NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task

Muhammad Abdul-Mageed,$^{\lambda\xi}$ **Amr Keleg,**$^{\sigma}$ **AbdelRahim Elmadany,**$^{\lambda}$ **Chiyu Zhang,**$^{\lambda}$
Injy Hamed,$^{\xi\mu}$ **Walid Magdy,**$^{\sigma}$ **Houda Bouamor,**$^{\delta}$ **Nizar Habash**$^{\mu\xi}$
$^{\lambda}$The University of British Columbia $^{\xi}$MBZUAI
$^{\sigma}$The University of Edinburgh $^{\delta}$Carnegie Mellon University in Qatar
$^{\mu}$New York University Abu Dhabi
{muhammad.mageed@,a.elmadany@,chiyuzh@mail.}ubc.ca
{a.keleg@sms,wmagdy@inf}.ed.ac.uk
injy.hamed@mbzuai.ac.ae hbouamor@cmu.edu nizar.habash@nyu.edu

Abstract

We describe the findings of the fifth Nuanced Arabic Dialect Identification Shared Task (NADI 2024). NADI's objective is to help advance SoTA Arabic NLP by providing guidance, datasets, modeling opportunities, and standardized evaluation conditions that allow researchers to collaboratively compete on prespecified tasks. NADI 2024 targeted both dialect identification cast as a multi-label task (Subtask 1), identification of the Arabic level of dialectness (Subtask 2), and dialect-to-MSA machine translation (Subtask 3). A total of 51 unique teams registered for the shared task, of whom 12 teams have participated (with 76 valid submissions during the test phase). Among these, three teams participated in Subtask 1, three in Subtask 2, and eight in Subtask 3. The winning teams achieved 50.57 F_1 on Subtask 1, 0.1403 RMSE for Subtask 2, and 20.44 BLEU in Subtask 3, respectively. Results show that Arabic dialect processing tasks such as dialect identification and machine translation remain challenging. We describe the methods employed by the participating teams and briefly offer an outlook for NADI.

1 Introduction

Arabic is a collection of languages, language varieties, and dialects that can be classified into three main categories. *Classical Arabic* (CA) is the language of the Qur'an, old literature, and old scientific writing. CA has played a significant role in the spread of Islamic culture and continues to be crucial for scholarship in language and religious institutes. *Modern Standard Arabic* (MSA) is a simplified and 'modernized' descendent of CA that is employed in formal education and pan-Arab media as well as governmental circles in most Arab countries. *Dialectal Arabic* (DA) refers to the various forms of Arabic spoken in different parts of the Arab world in informal settings, TV shows, and everyday life. These three main categories of Arabic share lexica and grammatical structures to varying degrees, with some dialects being at one end of the continuum and CA at the other end.

The *Nuanced Arabic Dialect Identification (NADI)* shared task series was launched in 2020 as a venue for creating resources, affording modeling opportunities, and building a research community around the processing of dialectal Arabic. NADI 2024 is the fifth version of the shared task, hosted by the Second Arabic Natural Language Processing Conference (ArabicNLP 2024).

Dialect identification is the task of automatically detecting the source variety of a given text or speech segment. In previous versions of NADI, dialect identification was cast as a *single-label* classification task. That is, the input text can be assigned only one dialectal category (usually at the country level). Arabic dialects, however, can overlap significantly. This is especially the case for dialects spoken in geographically proximate areas where lexica, grammatical structures, as well as sound patterns are usually shared to notable degrees. While speech language identification models, e.g., (Kulkarni and Aldarmaki, 2023; Sullivan et al., 2023; Radhakrishnan et al., 2023), would typically have access to acoustic features that can help facilitate teasing apart these neighboring varieties, this is not the case for text-based systems as the input is intrinsically impoverished. Aligning with this observation, recent work by Keleg et al. (2023) analyzed the errors of a single-label dialect classification system and found that about 66% of these errors are not true errors. To accommodate these research findings and open up a space for further investigation of challenges faced by single-label models, we design a subtask in NADI 2024 (Subtask 1) as a *multi-label* classification task in which the given text can belong to more than a single Arabic dialect. Since Arabic dialects also overlap, sometimes significantly, with MSA, we also introduce a new subtask for estimating the

Proceedings of The Second Arabic Natural Language Processing Conference, pages 709–728
August 16, 2024

level of *dialectness* of text on a scale between zero and one (Subtask 2). Due to the challenges posed by dialects for machine translation (MT) systems, NADI 2024 continues to offer opportunities for advancing the translation of Arabic dialects through Subtask 3. We now review related literature.

2 Literature Review

2.1 Arabic Dialect Identification

Unlike CA and MSA, which have been studied rather extensively (Badawi, 1973; Brustad, 2000; Holes, 2004), DA received attention relatively recently. Most early efforts focused on creating regional or country-level dialect datasets (Diab et al., 2010; Harrat et al., 2014; Jarrar et al., 2016; Khalifa et al., 2016; Alsarsour et al., 2018; El-Haj, 2020) and region-level dialect identification models from text (Zaidan and Callison-Burch, 2011; Elfardy et al., 2014; Meftouh et al., 2015; Bouamor et al., 2018; Humayun et al., 2023). Several works also introduced larger Twitter datasets covering dialects from 10-21 countries (Mubarak and Darwish, 2014; Abdul-Mageed et al., 2018; Zaghouani and Charfi, 2018; Abdelali et al., 2021; Issa et al., 2021; Baimukan et al., 2022; Althobaiti, 2022), with some works such as Abdul-Mageed et al. (2020b) introducing models targeting country, province, and city levels. Several benchmarks, e.g., ORCA (Elmadany et al., 2023b) and DOLPHIN (Nagoudi et al., 2023), involve several dialectal datasets.

The NADI shared task continues to build on these previous efforts in offering datasets and affording modeling opportunities for identifying Arabic dialects (Abdul-Mageed et al., 2020a, 2021b, 2022, 2023). This year, we employ a multi-label setting to take into account (i) suggestions by Keleg and Magdy (2023) who highlight the limitations of addressing dialect identification as a single-label classification problem and propose defining it in a multi-label setting and (ii) issues of overlap in identical sentences across different dialects in the MADAR-26 test set (Bouamor et al., 2018) identified by Olsen et al. (2023).

2.2 Dialectness Level of Arabic

DA does not exist in isolation from MSA. DI identification on the sentence level takes a binary view in distinguishing between MSA and DA. Badawi (1973) identified five different levels of spoken Arabic in Egypt ranging from *Heritage Classical Arabic* to *Illiterate Colloquial Arabic*. He identified some phonological, morphological, lexical, and syntactic features of each of these levels. The same Arabic speaker employs different levels according to different sociolinguistic factors.

In an early work by Zaidan and Callison-Burch (2011), they asked crowdsourced annotators to identify the dialect and the level of dialectness of online comments to newspaper articles, forming the *AOC* dataset. They have only provided four short descriptive labels for the levels of dialectness (*No dialect, A bit of dialect, Mixed, Mostly Dialect*) and relied on the annotators' perceptions of these labels. Notably, their guidelines allow for assigning a high level of dialectness to a sentence having a highly dialectal word even if the remaining words are perceived to be less dialectal or in MSA, which is not the case for previous guidelines.

Recently, Keleg et al. (2023) recycled the discrete level of dialectness labels from the *AOC* dataset, transforming them into a continuous variable, Arabic Level of Dialectiness (*ALDi*), taking values between 0 and 1, to form the *AOC-ALDi* dataset. We decided to use the same operationalization of ALDi, while providing more elaborate guidelines to reduce the variability of the assigned ALDi levels for the same sentences.

2.3 Arabic Machine Translation

Several studies have addressed dialectal Arabic machine translation (MT), covering multiple dialects and translation directions (Abdul-Mageed et al., 2023). With regards to translation between Arabic variants, throughout the last year, numerous initiatives have focused on this task. Notably, the OSACT Dialect to MSA MT shared-task (Elneima et al., 2024) focused on translating Arabic dialects (Gulf, Egyptian, Levantine, Iraqi, and Maghrebi) into MSA. The proposed approaches mostly involved utilizing pretrained large language models (LLMs) with experimental designs incorporating zero-shot, few-shot, and fine-tuning setups, as well as data augmentation. In line with benchmarking LLMs, Alam et al. (2024) presented a benchmark with a focus on dialectal languages, including translating 25 Arabic dialects to MSA. Other researchers have focused on translation between specific Arabic dialects and MSA, including Egyptian (Faheem et al., 2024), Tunisian (Kchaou et al., 2023), and Algerian (Babaali et al., 2024). Despite the growing interest in this MT task, it received less attention compared to translation between Arabic and foreign languages, where several efforts involved evaluating and benchmarking LLMs (Kadaoui et al., 2023; Nagoudi et al., 2023; Banimelhem and Amayreh, 2023; Abdelali et al., 2024; Enis and Hopkins, 2024; Alkhawaja, 2024). To further advance research in MT across Arabic variants and build on our previous efforts, we include DA to MSA MT task again this year.

2.4 The History of NADI Shared Task

NADI 2020, the first NADI shared task (Abdul-Mageed et al., 2020a) involved two subtasks, one targeting country level (21 countries) and another focusing on province level (100 provinces), both exploiting Twitter data. NADI 2020 was the first shared task to exploit naturally occurring fine-grained dialectal text at the sub-country level. *NADI 2021*, the second version (Abdul-Mageed et al., 2021b) targeted the same 21 Arab countries and 100 corresponding provinces as NADI 2020, also using Twitter data. However, it improved upon the previous version by removing non-Arabic data and distinguishing between MSA and DA. It introduced four subtasks: MSA-country, DA-country, MSA-province, and DA-province.

NADI 2022 (Abdul-Mageed et al., 2022) continued the focus on studying Arabic dialects at the country level, but also included dialectal sentiment analysis with an objective to explore variation in socio-geographical regions that had not been extensively studied before. Finally, *NADI 2023*, the fourth edition (Abdul-Mageed et al., 2023), proposed new MT subtasks from four dialectal Arabic varieties to MSA, in two themes: open-track and closed-track, as well as a dialect identification subtask at the country level.

In this paper, we introduce the fifth edition of NADI by remodeling the dialect identification task into a multi-label classification task, introducing the ALDi estimation subtask, and continuing our MT subtask.

3 Task Description

NADI 2024 maintains the focus on processing dialectal Arabic. More concretely, we target both dialect identification (DID) and dialectal machine translation through three subtasks. **Subtask 1** focuses on DID, cast as a multi-label classification task, and **Subtask 2** aims at capturing the Arabic level of dialectness in texts (ALDi). As translation of Arabic dialects remains particularly challenging, we devote **Subtask 3** to dialect MT. We now describe each subtask in detail.

3.1 Subtask 1 – Multi-Label Dialect ID

In this subtask, we propose multi-label dialect identification (MDID) at the country level. The objective is to evaluate the feasibility of using single-label Arabic dialect identification datasets to train a multi-label system that can predict *all* dialects in which a given sentence is valid.

Tranining Data We provide participants with the *training* splits of following datasets: `MADAR-2018` (Bouamor et al., 2019) `NADI-2020-TWT`, `NADI-2021-TWT`, and `NADI-2023-TWT` (Abdul-Mageed et al., 2020a, 2021b, 2023).

