Weakly and Semi-Supervised Learning for Arabic Text Classification using Monodialectal Language Models

Reem AlYami[2,3] and **Rabeah Al-Zaidy**[1,3]
[1]Center for Integrative Petroleum Research (CIPR), [2]Preparatory Year Program,
[3]Information and Computer Science Department
King Fahd University of Petroleum and Minerals
Saudi Arabia
reem.yami@kfupm.edu.sa, rabeah.alzaidy@kfupm.edu.sa

Abstract

The lack of resources such as annotated datasets and tools for low-resource languages is a significant obstacle to the advancement of Natural Language Processing (NLP) applications targeting users who speak these languages. Although learning techniques such as semi-supervised and weakly supervised learning are effective in text classification cases where annotated data is limited, they are still not widely investigated in many languages due to the sparsity of data altogether, both labeled and unlabeled. In this study, we deploy both weakly, and semi-supervised learning approaches for text classification in low-resource languages and address the underlying limitations that can hinder the effectiveness of these techniques. To that end, we propose a suite of language-agnostic techniques for large-scale data collection, automatic data annotation, and language model training in scenarios where resources are scarce. Specifically, we propose a novel data collection pipeline for under-represented languages, or dialects, that is language and task agnostic and of sufficient size for training a language model capable of achieving competitive results on common NLP tasks, as our experiments show. The models will be shared with the research community [1].

1 Introduction

In recent years, the emergence of social media platforms allowed the increased use of the informal form of a language in online user-generated content. As a result, more languages are present in online content, introducing a challenge to language processing tools that are developed to improve user experience. This is evident in the discrepancy in the levels of support for many tasks in language technologies for different languages, such as the lack of keyboard support and spell checking extensions for low resource languages, even those with a large online user base (Soria et al., 2018).

Supervised learning models for text classification are ubiquitous in natural language processing tasks (Minaee et al., 2021). For high-resource languages such as English, Chinese, and German, a variety of annotated datasets are constantly made available by both industry and academia (Wang et al., 2019a; Xu et al., 2020; Schabus et al., 2017). On the other hand, low-resource languages such as many Asian languages still suffer from a shortage of annotated datasets for fundamental NLP tasks, including text classification (Joshi et al., 2020). Given that many NLP applications, whether speech or text, heavily rely on classification, this shortage can negatively impact the accessibility of AI-enabled services to speakers of these languages (Minaee et al., 2020). To assist in reducing this gap of opportunity, a large body of studies in the NLP community is dedicated to facing challenges with low-resource languages using several approaches.

One approach is to focus on developing multilingual models that are capable of learning language-agnostic representations of data (Wang et al., 2020). Another approach uses meta-learning and few-shot learning models to improve results on tasks with small sets of annotated data (Pires et al., 2019; Artetxe et al., 2017). Adapting to small sets of data can also be achieved using semi-supervised models where a seed of annotated data is used to bootstrap a supervised model using only a relatively small set of labeled data (Van Engelen and Hoos, 2020). Weakly supervised models fall into this class of approaches as well, where primary external knowledge sources are incorporated to provide larger sets of annotated data for the model (Elnagar et al., 2019; Guellil et al., 2020). For extremely low-resourced languages, these techniques are difficult to apply due to the lack representative datasets whether labeled or unlabeled (Joshi et al., 2020).

In this work, we address the challenges facing incorporating learning techniques designed for scenarios where annotated data is scarce. Specifically,

[1]https://huggingface.co/reemalyami

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 260 - 272
December 8, 2022

for Arabic dialects, the main challenge is that in data sources where dialectal data in a raw form is abundant, it is rarely distinguished from other Arabic dialects, posing a challenge when the goal is to target a specific dialect. To that end, we curate and construct datasets and dictionaries, develop an automatic annotation scheme, develop multiple Pre-trained Language Models (PLMs) and conduct an empirical study to examine the performance of the text classification task under the learning paradigms of semi, weak and full supervision. Although Arabic is a widely spoken language, with over 400 million speakers, it still remains a low-resource language, especially in terms of the availability of annotated datasets for emerging NLP tasks (Althobaiti, 2020). Thus, the approaches proposed in this work, although testing on Arabic, are applicable to any similarly low-resourced language.

In summary, the contributions of this paper are:

1. Propose a novel data collection pipeline from Twitter that is language and task agnostic.
2. Construct seven Arabic dialect-specific dictionaries.
3. Develop an automatic annotation technique for Arabic dialects.
4. Train seven Arabic dialect-specific language models.
5. Propose a novel technique for Arabic dialect classification that improves over conventional semi-supervised methods.
6. Evaluate the performance of Arabic dialect identification in supervised, weakly supervised, and semi-supervised settings.

The remainder of this paper is organized as follows. In the next section we present related work. Section 3 presents the data collection and annotation pipeline. In Section 4 we describe the proposed language models. Section 5 describes the classification models. In Section 6 we describe the experimental setup and evaluation. In Section 7 we provide a discussion. In Section 8 we conclude and describe future directions for the work.

