Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification

Amr Keleg and **Walid Magdy**

Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
akeleg@sms.ed.ac.uk, wmagdy@inf.ed.ac.uk

Abstract

Automatic Arabic Dialect Identification (ADI) of text has gained great popularity since it was introduced in the early 2010s. Multiple datasets were developed, and yearly shared tasks have been running since 2018. However, ADI systems are reported to fail in distinguishing between the micro-dialects of Arabic. We argue that the currently adopted framing of the ADI task as a single-label classification problem is one of the main reasons for that. We highlight the limitation of the incompleteness of the *Dialect* labels and demonstrate how it impacts the evaluation of ADI systems. A manual error analysis for the predictions of an ADI, performed by 7 native speakers of different Arabic dialects, revealed that $\approx 66\%$ of the validated errors are not true errors. Consequently, we propose framing ADI as a multi-label classification task and give recommendations for designing new ADI datasets.

1 Introduction

ADI of text is an NLP task meant to determine the Arabic Dialect of the text from a predefined set of dialects. Arabic dialects can be grouped according to different levels (1) major regional level: Levant, Nile Basin, Gulf, Gulf of Aden, and Maghreb (2) country level: more than 20 Arab countries, and (3) city level: more than 100 micro-dialects (Cotterell and Callison-Burch, 2014; Baimukan et al., 2022).

Different datasets were built curating data from various resources with labels of different degrees of granularities: (1) regional-level (Zaidan and Callison-Burch, 2011; Alsarsour et al., 2018), (2) country-level (Abdelali et al., 2021; Abdul-Mageed et al., 2022, 2023), or (3) city-level (Bouamor et al., 2019; Abdul-Mageed et al., 2020a, 2021b). Despite attracting lots of attention and effort for over a decade, ADI is still considered challenging, especially for the fine-grained distinction of micro-Arabic dialects on the country and city levels. This

Dialects	Sentence
Iraq, Jordan, Lebanon, Libya, Oman, Palestine Qatar, Saudi Arabia, Sudan, Syria, Tunisia, Yemen	وين المحطة؟ **Where is the station?**
Iraq, Morocco, Qatar	شنو رقم الرحلة؟ **What is the flight/trip number?**

Table 1: The MADAR corpus (Bouamor et al., 2018) has English/French sentences manually translated into different Arabic dialects. The table shows two sentences having the same translation across multiple country-level dialects.

is generally demonstrated by the inability of ADI models to achieve high macro-F1 scores.

We believe that framing ADI as a single-label classification problem is a major limitation, especially for short sentences that might not have enough distinctive cues of a specific dialect as per Table 1. Therefore, assigning a **single dialect label** to each sentence either automatically (e.g.: using geotagging) or manually makes the labels incomplete, which in turn affects the fairness of the evaluation process. The single-label limitation for DI was also discussed for other languages such as French (Bernier-colborne et al., 2023).

The need for improving the framing of ADI and consequently the ADI resources was previously noted by Althobaiti (2020), who concluded the *Future Directions* section of her survey of Arabic Dialect Identification (ADI) with the following:

> "There is also a need to criticize the available resources and analyze them in order to find the gaps in the available ADI resources."

In this paper, we introduce the concept of *Maximal Accuracy* for ADI datasets having single labels. We then provide recommendations for how to build new ADI datasets in a multi-label setup to alleviate the limitations of single-label datasets. We hope that our study will spark discussions among

the Arabic NLP community about the modeling of the ADI task, which would optimally lead to the creation of new datasets of more complete labels, and help in improving the quality of the ADI models. The main contributions of the paper can be summarized as follows:

1. Criticizing the current modeling of the ADI task as a single-label classification task by empirically estimating the *Maximal Accuracy* for multiple existing ADI datasets.

2. Performing an error analysis for an ADI model by recruiting native speakers of seven different country-level Arabic dialects.

3. Presenting a detailed proposal for how multi-label classification can be used for ADI.

2   How are Current ADI Datasets Built?

There have been multiple efforts to build several datasets for the ADI task using multiple techniques. We recognize four main techniques: (1) Manual Human Annotation, (2) Translating sentences into predefined sets of dialects, (3) Automatic labeling of data using distinctive lexical cues, and (4) Automatic labeling using geo-tagging.