Dev and Test Data We provide a new multi-label development set: `MDID-DEV`, henceforth `MDID-DEV` for brevity, as explained in §4. This dataset has 120 samples with manually assigned validity labels of eight different Arab countries: *Algeria, Egypt, Jordan, Palestine, Sudan, Syria, Tunisia,* and *Yemen*. Examples of those sentences are provided in Table 1. We do not restrict systems to these eight dialects. Hence we include two undisclosed dialects in our test data and ask participants to develop their models such that they can predict all valid dialects out of the 18 country-level ones from NADI 2023. The undisclosed dialects are *Iraq* and *Morocco*. Accordingly, the `MDID-TEST` set contains 1,000 samples covering nine dialects.[1]

Restrictions Subtask 1 operates under a ***closed-track*** policy where participants are allowed to *only* use the datasets we provide for system training.

3.2 Subtask 2 – ALDi Estimation

Keleg et al. (2023) define the Level of Dialectness as the extent by which a sentence diverges from the Standard Language. We use their operationalization to estimate the ALDi of sentences as a continuous value in the range [0, 1]; where 0 means MSA and 1 implies high divergence from MSA.

Training Data We provided the teams with AOC-ALDi dataset's training split (Keleg et al., 2023).

Dev and Test Data The dev and test sets collected for Subtask 1 were extended with a second layer of annotation for manual ALDi levels, forming `ALDi-DEV` and `ALDi-TEST` sets. The annotation process is outlined in §4.

[1] We also note that one of the Jordanian annotators did not complete the labeling process on time, and so we did not include the labels from Jordanian annotators in the test sets.

Sentence	GEO	Valid in	ALDi
اللهم انت ربي لا اله الا انت خلقتني وانا عبدك وانا علي عهدك ووعدك ما استطعت اعوذ بك من شر ما صنعت #اذكار_الصباح_و_المساء	**SA**	**DZ, EG, JO, PS, SD, SY, TN, YE**	0.00
#تريكه_في_كاس_العالم شاهد ماذا قال #الغندور علي المطالبه برجوع ابو تريكه للعب مع مصر في كاس العالم	**EG**	**DZ, EG, JO, PS, SD, SY, TN, YE**	0.15
لمن الحياه ترسل ليك رساله	**SD**	**PS, SD, YE**	0.58
وين يلعب هذا ما شفته	**AE**	**DZ, PS, YE**	0.64
الحمد لله الجو برد الايمات يلي فاتوا الواحد مغموم اتقول مكسد بين فرادي تينه شمام	**LY**	**DZ**	0.83
ايش دخل عارك ياحسن زميطه هذي العينات اللي يشتي له قيرعي	**YE**	**YE**	1.00

Table 1: Sample sentences from MDID-DEV with their geolocated country (**GEO**), valid dialect labels (Subtask 1), and ALDi scores (Subtask 2). **DZ**: Algeria, **EG**: Egypt, **JO**: Jordan, **LY**: Libya, **PS**: Palestine, **SA**: Saudi Arabia, **SD**: Sudan, **SY**: Syria, **TN**: Tunisia, **AE**: UAE, **YE**: Yemen.

Dialect	Source (Dialect)	Target (MSA)
Egyptian	أنا بس عايزك ترتبي نفسك من دلوقتي إنك تبقي زي ستات عيلة الدالي.	أنا أريدك أن ترتبي نفسك من الآن أن تكوني مثل سيدات عائلة الدالي.
	طب ايه بقى اللي انت قولته و خليت الريس يقلب عليك بالشكل ده؟	حسنا، ما هذا الذي قلته، وجعلت الرئيس ينقلب عليك بهذا الشكل؟
	بلا مسقعة بلا رز بلبن اقعد ساكت بلاش كلام فارغ.	لا مسقعة، ولا أرز بلبن. اسكت، وبلا كلام فارغ.
Emirati	أنا يا بنيتيه، حيث إنه ريولي عورتني من كثر اليلسة	أنا يا ابنتي، رجلي تؤلمني من الجلوس لمدة طويلة
	خله يولي، هذا كله مغربلنا و لا مبهدلنا، ما يشبع، و بعدين ليش خايفين ؟	دعه وشأنه، لقد ضرنا وعذبنا، لا يشبع، ومما أنتم خائفون؟
	يعني ألحين إنتي ما تعرفين شي عن سالفة الرضاعة ؟	معنى ذلك أنك لا تعلمين شيئًا عن قصة الرضاعة؟
Jordanian	طمعانين! الشاطر ينهب، و هو راخي لهم الحبل و ساكت!	إنهم طماعون! الكل يسرق، وهو متهاون ولا يتكلم.
	مبدئيًا رح أكتب لك هاظ العلاج مشان يخفف عليها، و رح نشوف شو بصير معانا	في البداية سأصف لك هذا الدواء لتخفيف الآلام، وسنرى النتيجة.
	ما بدنا اياك تتورط، و افهم من الدكتور ايش اللي بده اياه منك	لا نريد أن تتورط، وافهم من الطبيب ما الذي يريده منك.
Palestinian	يم محيتش تمرظ إلا هسا؟ ترا انته لا للسدي و لا للهدي، مثلك مثل مسعود الحذر	يعني لم تمرض إلا الآن؟ بالمناسبة أنت عديم الجدوى، مثلك مثل مسعود الحذر.
	بس انا بدي اياك تروحي عند ابو العبد تخليه يشوف ابو يزن يتنازل عن حقو	لكن أنا أريدك أن تذهبي عند أبي العبد وتجعليه يتنازل عن حقه.
	فش عندهم شرش هالذكاء، من وين جايبيته اه قوليلي؟ اكيد مني يعني طالعة علي!	ليس عندهم ذكاء، من أين أحضرته؟ أخبريني. من المؤكد أنه مني، يعني أنت مثلي!

Table 2: Random examples from MT-2024-DEV dataset spanning the four covered dialects.

Restrictions Subtask 2 operates under an ***open-track*** policy, allowing participants to train their systems on any additional datasets of their choice, provided that they explain the sources of the data and how it is used and that these additional training datasets are public at the time of submission.

3.3 Subtask 3 – Machine Translation

Similar to NADI 2023, this subtask is focused on machine translation from four Arabic dialects (i.e., *Egyptian*, *Emirati*, *Jordanian*, and *Palestinian*) to MSA at the sentence level. Unlike NADI 2023 where we had a close-track version of the MT task, we exclusively offer an open-track theme this year.

Training Data We do not provide direct training data. However, to facilitate Subtask 3, we point participants to the MADAR parallel dataset[2] (Bouamor et al., 2019) for system training and a monolingual dataset[3] that participants can manually translate and use for training.

Dev and Test Data For Subtask 3, we manually curated and translated completely new development and test data that were not used in NADI-2023. The development split, MT-2024-Dev, comprises 400 sentences, with 100 sentences representing each of the four dialects, while test split, MT-2024-Test, totals 2,000 sentences, with 500 from each dialect. Table 2 shows example sentences from MT-2024-Dev for each of the four countries. During the competition, we intentionally kept the source domain of these datasets undisclosed.[4]

Restrictions Subtask 3 operates under an ***open-track*** policy, allowing participants to train their systems on any additional datasets of their choice, provided these additional training datasets are public at the time of submission. For example, par-

[2]MADAR dataset can be acquired directly at MADAR Parallel Corpus. It comprises parallel sentences encompassing the dialects of 25 cities from the Arab world, as well as English, French, and MSA. Participants are permitted to use only the Train split of the MADAR parallel data for this subtask and must report on the Dev and Test sets we provide. The use of MADAR Dev and Test sets is not allowed in this subtask.

[3]The monolingual dataset is composed of the training splits of NADI 2020, NADI 2021, and NADI 2023, comprising 20k, 20k, and 18K tweets, repectively.

[4]Since we typically maintain a live leaderboard for post-competition evaluation, we not disclose the `MT-2024` data domain here either.

ticipants were permitted to manually create new parallel datasets. For transparency and the benefit of the wider community, we required participants to submit the datasets they created along with their test set submissions.

3.4 Evaluation Metrics

The official evaluation metric for Subtask 1 is the macro-averaged F_1 score. More specifically, we compute the F_1 score independently for each country in the evaluation dataset (eight for the development set and nine for the test set), then compute the average of these individual-country F_1 scores.[5] Additionally, we report system performance in terms of `Precision`, `Recall`, and `Accuracy` for submissions to Subtask 1. The metric for Subtask 2 is the `Root Mean Square Error (RMSE)`. For Subtask 3, we use the BLEU score (Papineni et al., 2002) as the official metric.[6] We calculate the overall BLEU score over all the samples (i.e., across all dialects) to rank the submitted systems for Subtask 3. We also present BLEU scores calculated separately for each of the four dialects (i.e., *Egyptian*, *Emirati*, *Jordanian*, and *Palestinian*).

3.5 Submission Rules

We allowed participant teams to submit up to *five* runs for each test set, for each of the three subtasks. For each team, only the submission with the highest score was retained. While the official results were exclusively based on a blind test set, we requested participants to include their results on Dev splits in their papers. To facilitate the evaluation of participant systems, we established a CodaLab competition for scoring each subtask (i.e., a total of three CodaLabs).[7] Similar to previous NADI editions, we are keeping the CodaLab for each subtask active even after the official competition has concluded. This is to encourage researchers interested in training models and assessing systems using the shared task's blind test sets. A nuance is that since subtasks 1 and 2 are new to NADI with limited training data available publicly, we share the individual labels of the development/test sets for these two subtasks.[8]

4 Evaluation Data for Subtasks 1 and 2

4.1 Samples Curation

We employ the same methods as in Abdul-Mageed et al. (2023) to collect geolocated tweets, then randomly sample 80 data points for the following ten countries from which we could recruit annotators:[9] *Algeria, Egypt, Iraq, Jordan, Morocco, Palestine, Sudan, Syria, Tunisia,* and *Yemen*, in addition to 80 data points from four other Arab countries *Lebanon, Libya, Saudi Arabia, UAE*. These additional samples are expected to be labeled as invalid in the dialects of the ten countries from which we recruited the annotators. Including them ensures the dataset's samples cover a wider range of dialects.

We use an in-house MSA/DA classification model (acc=89.1%, F_1 score=88.6) introduced in Abdul-Mageed et al. (2021a) to ensure that for each country's 80 geolocated samples, five are in MSA, and 75 are in DA. The overall dataset size for the shared task is 1,120 samples. Each annotator labeled the whole dataset. We remove user mentions, URLs, and emojis from the data, but retain the hashtags, before labeling the samples. We annotate our dataset on Upwork, incurring a total cost of $\$1,700$.

4.2 Annotation Process

For Subtask 1, we follow Keleg and Magdy (2023)'s proposal for building multi-label ADI datasets, mainly by asking native speakers of different Arabic dialects (on the country level) to check if each sentence is valid in one of the dialects spoken in their countries or not. For Subtask 2, we decided to provide more elaborate definitions for the different levels of dialectness than those in the guidelines of Zaidan and Callison-Burch (2011).[10] For each tweet, we ask two questions:

Q1) *Is it possible that the tweet is authored by someone who speaks one of your country's dialects?*

Options: (a) *Yes*, (b) *Not Sure/Maybe*, or (c) *No*.