2 Related Work

2.1 Arabic Dialect Datasets

Arabic belongs to the group of *diglossic* languages, where different variations of the language are spoken in the community sharing the language. Arabic has two general forms, Modern Standard Arabic (MSA) the form used in written and formal communication among all speakers, and dialectal Arabic (DA), which are local variants of the language used in day-to-day communication varying based on region. In Arabic, there are multiple dialects in different regions of the Arab world: Gulf, Levantine and North Africa. Users commonly communicate in informal contexts using their local dialect rather than the formal MSA, more so in spoken than written. This introduces a challenge for Arabic-based applications. As a consequence of the scarcity of dialectal resources for Arabic, many studies focus on building Arabic dialectal corpora to investigate various NLP tasks in Arabic (Einea et al., 2019; Abdul-Mageed et al., 2020; Bouamor et al., 2018; Haouari et al., 2020; Elnagar et al., 2018; Hasanain et al., 2018) (Alyami and Olatunji, 2020; Al-Twairesh et al., 2018; Baly et al., 2019; Abdul-Mageed et al., 2018a; Abidi et al., 2017; Itani et al., 2017; Elnagar and Einea, 2016). Several of these datasets are publicly available (Haouari et al., 2020; Bouamor et al., 2018; Abdul-Mageed et al., 2020; Elnagar et al., 2018; Einea et al., 2019) and have greatly assisted both the research community and industry in tackling Arabic NLP challenges.

Datasets for the Arabic Dialect Identification (ADI) task vary in size, variety, granularity level, and the domain of the text. As seen in early work, datasets that investigate specific dialects on a specific domain, namely, news domain, do so on a certain granularity level that is the regional level (Zaidan and Callison-Burch, 2011, 2014; Malmasi et al., 2016). Other work developed dialectal datasets at the city and country levels. The first focuses on the dialects in specific cities in a country (Bouamor et al., 2018, 2019a; Abdul-Mageed et al., 2018b). Country-level studies focus on a specific country and all the sub-dialects spoke in that country. More recent works on the country level dialect focus on a specific task (Yang et al., 2020; Farha and Magdy, 2019; Habash et al., 2019) or investigate the combination of MSA data with other dialects (AlYami and AlZaidy, 2020; Alshargi et al., 2019; Khalifa et al., 2016). In many works, the collected data is based on crawling data from user-profile content, resulting in data samples that, semantically, represent the content discussed by specific users around a specific set of *seed words* (Abdul-Mageed et al., 2020; Bouamor et al., 2018, 2019a). In regards to automatic annotation of Arabic datasets, the existing tools focus specifically on linguistic annotation for limited Arabic varieties, especially MSA, which in turn cannot readily be used to annotate other dialects (Habash et al., 2009).

2.2 Arabic Dialect Identification

In many cases, it is beneficial to identify the specific dialect prior to performing core NLP tasks such as parsing, tokenizing or other downstream tasks such as semantic inference (Abdelali et al., 2016). For this reason, we conduct our study on the specific problem of Arabic dialect classification. Many ADI studies use n-gram based Language Model (LM) where they adopt different character level n-gram representations due to the Out Of Vocabulary (OOV) problem (Malmasi and Zampieri, 2017; Mishra and Mujadia, 2019; Ragab et al., 2019). Other features for classification such as Term Frequency — Inverse Document Frequency (TF-IDF) are used as well (Ragab et al., 2019; Bouamor et al., 2019b; Abdelali et al., 2021; Talafha et al., 2020; Gaanoun and Benelallam, 2020). Since many of these techniques lead to producing sparse representations, other work proposed utilizing static dense vectors (Elaraby and Abdul-Mageed, 2018; Meftouh et al., 2019).

Although dense vectors tend to improve classification performance in general, their adaptations in ADI yield results comparable to those of the n-gram models (Abu Farha and Magdy, 2019). Additionally, a key aspect to consider in Arabic dialects is *polysemous* words due to Arabic dialects having a shared vocabulary among them, yet the words in many cases have different meanings from one specific dialect to another (Zampieri and Nakov, 2021). Recent studies building on contextual features demonstrated promising results on a range of token and sequence classification tasks, including the dialect identification task (Zhang and Abdul-Mageed, 2019; Abdelali et al., 2021; Gaanoun and Benelallam, 2020; Abdelali et al., 2021).

Due to the shortage in datasets for many individual Arabic dialects, few efforts have utilized semi-supervised learning (SSL) in classifying Arabic dialects that showed promising results and some outperformed supervised learning approach (Zhang and Abdul-Mageed, 2019; Beltagy et al., 2020; Althobaiti, 2021). In recent years weak-supervision is utilized in text classification problems such as Arabic dialect identification, sentiment analysis and document classification as seen in the case of clinical text classification (Huang, 2015; Deriu et al., 2017; Meng et al., 2018; Wang et al., 2019b).

3 Data collection and annotation for low-resource languages

In this section we describe our proposed approaches for large data collection for specific languages and dialects and our automatic annotation approach for large data.

3.1 Large Data Collection

In order to build large datasets for low-resource languages we propose two approaches used to develop two datasets, Arabic Dialect Short Text dataset (ADST) and the Arabic Dialect Dictionary dataset (ADD). The collection approach for each is described below.

Arabic Dialect Short Text (ADST) is collected from Twitter, since many Arab countries are among the top 20 countries to use Twitter (Twi), in addition to the Twitter's feature that allows retrieving tweets given specific keywords. We use Tweepy API that permits data collection for research purposes under the digital millennium copyright act [2]. Our approach for language or dialect specific data, defines two parameters: keywords and the location of the dialect, defined using country geo-coordinates (latitude and longitude) via Free map online tool [3] (loc).

