NADI 2021: The Second Nuanced Arabic Dialect Identification Shared Task

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor,[†] Nizar Habash[‡]
The University of British Columbia, Vancouver, Canada
[†]Carnegie Mellon University in Qatar, Qatar
[‡]New York University Abu Dhabi, UAE
{muhammad.mageed, a.elmadany}@ubc.ca chiyuzh@mail.ubc.ca
hbouamor@cmu.edu nizar.habash@nyu.edu

Abstract

We present the findings and results of the Second Nuanced Arabic Dialect Identification Shared Task (NADI 2021). This Shared Task includes four subtasks: country-level Modern Standard Arabic (MSA) identification (Subtask 1.1), country-level dialect identification (Subtask 1.2), province-level MSA identification (Subtask 2.1), and province-level sub-dialect identification (Subtask 2.2). The shared task dataset covers a total of 100 provinces from 21 Arab countries, collected from the Twitter domain. A total of 53 teams from 23 countries registered to participate in the tasks, thus reflecting the interest of the community in this area. We received 16 submissions for Subtask 1.1 from five teams, 27 submissions for Subtask 1.2 from eight teams, 12 submissions for Subtask 2.1 from four teams, and 13 Submissions for subtask 2.2 from four teams.

1 Introduction

Arabic is the native tongue of $\sim$ 400 million people living the Arab world, a vast geographical region across Africa and Asia. Far from a single monolithic language, Arabic has a wide number of varieties. In general, Arabic could be classified into three main categories: (1) Classical Arabic, the language of the Qur'an and early literature; (2) Modern Standard Arabic (MSA), which is usually used in education and formal and pan-Arab media; and (3) dialectal Arabic (DA), a collection of geopolitically defined variants. Modern day Arabic is usually referred to as *diglossic* with a so-called 'High' variety used in formal settings (MSA), and a 'Low' variety used in everyday communication (DA). DA, the presumably 'Low' variety, is itself a host of variants. For the current work, we focus on geography as an axis of variation where people from various sub-regions, countries, or even provinces within the same country, may be using Arabic differently.

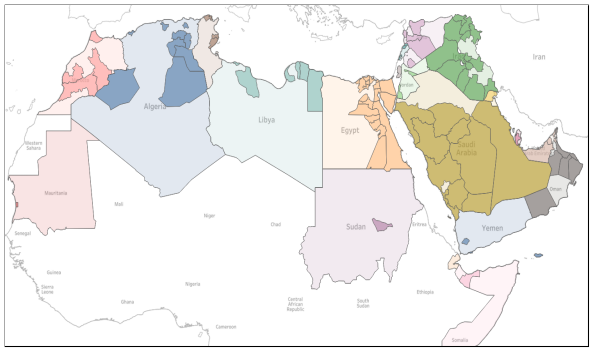

Figure 1: A map of the Arab World showing thr 21 countries and 100 provinces in the NADI 2021 datasets. Each country is coded in color different from neighboring countries. Provinces within each country are coded in a more intense version of the same color as the country.

The Nuanced Arabic Dialect Identification (NADI) series of shared tasks aim at furthering the study and analysis of Arabic variants by providing resources and organizing classification competitions under standardized settings. The First Nuanced Arabic Dialect Identification (NADI 2020) Shared Task targeted 21 Arab countries and a total of 100 provinces across these countries. NADI 2020 consisted of two subtasks: *country-level* dialect identification (Subtask 1) and *province-level* detection (Subtask 2). The two subtasks depended on Twitter data, making it the first shared task to target naturally-occurring fine-grained dialectal text at the sub-country level. The Second Nuanced Arabic Dialect Identification (NADI 2021) is similar to NADI 2020 in that it also targets the same 21 Arab countries and 100 corresponding provinces and is based on Twitter data. However, NADI 2021 has four subtasks, organized into country level and

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 244–259
Kyiv, Ukraine (Virtual), April 19, 2021.

province level. For each classification level, we afford both MSA and DA datasets as Table 1 shows.

Variety	Country	Province
MSA	Subtask 1.1	Subtask 2.1
DA	Subtask 1.2	Subtask 2.2

Table 1: NADI 2021 subtasks.

We provided participants with a new Twitter labeled dataset that we collected exclusively for the purpose of the shared task. The dataset is publicly available for research.[1] A total of 53 teams registered for the shard task, of whom 8 unique teams ended up submitting their systems for scoring. We allowed a maximum of five submissions per team. We received 16 submissions for Subtask 1.1 from five teams, 27 submissions for Subtask 1.2 from eight teams, 12 submissions for Subtask 2.1 from four teams, and 13 Submissions for subtask 2.2 from four teams. We then received seven papers, all of which we accepted for publication.

This paper is organized as follows. We provide a brief overview of the computational linguistic literature on Arabic dialects in Section 2. We describe the two subtasks and dataset in Sections 3 and Section 4, respectively. And finally, we introduce participating teams, shared task results, and a high-level description of submitted systems in Section 5.

2 Related Work

As we explained in Section 1, Arabic has three main categories: CA, MSA, and DA. While CA and MSA have been studied extensively (Harrell, 1962; Cowell, 1964; Badawi, 1973; Brustad, 2000; Holes, 2004), DA has received more attention only in recent years.

One major challenge with studying DA has been rarity of resources. For this reason, most pioneering DA works focused on creating resources, usually for only a small number of regions or countries (Gadalla et al., 1997; Diab et al., 2010; Al-Sabbagh and Girju, 2012; Sadat et al., 2014; Smaïli et al., 2014; Jarrar et al., 2016; Khalifa et al., 2016; Al-Twairesh et al., 2018; El-Haj, 2020). A number of works introducing multi-dialectal data sets and regional level detection models followed (Zaidan and Callison-Burch, 2011; Elfardy et al., 2014; Bouamor et al., 2014; Meftouh et al., 2015).

[1]The dataset is accessible via our GitHub at: https://github.com/UBC-NLP/nadi.

Arabic dialect identification work as further sparked by a series of shared tasks offered as part of the VarDial workshop. These shared tasks used speech broadcast transcriptions (Malmasi et al., 2016), and integrated acoustic features (Zampieri et al., 2017) and phonetic features (Zampieri et al., 2018) extracted from raw audio. Althobaiti (2020) is a recent survey of computational work on Arabic dialects.

The Multi Arabic Dialects Application and Resources (MADAR) project (Bouamor et al., 2018) introduced finer-grained dialectal data and a lexicon. The MADAR data were used for dialect identification at the city level (Salameh et al., 2018; Obeid et al., 2019) of 25 Arab cities. An issue with the MADAR data, in the context of DA identification, is that it was commissioned and not naturally occurring. Several larger datasets covering 10-21 countries were also introduced (Mubarak and Darwish, 2014; Abdul-Mageed et al., 2018; Zaghouani and Charfi, 2018). These datasets come from the Twitter domain, and hence are naturally-occurring.