Q2) *What is the Arabic Level of Dialectness (ALDi) of the tweet?* We define the following levels:

0. **Sound MSA**: Tweets written in fluent MSA.
1. **Formal Colloquial or Colloquial-influenced MSA**: Tweets written in a language close to MSA but using some colloquial expressions (lexemes/ morphemes).
2. **Natural/Ordinary Colloquial**: Tweets written in a colloquial language that is accepted

[5] Pariticiating teams submitted validity predictions for the 18 countries of the training sets. We plan to rerun the evaluation upon collecting labels for more country-level dialects.

[6] We plan to consider other MT evaluation metrics in future versions of the shared task.

[7] Our CodaLabs are available at the following links: `Subtask 1`, `Subtask 2`, and `Subtask 3`.

[8] We will be glad to consider collaborations on extending the datasets for all our subtasks to other country-level dialects.

[9] As per §3.1, labels from Jordan are only in the Dev set.

[10] Refer to §A of the Appendix for further details.

Country	Subtask 1			Subtask 2
	Fleiss κ	N valid	N ¬valid	Krip. α
Algeria	0.56	333 (205)	787 (666)	0.66
Morocco	0.62	230 (152)	890 (784)	0.74
Tunisia	0.67	189 (129)	931 (879)	0.75
Egypt	0.69	353 (257)	767 (682)	0.82
Sudan	0.67	393 (283)	727 (619)	0.66
Palestine	0.59	375 (245)	745 (587)	0.68
Syria	0.54	475 (305)	645 (543)	0.79
Iraq	0.61	271 (171)	849 (738)	0.73
Yemen	0.52	454 (291)	666 (477)	0.50

Table 3: Interannotator agreement scores – Fleiss' Kappa (κ) for Subtask 1 and Krippendorff's Alpha-interval method (α) for Subtask 2 – for the full dataset. We also report the number of valid, not valid sentences out of the 1,120 according to majority voting, while showing the number of sentences with complete agreement (between brackets). **Note:** The country-level Krippendorff's Alpha scores are computed for their respective country's valid samples.

and understood by all members of society of all ages and social/educational levels.

3. **Informal (or Vulgar) Colloquial**: Tweets written in a colloquial language having expressions that are not accepted or understood by all members of society. It does not have to be vulgar or weak.

We believe an Arabic speaker identifying the ALDi of a tweet needs to be familiar with the dialect in which the tweet is written. For this reason, an annotator is allowed to identify ALDi only if their answer to the first question (validity of tweet in one of their country-level dialects) is either *Yes* or *Not Sure/Maybe*.[11]

For each of the ten specified countries, we managed to recruit three native speakers through Upwork to label all the 1,120 tweets. Before inviting the annotators to the main task, we ask them to complete an onboarding task to get them acquainted with our objectives and clarify any misunderstandings. Afterward, the main task annotation process is split into five batches, 224 samples each, where feedback is provided to the annotators after each batch to ensure high quality. Annotators were paid \$8 after successfully completing each of the six tasks in addition to a bonus value between \$8 and \$12 after completing the whole process. After accounting for the platform fees, annotating the dataset cost about \$1,700.

Interannotator Agreement (IAA) Scores We use Fleiss' Kappa (κ) (Fleiss, 1971), and Kripendorff's Alpha (α) (Krippendorff, 2004) for measuring the IAA. The country-level scores in Table 3 indicate moderate to substantial agreement between the annotators for both subtasks. Moreover, there is not a noticeable variation among the scores across the countries, except for the α score for the Yemen annotators which is slightly lower than those of the other countries. We noticed the the Yemeni annotators had different perceptions of what counts as *Sound MSA (Level 0)* and what counts as *Natural/Ordinary Colloquial (Level 2)*.

[11] See §4.2 for information about the annotation process.

4.3 Label Aggregation Techniques

Subtask 1 A sentence is considered valid in a country-level dialect if among the three annotators from the respective countries: a) one of them answered *Yes*, and b) another answered *Yes* or *Maybe*. On average, the same-country annotators fully agreed on the validity of more than 66% of the valid samples, and the invalidity of more than 85% of the invalid samples, as per Table 3.

Subtask 2 For each sentence, the ordinal ALDi levels assigned by the annotators from the different countries are aggregated into a single numeric value $\in [0, 1]$. Discrete ALDi levels (0, 1, 2, 3) are transformed into the following numeric values $(0, \frac{1}{3}, \frac{2}{3}, 1)$. The mean of these numeric values is used as the overall ALDi score for the sentence.

As mentioned in §4.2, annotators only assigned ALDi levels to sentences they rated as valid in their country-level dialect. Consequently, the number of ALDi annotations per sentence can range from 0 to 3*N where N is the number of countries from which annotators are recruited. If a sentence is deemed invalid according to the majority vote label (Subtask 1) for a country-level dialect, we discard the resective ALDi annotations (if any) assigned by the annotators' of this country.

4.4 Formation of Development/Test Sets

We used 120 samples from the first batch as the development sets (`MDID-DEV`, `ALDi-DEV`) shared with the participating teams. The first batch's remaining samples and the samples of the 4 succeeding batches form the test sets (`MDID-TEST`, `ALDi-TEST`). For `ALDi-DEV` and `ALDi-TEST`, samples that are not valid in the considered dialects of the corresponding set have no assigned ALDi scores, and thus are not released as part of the dataset.

Analysis of the Development Sets Figure 1 shows that 13 samples are labeled as invalid in

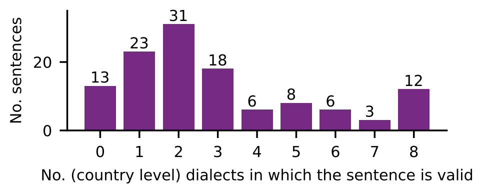


Figure 1: Distribution of the number of valid (country level) dialects out of 8 countries.

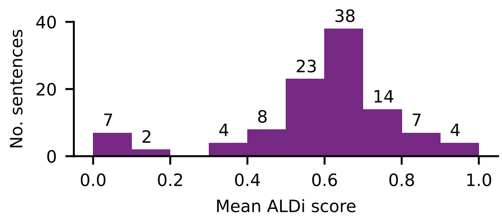


Figure 2: ALDi-DEV's scores.

all of the 8 considered countries in MDID-DEV, and 23 samples are valid in only one dialect. We also report that 84 samples are valid in 2 or more country-level dialects (i.e., 70% of the samples). This percentage is only expected to increase when validity annotations from other country-level Arabic dialects are collected.

The aggregated ALDi scores have a multimodal distribution as per Figure 2. The first mode is related to the automatically identified MSA samples in the development set (8 in total). All of these samples are assigned ALDi scores <0.2, and are judged as valid in all the considered country-level dialects. Conversely, the ALDi scores for the automatically identified DA samples are distributed around a score of 0.66 (the numeric value corresponding to Level 2 (*Natural/Ordinary Colloquial*).

5 Shared Task Teams & Results

5.1 Participating Teams

We received a total of 51 unique team registrations. At the testing phase, a total of 76 valid entries were submitted by 12 unique teams. The breakdown across the subtasks is as follows: *ten* submissions for Subtask 1 from *three* teams, *seven* submissions for Subtask 2 from *three* teams, and 16 submissions for Subtask 3 from *eight* teams. Table 4 lists the 12 teams. We received eight description papers, all of which were accepted for publication.

5.2 Baselines

We developed baseline (BL) models for each subtask for comparison against the teams' systems, as described below. These models were not shared with participating teams during the competition.

Team	Affiliation	Tasks
AlexUNLP-STM (Sakr et al., 2024)	Alexandria University, Egypt	2
Alson (AlMusallam and Ahma, 2024)	-, KSA	3
Arabic Train (Demidova et al., 2024)	MBZUAI, UAE	3
ASOS (Nacar et al., 2024)	Prince Sultan University, KSA	2, 3
CUFE (Ibrahim, 2024)	Cairo University, Egypt	2, 3
dzNLP (Lichouri et al., 2024)	USTHB, Algeria	1
Elyadata (Karoui et al., 2024)	Elyadata, Tunisia	1
MBZUAI BADG	MBZUAI, UAE	3
MBZUAI BLEU	MBZUAI, UAE	3
NLP_DI (Kanjirangat et al., 2024)	Dalle Molle Ins. of A.I., Switzerland	1
Shaheen	MBZUAI, UAE	3
VBNN	MBZUAI, UAE	3

Table 4: List of teams that participated in NADI-2024 shared task. Teams with accepted papers are cited.

Rank	System	Macro-average			
		Accuracy (↑)	Precision (↑)	Recall (↑)	F_1 score (↑)
1	**Elyadata**	$67.50_{\pm 3.7}$	$46.48_{\pm 10.1}$	$\mathbf{57.09}_{\pm 5.1}$	$\mathbf{50.57}_{\pm 7.1}$
BL I	Top 90%	$73.40_{\pm 6.1}$	$60.67_{\pm 14.5}$	$39.22_{\pm 14.6}$	$45.09_{\pm 11.3}$
2	**NLP_DI**	$71.88_{\pm 5.6}$	$53.64_{\pm 10.2}$	$37.42_{\pm 11.0}$	$43.27_{\pm 9.4}$
BL II	Random	$50.14_{\pm 1.6}$	$30.43_{\pm 8.8}$	$50.15_{\pm 2.1}$	$37.15_{\pm 7.2}$
BL III	Top 1	$\mathbf{73.42}_{\pm 7.6}$	$\mathbf{76.82}_{\pm 10.6}$	$17.77_{\pm 10.8}$	$27.30_{\pm 12.6}$
3	**dzNlp**	$71.38_{\pm 7.2}$	$63.22_{\pm 10.7}$	$12.87_{\pm 3.8}$	$20.98_{\pm 5.2}$

Table 5: Systems' performance on the test set of Subtask 1. See Appendix §C for a more detailed analysis.

Subtask 1 Baselines We use the softmax of a fine-tuned single-label DI system's logits to develop two baselines.[12] The first predicts the most probable labels such that their cumulative probability is $> 90\%$. The second assumes the sentence is only valid in the most probable prediction. Lastly, we implement a Random baseline that generates random binary predictions for the validity of the sentences in the considered dialects.

Subtask 2 Baselines We first use the Sentence ALDi model developed by Keleg et al. (2023) as our supervised baseline. The second baseline is based on the distribution of the ALDi scores for the development set (Figure 2), where we implement a model that generates a constant score of 0.67 for all the sentences. In the third baseline, we use a Random ALDi generator ($\in [0, 1]$).

Subtask 3 Baselines We extract parallel dialectal-to-MSA data of four dialects from MADAR-18 for training MT baselines for Subtask 3. We then fine-tune three baselines on the extracted data. These are *AraT5*$_{v2}$ (Elmadany et al., 2023a; Nagoudi et al., 2022), *mT5* (Xue et al., 2021), and *AraBART* (Kamal Eddine et al., 2022).

[12]The fine-tuned baseline model can be accessed through huggingface.co/AMR-KELEG/NADI2024-baseline.

Rank	System	RMSE (↓)
1	**ASOS**	**0.1403**
2	**AlexUNLP-STM**	0.1406
3	**CUFE**	0.2001
BL I	Sentence ALDi	0.2178
BL II	Constant (0.67)	0.2361
BL III	Random	0.3521

Table 6: Systems performance on Subtask 2 test set.