In contrast to studies where keywords are static, which limits dialect diversity and coverage (Bouamor et al., 2019a; Abdul-Mageed et al., 2020), we propose to collect keywords *dynamically*, i.e. collected from Twitter on a daily basis. Keyword are obtained from the *trending keywords* feature in Twitter for each of the targeted countries to capture words related to the speakers of a given dialect.

In order to collect country coordinates for Twitter Data Collection we divide this into two sub-components. These components are as follows:

1. **Country Centric Point:** To ensure collecting dialectal tweets from the specified countries. One of the parameters that can be passed to the Twitter query is the latitude and longitude of the targeted point to collect tweets from

[2] https://help.twitter.com/en/rules-and-policies/copyright-policy
[3] https://www.freemaptools.com/radius-around-point.htm

the selected geographical location on the map. Since Twitter permits that a geometric centering point on the country's map is specified using latitude and longitude and curating all the tweets in the circle radius inside each country using an online tool to obtain these data points as illustrated in the Figure 1.

Figure 1: Specifying a centring geographical point in Saudi Arabia.

2. **Coordinates:** After defining a centring point the countries coordinates were retrieved along with area of the circle radius. In order to verify the retrieved coordinates another online tool is utilized were it yielded identical results [4].

Data Preprocessing

The preprocessing step includes de-duplication, Arabic letter normalization, removal of digits, character elongation, and samples with less than seven tokens in order to have richer representation. The effect of preprocessing on ADST size is shown in Table 1.

Country	Retrieved Tweets	Unique Tweets	7+ Tokens Tweets
Saudi Arabia (SA)	4,693,533	3,614,590	**2,415,622**
Egypt (EG)	5,677,800	3,313,610	**2,099,977**
Kuwait (KU)	4,047,308	823,546	**477,973**
Oman (OM)	665,463	316,500	**200,384**
Lebanon (LB)	670,715	294,275	**204,430**
Jordan (JO)	657,472	232,124	**97,400**
Algeria (DZ)	245,480	115,564	**103,488**

Table 1: ASTD size and the effect of the preprocessing on the tweets

Arabic Dialect Dictionary (ADD)

In this study a dictionary refers to a list of words and symbols that is usually used to automatically label data in case human annotation is unavailable as it is a cost effective method (Jurafsky and Martin, 2009). In our work seven Arabic dialectal dictionaries are built from different Arabic dialect sources.

[4] https://latitude.to/lat/23.48690/lng/44.82030

A dictionary for each country is built by collecting popular dialect-specific terms from public websites *Mo3jam* [5] and *Atlas Allhajaat* [6], where both sources provide a list of dialectal terms. The ADD is normalized using a similar process to ASTD in addition to stopword removal. Stopwords are collected from an online linguistic repository (El-Khair, 2017; ASW) of **1,614** stopwords. Finally, the ADD is reviewed by a human reviewer for final cleaning; the resulting dictionary description is shown in Table 2.

Country	SA	DZ	EG	JO	LB	KU	OM
#ADD	7,045	3,869	2,227	1,453	1,195	2,066	1,550

Table 2: The ADD Size

3.2 Automatic Data Annotation

Annotating a large dataset of Arabic dialects for the ADI task manually is costly, which introduces the need for an automatic annotation approach.

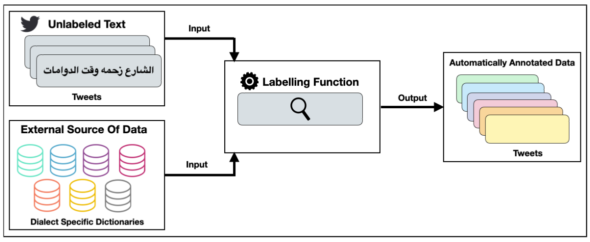
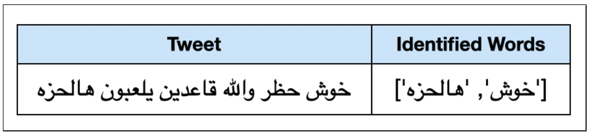


Figure 2: Tweets Automatic Annotation Process.

Our proposed automatic annotation process is shown in Figure 2. The annotation is performed through a labeling function that utilizes ADD as an external source to generate automatic labels. The data is annotated automatically using the dialect-specific dictionary (ADD), where the tweets curated from a particular country are labeled as a positive sample of the country dialect only if the tweet contains n or more tokens from the corresponding country's dictionary, as illustrated in Figure 3. In our work we set $n = 2$ after an empirical assessment. After annotating the dialect, each dialect has its own automatically annotated dataset. Each dataset contains the positive dialect instances, and for the negative samples, the other automatically labeled dialect samples from other dialects are incorporated, producing a balanced dataset. The size of the resulting dataset is shown in Table 3.

[5] https://ar.mo3jam.com/
[6] http://www.atlasallhajaat.com/



Tweet	Identified Words
خوش حظر والله قاعدين يلعبون هالحزه	['خوش', 'هالحزه']

Figure 3: Sample tweet that is automatically labeled

Dialect	SA	DZ	EG	JO	LB	KU	OM
Total	104,976	61,860	104,976	17,496	20,304	53,052	29,664

Table 3: Automatically Annotated Data

4 AraRoBERTa

This section provides a description of the dialect-specific language models developed using the large datasets we collected. To obtain the Arabic RoBERTa (AraRoBERTa) models, we train 7 BERT-based models using the RoBERTa-base configuration with Masked Language Modeling (MLM) pre-training objective (Devlin et al., 2018; Liu et al., 2019). It consists of *12* encoder layers/blocks, *768* hidden dimensions, *12* attention heads, and *512* maximum sequence length (Devlin et al., 2018; Wolf et al., 2020). The batch size is 32 with 10 epochs after initial experimentation based on the loss. Although initial experimentation is done on the hyperparameter, the adopted values are similar to the literature.