Several works have also focused on sociopragmatics meaning exploiting dialectal data. These include sentiment analysis (Abdul-Mageed et al., 2014), emotion (Alhuzali et al., 2018), age and gender (Abbes et al., 2020), offensive language (Mubarak et al., 2020), and sarcasm (Abu Farha and Magdy, 2020). Concurrent with our work, (Abdul-Mageed et al., 2020c) also describe data and models at country, province, and city levels.

The first NADI shared task, NADI 2020 (Abdul-Mageed et al., 2020b), comprised two subtasks, one focusing on 21 Arab countries exploiting Twitter data, and another on 100 Arab provinces from the same 21 countries. As is explained in (Abdul-Mageed et al., 2020b), the NADI 2020 datasets included a small amount of non-Arabic and also a mixture of MSA and DA. For NADI 2021, we continue to focus on 21 countries and 100 provinces. However, we breakdown the data into MSA and DA for a stronger signal. This also gives us the opportunity to study each of these two main categories independently. In other words, in addition to dialect and sub-dialect identification, it allows us to investigate the extent to which MSA itself can be teased apart at the country and province levels. Our hope is that NADI 2021 will support exploring variation in geographical regions that have not been studied before.

3 Task Description

The NADI shared task consists of four subtasks, comprising two levels of classification–country and province. Each level of classification is carried out for both MSA and DA. We explain the different subtasks across each classification level next.

3.1 Country-level Classification

- **Subtask 1.1: Country-level MSA.** The goal of Subtask 1.1 is to identify country level MSA from short written sentences (tweets). NADI 2021 Subtask 1.1 is novel since no previous works focused on teasing apart MSA by country of origin.

- **Subtask 1.2: Country-level DA.** Subtask 1.2 is similar to Subtask 1.1, but focuses on identifying country level *dialect* from tweets. Subtask 1.2 is similar to previous works that have also taken country as their target (Mubarak and Darwish, 2014; Abdul-Mageed et al., 2018; Zaghouani and Charfi, 2018; Bouamor et al., 2019; Abdul-Mageed et al., 2020b).

We provided labeled data to NADI 2021 participants with specific training (TRAIN) and development (DEV) splits. Each of the 21 labels corresponding to the 21 countries is represented in both TRAIN and DEV. Teams could score their models through an online system (codalab) on the DEV set before the deadline. We released our TEST set of unlabeled tweets shortly before the system submission deadline. We then invited participants to submit their predictions to the online scoring system housing the gold TEST set labels. Table 2 shows the distribution of the TRAIN, DEV, and TEST splits across the 21 countries.

3.2 Province-level Classification

- **Subtask 2.1: Province-level MSA.** The goal of Subtask 2.1 is to identify the specific state or province (henceforth, *province*) from which an MSA tweet was posted. There are 100 province labels in the data, and provinces are unequally distributed among the list of 21 countries.

- **Subtask 2.2: Province-level DA.** Again, Subtask 2.2 is similar to Subtask 2.1, but the goal is identifying the province from which a *dialectal* tweet was posted.

While the MADAR shared task (Bouamor et al., 2019) involved prediction of a small set of cities, NADI 2020 was the first to propose automatic dialect identification at geographical regions as small as provinces. Concurrent with NADI 2020, (Abdul-Mageed et al., 2020c) introduced the concept of *microdialects*, and proposed models for identifying language varieties defined at both province and city levels. NADI 2021 follows these works, but has one novel aspect: We introduce province-level identification for MSA and DA independently (i.e., each variety is handled in a separate subtask). While province-level sub-dialect identification may be challenging, we hypothesize province-level MSA might be even more difficult. However, we were curious to what extent, if possible at all, a machine would be successful in teasing apart MSA data at the province-level.

In addition, similar to NADI 2020, we acknowledge that province-level classification is somewhat related to geolocation prediction exploiting Twitter data. However, we emphasize that geolocation prediction is performed at the level of *users*, rather than tweets. This makes our subtasks different from geolocation work. Another difference lies in the way we collect our data as we will explain in Section 4. Tables 11 and 12 (Appendix A) show the distribution of the 100 province classes in our MSA and DA data splits, respectively. ***Importantly, for all 4 subtasks, tweets in the TRAIN, DEV and TEST splits come from disjoint sets.***

3.3 Restrictions and Evaluation Metrics

We follow the same general approach to managing the shared task as our first NADI in 2020. This includes providing participating teams with a set of restrictions that apply to all subtasks, and clear evaluation metrics. The purpose of our restrictions is to ensure fair comparisons and common experimental conditions. In addition, similar to NADI 2020, our data release strategy and our evaluation setup through the CodaLab online platform facilitated the competition management, enhanced timeliness of acquiring results upon system submission, and guaranteed ultimate transparency.[2]

Once a team registered in the shared task, we directly provided the registering member with the data via a private download link. We provided the data in the form of the actual tweets posted to the Twitter platform, rather than tweet IDs. This

[2] https://codalab.org/

Country	Provinces	MSA (Subtasks 1.1 & 2.1)					DA (Subtasks 1.2 & 2.2)				
		Train	DEV	TEST	Total	%	Train	DEV	TEST	Total	%
Algeria	9	1,899	427	439	2,765	8.92	1,809	430	391	2,630	8.48
Bahrain	1	211	51	51	313	1.01	215	52	52	319	1.03
Djibouti	1	211	52	51	314	1.01	215	27	7	249	0.80
Egypt	20	4,220	1,032	989	6,241	20.13	4,283	1,041	1,051	6,375	20.56
Iraq	13	2,719	671	652	4,042	13.04	2,729	664	664	4,057	13.09
Jordan	2	422	103	102	627	2.02	429	104	105	638	2.06
Kuwait	2	422	103	102	627	2.02	429	105	106	640	2.06
Lebanon	3	633	155	141	929	3.00	644	157	120	921	2.97
Libya	6	1,266	310	307	1,883	6.07	1,286	314	316	1,916	6.18
Mauritania	1	211	52	51	314	1.01	215	53	53	321	1.04
Morocco	4	844	207	205	1,256	4.05	858	207	212	1,277	4.12
Oman	7	1,477	341	357	2,175	7.02	1,501	355	371	2,227	7.18
Palestine	2	422	102	102	626	2.02	428	104	105	637	2.05
Qatar	1	211	52	51	314	1.01	215	52	53	320	1.03
KSA	10	2,110	510	510	3,130	10.10	2,140	520	522	3,182	10.26
Somalia	2	346	63	102	511	1.65	172	49	55	276	0.89
Sudan	1	211	48	51	310	1.00	215	53	53	321	1.04
Syria	6	1,266	309	306	1,881	6.07	1,287	278	288	1,853	5.98
Tunisia	4	844	170	176	1,190	3.84	859	173	212	1,244	4.01
UAE	3	633	154	153	940	3.03	642	157	158	957	3.09
Yemen	2	422	88	102	612	1.97	429	105	106	640	2.06
Total	**100**	**21,000**	**5,000**	**5,000**	**31,000**	**100**	**21,000**	**5,000**	**5,000**	**31,000**	**100**

Table 2: Distribution of classes and data splits over our MSA and DA datasets for the four subtasks.

guaranteed comparison between systems exploiting identical data. For all four subtasks, we provided clear instructions requiring participants not to use any external data. That is, teams were required to only use the data we provided to develop their systems and no other datasets regardless how these are acquired. For example, we requested that teams do not search nor depend on any additional user-level information such as geolocation. To alleviate these strict constraints and encourage creative use of diverse (machine learning) methods in system development, we provided an unlabeled dataset of 10M tweets in the form of tweet IDs. This dataset is in addition to our labeled TRAIN and DEV splits for the four subtasks. To facilitate acquisition of this unlabeled dataset, we also provided a simple script that can be used to collect the tweets. We encouraged participants to use these 10M unlabeled tweets in any way they wished.