Rank	System	BLEU(↑)				
		Overall	Egy.	Emi.	Jor.	Pal.
1	**Arabic Train**	**20.44**	16.57	**23.38**	**21.37**	**20.62**
2	**Alson**	17.46	**16.76**	12.53	20.94	18.43
3	**ASOS**	17.13	14.82	19.39	15.80	18.38
4	**CUFE**	16.09	14.86	17.35	15.98	16.20
5	**MBZUAI BLEU**	10.54	8.53	7.61	15.72	11.08
6	**VBNN**	9.24	8.62	6.30	11.79	10.54
BL I	AraT5_{v2}	6.87	9.38	4.61	4.90	8.13
7	**MBZUAI BADG**	2.78	3.03	1.74	3.91	2.48
BL II	mT5	2.81	3.08	2.23	3.11	2.95
BL III	AraBART	0.87	0.77	0.81	1.11	0.88
8	**Shaheen**	0.00	0.00	0.00	0.00	0.00

Table 7: Performance of the systems on the test set of Subtask 3. Results are sorted by overall BLEU score.

5.3 Shared Task Results

Subtask 1 ***Elyadata*** came first with a macro-averaged F_1 score of 50.57%, being the only team to beat the *Top 90% baseline* model as per Table 5.

Subtask 2 ***ASOS***, the top-performing team, achieved the lowest RMSE of 0.1403, while ***AlexUNLP-STM*** achieved a similar RMSE of 0.1406 coming second in the ranking. As shown in Table 6, all the teams managed to improve over our baselines, including systems trained on the AOC-ALDi dataset which has a different nature (comments on news not tweets) and was annotated based on less nuanced guidelines than ours.

Subtask 3 Table 7 shows the leaderboard of Subtask 3. ***ArabicTrain*** won first place, achieving a BLEU score of 20.44. We observe that *six* teams outperform our best baseline on this subtask.

5.4 General Description of Submitted Systems

A summary of approaches employed by the various teams is provided in Table 8. We briefly describe the top systems for each subtask here.

Subtask 1 The winning team, ***Elyadata***, extracted dialectal vocabularies from the training data, and used them to augment the labels of the single-label training dataset. They then used a max pooling layer to merge the predictions of a MARBERT-based ensemble model forming an array of logit predictions. Lastly, they optimized a threshold using the development set to convert the logits into multi-label predictions.

Subtask 2 ***ASOS*** fine-tuned a regression head of multiple layers on top of MARBERT's [CLS] embedding. ***AlexUNLP-STM*** used the median of an ensemble of regression heads with sigmoid activation on top of AraBERT, trained to minimize contrastive and RMSE losses. Noticeably, their model's performance dropped when non-Arabic letters were discarded. We observed that code-switching affected the annotators' ALDi judgments differently, which is in-line with the team's justification for the performance drop.

Subtask 3 The winning team, ***Arabic Train***, utilized samples from MADAR (Bouamor et al., 2019) training set as the one-shot example to prompt LLM Jais (Sengupta et al., 2023) for translating Arabic dialects to MSA. Team ***Alson*** exploited ChatGPT to generate parallel data for translating Jordanian and Palestinian dialects to MSA and then fine-tuned AraT5 with generated samples and MADAR dataset.

6 Discussion

Precision of Geolocated Labels Although geolocation can alleviate the need for manually annotating the samples (Abdul-Mageed et al., 2021b), it can be error-prone (Abdul-Mageed et al., 2020a; Abdelali et al., 2021). For the 1,050 DA samples of the development and test set, we can estimate the precision of the geolocated labels by comparing them to the manual validity labels as demonstrated in Figure 3. Based on this method, we find that the precision of the geolocated labels could be as high as 94.6% (71 out of 75 samples) for Egypt, and as low as 49.3% (37/75) for Tunisia.

Impact of Named Entities ADI models, trained on single-label data, can make spurious connections between named entities (e.g.: specific locations) and country-level labels (Abdul-Mageed et al., 2020b). In `NADI-2021-TWT` for example, 52 samples out of the 66 mentioning لبنان (Lebanon) are geolocated to and labeled as *Lebanon*. Such spurious connections might be the reason why the following n-grams عراق, لبنان, تونس, يمن are among the most discriminative for the dialects of *Iraq*, *Lebanon*, *Tunisia*, and *Yemen*, respectively (AAlAbdulsalam, 2022). Manual annotation can alleviate this limitation.

Task	Team	Metric	Features		Techniques								
			N-gram	TFIDF	C-ML	NNs	PLM	LLM	Ensemble	Adapter	Post-Poc.	Cont. L.	D-Aug.
1	Elyadata	50.57				✓	✓		✓				
	NLP_DI	43.27				✓	✓		✓		✓		
	dzNLP	20.98	✓	✓	✓				✓				
2	ASOS	0.1403				✓	✓		✓				✓
	AlexUNLP-STM	0.1406				✓	✓		✓			✓	
3	Arabic Train	20.44						✓		✓			✓
	Alson	17.46					✓	✓					✓
	ASOS	17.13					✓	✓					✓
	CUFE	16.09					✓	✓		✓			

Table 8: Summary of approaches used by participating teams NADI 2024 shared task. Teams are sorted by their performance on the official metric of each subtask. *C-ML* (Classifcal ML) indicates any non-neural machine learning methods such as naive Bayes and support vector machines. The term *NNs* refers to any model based on neural networks (e.g. RNN, CNN, and Transformer) trained from scratch. *PLM* refers to neural networks pretrained with unlabeled data such as MARBERT and has less than 1B parameters. *LLM* refers to neural networks containing more than 1B parameters. Approaches also included contrastive loss (*Cont. L.*) and data augmentation (*D-Aug.*)

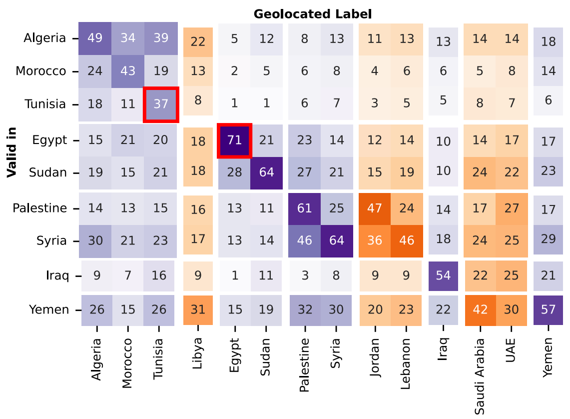


Figure 3: The number of DA samples valid in the annotators' country-level dialects (rows) across the 14 countries to which the samples are geolocated (columns). Each row represents the distribution of the geolocated labels for the sentences valid in the row's country-level dialect. Orange columns indicate the countries not represented by the annotators. The max cell value is 75.

6.1 Lessons Learned

We share our reflections on the creation of evaluation datasets for Subtasks 1 and 2 as per §4.

Subtask 1 Complexity Previous research asking Arabic speakers to check the validity of sentences in their native dialects (See Table A1) reported moderate to high agreement between the annotators (only two per country) for *most* of the considered regional-level and country-level dialects. Unlike previous works, we recruited three annotators per country and asked them to judge all samples, rather than those geolocated to their own respective countries. Therefore, our annotation task is possibly harder than previous ones, which is reflected in the IAA scores in Table 3.

Subtask 1 Labels From a task design perspective, we observe that the frequency of usage of the *Maybe (Not sure)* label varies across annotators. For this reason, including this particular label (rather than using a binary *Valid/Not Valid* setup), needs to be further investigated to understand its implications on the aggregated validity labels.

Annotation Quality Monitoring Two authors who are speakers of Egyptian Arabic were responsible for monitoring the quality of the annotations, providing feedback, and marking the samples with high disagreement for reannotation. We believe that having dialect leads who are native speakers of the different Arabic dialects would allow for better monitoring of the annotation process. We hope that our shared task will inspire future collaborative research to extend the labels of our evaluation dataset to include more country-level dialects.

7 Conclusion

This year, we organized NADI 2024, the fifth edition of the shared task, having three subtasks: multi-label dialect identification (MDID), Arabic level of dialectness (ALDi) estimation, and DA-to-MSA machine translation. We had 51 registered teams, out of which 12 submitted their systems' predictions with eight accepted system description papers. The results indicate that there is still room for improvement across the various tasks. In the future, we intend to cover more Arabic dialects in NADI and propose novel ways of modeling that involve the use of large language models.

Limitations

Our work has a number of limitations, as follows:

- This edition of NADI focused on only 10 country-level dialects for Subtasks 1 and 2. This is due to challenges with recruiting annotators as well as the lack of high-quality datasets for countries such as *Comoros*, *Djibouti*, *Mauritania*, and *Somalia*.

- NADI continues to use short texts for the Arabic dialects. That is, due to the shortage of dialectal data from other sources, we depend on posts from Twitter. Although these data have thus far empowered the development of effective dialect identification models, it is desirable to afford data from other domains that have longer texts. This will allow the development of more widely applicable models.

- The label aggregation techniques (See §4.3) used for the evaluation sets of Subtask 1, 2 attempts to reduce the impact of the few inevitable inaccurate annotations. However, they could also inhibit interannotator disagreement that is caused by having different perceptions (i.e., what sentences are valid in their country-level dialects, or what the level of dialectness of sentences are) (Ovesdotter Alm, 2011; Mostafazadeh Davani et al., 2022).

- Our machine translation subtask focuses only on four dialects without offering a training dataset. Modern MT systems need much larger data to perform well. Again, in spite of our best efforts, parallel datasets involving dialects remain limited.

- Due to limited resources, we were able to provide only a single reference annotation for Subtask 3 test samples. We acknowledge that machine translation requires multiple evaluation references to ensure a more reliable assessment.

- We acknowledge that the BLEU score for evaluating machine translation output has its limitations (Popović, 2017; Kocmi et al., 2021; Rei et al., 2022). We expect that using more diverse metrics, such as ChrF (Popović, 2017) and COMET (Rei et al., 2022), can enhance the reliability of evaluation results.

Ethical Considerations

The NADI-2024 Subtask 1, 2 datasets are sourced from the public domain (i.e., X former Twitter), with user personal information and identity carefully concealed. Similarly, the NADI-2024 Subtask 3 dataset is manually created. Again, we take meticulous measures to remove user identities and personal information from our datasets. As a result, we have minimal concerns about the retrieval of personal information from our data. However, it is crucial to acknowledge that the datasets we collect to construct NADI-2024 Subtask 1, 2 may contain potentially harmful content. Additionally, during model evaluation, there is a possibility of exposure to biases that could unintentionally generate problematic content.

Finally, we note that the annotation process we followed for creating the evaluation dataset of the first two subtasks was approved by the research ethics committee of the University of Edinburgh, School of Informatics with reference number 839548.

Acknowledgments

Muhammad Abdul-Mageed acknowledges support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[13] and UBC ARC-Sockeye.

Amr Keleg's work is supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics.

References

Abdulrahman Khalifa AAlAbdulsalam. 2022. SQU-CS @ NADI 2022: Dialectal Arabic identification using one-vs-one classification with TF-IDF weights computed on character n-grams. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 436–441, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel

[13] https://alliancecan.ca

Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, Majd Hawasly, Nadir Durrani, and Firoj Alam. 2024. LAraBench: Benchmarking Arabic AI with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520, St. Julian's, Malta. Association for Computational Linguistics.