The optimization is similar to the adopted BERT optimization (Liu et al., 2019), using the Adam optimizer (Kingma and Ba, 2017) with similar parameters. The collected tweets described in Section 3.1 from each dialect are utilized for pre-training the corresponding AraRoBERTa dialectal language model as shown in Table 4. We use the Byte Per Encoding (BPE) tokenizer using HuggingFace implementation [7]. BPE resolves the OOV problem, making it simpler, more efficient, and provides a small vocabulary size that is 52K (Sennrich et al., 2016). The developed AraRoBERTa models and the selected contextual baselines are described in Table 4 in term of the Arabic training data, the vocabulary size and the model configuration. In this work AraRoBERTa is built using HuggingFace Transformers API (Wolf et al., 2020) on (1x16GB NVIDIA Tesla P100) GPU.

Also, other contextual baselines are used to compare the performance of AraRoBERTa variations against as shown in Table 4. These models are: 1) **mBERT**: The multilingual version of BERT that is trained on 100 languages including Arabic (Devlin et al., 2018). 2) **XLM-R** The multilingual version of RoBERTa that is trained on 100 languages (Conneau et al., 2020). 3) **AraBERT** A monolingual model developed on Arabic specifically MSA (Antoun et al., 2020).

[7] https://huggingface.co/docs/transformers/tokenizer_summary#byte-pair-encoding

5 ADI Models

The ADI task is formed as a classification task. We adopt three classification models using semi and weak supervision paradigms. In these models, we build on a transformer-based classifier. In this section, we provide an overview of our proposed models.

5.1 Dialect Classification Problem

The Arabic dialect classification problem is defined as follows. Given a set of short texts,

$$D = \{(t_1, y_1), (t_2, y_2), ...(t_n, y_n)\}$$

where t denotes the short text instances, n denotes the number of instances and the label is denoted by $Y = \{P, N\}$ where P represent a specific Arabic dialect and N represent the negative samples that does not belong to the dialect, the model performs binary classification to assign each t_i a y_j label.

5.2 Semi-Supervised Model

The conventional SSL approach known as *self-training* illustrated in Figure 4 does not ensure having negative samples in the training data since the data is collected from a specific country affecting the performance of the model. Hence, another semi-supervised approach is proposed to mitigate the limitation of the conventional SSL approach.

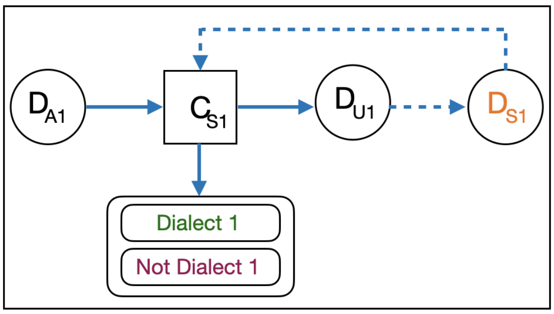


Figure 4: The pipeline for conventional semi-supervised classification model.

The proposed SSL task learns from both the labeled and unlabeled data. For the labeled data, we manually annotated dataset as follows. A human expert labels each tweet as belonging to one of

Model	Training Data			Vocabulary		Configuration	
	Source	Variant	#Tokens	Tokenizer	Size	Arch.	#Params.
mBERT	Wikipedia	MSA/Multi-Lang	Ar(153M)/All(1.5B)	WP	Ar(5K)/All(110K)	base	110M
XLM-R_B	CommonCrawl	MSA/Multi-Lang	Ar(2.9B)/All(295B)	SP	Ar(14K)/All(250K)	base	270M
AraBERT	Several (3 sources)	MSA	2.5B	SP	Ar(60K)/All(64K)	base	135M
AraRoBERTa-SA	Arabic Twitter	SA DA	45.4M	BPE	52K	base	126M
AraRoBERTa-EG		EG DA	37.2M	BPE	52K	base	126M
AraRoBERTa-KU		KU DA	8.9M	BPE	52K	base	126M
AraRoBERTa-OM		OM DA	3.8M	BPE	52K	base	126M
AraRoBERTa-LB		LB DA	3.6M	BPE	52K	base	126M
AraRoBERTa-JO		JO DA	2.6M	BPE	52K	base	126M
AraRoBERTa-DZ		DZ DA	1.9M	BPE	52K	base	126M

Table 4: Configurations of existing models and AraRoBERTa models. WP is WordPiece and SP is SentencePiece tokenizers.

seven pre-defined dialects which is then reviewed by another expert. Both annotators are either native speakers or closely familiar with the dialect. The seven dialects we consider are: Saudi Arabia, Kuwait, Oman, and Egypt, Algeria, Jordan, and Lebanon. For the last 3 countries, native speakers are recruited to label the data from a freelance service website [8]. The annotators are compensated based on their offer in the platform. A request explaining the required task is raised, then each freelancer offers her/his services with the price defined by the freelancer. If a mutual agreement is reached, the freelancer is paid before performing the task.