A common limitation to all those techniques is modeling the task as a single-label classification task, where each sentence in the datasets is assigned to only one dialect while ignoring the fact that the same sentence can be valid in multiple dialects. Furthermore, each of these techniques has its own additional limitations that affect the quality of the labels as follows:

(1) Manual Human Annotation where annotators categorize Arabic sentences into one dialect from a predefined list of dialects (Zaidan and Callison-Burch, 2011; Huang, 2015; Malmasi et al., 2016; Zampieri et al., 2017, 2018).

Limitations: It was found that annotators over-identify their own native dialects (Zaidan and Callison-Burch, 2014; Abu Farha and Magdy, 2022). Therefore, the annotations for sentences that are valid in multiple dialects might be skewed toward the countries from which most of the annotators originate, causing a representation bias. Moreover, accurately determining the Arabic dialect of a sentence requires exposure to the different dialects of Arabic, which might not be a common case for Arabic speakers.

(2) Translation in which participants are asked to translate sentences into their native Arabic dialects (Ho, 2006-; Bouamor et al., 2014; Meftouh et al., 2015; Bouamor et al., 2018; Mubarak, 2018). If all the participants are asked to translate the same source sentences, then the dataset is composed of parallel sentences in various dialects. The main application of these datasets is to help develop machine translation systems, however, they are sometimes used for ADI. Figure 1 demonstrates how a corpus of parallel sentences is transformed into a corresponding DI dataset.

Limitations: While the labels of the corresponding DI dataset are correct, a source sentence might have the same translation in multiple Arabic dialects, Table 1. In such cases, a single-label classifier is asked to predict different *Dialect* labels despite the input sentence being the same.

Moreover, the syntax, and lexical items in the translated sentences might be affected by the corresponding syntactic and lexical features of the source sentences, especially if the source sentence is MSA or a variant of DA (Bouamor et al., 2014; Harrat et al., 2017). Such effects might make the translated sentences sound unnatural to native speakers of these dialects.

(3) Distinctive Dialectal Terms where text is curated based on the appearance of a term from a seed list of distinctive dialectal terms. These terms are used to automatically determine the dialect of the text (Alsarsour et al., 2018; Althobaiti, 2022).

Limitations: The curated data is constrained by the diversity of the terms used to collect it.

(4) Geo-tagging where the text is automatically labeled using information about the location or the nationality of its writer (Mubarak and Darwish, 2014; Salama et al., 2014; Al-Obaidi and Samawi, 2016; Al-Moslmi et al., 2018; Zaghouani and Charfi, 2018; Charfi et al., 2019; El-Haj, 2020; Abdelali et al., 2021; Abdul-Mageed et al., 2020a, 2021b, 2022).

Limitations: While this technique allows for curating data from different Arab countries, it does not consider that speakers of a variant of DA might be living in an Arab country that speaks another variant (e.g.: An Egyptian living in Kuwait) (Charfi et al., 2019; Abdul-Mageed et al., 2020a). Moreover, some of the curated sentences might be written in MSA, so the curated sentences need to be split into DA sentences and MSA ones (Abdelali et al., 2021; Abdul-Mageed et al., 2021b, 2022).