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. 2018. You tweet what you speak: A city-level dataset of Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021a. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020a. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021b. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The third nuanced Arabic dialect identification shared task. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020b. Toward 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, Online. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024. CODET: A benchmark for contrastive dialectal evaluation of machine translation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1790–1859, St. Julian's, Malta. Association for Computational Linguistics.

Linda Alkhawaja. 2024. Unveiling the new frontier: ChatGPT-3 powered translation for Arabic-English language pairs. *Theory and Practice in Language Studies*, 14(2):347–357.

Manan AlMusallam and Samar Ahma. 2024. Alson at NADI 2024 shared task: Alson - A fine-tuned model for Arabic Dialect Translation. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

Israa Alsarsour, Esraa Mohamed, Reem Suwaileh, and Tamer Elsayed. 2018. DART: A large dataset of dialectal Arabic tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Maha J. Althobaiti. 2022. Creation of annotated country-level dialectal Arabic resources: An unsupervised approach. *Natural Language Engineering*, 28(5):607–648.

Baligh Babaali, Mohammed Salem, and Nawaf R Alharbe. 2024. Breaking language barriers with ChatGPT: enhancing low-resource machine translation between Algerian Arabic and MSA. *International Journal of Information Technology*, pages 1–10.

As-Said Muhámmad Badawi. 1973. *Mustawayat al-arabiyya al-muasira fi Misr*. Dar al-maarif.

Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. Hierarchical aggregation of dialectal data for Arabic dialect identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4586–4596, Marseille, France. European Language Resources Association.

Omar Banimelhem and Wlla Amayreh. 2023. Is ChatGPT a good English to Arabic machine translation tool? In *2023 14th International Conference on Information and Communication Systems (ICICS)*, pages 1–6.

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 Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

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

Anastasiia Demidova, Hanin Atwany, Nour Rabih, and Sanad Shaban. 2024. Arabic Train at NADI 2024 shared task: LLMs' Ability to Translate Arabic Dialects into Modern Standard Arabic. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

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 Twelfth Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France. European Language Resources Association.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2014. AIDA: Identifying code switching in informal Arabic text. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 94–101, Doha, Qatar. Association for Computational Linguistics.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023a. Octopus: A multitask model and toolkit for Arabic natural language generation. In *Proceedings of ArabicNLP 2023*, pages 232–243, Singapore (Hybrid). Association for Computational Linguistics.

AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023b. ORCA: A challenging benchmark for Arabic language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586, Toronto, Canada. Association for Computational Linguistics.

Ashraf Hatim Elneima, AhmedElmogtaba Abdelmoniem Ali Abdelaziz, and Kareem Darwish. 2024. OSACT6 dialect to MSA translation shared task overview. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 93–97, Torino, Italia. ELRA and ICCL.

Maxim Enis and Mark Hopkins. 2024. From LLM to NMT: Advancing low-resource machine translation with Claude. *arXiv preprint arXiv:2404.13813*.

Mohamed Atta Faheem, Khaled Tawfik Wassif, Hanaa Bayomi, and Sherif Mahdy Abdou. 2024. Improving neural machine translation for low resource languages through non-parallel corpora: a case study of Egyptian dialect to modern standard Arabic translation. *Scientific Reports*, 14(1):2265.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Salima Harrat, Karima Meftouh, Mourad Abbas, and Kamel Smaïli. 2014. Building resources for algerian arabic dialects. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 2123–2127. ISCA.

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

Mohammad Ali Humayun, Hayati Yassin, Junaid Shuja, Abdullah Alourani, and Pg Emeroylariffion Abas. 2023. A transformer fine-tuning strategy for text dialect identification. *Neural Computing and Applications*, 35(8):6115–6124.

Michael Ibrahim. 2024. CUFE at NADI 2024 shared task: Fine-Tuning Llama-3 To Translate From Arabic Dialects To Modern Standard Arabic. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

Elsayed Issa, Mohammed AlShakhori1, Reda Al-Bahrani, and Gus Hahn-Powell. 2021. Country-level Arabic dialect identification using RNNs with and without linguistic features. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 276–281, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

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.

Karima Kadaoui, Samar Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. TARJAMAT: Evaluation of bard and ChatGPT on machine translation of ten Arabic varieties. In *Proceedings of ArabicNLP 2023*, pages 52–75, Singapore (Hybrid). Association for Computational Linguistics.

Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022. AraBART: a pretrained Arabic sequence-to-sequence model for abstractive summarization. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 31–42, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2024. NLP_DI at NADI 2024 shared task: Multi-label Arabic Dialect Classifications with an Unsupervised Cross-Encoder. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

Amira Karoui, Rami Kammoun, Farah Gharbi, Imen Laouirine, and Fethi Bougares. 2024. ELYADATA at NADI 2024 shared task: Arabic Dialect Identification with Similarity-Induced Mono-to-Multi Label Transformation. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

Saméh Kchaou, Rahma Boujelbane, and Lamia Hadrich. 2023. Hybrid pipeline for building Arabic Tunisian dialect-standard Arabic neural machine translation model from scratch. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(3):1–21.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic level of dialectness of text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

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).

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Klaus Krippendorff. 2004. *Content Analysis, an Introduction to Its Methodology*, second edition. Thousand Oaks, CA: Sage Publications.

Ajinkya Kulkarni and Hanan Aldarmaki. 2023. Yet another model for Arabic dialect identification. In *Proceedings of ArabicNLP 2023*, pages 435–440, Singapore (Hybrid). Association for Computational Linguistics.

Mohamed Lichouri, Khaled Lounnas, Boualem Nadjib Zahaf, and Mehdi Ayoub Rabiai. 2024. dzNLP at NADI 2024 Shared Task: Multi-Classifier Ensemble with Weighted Voting and TF-IDF Features. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

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 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34, Shanghai, China.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Hamdy Mubarak and Kareem Darwish. 2014. Using Twitter to collect a multi-dialectal corpus of Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 1–7, Doha, Qatar. Association for Computational Linguistics.

Omer Nacar, Serry Sibaee, Abdullah I. Alharbi, Lahouari Ghouti, and Anis Koubaa. 2024. ASOS at NADI 2024 shared task: Bridging Dialectness Estimation and MSA Machine Translation for Arabic Language Enhancement. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Ahmed El-Shangiti, and Muhammad Abdul-Mageed. 2023. Dolphin: A challenging and diverse benchmark for Arabic NLG. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1404–1422, Singapore. Association for Computational Linguistics.

Helene Olsen, Samia Touileb, and Erik Velldal. 2023. Arabic dialect identification: An in-depth error analysis on the MADAR parallel corpus. In *Proceedings of ArabicNLP 2023*, pages 370–384, Singapore (Hybrid). Association for Computational Linguistics.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Narsis A Kiani, David Gomez-Cabrero, and Jesper N Tegner. 2023. A parameter-efficient learning approach to Arabic dialect identification with pre-trained general-purpose speech model. *arXiv preprint arXiv:2305.11244*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Abdelrahman Sakr, Marwan Torki, and Nagwa El-Makky. 2024. AlexUNLP-STM at NADI 2024 shared task: Quantifying the Arabic Dialect Spectrum with Contrastive Learning, Weighted Sampling, and BERT-based Regression Ensemble. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models.

Peter Sullivan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. On the robustness of Arabic speech dialect identification. *arXiv preprint arXiv:2306.03789*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wajdi Zaghouani and Anis Charfi. 2018. Arap-tweet: A large multi-dialect Twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Appendices

A Design of the Annotation Guidelines

In this year's edition of NADI, we introduced two new subtasks, MDID, and ALDi Estimation. As explained in §4, we annotated 1,120 tweets for both subtasks to build the development and test splits.

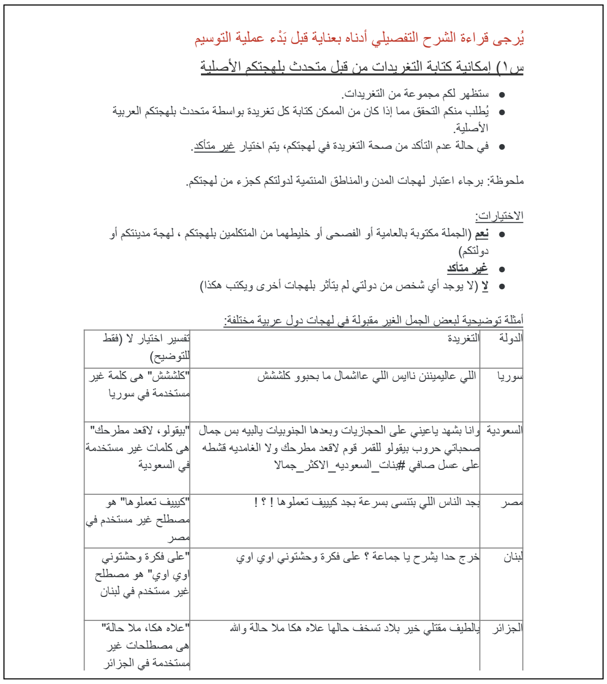
Subtask 1 There has been multiple attempts asking annotators to check if sentences are written in their Arabic dialect. This was done mainly to validate dialect labels that are automatically assigned using geolocating methods/distinctive dialectal cues (Alsarsour et al., 2018; Abdelali et al., 2021; Althobaiti, 2022) or to perform error analysis for the predictions of DI systems (Keleg and Magdy, 2023). Arabic speakers have different perceptions of their country-level dialects that depend on their backgrounds and exposure to different speaking communities. Such differences could impact their understanding of the validity of sentences in their country-level dialects. Moreover, previous wordings used to check the validity of these sentences shown in Table A1 were used to validate the labels for sentences that are more probable to be in the annotator's native dialect (i.e., not in another dialect nor in MSA).

Conversely, we asked the annotators to label 1120, uniformly representing 14 country-level dialects. Moreover, we are interested in checking the validity of the sentences in any of the dialects spoken in the annotator's country of origin, and not just in their native Arabic dialect. Consequently, we used the wording mentioned in §4.2, in addition to providing some examples as shown in Figure A1.

Wording
(Alsarsour et al., 2018) Asked annotators to label each tweet as either in: (their native dialect, MSA, or other).
(Abdelali et al., 2021) *Is this tweet consistent with the dialect spoken in your country?* (Yes, No).
(Althobaiti, 2022) *Is this sentence written in* `Dialect` *dialect?* (Yes, No); `Dialect` is the demonymic form of the annotator's country.
(Keleg and Magdy, 2023) *Is this sentence valid in your dialect?* (Yes, Not Sure, No).

Table A1: Previous wordings for checking the validity of sentences in an Arabic dialect.

Subtask 2 We follow Zaidan and Callison-Burch (2011)'s setup in which they asked the annotators to assign a discrete ALDi level to each sentence as shown in Figure A2. In their setup, we noticed that Level 3 (*Mostly Dialectal* (معظمها عامية)) could not fully separate between sentences having a word perceived as highly dialectal, and sentences having a majority of dialectal words that are not perceived as highly dialectal on the word level. Therefore, we provide descriptive labels to the 4 levels, and short descriptions for the sentences of each label (See 4.2 for the English translation of these labels/descriptions). Moreover, we use two examples to further explain the concept of ALDi on the word and sentence level, as per Figure A3. This would allow for better separation between the different ALDi levels, and higher IAA scores.