Only annotators with the location corresponding to the needed dialect were hired. A meeting with each freelancer is conducted to explain the task then an initial sample of 10 tweets is annotated by the annotator to ensure the task is understood by the annotator. In addition to this data, the dataset from the NADI shared task, released under the creative commons license, is used (Abdul-Mageed et al., 2020). The proposed semi-supervised model is illustrated in Figure 5. For dialect i the classifier C_{Si} takes as an input the annotated data D_A and after initial training it is utilized to produce the pseudo-labels: $Y_S = \{P_S, N_S\}$ on the unlabeled data D_U. In the pseudo-labeled data D_{Si} the negative samples are denoted by $N_{S_i} = P_{S_1}, \ldots P_{S_{m-1}}$ where $P_{S_i} \notin N_{S_i}$ and $|P_{S_i}| == |N_{S_i}|$ as illustrated in the figure where the colors denote the negative sample that corresponds to the positive sample for each dialect. That is then augmented with the labeled data for the model to train on both data until the defined termination criteria is reached.

[8] https://khamsat.com/

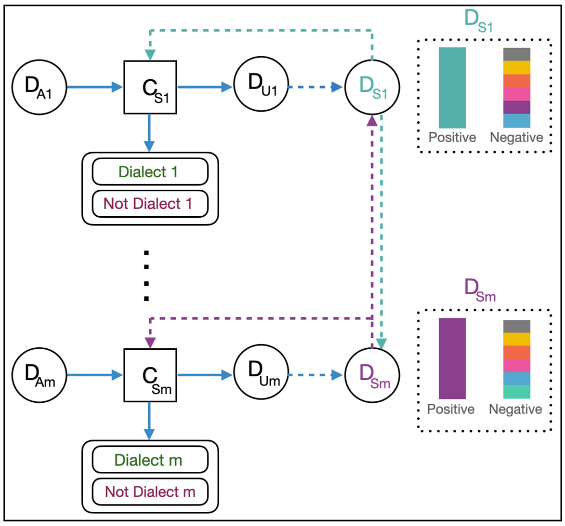


Figure 5: The pipeline for the proposed semi-supervised classification model.

5.3 Weakly-Supervised Model

This learning task learns from unlabeled data by providing an approximate label. The set of weak labels (class) are assigned using a labelling function g that utilizes an external source of information to annotate the unlabeled instances D_U producing Y_W, where $Y_W = \{P_w, N_w\}$, denoting a weak label. This is performed on all unlabeled data to create a new training set $D_W = \{(t_{w_1}, y_{w_1}), (t_{w_2}, y_{w_2}), \ldots..(t_{w_m}, y_{w_m})\}$, where m denotes the number of samples and $w_i \in Y_W$. Here the labels Y_W are produced automatically, as illustrated in Figure 6. The weakly labeled data D_W produced by the automatic annotator for dialect i is subsequently used to train a binary classifier C_{Wi} to predict dialect i.

6 Experiments and Evaluation

Here we describe the evaluation experiments for fully, semi and weakly supervised learning models

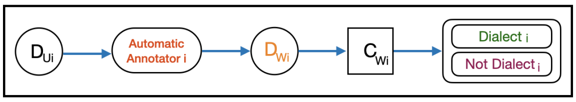


Figure 6: The pipeline for the weakly-supervised classification model.

for the ADI task. The performance is evaluated using the F-1 measure, following existing literature.

6.1 Supervised ADI

We follow an experimental setup similar to the pre-training task as described in Section 4 except for the number of epochs, which is five. This experiment evaluates the performance of AraRoBERTa variations on the dialect classification task using the manually annotated data described earlier with a train/validation/test split of 70/10/20 respectively. Additionally, the results are compared with other contextual baselines described earlier and with a traditional machine learning model, namely, Logistic Regression (LR) as it yielded the best results on the same task in a previous study (AlYami and AlZaidy, 2020). The training data for LR is similar to the ones described above and TF-IDF is used to represent text. The experiment is preformed with 10-fold cross validation and a train/test split of 80/20.

Experimental Results The results for the supervised experiments are shown in Table 5. Larger AraRoBERTa models, namely, AraRoBERTa-SA and AraRoBERTa-EG, outperform other models. AraRoBERTa-KU model outperforms its multilingual counterparts and is slightly lower than AraBERT. In other cases, both AraRoBERTa and AraBERT yielded similar results, and the other multilingual models outperformed them. Except for AraRoBERTa-OM yielding the lowest performance among other models. Although AraRoBERTa models are trained on maximum 1.8% of the data that AraBERT is trained on, it yields very competitive results. In five out of seven AraRoBERTa flavors, it outperformed the contextual baseline models as shown in Table 5.

For the remaining two, although trained an even smaller fraction, it yielded a similar performance to AraBERT and multilingual models. This encourages training other models on a specific content even if the available data size is smaller compared to other training data in the literature. Additionally, when comparing AraRoBERTa against LR the two largest AraRoBERTa models outperform it. Also, AraRoBERTa-KU yields a slightly lower result. However, from the results, when having access to small dataset size, traditional ML performs better.