Dataset	Ct/Cn/Re	Description
(1) Manual Labeling		
AOC (Zaidan and Callison-Burch, 2011)	- / - / 5 *	- Online comments to news articles, manually labeled three times by crowd-sourced human annotators.
Facebook test set (Huang, 2015) Note: Data attached to the paper on ACL Anthology.	- / - / 3	- 2,382 public Facebook posts manually annotated into Egyptian, Levantine, Gulf Arabic, and MSA.
VarDial 2016 (Malmasi et al., 2016) Note: The link provided is not working.	- / - / 4	- Sentences sampled from transcripts of broadcast, debate and discussion programs from AlJazeera. The dialects of these recorded
VarDial 2017 (Zampieri et al., 2017)	- / - / 4	programs were manually labeled. MSA is included as a 5th dialect
VarDial 2018 (Zampieri et al., 2018) Note: VarDial 2018 used the same data as VarDial 2017.	- / - / 4	class for the models. Audio features were used in the 2017 and 2018 editions to allow for building multimodal models.
ArSarcasm-v2 (Abu Farha et al., 2021)	- / - / 4 *	- 15,548 tweets sampled from previous sentiment analysis datasets, annotated for their dialect (including MSA).
(2) Translation		
Tatoeba (Ho, 2006-)	- / 8 / 4	- An ever-growing crowdsourced corpus of multilingual translations, that include MSA and 8 different Arabic dialects.
MPCA (Bouamor et al., 2014)	- / 5 / 3	- 2,000 Egyptian Arabic sentences from a pre-existing corpus, manually translated into 4 other country-level dialects in addition to MSA.
PADIC (Meftouh et al., 2015)	5 / 4 / 2	- 6,400 sentences sampled from the transcripts of recorded conversations and movie/TV shows in Algerian Arabic and manually translated into 4 other dialects and MSA.
DIAL2MSA (Mubarak, 2018)	- / - / 4	- Dialectal tweets manually translated into MSA.
MADAR6 (Bouamor et al., 2019)	5 / 5 / 4	- 10,000 sentences manually translated into 5 city-level Arabic dialects in addition to MSA.
MADAR26 (Bouamor et al., 2019)	25 / 15 / 5	- 2,000 sentences manually translated into 25 city-level Arabic dialects in addition to MSA.
(3) Distinctive Lexical Cues		
DART (Alsarsour et al., 2018)	- / - / 5 *	- Tweets streamed using a seed list of distinctive dialectal terms, which are used to initially assign a dialect to each tweet, before having them manually verified by crowdsourced annotators.
Twt15DA (Althobaiti, 2022) Note: Data shared as (tweet IDs, labels) only.	- / 15 / 5	- Tweets curated by iteratively augmenting lists of distinctive dialectal cues, starting with a seed list for each dialect.
(4) Geo-tagging		
(Mubarak and Darwish, 2014) Note: Not publicly available.	- / ? / ?	- Arabic tweets streamed from Twitter, then automatically annotated using the reported user locations of the tweets' authors.
YouDACC (Salama et al., 2014) Note: Not publicly available.	- / 8 / 5 *	- Comments to youtube videos labeled using the videos' countries of origin, and the authors' locations.
OMCCA (Al-Obaidi and Samawi, 2016)	5 / 2 / 2	- 27,912 reviews scrapped from Jeeran.com, and automatically labeled using the location of the reviewer.
MASC (Al-Moslmi et al., 2018)	- / 6 / 4	- 9,141 reviews curated from online reviewing sites, Google Play, Twitter, and Facebook. The country of the reviewer is used as a proxy for the dialect of the review.
Shami (Abu Kwaik et al., 2018)	- / 4 / 1	- Sentences in one of the 4 Levantine dialects: (1) manually collected from discussions about public figures on online fora; (2) automatically collected from the Twitter timelines of public figures.
ARAP-Tweet (Zaghouani and Charfi, 2018) Note: No download link on their site.	- / 16 / 5 *	- A corpus of tweets from 1100 users, annotated at the user level for the dialect, age, and gender.
ARAP-Tweet 2.0 (Charfi et al., 2019) Note: No download link on their site.	- / 17 / 5 *	- A corpus of tweets from about 3000 users, annotated at the user level for the dialect, age, and gender.
Habibi (El-Haj, 2020)	- / 18 / 6 *†	- Songs' lyrics labeled by the country of origin of their singers.
QADI (Abdelali et al., 2021) Note: Training data shared as (tweet IDs, labels) only.	- / 18 / 5	- Tweets automatically labeled based on the locations of the authors in the user description field. The labels of the testing set of each country were validated by a native speaker of each country's dialect.
NADI2020 (Abdul-Mageed et al., 2020a)	100 / 21 / 5	- Tweets of users staying in the same province for 10 months,
NADI2021 (Abdul-Mageed et al., 2021b)	100 / 21 / 5	automatically labeled by geotagging the tweets of the selected users.
NADI2022 (Abdul-Mageed et al., 2022)	- / 18 / 5	
NADI2023 (Abdul-Mageed et al., 2023)	- / 18 / 5	- Currently not disclosed
(5) Miscellaneous		
Arabic Dialects Dataset (El-Haj et al., 2018)	- / - / 4 *	- 12,801 sentences sampled from the AOC dataset, in addition to 3,693 sentences sampled from the *Internet Forums* category of the Tunisian Arabic Corpus (McNeil and Faiza, 2010-).