يُرجى قراءة الشرح التفصيلي أدناه بعناية قبل بَدْء عملية التوسيم

س١) إمكانية كتابة التغريدات من قبل متحدث بلهجتكم الأصلية

- ستظهر لكم مجموعة من التغريدات.
- يُطلب منكم التحقق مما إذا كان من الممكن كتابة كل تغريدة بواسطة متحدث بلهجتكم العربية الأصلية.
- في حالة عدم التأكد من صحة التغريدة في لهجتكم، يتم اختيار غير متأكد.

ملحوظة: برجاء اعتبار لهجات المدن والمناطق المنتمية لدولتكم كجزء من لهجتكم.

الاختيارات:

- **نعم** (الجملة مكتوبة بالعامية أو الفصحى أو خليطهما من المتكلمين بلهجتكم ، لهجة مدينتكم أو دولتكم)
- **غير متأكد**
- **لا** (لا يوجد أي شخص من دولتي لم يتأثر بلهجات أخرى ويكتب هكذا)

أمثلة توضيحية لبعض الجمل الغير مقبولة في لهجات دول عربية مختلفة:

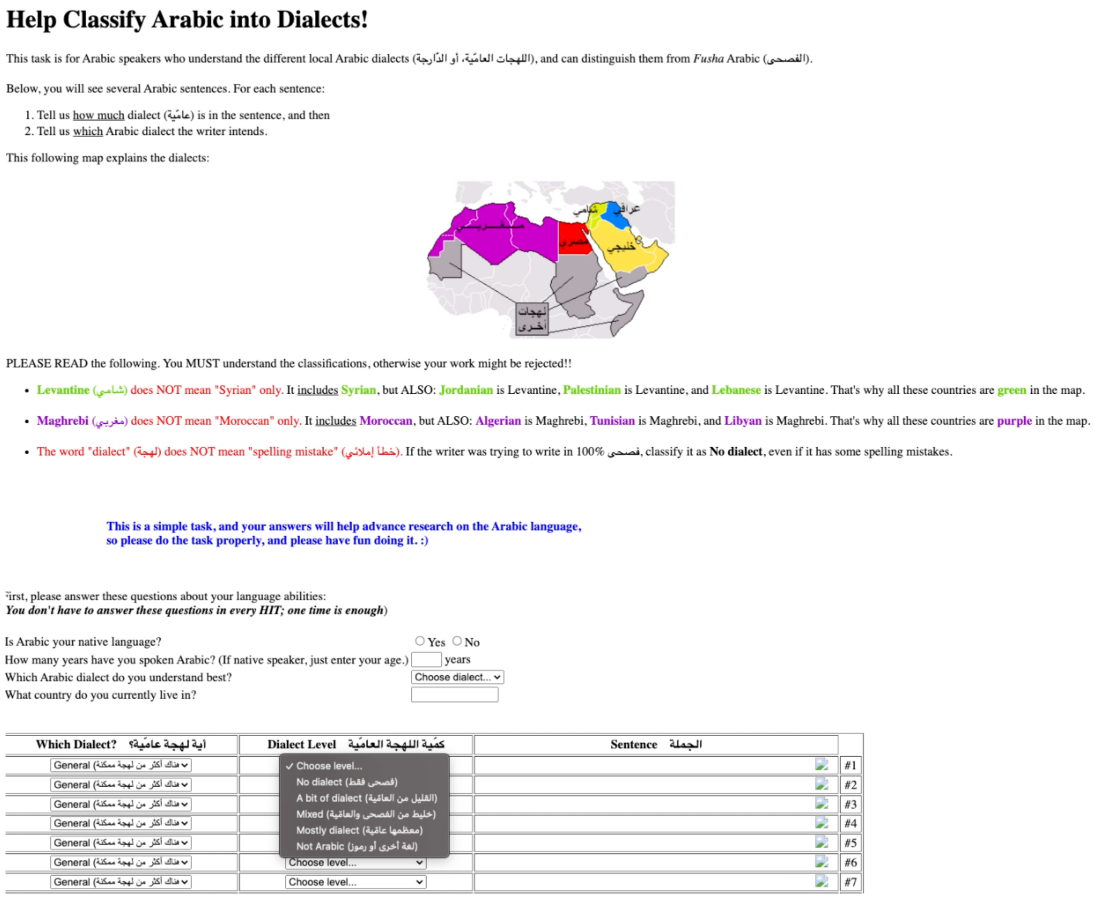
الدولة	التغريدة	تفسير اختيار لا (فقط للتوضيح)
سوريا	اللي عاليميننن نايس اللي عاشمال ما بحبوو كلششش	"كلششش" هى كلمة غير مستخدمة في سوريا
السعودية	وانا بشهد ياعيني على الحجازيات وبعدها الجنوبيات ياليبه بس جمال صحباتي حروب بيقولو للقمر قوم لاقعد مطرحك ولا الغامديه قشطه على عسل صافي #بنات_السعوديه_الاكثر_جمالا	"بيقولو، لاقعد مطرحك" هى كلمات غير مستخدمة في السعودية
مصر	بجد الناس اللي بتنسى بسرعة بجد كيييف تعملوها ! ؟ !	"كيييف تعملوها" هو مصطلح غير مستخدم في مصر
لبنان	خرج حدا يشرح يا جماعة ؟ على فكرة وحشتوني اوي اوي	"على فكرة وحشتوني اوي اوي" هو مصطلح غير مستخدم في لبنان
الجزائر	يالطيف مقتلي خير بلاد تسخف حالها علاه هكا ملا حالة والله	"علاه هكا، ملا حالة" هى مصطلحات غير مستخدمة في الجزائر

Figure A1: The guidelines used for annotating the validity of the sentences (Subtask 1 - MDID).

B Evaluation Data Annotation Process

As summarized in §4.2, we asked the annotators to complete a short onboarding task before joining the 5 main annotation tasks. In this section, we provide further details about the annotation process.

Onboarding Tasks QADI's test set (Abdelali et al., 2021) has 3,303 tweets geolocated to 18 different Arab countries (including the 14 countries represented in the samples of our dataset). The geolocated label for each tweet was then validated by a native speaker from that country, who checked if the "tweet is consistent with the dialect spoken in their country". Additionally, the test set has 200 tweets automatically classified as written in MSA.



Help Classify Arabic into Dialects!

This task is for Arabic speakers who understand the different local Arabic dialects (اللهجات العاميّة، أو الدّارجة), and can distinguish them from *Fusha* Arabic (الفصحى).

Below, you will see several Arabic sentences. For each sentence:

1. Tell us how much dialect (عاميّة) is in the sentence, and then
2. Tell us which Arabic dialect the writer intends.

This following map explains the dialects:

PLEASE READ the following. You MUST understand the classifications, otherwise your work might be rejected!!

- Levantine (شامي) does NOT mean "Syrian" only. It includes Syrian, but ALSO: Jordanian is Levantine, Palestinian is Levantine, and Lebanese is Levantine. That's why all these countries are green in the map.
- Maghrebi (مغربي) does NOT mean "Moroccan" only. It includes Moroccan, but ALSO: Algerian is Maghrebi, Tunisian is Maghrebi, and Libyan is Maghrebi. That's why all these countries are purple in the map.
- The word "dialect" (لهجة) does NOT mean "spelling mistake" (خطأ إملائي). If the writer was trying to write in 100% فصحى, classify it as **No dialect**, even if it has some spelling mistakes.

This is a simple task, and your answers will help advance research on the Arabic language, so please do the task properly, and please have fun doing it. :)

First, please answer these questions about your language abilities:
You don't have to answer these questions in every HIT; one time is enough)

Is Arabic your native language? Yes No
How many years have you spoken Arabic? (If native speaker, just enter your age.) years
Which Arabic dialect do you understand best? Choose dialect...
What country do you currently live in?

Which Dialect? أية لهجة عاميّة؟	Dialect Level كميّة اللهجة العاميّة	Sentence الجملة	
General (هناك أكثر من لهجة ممكنة)	✓ Choose level...		#1
General (هناك أكثر من لهجة ممكنة)	No dialect (فصحى فقط)		#2
General (هناك أكثر من لهجة ممكنة)	A bit of dialect (القليل من العاقية)		#3
General (هناك أكثر من لهجة ممكنة)	Mixed (خليط من الفصحى والعاقية)		#4
General (هناك أكثر من لهجة ممكنة)	Mostly dialect (معظمها عاقية)		#5
General (هناك أكثر من لهجة ممكنة)	Not Arabic (لغة أخرى أو رموز) / Choose level...		#6
General (هناك أكثر من لهجة ممكنة)	Choose level...		#7

Figure A2: A screenshot of the annotation interface of the AOC dataset (Zaidan and Callison-Burch, 2011).

س٢) تحديد مستوى العامية/الدارجة

- بالنسبة للتغريدات التي كان من الممكن أن يكتبها متحدث بلهجتكم الأصلية، يرجى تقييم مستوى اللهجة (العامية) في كل تغريدة.

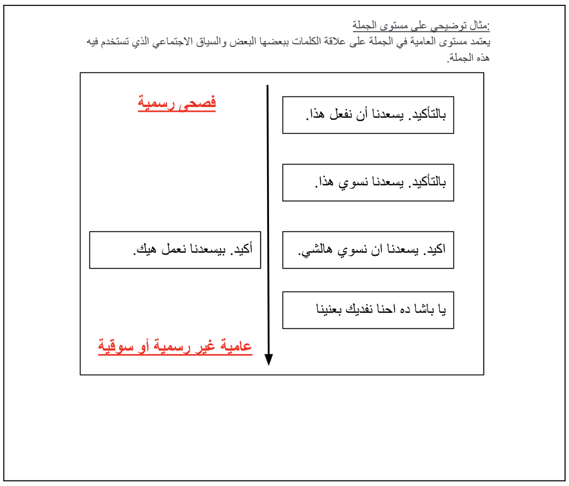
توضيح لبعض من خصائص المستويات المختلفة

- **فصحى سليمة:** تغريدات مكتوبة بلغة عربية فصيحة.
- **عامية رسمية/شبه فصيحة أو فصحى متأثرة بالعامية:** تغريدات مكتوبة بلغة تقترب من الفصحى ولكن تستخدم بعض التعبيرات العامية (مفردات وتصاريف).
- **عامية طبيعية (عامية عادية):** تغريدات مكتوبة بلغة عامية مقبولة ومفهومة من كافة أفراد المجتمع بمختلف أعمارهم ومستوياتهم الاجتماعية والتعليمية.
- **عامية غير رسمية (أو سوقية):** تغريدات مكتوبة بلغة عامية فيها تعبيرات غير مقبولة أو غير مفهومة من كافة أفراد المجتمع. لا يشترط أن تكون مبتذلة أو ركيكة.

كمتحدث باللغة العربية علي منصة X (تويتر سابقًا)، قد استخدم الفصحى للتحدث مع حساب رسمي لسفارة دولة عربية، مستوى عامية شبه فصيحة لتقديم التعازي لصديق، مستوى عامية طبيعية (عادية) لإبداء الرأى في قضية مجتمعية، ومستوي عامية غير رسمية للمزاح مع صديق مقرب.

مثال توضيحي فقط على مستوى الكلمة الواحدة:
الأفعال الآتية تستخدم للتعبير عن الشعور بالسعادة والسرور.

(a) The guidelines for the ALDi Estimation subtask.