Dialect	AraRoBERTa	AraBERT	mBERT	XLMR	LR
SA	**0.836**	0.806	0.823	0.784	0.791
EG	**0.934**	0.898	0.872	0.879	0.862
KU	0.916	0.913	0.883	0.886	**0.921**
OM	0.718	0.845	0.839	**0.896**	0.883
LB	0.849	0.849	0.879	0.866	**0.892**
JO	0.848	0.856	0.872	0.833	**0.881**
DZ	0.859	0.855	0.873	0.908	**0.923**

Table 5: The supervised classification results. The best results are in bold.

6.2 Semi-supervised ADI

The performance of semi-supervised classifiers is evaluated on the same test set used in the supervised baseline. Then, it is compared against it. The sample size for the unlabeled data is reduced due to computational limitations where a random sample of 16,000 training samples are selected to perform the semi-supervised experiments with a 0.95 threshold for the prediction confidence for the pseudo-labeled instances. The training stops when the remaining unlabeled data points are less than 5% .

Experimental Results The results of the SSL classifier are shown in Table 6. We can notice it outperforms the performance of the supervised models in multiple dialects. Also, we can notice that AraRoBERTa-Om and AraRoBERTa-LB that were built on the lower end in terms of training data, yield better performance than its supervised AraRoBERTa counterparts.

Dialect	Supervised	SSL
SA	**0.84**	0.83
EG	**0.93**	**0.93**
KU	**0.92**	0.89
OM	0.72	**0.80**
LB	0.85	**0.88**
JO	**0.85**	0.83
DZ	0.86	**0.87**

Table 6: The semi-supervised classification results. The best results are in bold.

6.3 Weak-supervised Dialect Classification

The performance of weak-supervised classifiers is evaluated on the same test set used in the supervised baseline. Then, it is compared against it. This setup follows the supervised setup, however, the number of epochs is different since initial experiments showed that three epochs are suitable as the training data is larger and the training loss flattens before reaching three epochs.

Experimental Results The results for the weak-supervised experiments are shown in Table 7 in general for all models across dialects yield lower performance compared to AraRoBERTa supervised classifiers as shown by the performance change. Although the classification data size is larger by around *6x* for the Jordan dialect and up to *33x* for Saudi dialect. However, the degrade in performance is noticeable in AraRoBERTa models trained on smaller data size like AraRoBERTa-JO rather than larger models like AraRoBERTa-SA.

Dialect	Supervised	WSL
SA	**0.84**	0.81
EG	**0.93**	0.86
KU	**0.92**	0.61
OM	**0.72**	0.40
LB	**0.85**	0.78
JO	**0.85**	0.71
DZ	**0.86**	0.78

Table 7: The weak-supervised classification results. The best results are in bold.

7 Discussion

This section provides an analysis for the experimental results and discusses the significant findings.

7.1 Supervised Classification Model

As shown in the experiments above, we note that the least performing model on the supervised classification task is AraRoBERTa-OM. The model has a false-negative rate of 20.75%, whereas the false-positive rate is only 2.25%, indicating a bias towards rejecting Omani texts although the model is balanced for positive and negative samples. To probe this further, the model was tested again on a slightly-modified version of the test set, where we replaced positive samples that were misclassified by the model, with different positive samples that contained more Omani-specific terms. The amount of replaced samples is around 10% of the test data. As a result, the ability of the model to identify the Oman dialect increased, reflected by an 3% increase in the true-positive rate and a decrease in the false-negatives from the previous 20.75% to 18.12%. This can be due to the training set of AraRoBERTa-OM, which could have contained a larger portion of utterances with majority of tokens are Omani specific terms and did not account for ones with majority of tokens that are common with other dialects.

In other cases, the classification inaccuracies may not be a result of the training set for the language model but rather be due to the dialect itself. For instance, AraRoBERTa-SA and AraRoBERTa-LB both exhibit a more inclusive bias, i.e. labeling other dialects as positive, with false-positive rates of 11.38% and 11.62%, respectively, compared to low false-negatives of around 4% for each. To probe this further we examine missclassified samples in the test set, where we show some examples in Figures 7 and 8. For the examples in Figure 7 , although the full tweet belongs to another dialect, Jordan dialect, we can see all of the words in the tweet can be used by Saudi speakers in regions near the Saudi/Jordan border.

On the other hand, in Figure 8, the first sample is Egyptian dialect where the second is Saudi, using words that are specific to these dialects. This contrast indicates that a bias towards false-positives can be attributed to either a training set for the language model that is not sufficiently representative of the dialect, or to the approach with which Arabic dialects are generally defined, i.e. by country. Typically, regions along the borders of countries commonly share a similar dialect, which in certain datasets becomes more pronounced in cases of large and centrally located countries such as Saudi Arabia.

اخوي جاب ايفون برو حكالي بطاريته احسن من الايفونات الي قبل
علي الاقل هي خلصت توجيهي معها حق شوي انتي بشو مريتي

Figure 7: A sample of the misclassified tweets by AraRoBERTa-SA, these samples are negative samples. However, the model classified them as Saudi.

7.2 Semi-supervised Learning

The results of the SSL classifier are shown in Table 8. Note that the performance at iteration-0 is supervised and semi-supervised at iteration 1 and 2.

اذا بفتح كمان سناب حد وبلاقيه حاطط لقيت الطبطبه رح اسويله ديليت ماصارت اغنيه ترا
تو جالسين نتفق ف سناب انتي اول وحده تروحي

Figure 8: A sample of the misclassified tweets by AraRoBERTa-LB, these samples are negative samples. However, the model classified them as Lebanese.