Table 2: The list of single-labeled ADI datasets categorized by the labeling techniques. We follow the regional categorization of Baimukan et al. (2022). **Ct/Cn/Re**: the number of cities (provinces), countries, and regions respectively. *: The regional dialects are defined as Egypt, Iraq, Levant, Gulf, and Maghreb (Cotterell and Callison-Burch, 2014). †: Sudanese Arabic is considered as another regional dialect. **?**: Missing information.

Corpus of parallel sentences				
Egypt	**Tunisia**	**Syria**	**Jordan**	**Palestine**
ازيك يا جومانا وحشاني	شحوالك يا جومانا توحشتك	كيفك يا جومانا اشتقتلك	كيفك جومانا اشتقتلك كثير	كيف حالك يا جمانه مشتاقلك

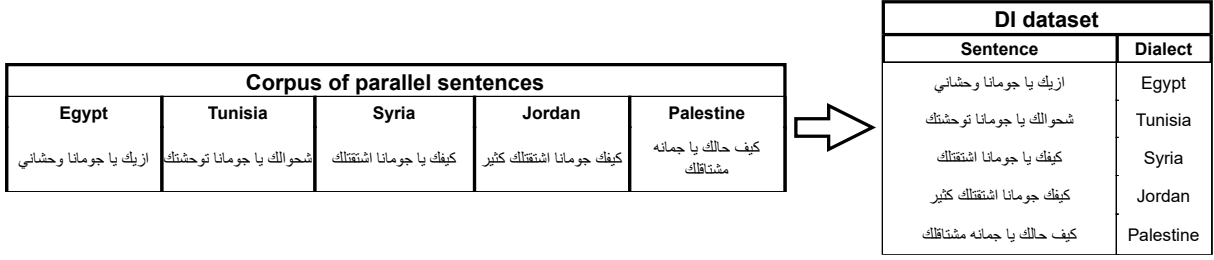| DI dataset	
Sentence	**Dialect**
ازيك يا جومانا وحشاني	Egypt
شحوالك يا جومانا توحشتك	Tunisia
كيفك يا جومانا اشتقتلك	Syria
كيفك جومانا اشتقتلك كثير	Jordan
كيف حالك يا جمانه مشتاقلك	Palestine

Figure 1: A demonstration of how parallel dialectal sentences are transformed into DI samples. The parallel sentences are sampled from the MPCA corpus (Bouamor et al., 2014)

3 Maximal Accuracy of Single-label ADI Datasets

For a single-label ADI dataset consisting of sentences where each is assigned one dialect label, assume that a percentage $\mathbf{Perc_2}$ of those sentences is valid in $\mathbf{2}$ different dialects. For those sentences, only one of the valid dialects is listed as their label. An effective model trained to predict a single label will randomly assign each of these sentences to one of its two valid labels. Thus, the expected maximal accuracy on the dataset $\mathbf{E[Accuracy_{max}]}$ that the model can achieve would then be:

$$\mathbf{E[Accuracy_{max}]} = (100 - \mathbf{Perc_2}) + \frac{\mathbf{Perc_2}}{2} \quad (1)$$

For example, if 40% of the sentences are valid in two dialects (i.e.: $\mathbf{Perc_2} = 40\%$), then the $\mathbf{E[Accuracy_{max}]}$ of the dataset would be 80%. This becomes worse when a sentence is valid in more dialects, exceeding ten valid dialects in some cases (as shown in Table 1). Thus, for a total number of dialects $N_{dialects}$, the equation above can then generalized to:

$$\mathbf{E[Accuracy_{max}]} = \mathbf{Perc_1} + \sum_{n=2}^{n=N_{dialects}} \frac{\mathbf{Perc_n}}{n} \quad (2)$$

where $\mathbf{Perc_1}$ is the percentage of samples that are only valid in one dialect, $\mathbf{Perc_n}$ is the percentage of samples valid in n dialects, $N_{dialects}$ represent the total number of dialects considered, and $\sum_{n=1}^{n=N_{dialects}} \mathbf{Perc_n} = 100\%$.