مثال توضيحي على مستوى الجملة:
يعتمد مستوى العامية في الجملة على علاقة الكلمات ببعضها البعض والسياق الاجتماعي الذي تستخدم فيه هذه الجملة.

(b) An example of different-ALDi variants of a sentence.

Figure A3: Screenshots of the guidelines that were provided to the annotators to determine the ordinal ALDi level (Subtask 2) for the 1120 sentences of the development and test sets.

In order to get the annotators acquainted with our annotation guidelines shown in Figure A1, and Figure A3, we asked them to label 35 tweets from QADI as an initial onboarding task. Each country's onboarding task had 10 tweets labeled as consistent with the dialect(s) spoken in the country according to QADI's annotations and 5 MSA samples.[14] We included MSA samples to ensure that the onboarding tasks contain tweets of potentially different levels of dialectness. Additionally, we had 2 DA samples from 10 other country-level dialects, that would act as negatives for the first subtask (i.e., some of these samples are expected not to be valid in the considered country of the onboarding task). For each country's onboarding task, the composing samples were randomly shuffled. The annotators were not given information about the samples' geolocated labels or their distribution across the different countries/labels.

Quality Assurance and Feedback For each country's onboarding task, and thanks to the labels from QADI, we could perform two automatic checks for assessing the quality of the annotations:

- **Check (1)** The 10 samples geolocated to an annotator's country are expected to (a) be labeled as valid, and (b) with an ALDi level > 0.
- **Check (2)** The 5 MSA samples are expected to be marked as valid by all the annotators, with Level (0) *Sound MSA* as their ALDi.

On inspecting the annotators' performance on the onboarding task, we found that these automated checks are generally satisfied by the annotators, as per Table B2. The checks also helped us identify any misinterpretation of the guidelines, and provide feedback to our annotators before labeling the samples of the main task. For instance, one of the Algerian annotators *Algeria (C)* interpreted question (1) as identifying if the tweet matches how an Algerian typically writes, and consequently marked all the MSA samples as not valid in Algerian Arabic. For these sentences, he provided an alternative translation that sounds more natural to him (e.g., for the sample واحلى اخت انتي . . نفتخر و نعتز بيك اكيد ربي يحفظك, he provided the following alternative انتي اخت مخيرة .. نزوخ و نستعرف بيك ربي يحفظك). Another Syrian annotator *Syria (C)* chose *Maybe/ Not Sure* for 4 out of the 5 MSA samples, as she understood the first question as if the sentence could have <u>only</u> been written by a speaker of Syrian Arabic. Her explanation for her choice is: *For the MSA sentences, it is hard to accurately identify the speaker's dialect, so I chose unsure*.

Moreover, we noticed that for the onboarding task of Algeria, 3 samples geolocated to Algeria (Table B3), are labeled as not valid in Algerian Arabic by all the three Algerian annotators. Given that (a) the members of our team responsible for the annotation process are native speakers of Egyptian Arabic only, and (b) the labels for the main task are not validated (i.e., assigned based on geolocation only), we could not judge whether these are errors in the labels provided by our annotators or if QADI's validated labels for these samples were incorrect. Therefore, we decided not to use these checks as automatic measures for accepting or rejecting the labels provided by the annotators for the main task batches, and resorted to manually inspecting the annotations by the end of each annotation batch, as elaborated next.

Main Task Batches Following the onboarding task, we invited the annotators to label the task's data, split into 5 batches, of 224 samples each. We ran the annotation batches over 5 weeks (1 batch per week), to ensure a higher annotation quality.

By the end of each batch, and as done for the onboarding task, we used the 2 aforementioned checks to inspect the quality of the annotations. Moreover, we compared the labels provided by the annotators of each country against each other. We also kept track of the quality using automatic IAA metrics namely Fleiss' Kappa (κ) for Subtask 1, and Kripendorff's alpha (α) for Subtask 2. For the first batch, we flagged all the instances of disagreement, asked the annotators to relabel them and write comments in case these flagged instances were deemed as not valid. This allowed us to have a better assessment of the reasons for disagreement, and provide the annotators with tailored feedback accordingly.

For the following three batches, we tried to categorize clear patterns of disagreement between the annotators (e.g., an annotator systematically disagreeing with the other annotators) and discussed them individually with the annotators to rectify them in future batches. We have only asked them to relabel the samples of high disagreement in case we could not determine a pattern for the disagree-

[14] We noticed that some non-MSA samples of QADI were flagged as being in MSA by our in-house MSA/DA classifier, so we relied on the predictions of our model for considering a sample as MSA.

Annotator #	Algeria (y/m/n)	Egypt (y/m/n)	Iraq (y/m/n)	Jordan (y/m/n)	Lebanon (y/m/n)	Libya (y/m/n)	Morocco (y/m/n)	Palestine (y/m/n)	Saudi Arabia (y/m/n)	Sudan (y/m/n)	Syria (y/m/n)	Tunisia (y/m/n)	UAE (y/m/n)	Yemen (y/m/n)	MSA (y/m/n)
	Sample Geolocated Label														
Algeria (A)	**6**/0/4 *	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	1/0/**1**	1/1/**0**	-	-	**5**/0/0
Algeria (B)	**7**/0/3 *	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	-	1/0/**1**	0/0/**2**	0/0/**2**	1/0/**1**	1/0/**1**	2/0/**0**	-	-	**5**/0/0
Algeria (C)	**5**/0/5 *	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	1/0/**1**	1/0/**1**	-	-	**0**/0/5 *
Egypt (A)	0/0/**2**	**8**/0/2	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	1/0/**1**	0/0/**2**	0/0/**2**	-	-	**5**/0/0
Egypt (B)	0/1/**1**	**10**/0/0	1/0/**1**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	-	**5**/0/0
Egypt (C)	0/0/**2**	**10**/0/0	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/1/**1**	0/0/**2**	1/0/**1**	0/0/**2**	0/0/**2**	-	-	**4**/0/1
Iraq (A)	1/1/**0**	1/0/**1**	**6**/1/3	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	-	**5**/0/0
Iraq (B)	0/2/**0**	0/1/**1**	**7**/3/0	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	-	**3**/2/0
Iraq (C)	2/0/**0**	1/0/**1**	**10**/0/0	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/1/**1**	-	-	**5**/0/0
Morocco (A)	2/0/**0**	2/0/**0**	1/0/**1**	2/0/**0**	2/0/**0**	-	**10**/0/0	1/1/**0**	1/1/**0**	2/0/**0**	2/0/**0**	2/0/**0**	-	-	**5**/0/0
Morocco (B)	2/0/**0**	2/0/**0**	2/0/**0**	2/0/**0**	2/0/**0**	-	**10**/0/0	2/0/**0**	2/0/**0**	1/1/**0**	2/0/**0**	2/0/**0**	-	-	**5**/0/0
Morocco (C)	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	**8**/0/2	0/0/**2**	0/0/**2**	0/0/**2**	1/0/**1**	1/0/**1**	-	-	**4**/0/1
Palestine (A)	1/0/**1**	1/0/**1**	0/0/**2**	1/0/**1**	1/0/**1**	-	0/0/**2**	**8**/0/2	0/0/**2**	0/0/**2**	1/0/**1**	1/0/**1**	-	-	**4**/0/1
Palestine (B)	2/0/**0**	2/0/**0**	1/0/**1**	2/0/**0**	2/0/**0**	-	1/0/**1**	**7**/0/3	0/0/**2**	0/0/**2**	2/0/**0**	1/0/**1**	-	-	**4**/0/1
Palestine (C)	1/0/**1**	1/1/**0**	0/0/**2**	1/0/**1**	2/0/**0**	-	0/0/**2**	**9**/1/0	0/0/**2**	1/0/**1**	1/0/**1**	1/0/**1**	-	-	**5**/0/0
Sudan (A)	0/1/**1**	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	**9**/1/0	0/0/**2**	0/0/**2**	-	-	**4**/0/1
Sudan (B)	2/0/**0**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	**9**/0/1	0/0/**2**	0/0/**2**	-	-	**5**/0/0
Sudan (C)	2/0/**0**	0/0/**2**	0/1/**1**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	**9**/1/0	1/0/**1**	1/0/**1**	-	-	**5**/0/0
Syria (A)	0/0/**2**	1/0/**1**	0/0/**2**	1/0/**1**	2/0/**0**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	**9**/0/1	0/0/**2**	-	-	**5**/0/0
Syria (B)	2/0/**0**	1/0/**1**	1/0/**1**	1/0/**1**	1/1/**0**	-	0/1/**1**	0/0/**2**	0/0/**2**	0/0/**2**	**9**/1/0	1/1/**0**	-	-	**5**/0/0
Syria (C)	0/1/**1**	1/0/**1**	0/0/**2**	0/0/**2**	1/0/**1**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	**9**/1/0	0/0/**2**	-	-	**1**/4/0 *
Tunisia (A)	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	1/0/**1**	0/0/**2**	0/0/**2**	1/0/**1**	**9**/0/1	-	-	**2**/0/3 *
Tunisia (B)	1/0/**1**	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	**9**/0/1	-	-	**5**/0/0
Tunisia (C)	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	1/0/**1**	**8**/1/1	-	-	**4**/1/0
Yemen (A)	1/1/**0**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/1/**1**	0/0/**2**	0/0/**2**	1/0/**1**	1/0/**1**	-	**8**/1/1	**5**/0/0
Yemen (B)	2/0/**0**	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	1/0/**1**	0/0/**2**	1/0/**1**	1/0/**1**	-	**10**/0/0	**4**/0/1
Yemen (C)	2/0/**0**	1/0/**1**	0/0/**2**	0/0/**2**	1/0/**1**	-	0/0/**2**	0/0/**2**	0/1/**1**	0/0/**2**	1/0/**1**	0/1/**1**	-	**8**/1/1	**5**/0/0

Table B2: The distribution of the validity labels for the samples of the onboarding tasks presented as the number of each of the following decisions (Yes/Maybe/No), split into columns according to QADI's geolocated label of the samples. **Note #1:** The bolded value in each column represents the expected decision. **Note #2**: We initially discarded Libya, UAE, and Yemen from our dataset, and thus the onboarding datasets of the other countries do not have samples from these three countries. **Note #3**: We marked the unexpected patterns with *.

Samples from Algeria's Onboarding Task deemed invalid by the three Algerian annotators
* ههههههههههههههههههههههههههههههههه ياي وبتصحح كمان مصدقة نفسها يابنتي مش هيعبرك برضو سايكو بجد هموت
* بسم الله الرحمن راح نبدأ نعد واجيب البت دي تعدي معايا
* عباس لسه ما فيه وسم لظاهرة اليوم نزلو وقول وزززززع وخليه يوصل ترند عالمي

Table B3: Three samples categorically annotated as invalid by the three Algerian annotators, yet are geolocated to Algeria according to QADI's test set.