For all four subtasks, the official metric is macro-averaged F_1 score obtained on blind test sets. We also report performance in terms of macro-averaged precision, macro-averaged recall and accuracy for systems submitted to each of the four subtasks. Each participating team was allowed to submit up to five runs for each subtask, and only the highest scoring run was kept as representing the team. Although official results are based only on a blind TEST set, we also asked participants to report their results on the DEV set in their papers. We setup four CodaLab competitions for scoring participant systems.[3] We will keep the Codalab competition for each subtask live post competition, for researchers who would be interested in training models and evaluating their systems using the shared task TEST set. For this reason, we will not release labels for the TEST set of any of the subtasks.

4 Shared Task Datasets

We distributed two Twitter datasets, one in MSA and another in DA. Each tweet in each of these two datasets has two labels, one label for country level and another label for province level. For example, for the MSA dataset, the same tweet is assigned one out of 21 country labels (Subtask 1.1) and one out of 100 province labels (Subtask 2.1). The same applies to DA data, where each tweet is assigned a country label (Subtask 1.2) and a province label (Subtask 2.2). Similar to MSA, the tagset for DA data has 21 country labels and 100 province labels.

[3]Links to the CodaLab competitions are as follows: Subtask 1.1: https://competitions.codalab.org/competitions/27768, Subtask 1.2: https://competitions.codalab.org/competitions/27769, Subtask 2.1: https://competitions.codalab.org/competitions/27770, Subtask 2.2: https://competitions.codalab.org/competitions/27771.

In addition, as mentioned before, we made available an unlabeled dataset for optional use in any of the four subtasks. We now provide more details about both the labeled and unlabeled data.

4.1 Data Collection

Similar to NADI 2020, we used the Twitter API to crawl data from 100 provinces belonging to 21 Arab countries for 10 months (Jan. to Oct., 2019).[4] Next, we identified users who consistently and *exclusively* tweeted from a single province during the whole 10 month period. We crawled up to 3,200 tweets from each of these users. We select only tweets assigned the Arabic language tag (`ar`) by Twitter. We lightly normalize tweets by removing usernames and hyperlinks, and add white space between emojis. Next, we remove retweets (i.e., we keep only tweets and replies). Then, we use character-level string matching to remove sequences that have < 3 Arabic tokens.

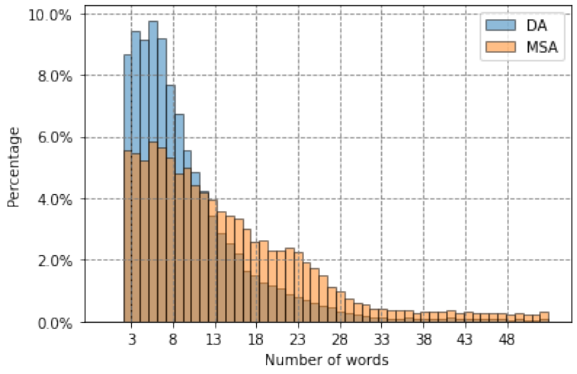

Figure 2: Distribution of tweet length (trimmed at 50) in words in NADI-2021 labeled data.

Since the Twitter language tag can be wrong sometimes, we apply an effective in-house language identification tool on the tweets and replies to exclude any non-Arabic. This helps us remove posts in Farsi (`fa`) and Persian (`ps`) which Twitter wrongly assigned an Arabic language tag. Finally, to tease apart MSA from DA, we use the dialect-MSA model introduced in Abdul-Mageed et al. (2020a) (acc= 89.1%, F1= 88.6%).

4.2 Data Sets

To assign labels for the different subtasks, we use user *location* as a proxy for *language variety labels* at both country and province levels. This applies to both our MSA and DA data. That is, we label tweets from each user with the country and province from which the user consistently posted for the *whole* of the 10 months period. Although this method of label assignment is not ideal, it is still a reasonable approach for easing the bottleneck of data annotation. For both the MSA and DA data, across the two levels of classification (i.e., country and province), we randomly sample 21K tweets for training (TRAIN), 5K tweets for development (DEV), and 5K tweets for testing (TEST). These three splits come from three disjoint sets of users. We distribute data for the four subtasks directly to participants in the form of actual tweet text. Table 2 shows the distribution of tweets across the data splits over the 21 countries, for all subtasks. We provide the data distribution over the 100 provinces in Appendix A. More specifically, Table 11 shows the province-level distribution of tweets for MSA (Subtask 2.1) and Table 12 shows the same for DA (Subtask 2.2). We provide examples DA tweets from a number of countries representing different regions in Table 3. For each example in Table 3, we list the province it comes from. Similarly, we provide example MSA data in Table 4.

Unlabeled 10M. We shared 10M Arabic tweets with participants in the form of tweet IDs. We crawled these tweets in 2019. Arabic was identified using Twitter language tag (`ar`). This dataset does not have any labels and we call it UNLABELED 10M. We also included in our data package released to participants a simple script to crawl these tweets. Participants were free to use UNLABELED 10M for any of the four subtasks in any way they they see fits.[5] We now present shared task teams and results.

5 Shared Task Teams & Results

5.1 Our Baseline Systems

We provide two simple baselines, Baseline I and Baseline II, for each of the four subtasks. **Baseline I** is based on the majority class in the TRAIN data for each subtask. It performs at $F_1 = 1.57\%$ and $accuracy = 19.78\%$ for Subtask 1.1, $F_1 = 1.65\%$ and $accuracy = 21.02\%$ for Subtask 1.2, $F_1 = 0.02\%$ and $accuracy = 1.02\%$ for Subtask

[4] Although we tried, we could not collect data from Comoros to cover all 22 Arab countries.

[5] Datasets for all the subtasks and UNLABELED 10M are available at https://github.com/UBC-NLP/nadi. More information about the data format can be found in the accompanying README file.