The higher the percentages $\mathbf{Perc_n}$ where $n \in [2, N_{dialects}]$, the lower the maximal accuracy would be. The same pattern would apply to F1 scores. Therefore, a model might be achieving low F1 scores as a consequence of framing DI as a single-label classification task, which might result in high $\mathbf{Perc_n}$ values.

Our objective in this paper is to estimate the value of $\mathbf{E[Accuracy_{max}]}$ for the existing datasets, which should examine the validity of our hypothesis that modeling ADI task as a single-label classification can be highly sub-optimal.

4 Estimating the Maximal Accuracy of Datasets

In our study, we focus on the country-level ADI for which multiple shared tasks have been organized since 2019 (Bouamor et al., 2019; Abdul-Mageed et al., 2020a, 2021b, 2022).

In order to quantify the percentages $Perc_n$, each sample of a dataset needs to be assessed by native speakers from all the Arab countries. Given our inability to recruit participants from all the Arab countries, we will estimate the percentages using two methods that provide lower bounds $\widetilde{\mathbf{Perc}}_n$ for the actual values $\mathbf{Perc_n}$ (i.e.: $\widetilde{\mathbf{Perc}}_n \leq \mathbf{Perc_n}$). Consequently, the estimated maximal accuracy is an upper bound for its true value.

4.1 Datasets Derived from Parallel Corpora

Initially, we examine the possibility of having Arabic sentences valid in multiple dialects by examining parallel corpora of Arabic dialects, which have sentences translated into multiple dialects. While a manual translation of a sentence can be phrased in different forms within the same dialect, we still examine if by chance we can find identical manually-translated sentences in different dialects by different translators.

For the four parallel corpora **Multidialectal Parallel Corpus of Arabic (MPCA)** (Bouamor et al., 2014), **PADIC** (Meftouh et al., 2015), **MADAR6**, and **MADAR26** (Bouamor et al., 2018), we transformed the parallel sentences into *(sentence, dialect)* pairs as in subtask (1) of the MADAR shared task (Bouamor et al., 2019). We then mapped the dialect labels for **PADIC**, **MADAR6**, and **MADAR26** from city-level dialects to country-level ones. In case the same sentence is used in

Dataset	$N_{dialects}$	$N_{samples}$	$\sum_{n=2}^{n=N_{dialects}} \widetilde{\text{Perc}}_n$	$\widetilde{\mathbb{E}}[\text{Accuracy}_{max}]$
PADIC	4	29,138	5.2%	97.1%
MPCA	5	4,960	7.8%	95.4%
MADAR6	5	49,476	2.3%	98.7%
MADAR26	15	48,624	9.6%	93.9%

Table 3: The estimated percentages and the corresponding expected maximal accuracy for the DI datasets formed using the four parallel corpora. The estimated maximal accuracies are upper bounds for the true maximal accuracies, and we expect the true values to be significantly lower than these estimates.
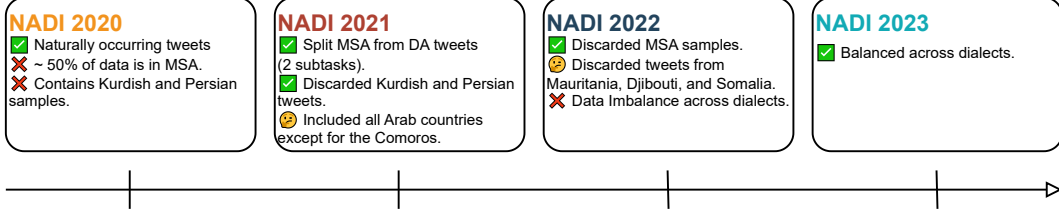

Figure 2: The evolution of the NADI datasets used for the shared tasks run between 2020 and 2023.

different cities within the same country, a single copy is kept. The sentences are then preprocessed by discarding Latin and numeric characters in addition to diacritics and punctuation. Lastly, we estimated the percentages $\widetilde{\text{Perc}}_n$ by computing the percentages of sentences that have the exact same translation in n dialects.

The upper bound for the maximal accuracies of the four corpora lies in the range $[93.9\%, 98.7\%]$ as per Table 3. The fact that the maximal accuracy for **MADAR26** is lower than that for **MADAR6** demonstrates that the probability that a sentence is valid in multiple dialects increases as more translations in other country-level dialects are considered.