Country	Fleiss' Kappa (κ)				
	Batch 1	**Batch 2**	**Batch 3**	**Batch 4**	**Batch 5**
Algeria	0.49 (59) 0.58 (59)	0.49 (80) -	0.54 (65) -	0.51 (67) 0.61 (70)	0.48 (59) 0.59 (59)
Morocco	0.62 (34) -	0.42 (28) 0.81 (53)	0.27 (49) 0.5 (50)	0.36 (41) 0.53 (46)	0.48 (43) 0.62 (47)
Tunisia	0.43 (28) 0.56 (41)	0.67 (47) -	0.64 (33) 0.71 (40)	0.53 (31) 0.7 (31)	0.46 (32) 0.71 (30)
Egypt	0.58 (62) 0.7 (61)	0.63 (69) -	0.56 (68) -	0.64 (81) 0.74 (81)	0.69 (74) 0.79 (74)
Sudan	0.57 (68) 0.72 (74)	0.53 (76) -	0.58 (84) -	0.56 (81) 0.74 (81)	0.67 (77) 0.79 (78)
Palestine	0.4 (114) 0.58 (111)	0.41 (59) -	0.52 (72) -	0.51 (61) 0.69 (67)	0.54 (71) 0.74 (66)
Syria	0.39 (83) 0.56 (89)	0.49 (92) -	0.49 (87) -	0.59 (109) -	0.53 (90) 0.57 (98)
Iraq	0.59 (58) -	0.52 (44) -	0.59 (50) -	0.61 (51) 0.69 (62)	0.59 (53) 0.64 (57)
Yemen	0.46 (101) 0.55 (104)	0.57 (99) -	0.52 (81) -	0.45 (76) -	0.45 (94) 0.5 (94)

(a) Subtask 1 - Validity Labels.

Krippendorff's Alpha (α)				
Batch 1	**Batch 2**	**Batch 3**	**Batch 4**	**Batch 5**
0.745 (59) 0.663 (59)	0.615 (80) -	0.708 (65) -	0.513 (67) 0.536 (70)	0.715 (59) 0.666 (59)
0.823 (34) -	0.691 (28) 0.76 (53)	0.767 (49) 0.742 (50)	0.768 (41) 0.811 (46)	0.687 (43) 0.742 (47)
0.798 (28) 0.787 (41)	0.738 (47) -	0.8 (33) 0.808 (40)	0.71 (31) 0.722 (31)	0.664 (32) 0.698 (30)
0.845 (62) 0.796 (61)	0.828 (69) -	0.791 (68) -	0.791 (81) 0.788 (81)	0.862 (74) 0.82 (74)
0.765 (68) 0.746 (74)	0.537 (76) -	0.657 (84) -	0.696 (81) 0.727 (81)	0.624 (77) 0.643 (78)
0.731 (114) 0.645 (111)	0.752 (59) -	0.673 (72) -	0.633 (61) 0.739 (67)	0.559 (71) 0.573 (66)
0.845 (83) 0.829 (89)	0.709 (92) -	0.866 (87) -	0.751 (109) -	0.774 (90) 0.796 (98)
0.677 (58) -	0.684 (44) -	0.724 (50) -	0.733 (51) 0.776 (62)	0.795 (53) 0.816 (57)
0.561 (101) 0.498 (104)	0.495 (99) -	0.457 (81) -	0.568 (76) -	0.397 (94) 0.433 (94)

(b) Subtask 2 - ALDi Labels.

Table B4: The detailed IAA scores for each of the 5 main annotation tasks, computed independently for each country's 3 annotators. The second line for each country represents the IAA scores after providing feedback to the annotators and asking them to reannotate the samples of high disagreement. **Note:** The number of sentences valid in each country-level dialect after applying majority voting is shown between (brackets).

ment. For the last batch, we resorted to asking the annotators to relabel the samples of disagreement, to get an approximate evaluation of the impact of this process on the aggregated labels.

Analysis of the IAA Scores Table B4 demonstrates how the IAA scores (Fleiss' Kappa for Subtask 1, and Krippendorff's Alpha for Subtask 2) changed as the annotation process progressed. First, the values hint at acceptable levels of agreement between the annotators for both subtasks. However, we notice that the range of the IAA scores differs from one country to another, especially for Subtask 1. The variation in the ranges of the IAA scores could be attributed to (a) the level of homogeneity between the dialects spoken in each country, and (b) the annotators' representativeness/knowledge of the different dialects spoken in their countries. Recruiting annotators from different regions within the same country (e.g., the case of the Algerian annotators), could increase the possibility of disagreement compared to when they all came from the same region (e.g., the case of the Egyptian annotators, where all are from Cairo).

Regarding the annotators' performance, we notice that the agreement between the annotators categorically increased by asking them to reannotate the high-disagreement sentences for their validity in their country-level dialects (Subtask 1). That said, the impact of this relabeling process on the number of valid sentences according to the majority voting is minimal for the last annotation batches. This increase in the agreement scores post-relabeling was not as consistent for the ALDi levels (Subtask 2), in which we sometimes notice insignificant decreases. This could be attributed to the subjectivity of the ALDi Estimation task, compared to the Validity task. Lastly, the annotators' performance, measured by the IAA scores, was consistent across the different annotation batches, showcasing the effectiveness of our process.

Rank	System	Macro-average			
		Acc. (↑)	Prec. (↑)	Recall (↑)	F_1 score (↑)
1	**Elyadata**	69.27$_{\pm 4.3}$	43.07$_{\pm 11.0}$	**62.17$_{\pm 5.6}$**	**49.85$_{\pm 8.3}$**
BL I	Top 90%	75.99$_{\pm 6.2}$	57.08$_{\pm 15.4}$	41.92$_{\pm 14.9}$	45.21$_{\pm 10.3}$
2	**NLP_DI**	74.41$_{\pm 6.2}$	49.56$_{\pm 11.0}$	39.70$_{\pm 11.4}$	43.02$_{\pm 9.3}$
BL II	Random	50.06$_{\pm 1.8}$	26.09$_{\pm 9.3}$	49.92$_{\pm 3.3}$	33.31$_{\pm 8.4}$
BL III	Top 1	**77.40$_{\pm 8.0}$**	**75.20$_{\pm 11.3}$**	20.52$_{\pm 11.7}$	30.37$_{\pm 12.8}$
3	**dzNlp**	75.42$_{\pm 7.7}$	61.21$_{\pm 11.7}$	15.17$_{\pm 5.3}$	23.61$_{\pm 6.7}$

(a) DA samples.

Macro-average			
Acc. (↑)	Prec. (↑)	Recall (↑)	F_1 score (↑)
40.68$_{\pm 9.4}$	95.92$_{\pm 3.4}$	39.02$_{\pm 10.6}$	54.58$_{\pm 9.1}$
34.23$_{\pm 15.3}$	96.43$_{\pm 4.2}$	30.75$_{\pm 16.2}$	44.65$_{\pm 17.8}$
33.51$_{\pm 15.2}$	97.40$_{\pm 3.1}$	30.56$_{\pm 15.6}$	44.48$_{\pm 16.1}$
51.43$_{\pm 5.6}$	94.36$_{\pm 4.5}$	**51.34$_{\pm 6.0}$**	**66.30$_{\pm 5.5}$**
13.26$_{\pm 7.7}$	**100.00$_{\pm 0.0}$**	7.81$_{\pm 7.4}$	13.71$_{\pm 11.3}$
10.22$_{\pm 5.3}$	85.19$_{\pm 31.9}$	5.22$_{\pm 4.4}$	9.49$_{\pm 7.7}$

(b) MSA samples.

Table C5: The performance of the systems submitted to Subtask 1 on the DA and MSA samples of the test set. The systems are ordered according to their macro-averaged F1 scores on the whole test set as indicated in Table 5.

Rank	System	Macro-average				Individual Region F_1 score (↑)				
		Acc. (↑)	Prec. (↑)	Recall (↑)	F_1 score (↑)	Maghreb$_3$	Nile$_2$	Levant$_2$	Gulf$_1$	Gulf of Aden$_1$
1	**Elyadata**	68.02$_{\pm 4.1}$	52.25$_{\pm 12.0}$	67.16$_{\pm 5.4}$	**58.18$_{\pm 8.9}$**	55.42	68.54	**67.81**	45.21	**53.89**
BL I	Top 90%	**73.07$_{\pm 4.7}$**	62.54$_{\pm 14.0}$	54.28$_{\pm 14.8}$	56.16$_{\pm 12.3}$	**61.08**	**69.58**	64.20	**51.17**	34.76
2	**NLP_DI**	71.73$_{\pm 4.6}$	57.50$_{\pm 12.2}$	49.65$_{\pm 13.3}$	53.09$_{\pm 12.7}$	54.71	67.97	65.65	36.69	40.43
BL II	Random	46.91$_{\pm 5.8}$	35.52$_{\pm 9.8}$	**69.01$_{\pm 15.3}$**	46.19$_{\pm 10.9}$	45.59	56.40	58.06	27.91	43.00
BL III	Top 1	72.52$_{\pm 7.5}$	**77.74$_{\pm 13.6}$**	28.87$_{\pm 12.5}$	40.25$_{\pm 14.8}$	50.59	60.45	31.84	40.55	17.80
3	**dzNlp**	69.94$_{\pm 7.6}$	68.39$_{\pm 10.0}$	21.85$_{\pm 9.1}$	32.42$_{\pm 11.5}$	44.14	44.27	34.09	25.55	14.03

Table C6: The performance of the systems submitted to Subtask 1, in predicting multi-label macro-regional dialects for the DA samples of the test set. In addition to the Macro-average F1 score, the individual f1 score for each region is reported. **Note:** the countries representing the regions are: *Maghreb* (Algeria, Tunisia, Morocco), *Nile* (Egypt, Sudan), *Levant* (Palestine, Syria), *Gulf* (Iraq), and *Gulf of Aden* (Yemen).

C Detailed Analysis of Subtask 1 Results

As described in §4.1, 75 out of the 1,120 samples used to form the development and test sets for Subtasks 1 and 2 are automatically identified as being in MSA. For Subtask 1, these MSA samples are expected to be labeled as valid in all the considered dialects. On checking the validity labels for these samples, we indeed found that they are mostly deemed valid in all the considered country-level dialects. The developed systems are expected to accordingly predict that these sentences are valid in all the considered dialects.

Consequently, we report their performance on the automatically identified DA, and MSA samples respectively in Table C5. Since the MSA samples represent a small proportion of the development and test sets, we find that the models' performance on the DA samples is not different from their overall performance reported in Table 5.

For the MSA samples, we notice that the macro-average Recall needs to be improved. A two-stage solution could be proposed in which a classifier first identifies if a sentence is in MSA or DA. MSA sentences can be predicted to be valid in all the considered dialects with high accuracy. Conversely, the validity labels for the DA samples could be identified using another multilabel dialect identification system.

Regional Level Performance The results in Tables 5, C5 indicate that there is room for improvement for the multi-label ADI systems to be reliably able to accurately operate on the country-level. Consequently, we group the nine country labels of the test set into macro-regional dialects according to (Baimukan et al., 2022) as follows: *Maghreb* (Algeria, Tunisia, Morocco), *Nile Basin* (Egypt, Sudan), *Levant* (Palestine, Syria), *Gulf* (Iraq), and *Gulf of Aden* (Yemen). For each region, a sample is considered valid in the region if it is valid in any of the region's countries for which we have validity labels. For example, a sample annotated as valid in Algeria, Tunisia, and Sudan will be considered valid in *Maghreb* and *Nile Basin*. We similarly consider the systems' predictions for the same nine countries and aggregate them into macro-regional dialects.

The models' performance predicting the macro-regional dialects is higher than that for country-level ones as per Table C6. However, the improvement is not as high as might have been expected, indicating that even multi-label macro-regional dialect identification is a challenging task. In the future, we plan to extend the labels in our test set to cover more countries, especially from the *Gulf* region.