Country	Province	Tweet
Algeria	Bouira	شحال راكي تبيعي فيه... نحي ندي كونتيتي
	Khenchela	شحال يقلبو النم هاذوك لي يقدسو الراپرز ،!
	Oran	راك زعفان مزية ربحتوا
Egypt	Alexandria	بس مش زي ما هنقدر نعدي حياتنا
	Minya	بص آنا كل مضميري هيبقي صاحي هجيلك تقتلهولي
	Sohag	أنا معنديش حد يفسحنى زي الوليه
KSA	Ar-Riyad	و يعدين تجلسون تلعبون سماش !!
	Ash-Sharqiyah	يابن الحلال اسمهم رجال فقراويه مايفهمون تبي حريمهم تفهم
	Tabuk	طيب ايش دخل الوزارة هذي وفاه طبيعيه
Morocco	Marrakech-Tensift-Al-Haouz	مافي ربح ولا كرش العصبية اشرب بنادول عشان ارتاح
	Meknes-Tafilalet	مراداون إنيخ أداخ إعفو ربي خ إمداكلن لحاجت !!!!
	Souss-Massa-Draa	أبيه نسيتهم كاع حتى هما متشيعيين بزاف
Oman	Ash-Sharqiyah	يحصلج انتي اصلاً حد يعلق لج.
	Dhofar	اسمعي قران ، ميبي تغفين عليه
	Musandam	ماني بقايل غلاتك هقوه وخابت بقول عين الحسود الله يجازيها
Palestine	Gaza-Strip	احنا اليوم عاملين مفتول وانتو ؟؟
	West-Bank	اتفقنا ع هيك
	West-Bank	اليوم كنت فيهم بجننو
Sudan	Khartoum	صرت عادي اشوف أشياء تقهرني واسكت. لأني ادري لو تكلمت م راح القي شي برضيني صرت انهي كلامي...
	Khartoum	من وين جبتي كلامك دا..... إستندتي علي شنو إنه لو قتلنا ح نكون قتلة...
	Khartoum	انا شكلي حأتبع سياسة قصي وابلك من طرف.
UAE	Abu-Dhabi	يحسبني لاهي عنه و أنا ملتهي به
	Dubai	يلي يبا خوتي بالطيب خاويته
	Ras-Al-Khaymah	ل موسم نخربها بالاخر وما تدري شو المشكلة الأساسية احسك مني مليت

Table 3: Randomly picked DA tweets from select provinces and corresponding countries.

2.1, and $F_1 = 0.02\%$ and $accuracy = 1.06\%$ for Subtask 2.2.

Baseline II is a fine-tuned multi-lingual BERT-Base model (mBERT)[6]. More specifically, we fine-tune mBERT for 20 epochs with a learning rate of $2e-5$, and batch size of 32. The maximum length of input sequence is set as 64 tokens. We evaluate the model at the end of each epoch and choose the best model on our DEV set. We then report the best model on the TEST set. Our best mBERT model obtains $F_1 = 14.15\%$ and $accuracy = 24.76\%$ on Subtask 1.1, $F_1 = 18.02\%$ and $accuracy = 33.04\%$ on Subtask 1.2, $F_1 = 3.39\%$ and $accuracy = 3.48\%$ on Subtask 2.1, and $F_1 = 4.08\%$ and $accuracy = 4.18\%$ on Subtask 2.2 as Tables 6, 7, 8, and 9, respectively.

[6]https://github.com/google-research/bert

5.2 Participating Teams

We received a total of 53 unique team registrations. After evaluation phase, we received a total of 68 submissions. The breakdown across the subtasks is as follows: 16 submissions for Subtask 1.1 from five teams, 27 submissions for Subtask 1.2 from eight teams, 12 submissions for Subtask 2.1 from four teams, and 13 submissions for Subtask 2.2 from four teams. Of participating teams, seven teams submitted description papers, all of which we accepted for publication. Table 5 lists the seven teams.

Country	Province	Tweet
Algeria	Biskra	عندك حق جلال .اصبح مستحيل
	Oran	أنتظرتك طويلا على عتبة اللقاء جهزت خطابا ... و عتابا .. و ما اردت غير عناقا .. اخبرتك كثيرا ... انني اريد ان ابكي... اوغى
	Ouargla	أجمل شيئ هو البساطة فألف تحية لعائلة أبو تريك
Egypt	Faiyum	الان يرحل عن ربوعك فارس مهزوم
	Minya	من الأذكار، أعوذ بكلمات الله التامة من كل شيطان وهامة ومن كل عين لامة
	Red-Sea	اللهم بشرني بالجنه ، اللهم ارزقني الفردوس الاعلى من غير حساب ولا سابقة عذاب
KSA	Al-Madinah	ومن جميل ما قيل في السّلام : سلاماً على من مَرَ على مُرِنَا فحَلّاه .
	Ar-Riyad	اللهُم إجعلني من الذين إذا أحسنوا إستبشروا وإذا أساءوا أستغفَرُ.
	Jizan	لا إله إلا أنت سُبحانك إنك على كل شئ وكيل
Morocco	Marrakech-Tensift-Al-Haouz	اصلا متى مر يوم بدون لا اتضايق ؟
	Meknes-Tafilalet	اللهم اجعل أيامنا كلها أعيادًا بطاعتك .. وامطر على قلوبنا فرحًا لا ينتهي
Oman	Ad-Dhahirah	النور نورك اخي
	Muscat	يحفظه ويطول لنا بعمره..اشتقنا لطلته الغالية عسى ربي يلبسه ثوب الصحة والعافية
Palestine	Gaza-Strip	معبرة جدا ..لا فض فوك
	West-Bank	بارك الله فيك ابو جود
	West-Bank	الصراحه هذا واقعنا العربي
Sudan	Khartoum	لا يا عزيزي الحرب بدأها الترابي .. وقبل وصول الترابي لرئاسة الوزراء كانت الحركة الاسلامية تطبخ الصراع وتخلق الصراع تجهيزا للحرب بقيادة الترابي
	Khartoum	اللهم إني استودعتك مستقبلي فأجعله أجمل مما تمنيت
UAE	Dubai	بعض مما عندكم يا فندم
	Ras-Al-Khaymah	لاجديد ميسي يجلد مدريد
	Ras-Al-Khaymah	نشاطنا وأفكارنا بالإيجاب أو بالسلب تشبه المغناطيس ، ونحن نحاول أن نتجنب المشاكل تستمر المشاكل في الحدوث وقد تزداد سوءاً

Table 4: Randomly picked MSA tweets from select provinces and corresponding countries.

Team	Affiliation	Tasks
AraDial_MJ (Althobaiti, 2021)	Taif Uni, KSA	1.2
Arizona (Issa, 2021)	Uni of Arizona, USA	1.2
CairoSquad (AlKhamiss et al., 2021)	Microsoft, Egypt	all
CS-UM6P (El Mekki et al., 2021)	Mohammed VI Polytech, Morocco	all
NAYEL (Nayel et al., 2021)	Benha Uni, Egypt	all
Phonemer (Wadhawan, 2021)	Flipkart Private Limited, India	all
Speech Trans (Lichouri et al., 2021)	CRSTDLA, Algeria	1.1, 1.2

Table 5: List of teams that participated in one or more of the four subtasks *and* submitted a system description paper.