4.2 Datasets of Geolocated Dialectal Sentences

The Nuanced Arabic Dialect Identification (NADI) shared tasks (Abdul-Mageed et al., 2020a, 2021b, 2022) used datasets that are built by collecting Arabic tweets authored by users who have been tweeting from the same location for 10 consecutive months. The geolocation of the users is then used as a label for their tweets. The creators of NADI have been improving the quality of the dataset from one year to another as summarized in Figure 2.

While the NADI shared tasks have been attracting active participation, the best-performing models in NADI 2022 achieved macro F1 scores of 36.48% and 18.95%, and accuracies of 53.05% and 36.84% on two test sets (Abdul-Mageed et al., 2022). The baseline MarBERT-based model (Abdul-Mageed et al., 2021a) fine-tuned on the training dataset achieves competitive results (macro F1 scores: 31.39% and 16.94%, accuracies:

47.77% and 34.06%).

Model Description Given the competitiveness of the baseline model, we fine-tuned the MarBERT model on the balanced training dataset of NADI 2023, and then we used the QADI dataset (Abdelali et al., 2021) as our test set. QADI's test set covers the same 18 countries as NADI 2023. We decided to analyze the errors of our model on QADI for two reasons: 1) At the time of writing the paper, the test set of NADI 2023 was not released (even for earlier NADIs, the labels of the test sets are not publicly released); 2) The dialect labels of the samples of QADI's test set were automatically assigned using geolocations similar to NADI, but the label of each sample was validated by a native speaker of the sample's label, which gives additional quality assurance for QADI over NADI.

The model achieves an accuracy of **50.74%** on QADI's test set with the full classification report in Table A2. Figure 3 visualizes how the predictions and labels are confused together.

Manual Error Analysis We recruited native-speaker participants of Algerian, Egyptian, Palestinian, Lebanese, Saudi Arabian, Sudanese, and Syrian Arabic to validate the False Positives (FPs) that the model makes for those dialects. Each participant is shown the FPs for their native dialect, one at a time, and is asked to validate them as indicated in Figure B1 [1]. If the participant found the FP sample to be valid in their native dialect, it means

---

[1] We release the judgments through: `github.com/AMR-KELEG/ADI-under-scrutiny/tree/master/data`

Figure 3: The confusion matrix for the predictions of a MarBERT model on QADI's test set. The model was fine-tuned using NADI 2023's training dataset.
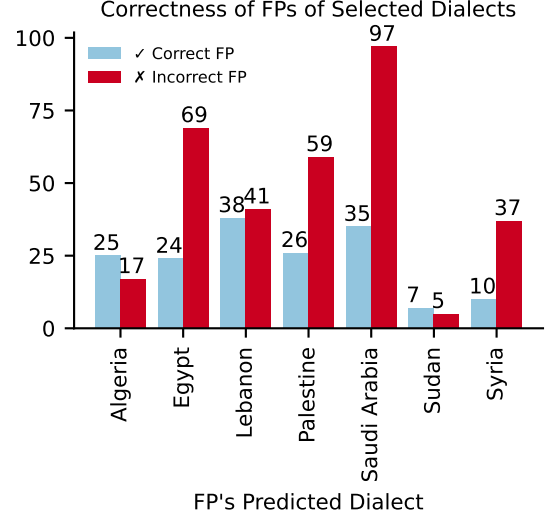

Figure 4: The distribution of the annotations for the validity of the False Positives (FPs) in 7 Arabic dialects. Correct FP represents the FP samples for which the model's prediction is invalid. Incorrect FP the FP samples for which the model's prediction is valid.

that this sample is valid in at least two different Arabic dialects (i.e.: the sample's original label, and the model's prediction) [2]. However, it can still be valid in additional dialects, which we did not check for due to the limited number of participants.

Validity of the Model's FPs Out of 490 validated FPs, 325 were found to be also valid in the other dialect they were classified to, which represents $\approx 66\%$ of the validated errors. Having such a great proportion of FPs that are not true errors hinders the ability to properly analyze and improve the ADI models. For Egyptian, Palestinian, Saudi Arabian, and Syrian Arabic, the majority of the FPs are incorrect as demonstrated in Figure 4 (i.e.: the model's prediction should be considered to be correct). As expected, dialects grouped in the same region are similar, and thus the FPs of a dialect would generally have labels of other dialects from the same region as in Figure 5.