The performance in later iterations outperforms the model's performance at iteration-0 in the majority of the models. Indicating the effectiveness of the proposed approach.

LM	Iteration	Training	F-1	Remaining %
AraRoBERTa-SA	0	2,800	0.818	9%
	1	28,528	**0.834**	7%
	2	30,028	0.83	<1%
AraRoBERTa-EG	0	2,800	0.933	63%
	1	17,020	**0.925**	34%
	2	23,608	0.911	2%
AraRoBERTa-KU	0	2,800	0.902	68%
	1	17,416	0.882	28%
	2	22,420	**0.886**	2%
AraRoBERTa-OM	0	2,800	0.84	51 %
	1	17,284	**0.802**	43%
	2	22,564	0.784	3%
AraRoBERTa-LB	0	2,800	0.876	72%
	1	23,440	**0.883**	25%
	2	27,928	0.864	1%
AraRoBERTa-JO	0	2,800	0.839	65%
	1	20,488	**0.832**	32%
	2	27,016	0.812	<1%
AraRoBERTa-DZ	0	2,800	0.859	84%
	1	27,016	0.854	13%
	2	29,608	**0.873**	<1%

Table 8: The semi-supervised classifiers results. The *Remaining %* equals the *remaining samples/original sample size (16K)*.

7.3 Weak-supervised Classification Model

In order to understand the results obtained by the AraRoBERTa models in weak-supervised setup, we looked at the performance of the models on the validation data as shown in Table 9. We can see the results obtained indicate the model learned from the automatically labeled data and obtained high results. However, the performance on the test data indicates that the models with lower results have learned from noisy samples, which can be one of the downsides of utilizing this approach. Here we can see this when comparing supervised AraRoBERTa-KU and the weak-supervised AraRoBERTa-KU, we can see the model is predicting the automatic positive sample as a negative sample. Indicating that these samples are noisy since the supervised version can identify the positive samples easily. On the other hand, we can see the effectiveness of weak-supervised on the same task but in different dialects like SA and EG. Providing a promising way of automatically labeling the dialect given a model trained on large data like SA and EG.

Dialect	Validation	Test	Performance Change
SA	0.9	0.812	-8.8%
EG	0.955	0.857	-9.8%
KU	0.948	0.744	-20.4%
OM	0.915	0.404	-51.1%
LB	0.966	0.783	-18.3%
JO	0.884	0.708	-17.6%
DZ	0.929	0.776	-15.3%

Table 9: The performance of AraRoBERTa in the weak-supervised setting on both the validation and test phases in all dialects based on the F-1 score.

8 Conclusion

This paper proposed different approaches for Arabic dialect text classification as a low-resource scenario and conducted an empirical study to evaluate the performance of the adopted approaches. The paper proposed a novel data collection pipeline from Twitter that is language and task agnostic.

Also, developed dialect-specific contextual language models to learn from unlabeled data that yield effective and stable performance across dialects, as seen in supervised classification. While AraRoBERTa models were pretrained on a fraction of the data size that other contextual baselines were trained on, the results showed that most of the supervised AraRoBERTa models outperformed these models. In addition, when compared to the traditional ML model, larger AraRoBERTa models outperform it as well.

Additionally, to the best of our knowledge, we constructed the first dialectal dictionary to utilize it in the automatic annotation in scenarios where labeled data are not available and then utilized in a weak-supervised task. Although the automatic function contains one hand-crafted rule, this approach is a promising technique for annotating large data and utilizing it in a text classification task. Also, the proposed SSL model can be adopted when only a few labeled examples are available where it shows its effectiveness and stability.

References

Arabic stop words. http://www.abuelkhair.net/index.php/en/arabic/arabic-stop-words. (Accessed on 10/22/2021).

Filtering tweets by location | docs | twitter developer platform. https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location. (Accessed on 01/02/2022).

Twitter: most users by country | statista. https://www.statista.com/statistics/242606/\number-of-active-twitter-users-in\-selected-countries/. (Accessed on 07/28/2021).

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.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018a. 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, Hassan Alhuzali, and Mohamed Elaraby. 2018b. 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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Karima Abidi, Mohamed Amine Menacer, and Kamel Smaili. 2017. Calyou: A comparable spoken algerian corpus harvested from youtube. In *18th Annual Conference of the International Communication Association (Interspeech)*.

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.

Nora Al-Twairesh, Rawan Al-Matham, Nora Madi, Nada Almugren, Al-Hanouf Al-Aljmi, Shahad Alshalan, Raghad Alshalan, Nafla Alrumayyan, Shams Al-Manea, Sumayah Bawazeer, et al. 2018. Suar: Towards building a corpus for the saudi dialect. *Procedia computer science*, 142:72–82.

Faisal Alshargi, Shahd Dibas, Sakhar Alkhereyf, Reem Faraj, Basmah Abdulkareem, Sane Yagi, Ouafaa Kacha, Nizar Habash, and Owen Rambow. 2019. Morphologically annotated corpora for seven Arabic dialects: Taizi, sanaani, najdi, jordanian, syrian, iraqi and Moroccan. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 137–147, Florence, Italy. Association for Computational Linguistics.

Maha J. Althobaiti. 2020. Automatic arabic dialect identification systems for written texts: A survey.

Maha J Althobaiti. 2021. Country-level arabic dialect identification using small datasets with integrated machine learning techniques and deep learning models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 265–270.