5.3 Shared Task Results

Table 6 presents the best TEST results for all 5 teams who submitted systems for Subtask 1.1. Based on the official metric, $macro - F_1$, CairoSquad obtained the best performance with 22.38% F_1 score. Table 7 presents the best TEST results of each of the eight teams who submitted systems to Subtask 1.2. Team CairoSquad achieved the best F_1 score that is 32.26%. Table 8 shows the best TEST results for all four teams who submitted systems for Subtask 2.1. CairoSquad achieved the best performance with 6.43% F_1 score.

Table 9 provides the best TEST results of each of the four teams who submitted systems to Subtask 2.2. CairoSquad also achieved the best perfor-

Team	F_1	Acc	Precision	Recall
CairoSquad	22.38(1)	35.72(1)	31.56(1)	20.66(1)
Phonemer	21.79(2)	32.46(3)	30.03(3)	19.95(2)
CS-UM6P	21.48(3)	33.74(2)	30.72(2)	19.70(3)
Speech Translation	14.87(4)	24.32(4)	18.95(4)	13.85(4)
Our Baseline II	14.15	24.76	20.01	13.21
NAYEL	12.99(5)	23.24(5)	15.09(5)	12.46(5)
Our Baseline I	1.57	19.78	0.94	4.76

Table 6: Results for Subtask 1.1 (country-level MSA). The numbers in parentheses are the ranks. The table is sorted on the $macro - F1$ score, the official metric.

Team	F_1	Acc	Precision	Recall
CairoSquad	32.26(1)	51.66(1)	36.03(1)	31.09(1)
CS-UM6P	30.64(2)	49.50(2)	32.91(2)	30.34(2)
Phonemer	24.29(4)	44.14(3)	30.24(3)	23.70(4)
Speech Translation	21.49(5)	40.54(5)	26.75(5)	20.36(6)
Arizona	21.37(6)	40.46(6)	26.32(6)	20.78(5)
AraDial_MJ	18.94(7)	35.94(8)	21.58(8)	18.28(7)
NAYEL	18.72(8)	37.16(7)	21.61(7)	18.12(8)
Our Baseline II	18.02	33.04	18.69	17.88
Our Baseline I	1.65	21.02	1.00	4.76

Table 7: Results for Subtask 1.2 (province-level MSA)

Team	F_1	Acc	Precision	Recall
CairoSquad	6.43(1)	6.66(1)	7.11(1)	6.71(1)
Phonemer	5.49(2)	6.00(2)	6.17(2)	6.07(2)
CS-UM6P	5.35(3)	5.72(3)	5.71(3)	5.75(3)
NAYEL	3.51(4)	3.38(4)	4.09(4)	3.45(4)
Our Baseline II	3.39	3.48	3.68	3.49
Our Baseline I	0.02	1.02	0.01	1.00

Table 8: Results for Subtask 2.1 (country-level DA).

Team	F_1	Acc	Precision	Recall
CairoSquad	8.60(1)	9.46(1)	9.07(1)	9.33(1)
CS-UM6P	7.32(2)	7.92(2)	7.73(2)	7.95(2)
NAYEL	4.55(3)	4.80(3)	4.71(3)	4.55(4)
Phonemer	4.37(4)	5.32(4)	4.49(4)	5.19(3)
Our Baseline II	4.08	4.18	4.54	4.22
Our Baseline I	0.02	1.06	0.01	1.00

Table 9: Results for Subtask 2.2 (province-level DA).

mance with 8.60%.[7]

5.4 General Description of Submitted Systems

In Table 10, we provide a high-level description of the systems submitted to each subtask. For each team, we list their best score of each subtask, the features employed, and the methods adopted/developed. As can be seen from the table, the majority of the top teams have used Transformers. Specifically, team CairoSquad and CS-UM6P developed their system utilizing MARBERT (Abdul-Mageed et al., 2020a), a pre-trained Transformer language model tailored to Arabic dialects and the domain of social media. Team Phonemer utilized AraBERT (Antoun et al., 2020a) and AraELECTRA (Antoun et al., 2020b). Team CairSquad apply adapter modules (Houlsby et al., 2019) and vertical attention to MARBERT fine-tuning. CS-UM6P fine-tuned MARBERT on country-level and province-level jointly by multi-task learning. The rest of participating teams have either used a type of neural networks other than Transformers or resorted to linear machine learning

[7]The full sets of results for Subtask 1.1, 1.2, 2.1, and 2.2 are in Tables 13, 14, 15 and 15, respectively, in Appendix A.

Team	F_1	Features						Techniques					
		N-gram	TF-IDF	Linguistics	Word embeds	PMI	Sampling	Classical ML	Neural nets	Transformer	Ensemble	Multitask	Semi-super
SUBTASK 1.1													
CairoSquad	22.38									✓	✓		
Phonemer	21.79									✓			
CS-UM6P	21.48									✓		✓	
Speech Trans	14.87	✓	✓				✓	✓	✓		✓		
NAYEL	12.99		✓					✓	✓				
SUBTASK 1.2													
CairoSquad	32.26									✓	✓		
CS-UM6P	30.64									✓		✓	
Phonemer	24.29									✓			
Speech Trans	21.49	✓	✓				✓	✓	✓		✓		
Arizona	21.37			✓	✓				✓				
AraDial_MJ	18.94	✓	✓		✓	✓		✓	✓		✓		✓
NAYEL	18.72		✓					✓	✓				
SUBTASK 2.1													
CairoSquad	6.43									✓	✓		
Phonemer	5.49									✓			
CS-UM6P	5.35									✓		✓	
NAYEL	3.51		✓					✓	✓				
SUBTASK 2.2													
CairoSquad	8.60									✓	✓		
CS-UM6P	7.32									✓		✓	
NAYEL	4.55		✓					✓	✓				
Phonemer	4.37									✓			

Table 10: Summary of approaches used by participating teams. PMI: poinwise mutual information. Classical ML refers to any non-neural machine learning methods such as naive Bayes and support vector machines. The term "neural nets" refers to any model based on neural networks (e.g., FFNN, RNN, and CNN) except Transformer models. Transformer refers to neural networks based on a Transformer architecture such as BERT. The table is sorted by official metric , $macro - F_1$. We only list teams that submitted a description paper. "Semi-super" indicates that the model is trained with semi-supervised learning.

models, usually with some form of ensembling.

6 Conclusion and Future Work

We presented the findings and results of the NADI 2021 shared task. We described our datasets across the four subtasks and the logistics of running the shared task. We also provided a panoramic description of the methods used by all participating teams. The results show that distinguishing the language variety of short texts based on small geographical regions of origin is possible, yet challenging. The total number of submissions during official evaluation (n=68 submissions from 8 unique teams), as well as the number of teams who registered and acquired our datasets (n=53 unique teams) reflects a continued interest in the community and calls for further work in this area.