Impact on Evaluation If we only consider the 725 samples that were correctly predicted by the model (TPs) in addition to the validated 490 FPs, then we know that 325 samples out of 1215 ones are at least valid in two different dialects. The $\tilde{\mathbf{P}}\mathbf{erc_2}$ for this subset is 26.7%, making the maximal accuracy $\mathbf{E}[\mathbf{Accuracy_{max}}]$ equal to 86.6%.

To further investigate the impact of the incorrect

FPs on the evaluation metrics, we computed the corrected True Positive value for each dialect $\mathbf{TP^*}$ as $\mathbf{TP^* = TP + Incorrect\ FP}$. Using these corrected $\mathbf{TP^*}$ values, we computed corrected precision, recall, and F1-scores. As per Table 4, the macro-averaged F1-score increased from 0.56 to 0.72. This clearly confirms our hypothesis that modeling ADI task as a single-label classification task leads to inaccurate evaluation of the systems.

5 Proposal for Framing the ADI Task

Given the limitations of using single-label classification for the ADI task, elaborated in §4, we propose alternative modeling for the task.

Zaidan and Callison-Burch (2014) asked crowd-sourced annotators to label dialectal sentences as being *Egyptian*, *Gulf*, *Iraqi*, *Levantine*, *Maghrebi*, *other dialect*, or *general dialect*. They used the *general dialect* for sentences that can be valid in multiple dialects. The *general dialect* is underspecified, and it is not clear whether it implies that a sentence is accepted in multiple dialects or in all of them. Therefore, the authors noticed that some of the annotators barely used the label, while others used it when they were not sure about the dialect of the underlying sentences. Moreover, they noticed that the annotators tend to over-identify their native dialects. Annotators might not realize that a sentence valid in their native dialect is also valid in

---

[2]Participants are given a third choice *Maybe / Not Sure*, which we count as *No* (i.e.: invalid in their dialect).

Figure 5: The distribution of the original labels for the False Positives (FPs) of the seven validated dialects. **Correct FP** represents the FP samples for which the model's prediction is invalid. **Incorrect FP** represents the FP samples for which the model's prediction is valid.

other dialects, and thus can end up choosing their native dialect as the label for this sentence, instead of the *general dialect* label.

Zampieri et al. (2023) focused on the binary distinction between two varieties of English, Portuguese, and Spanish. In addition to the two varieties of each language, the annotators are allowed to assign sentences to a third label *Both or Neither*. The evaluation results indicate that the *Both or Neither* label is harder to model computationally

than the other variety labels. The authors noted that there is room for improvement in the treatment and modeling of this third label.

Consequently, we believe that adding another label such as *general* or *Both or Neither* does not completely solve the limitations of single-label classification datasets. Conversely, framing the task as a multi-label classification would optimally alleviate the aforementioned limitations.

Dialect	TP	FP	TP*	FP*	FN	P	R	F1	P*	R*	F1*
Algeria	72	42	72 + 17 = 89	25	98	0.63	0.42	0.51	0.78	0.48	0.59
Egypt	170	93	170 + 69 = 239	24	30	0.65	0.85	0.73	0.91	0.89	0.90
Lebanon	134	79	134 + 41 = 175	38	60	0.63	0.69	0.66	0.82	0.74	0.78
Palestine	74	85	74 + 59 = 133	26	99	0.47	0.43	0.45	0.84	0.57	0.68
Saudi Arabia	88	132	88 + 97 = 185	35	111	0.40	0.44	0.42	0.84	0.62	0.72
Sudan	127	12	127 + 5 = 132	7	61	0.91	0.68	0.78	0.95	0.68	0.80
Syria	60	47	60 + 37 = 97	10	134	0.56	0.31	0.40	0.91	0.42	0.57
Macro-average						0.61	0.55	0.56	0.86	0.63	0.72

Table 4: The impact of the incorrect FPs on the precision **P**, recall **R**, and F1-score **F1**. Error samples for a specific predicted dialect (i.e.: FPs of this dialect) that are labeled as valid in this predicted dialect are counted as true positives in the corrected **TP*** score. The corrected **P***, **R*** and **F1*** are based on the corrected value of **TP***. $\mathbf{P^*} = \frac{\mathbf{TP^*}}{\mathbf{TP^*+FP^*}}, \mathbf{R^*} = \frac{\mathbf{TP^*}}{\mathbf{TP^*+FN}}, \mathbf{F1^*} = \frac{\mathbf{2*P^**R^*}}{\mathbf{P^*+R^*}}$
Note: P stands for Precision, R stands for Recall, and $F1$ stands for F1-score.

