervised learning of Arabic dialects. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 279–284, Florence, Italy, August. Association for Computational Linguistics.

Tong Zhang. 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 116, New York, NY, USA. Association for Computing Machinery.