In the future, we plan to host a third iteration of the NADI shared task that will use new datasets and encourage novel solutions to the set of problems introduced in NADI 2021. As results show all the fours subtasks remain very challenging, and we hope that encouraging further solutions will help advance work in this area.

Acknowledgments

We gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada, the Social Sciences Research Council of Canada, Compute Canada, and UBC Sockeye.

References

Ines Abbes, Wajdi Zaghouani, Omaima El-Hardlo, and Faten Ashour. 2020. Daict: A dialectal Arabic irony corpus extracted from twitter. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6265–6271.

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 Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. Samar: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech and Language*, 28(1):20–37.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020a. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. *arXiv preprint arXiv:2010.04900*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020b. NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, pages 97–110, Barcelona, Spain.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020c. Micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39.

Rania Al-Sabbagh and Roxana Girju. 2012. YADAC: Yet another Dialectal Arabic Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2882–2889.

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, Nourah Al-Mutlaq, Nada Almanea, Waad Bin Huwaymil, Dalal Alqusair, Reem Alotaibi, Suha Al-Senaydi, and Abeer Alfutamani. 2018. SUAR: Towards building a corpus for the Saudi dialect. In *Proceedings of the International Conference on Arabic Computational Linguistics (ACLing)*.

Hassan Alhuzali, Muhammad Abdul-Mageed, and Lyle Ungar. 2018. Enabling deep learning of emotion with first-person seed expressions. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 25–35.

Badr AlKhamiss, Mohamed Gabr, Muhammed El-Nokrashy, and Khaled Essam. 2021. Adapting MARBERT for Improved Arabic Dialect Identification: Submission to the NADI 2021 Shared Task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP 2021)*.

Maha J Althobaiti. 2020. Automatic Arabic dialect identification systems for written texts: A survey. *arXiv preprint arXiv:2009.12622*.

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 (WANLP 2021)*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020a. 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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020b. AraELECTRA: pre-training text discriminators for arabic language understanding. *arXiv preprint arXiv:2012.15516*.

MS Badawi. 1973. Levels of contemporary Arabic in Egypt. *Cairo: Dâr al Ma'ârif*.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

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.

Kristen Brustad. 2000. *The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects*. Georgetown University Press.

Mark W. Cowell. 1964. *A Reference Grammar of Syrian Arabic*. Georgetown University Press, Washington, D.C.

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. COLABA: Arabic dialect annotation and processing. In *LREC*

workshop on Semitic language processing, pages 66–74.

Mahmoud El-Haj. 2020. Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France.

Abdellah El Mekki, Abdelkader El Mahdaouy, Kabil Essefar, Nabil El Mamoun, Ismail Berrada, and Ahmed Khoumsi. 2021. CS-UM6P @ NADI'2021: BERT-based Multi-Task Model for Country and Province Level MSA and Dialectal Arabic Identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP 2021)*.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2014. Aida: Identifying code switching in informal Arabic text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 94–101, Doha, Qatar.

Hassan Gadalla, Hanaa Kilany, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi, Amr Shalaby, Krisjanis Karins, Everett Rowson, Robert MacIntyre, Paul Kingsbury, David Graff, and Cynthia McLemore. 1997. CALLHOME Egyptian Arabic transcripts LDC97T19. Web Download. Philadelphia: Linguistic Data Consortium.

R.S. Harrell. 1962. *A Short Reference Grammar of Moroccan Arabic: With Audio CD*. Georgetown classics in Arabic language and linguistics. Georgetown University Press.

Clive Holes. 2004. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Elsayed Issa. 2021. Country-level Arabic dialect identification using RNNs with and without linguistic features. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP 2021)*.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2016. Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, pages 1–31.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A Large Scale Corpus of Gulf Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.

Mohamed Lichouri, Mourad Abbas, Khaled Lounnas, Besma Benaziz, and Aicha Zitouni. 2021. Arabic Dialect Identification based on Weighted Concatenation of TF-IDF Transformers. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP 2021)*.

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.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on padic: A parallel Arabic dialect corpus. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation*.

Hamdy Mubarak and Kareem Darwish. 2014. Using Twitter to collect a multi-dialectal corpus of Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, Doha, Qatar.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of osact4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52.

Hamada Nayel, Ahmed Hassan, Mahmoud Sobhi, and Ahmed El-Sawy. 2021. Data-Driven Approach for Arabic Dialect Identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP 2021)*.

Ossama Obeid, Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2019. ADIDA: Automatic dialect identification for Arabic. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 6–11, Minneapolis, Minnesota. Association for Computational Linguistics.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic identification of Arabic language varieties and dialects in social media. *Proceedings of SocialNLP*, page 22.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1332–1344, Santa Fe, New Mexico, USA.

Kamel Smaïli, Mourad Abbas, Karima Meftouh, and Salima Harrat. 2014. Building resources for Algerian Arabic dialects. In *Proceedings of the Conference of the International Speech Communication Association (Interspeech)*.

Anshul Wadhawan. 2021. Dialect Identification in Nuanced Arabic Tweets Using Farasa Segmentation and AraBERT. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop (WANLP 2021)*.

Wajdi Zaghouani and Anis Charfi. 2018. ArapTweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

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: short papers-Volume 2*, pages 37–41. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the vardial evaluation campaign 2017.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardzic, Nikola Ljubešić, Jörg Tiedemann, et al. 2018. Language identification and morphosyntactic tagging: The second vardial evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 1–17.

Appendices

A Data

We provide the distribution Distribution of the NADI 2021 MSA data over provinces, by country (Subtask 2.1), across our our data splits in Table 11. Similarly, Table 12 shows the distribution of the DA data over provinces for all countries (Subtask 2.2) in our data splits.

B Shared Task Teams & Results

We provide full results for all four subtasks. Table 13 shows full results for Subtask 1.1, Table 14 for Subtask 1.2, Table 15 for Subtask 2.1, and Table 16 for Subtask 2.2.