R. AlYami and R. AlZaidy. 2020. Arabic dialect identification in social media. In *2020 3rd International Conference on Computer Applications Information Security (ICCAIS)*, pages 1–2.

Sarah N Alyami and Sunday O Olatunji. 2020. Application of support vector machine for arabic sentiment classification using twitter-based dataset. *Journal of Information & Knowledge Management*, 19(01):2040018.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Ramy Baly, Alaa Khaddaj, Hazem Hajj, Wassim El-Hajj, and Khaled Bashir Shaban. 2019. Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets. *arXiv preprint arXiv:1906.01830*.

Ahmad Beltagy, Abdelrahman Wael, and Omar ElSherief. 2020. Arabic dialect identification using bert-based domain adaptation. *arXiv preprint arXiv:2011.06977*.

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. 2019a. 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.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019b. 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jan Deriu, Aurélien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. 2017. Leveraging large amounts of weakly supervised data for multi-language sentiment classification. *Proceedings of the 26th International Conference on World Wide Web*.

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*.

Omar Einea, Ashraf Elnagar, and Ridhwan Al Debsi. 2019. Sanad: Single-label arabic news articles dataset for automatic text categorization. *Data in brief*, 25:104076.

Ibrahim Abu El-Khair. 2017. Effects of stop words elimination for arabic information retrieval: A comparative study.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ashraf Elnagar and Omar Einea. 2016. Brad 1.0: Book reviews in arabic dataset. In *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE.

Ashraf Elnagar, Omar Einea, and Ridhwan Al-Debsi. 2019. Automatic text tagging of arabic news articles using ensemble deep learning models. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*, pages 59–66.

Ashraf Elnagar, Yasmin S Khalifa, and Anas Einea. 2018. Hotel arabic-reviews dataset construction for sentiment analysis applications. In *Intelligent natural language processing: Trends and applications*, pages 35–52. Springer.

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.

Kamel Gaanoun and Imade Benelallam. 2020. Arabic dialect identification: An Arabic-BERT model with data augmentation and ensembling strategy. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 275–281, Barcelona, Spain (Online). Association for Computational Linguistics.

Imane Guellil, Faical Azouaou, and Francisco Chiclana. 2020. Arautosenti: automatic annotation and new tendencies for sentiment classification of arabic messages. *Social Network Analysis and Mining*, 10(1):1–20.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. Mada+ tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt*, volume 41, page 62.

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2020. Arcov-19: The first arabic covid-19 twitter dataset with propagation networks. *arXiv preprint arXiv:2004.05861*.

Maram Hasanain, Reem Suwaileh, Tamer Elsayed, Mucahid Kutlu, and Hind Almerekhi. 2018. Evetar: building a large-scale multi-task test collection over arabic tweets. *Information Retrieval Journal*, 21(4):307–336.

Fei Huang. 2015. Improved arabic dialect classification with social media data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2118–2126.

Maher Itani, Chris Roast, and Samir Al-Khayatt. 2017. Corpora for sentiment analysis of arabic text in social media. In *2017 8th international conference on information and communication systems (ICICS)*, pages 64–69. IEEE.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

D. Jurafsky and J.H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of Gulf Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4282–4289, Portorož, Slovenia. European Language Resources Association (ELRA).

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shervin Malmasi and Marcos Zampieri. 2017. Arabic dialect identification using iVectors and ASR transcripts. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 178–183, Valencia, Spain. Association for Computational Linguistics.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan. The COLING 2016 Organizing Committee.

Karima Meftouh, Karima Abidi, Salima Harrat, and Kamel Smaili. 2019. The SMarT classifier for Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 259–263, Florence, Italy. Association for Computational Linguistics.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-supervised neural text classification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 983–992.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. Deep learning based text classification: A comprehensive review. *CoRR*, abs/2004.03705.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3).

Pruthwik Mishra and Vandan Mujadia. 2019. Arabic dialect identification for travel and Twitter text. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 234–238, Florence, Italy. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: A data set of german online discussions. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1241–1244, New York, NY, USA. Association for Computing Machinery.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Claudia Soria, Valeria Quochi, and Irene Russo. 2018. The DLDP survey on digital use and usability of EU regional and minority languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein T Al-Natsheh. 2020. Multi-dialect arabic bert for country-level dialect identification. *arXiv preprint arXiv:2007.05612*.

Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Yanshan Wang, Sunghwan Sohn, Sijia Liu, Feichen Shen, Liwei Wang, Elizabeth J Atkinson, Shreyasee Amin, and Hongfang Liu. 2019b. A clinical text classification paradigm using weak supervision and deep representation. *BMC medical informatics and decision making*, 19(1):1–13.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qiang Yang, Hind Alamro, Somayah Albaradei, Adil Salhi, Xiaoting Lv, Changsheng Ma, Manal Alshehri, Inji Jaber, Faroug Tifratene, Wei Wang, Takashi Gojobori, Carlos M. Duarte, Xin Gao, and Xiangliang Zhang. 2020. Senwave: Monitoring the global sentiments under the covid-19 pandemic.

Omar F. Zaidan 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*, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

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

M. Zampieri and P. Nakov. 2021. *Similar Languages, Varieties, and Dialects: A Computational Perspective*. Studies in Natural Language Processing. Cambridge University Press.

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. Association for Computational Linguistics.