Province Name	TRAIN	DEV	TEST	Province Name	TRAIN	DEV	TEST
ae_Abu-Dhabi	211	51	51	kw_Jahra	211	52	51
ae_Dubai	211	52	51	lb_Akkar	211	52	39
ae_Ras-Al-Khaymah	211	51	51	lb_North-Lebanon	211	51	51
bh_Capital	211	51	51	lb_South-Lebanon	211	52	51
dj_Djibouti	211	52	51	ly_Al-Butnan	211	52	51
dz_Batna	211	52	51	ly_Al-Jabal-al-Akhdar	211	52	52
dz_Biskra	211	52	51	ly_Benghazi	211	51	51
dz_Bouira	211	12	51	ly_Darnah	211	52	51
dz_Béchar	211	52	31	ly_Misrata	211	52	51
dz_Constantine	211	51	51	ly_Tripoli	211	51	51
dz_El-Oued	211	52	51	ma_Marrakech-Tensift-Al-Haouz	211	51	51
dz_Khenchela	211	52	51	ma_Meknes-Tafilalet	211	52	52
dz_Oran	211	52	51	ma_Souss-Massa-Draa	211	52	51
dz_Ouargla	211	52	51	ma_Tanger-Tetouan	211	52	51
eg_Alexandria	211	51	51	mr_Nouakchott	211	52	51
eg_Aswan	211	52	51	om_Ad-Dakhiliyah	211	51	51
eg_Asyut	211	52	51	om_Ad-Dhahirah	211	32	51
eg_Beheira	211	52	51	om_Al-Batnah	211	51	51
eg_Beni-Suef	211	52	51	om_Ash-Sharqiyah	211	51	51
eg_Dakahlia	211	51	51	om_Dhofar	211	52	51
eg_Faiyum	211	52	51	om_Musandam	211	52	51
eg_Gharbia	211	52	51	om_Muscat	211	52	51
eg_Ismailia	211	52	51	ps_Gaza-Strip	211	51	51
eg_Kafr-el-Sheikh	211	52	20	ps_West-Bank	211	51	51
eg_Luxor	211	52	51	qa_Ar-Rayyan	211	52	51
eg_Minya	211	51	51	sa_Al-Madinah	211	51	51
eg_Monufia	211	52	51	sa_Al-Quassim	211	51	51
eg_North-Sinai	211	52	51	sa_Ar-Riyad	211	51	51
eg_Port-Said	211	51	51	sa_Ash-Sharqiyah	211	51	51
eg_Qena	211	51	51	sa_Asir	211	51	51
eg_Red-Sea	211	52	51	sa_Ha'il	211	51	51
eg_Sohag	211	51	51	sa_Jizan	211	51	51
eg_South-Sinai	211	51	51	sa_Makkah	211	51	51
eg_Suez	211	51	51	sa_Najran	211	51	51
iq_Al-Anbar	211	51	51	sa_Tabuk	211	51	51
iq_Al-Muthannia	211	52	51	sd_Khartoum	211	48	51
iq_An-Najaf	211	51	51	so_Banaadir	211	52	51
iq_Arbil	211	52	51	so_Woqooyi-Galbeed	135	11	51
iq_As-Sulaymaniyah	187	52	51	sy_Aleppo	211	51	51
iq_Babil	211	52	51	sy_As-Suwayda	211	51	51
iq_Baghdad	211	51	51	sy_Damascus-City	211	51	51
iq_Basra	211	51	51	sy_Hama	211	52	51
iq_Dihok	211	52	51	sy_Hims	211	52	51
iq_Karbala	211	52	40	sy_Lattakia	211	52	51
iq_Kirkuk	211	52	51	tn_Ariana	211	51	51
iq_Ninawa	211	52	51	tn_Bizerte	211	15	51
iq_Wasit	211	51	51	tn_Mahdia	211	52	23
jo_Aqaba	211	52	51	tn_Sfax	211	52	51
jo_Zarqa	211	51	51	ye_Aden	211	51	51
kw_Hawalli	211	51	51	ye_Ibb	211	37	51

Table 11: Distribution of the NADI 2021 MSA data over provinces, by country, across our TRAIN, DEV, and TEST splits (Subtask 2.1).

Province Name	TRAIN	DEV	TEST	Province Name	TRAIN	DEV	TEST
ae_Abu-Dhabi	214	52	52	kw_Jahra	215	53	53
ae_Dubai	214	53	53	lb_Akkar	215	53	14
ae_Ras-Al-Khaymah	214	52	53	lb_North-Lebanon	215	52	53
bh_Capital	215	52	52	lb_South-Lebanon	214	52	53
dj_Djibouti	215	27	7	ly_Al-Butnan	214	52	53
dz_Batna	215	34	10	ly_Al-Jabal-al-Akhdar	215	53	53
dz_Biskra	215	53	53	ly_Benghazi	214	52	52
dz_Bouira	215	26	53	ly_Darnah	215	53	53
dz_Béchar	215	53	11	ly_Misrata	214	52	53
dz_Constantine	215	52	53	ly_Tripoli	214	52	52
dz_El-Oued	215	53	52	ma_Marrakech-Tensift-Al-Haouz	214	52	53
dz_Khenchela	89	53	53	ma_Meknes-Tafilalet	215	50	53
dz_Oran	215	53	53	ma_Souss-Massa-Draa	215	53	53
dz_Ouargla	215	53	53	ma_Tanger-Tetouan	214	52	53
eg_Alexandria	214	52	52	mr_Nouakchott	215	53	53
eg_Aswan	214	52	52	om_Ad-Dakhiliyah	214	52	53
eg_Asyut	214	53	53	om_Ad-Dhahirah	215	40	53
eg_Beheira	214	52	52	om_Al-Batnah	214	52	53
eg_Beni-Suef	214	52	52	om_Ash-Sharqiyah	214	52	53
eg_Dakahlia	214	52	52	om_Dhofar	214	53	53
eg_Faiyum	214	52	53	om_Musandam	215	53	53
eg_Gharbia	214	52	53	om_Muscat	215	53	53
eg_Ismailia	214	52	53	ps_Gaza-Strip	214	52	52
eg_Kafr-el-Sheikh	215	52	53	ps_West-Bank	214	52	53
eg_Luxor	214	52	52	qa_Ar-Rayyan	215	52	53
eg_Minya	214	52	53	sa_Al-Madinah	214	52	52
eg_Monufia	215	52	53	sa_Al-Quassim	214	52	52
eg_North-Sinai	215	52	53	sa_Ar-Riyad	214	52	52
eg_Port-Said	214	52	52	sa_Ash-Sharqiyah	214	52	52
eg_Qena	214	52	53	sa_Asir	214	52	52
eg_Red-Sea	214	52	53	sa_Ha'il	214	52	52
eg_Sohag	214	52	52	sa_Jizan	214	52	53
eg_South-Sinai	214	52	53	sa_Makkah	214	52	52
eg_Suez	214	52	52	sa_Najran	214	52	53
iq_Al-Anbar	214	52	52	sa_Tabuk	214	52	52
iq_Al-Muthannia	215	53	53	sd_Khartoum	215	53	53
iq_An-Najaf	215	53	53	so_Banaadir	136	40	2
iq_Arbil	215	53	53	so_Woqooyi-Galbeed	36	9	53
iq_As-Sulaymaniyah	153	32	53	sy_Aleppo	215	52	23
iq_Babil	215	53	53	sy_As-Suwayda	214	53	53
iq_Baghdad	214	52	52	sy_Damascus-City	214	52	53
iq_Basra	214	52	53	sy_Hama	215	53	53
iq_Dihok	215	53	30	sy_Hims	214	53	53
iq_Karbala	215	53	53	sy_Lattakia	215	15	53
iq_Kirkuk	215	53	53	tn_Ariana	214	52	53
iq_Ninawa	215	53	53	tn_Bizerte	215	16	53
iq_Wasit	214	52	53	tn_Mahdia	215	52	53
jo_Aqaba	215	52	53	tn_Sfax	215	53	53
jo_Zarqa	214	52	52	ye_Aden	214	52	53
kw_Hawalli	214	52	53	ye_Ibb	215	53	53

Table 12: Distribution of the NADI 2021 DA data over provinces, by country, across our TRAIN, DEV, and TEST splits (Subtask 2.2).

Team	F_1	Acc	Precision	Recall
CairoSquad	22.38(1)	35.72(1)	31.56(3)	20.66(1)
CairoSquad	21.97(2)	34.90(2)	30.01(7)	20.15(2)
Phonemer	21.79(3)	32.46(6)	30.03(6)	19.95(4)
Phonemer	21.66(4)	31.70(7)	28.46(8)	20.01(3)
CS-UM6P	21.48(5)	33.74(4)	30.72(5)	19.70(5)
CS-UM6P	20.91(6)	33.84(3)	31.16(4)	19.09(6)
Phonemer	20.78(7)	32.96(5)	37.69(1)	18.42(8)
CS-UM6P	19.80(8)	31.68(8)	26.69(9)	19.04(7)
Speech Translation	14.87(9)	24.32(11)	18.95(14)	13.85(9)
Speech Translation	14.50(10)	24.06(12)	20.24(12)	13.24(10)
Speech Translation	14.48(11)	24.88(9)	22.88(10)	13.17(11)
NAYEL	12.99(12)	23.24(14)	15.09(15)	12.46(12)
NAYEL	11.84(13)	23.74(13)	19.42(13)	10.92(13)
NAYEL	10.29(14)	24.60(10)	33.11(2)	9.83(14)
NAYEL	10.13(15)	18.32(15)	11.31(16)	9.76(15)
NAYEL	7.73(16)	24.06(12)	21.07(11)	8.37(16)

Table 13: Full results for Subtask 1.1 (country-level MSA). The numbers in parentheses are the ranks. The table is sorted on the $macro\ F_1$ score, the official metric.

Team	F_1	Acc	Precision	Recall
CairoSquad	32.26(1)	51.66(1)	36.03(1)	31.09(1)
CairoSquad	31.04(2)	51.02(2)	35.01(2)	30.62(2)
CS-UM6P	30.64(3)	49.50(4)	32.91(6)	30.34(3)
CS-UM6P	30.14(4)	48.94(5)	33.20(4)	30.21(4)
CS-UM6P	29.08(5)	50.30(3)	34.99(3)	29.04(5)
IDC team	26.10(6)	42.70(9)	27.04(11)	25.88(6)
Phonemer	24.29(7)	44.14(6)	30.24(7)	23.70(7)
IDC team	24.00(8)	40.08(14)	25.57(15)	23.29(9)
Phonemer	23.56(9)	43.32(8)	28.05(10)	23.34(8)
Phonemer	22.72(10)	43.46(7)	28.13(9)	22.55(10)
Speech Translation	21.49(11)	40.54(10)	26.75(12)	20.36(12)
Arizona	21.37(12)	40.46(12)	26.32(13)	20.78(11)
Speech Translation	21.14(13)	40.32(13)	25.43(16)	20.16(14)
Speech Translation	21.09(14)	40.50(11)	26.29(14)	20.02(15)
Arizona	20.48(15)	40.04(15)	24.09(17)	20.22(13)
Arizona	19.85(16)	39.90(16)	22.89(18)	19.66(16)
AraDial_MJ	18.94(17)	35.94(22)	21.58(22)	18.28(17)
NAYEL	18.72(18)	37.16(20)	21.61(21)	18.12(18)
AraDial_MJ	18.66(19)	35.54(23)	21.45(23)	18.03(19)
AraDial_MJ	18.09(20)	37.22(19)	21.84(20)	17.55(20)
AraDial_MJ	18.06(21)	38.48(17)	22.70(19)	17.39(21)
IDC team	16.33(22)	29.82(25)	18.04(25)	16.10(22)
NAYEL	16.31(23)	38.08(18)	32.94(5)	15.91(23)
NAYEL	14.41(24)	32.78(24)	20.16(24)	14.11(24)
NAYEL	13.16(25)	36.96(21)	30.00(8)	13.83(25)
NAYEL	12.81(26)	26.48(26)	14.32(26)	12.66(26)
AraDial_MJ	4.34(27)	12.64(27)	4.33(27)	4.70(27)

Table 14: Full results for Subtask 1.2 (province-level MSA).

Team	F_1	Acc	Precision	Recall
CairoSquad	6.43(1)	6.66(1)	7.11(1)	6.71(1)
CairoSquad	5.81(2)	6.24(2)	6.26(2)	6.33(2)
Phonemer	5.49(3)	6.00(3)	6.17(3)	6.07(3)
Phonemer	5.43(4)	5.96(4)	6.12(4)	6.02(4)
CS-UM6P	5.35(5)	5.72(6)	5.71(7)	5.75(6)
Phonemer	5.30(6)	5.84(5)	5.97(6)	5.90(5)
CS-UM6P	5.12(7)	5.50(7)	5.24(8)	5.53(7)
CS-UM6P	4.72(8)	5.00(8)	5.97(5)	5.02(8)
NAYEL	3.51(9)	3.38(10)	4.09(9)	3.45(10)
NAYEL	3.47(10)	3.56(9)	3.53(10)	3.60(9)
NAYEL	3.16(11)	3.28(11)	3.38(12)	3.40(11)
NAYEL	3.15(12)	3.06(12)	3.43(11)	3.07(12)

Table 15: Full results for Subtask 2.1 (country-level DA).

Team	F_1	Acc	Precision	Recall
CairoSquad	8.60(1)	9.46(1)	9.07(1)	9.33(1)
CairoSquad	7.88(2)	8.78(2)	8.27(2)	8.66(2)
CS-UM6P	7.32(3)	7.92(4)	7.73(4)	7.95(3)
CS-UM6P	7.29(4)	8.04(3)	8.17(3)	7.90(4)
CS-UM6P	5.30(5)	6.90(5)	7.00(5)	6.82(5)
NAYEL	4.55(6)	4.80(10)	4.71(6)	4.55(10)
NAYEL	4.43(7)	4.88(9)	4.59(8)	4.62(9)
Phonemer	4.37(8)	5.32(6)	4.49(9)	5.19(6)
Phonemer	4.33(9)	5.26(7)	4.44(10)	5.14(7)
Phonemer	4.23(10)	5.20(8)	4.21(11)	5.08(8)
NAYEL	3.92(11)	4.12(12)	4.05(12)	4.00(12)
NAYEL	3.02(12)	3.10(13)	3.19(13)	3.19(13)
CS-UM6P	2.90(13)	4.20(11)	4.68(7)	4.13(11)

Table 16: Full results for Subtask 2.2 (province-level DA).