5.1 ADI as Multi-label Classification

Multi-label classification allows assigning one or more dialects to the same sample. Bernier-colborne et al. (2023) argued for using the multi-label classification setup after investigating a French DI corpus (FreCDo) (Gaman et al., 2022), covering four macro French dialects spoken in France, Switzerland, Belgium, and Canada. They found that the corpus has duplicated single-labeled sentences of different labels, and showed how these sentences impact the performance of DI models.

Labeling: Collecting multi-labels for a dataset requires the manual annotation of its samples. Dataset creators need to consider how they collect the annotations, and consequently who to recruit. An Arabic speaker of a specific dialect would be able to determine if a sentence is valid in their dialect or not (Salama et al., 2014; Abdelali et al., 2021). Althobaiti (2022) found that the average inter-annotator agreement score (Cohen's Kappa) is 0.64, where two native speakers of 15 different country-level Arabic dialects are asked to check the validity of tweets in their native dialects.

While human participants can sometimes infer the macro-dialect of a sentence that is not in their native dialect, it seems quite hard for them to predict the country-level dialects in which the sentence is valid (Abdul-Mageed et al., 2020b).
Recommendation: Ask Arabic speakers to identify if a sentence is valid in their native dialects or not as per (Salama et al., 2014; Abdelali et al., 2021;

Althobaiti, 2022). In order to include new dialects, speakers of these dialects need to be recruited.

Modeling: One way of building multi-label classification models is to use multiple binary classifiers. More specifically, a binary classifier is built to decide whether a sentence is valid in one dialect or not. For N dialects, N binary classifiers would be responsible for predicting the labels of a single sample.

Evaluation: For each supported dialect, evaluation metrics like accuracy, precision, recall, and F1-score can be used. Macro-averaging the metrics is a way to measure the average performance of the model across the different dialects.

Extensibility: The multi-label framing is extensible since more labels can be added to a previously annotated dataset. Adding a new dialect class does not invalidate the labels of the other dialect classes.

This does not apply to the single-label framing since an annotator would need to select a dialect out of a predefined set of dialects. Changing the set of dialects would require the reannotation of the whole dataset.

6 Conclusion

Single-label classification has been the defacto framing for Arabic Dialect Identification (ADI). We show that such framing implies that any model would have a maximal accuracy that is less than 100%, since some samples are valid in multiple dialects, and thus their labels are randomly assigned

from these dialects in which they are valid. For a set of 490 validated False Positives (FPs) of an ADI model, we found that the model's predicted dialects for 325 of them are also valid. The fact that about 66% of the FPs are not true errors hinders the ability to analyze and improve the ADI models, and hurts the reliability of the evaluation metrics.

Given this major limitation of single-label framing, we argue that ADI should be framed as a multi-label task. This follows the recommendation of Bernier-colborne et al. (2023) for French Dialect Identification. We hope that this paper will spark discussions across the Arabic NLP community about the current state of ADI, and encourage the creation of new datasets in a multi-label setup, with labels assigned manually by native speakers of the different Arabic dialects.

For future work, we will investigate the impact of the Arabic Level of Dialectness (ALDi) variable introduced by Keleg et al. (2023) on identifying the dialect of sentences. Intuitively, the dialect of a sentence with a high ALDi score is easier to identify since the sentence shows more features of dialectness than those of sentences having low ALDi scores. Therefore ALDi can be used to identify the samples that are more expected to be valid in multiple dialects, facilitating the annotation process of new DI datasets.

Limitations

Recruiting native speakers from the 18 Arab countries included in the NADI 2023 dataset proved to be hard. Moreover, we opted to only annotate the sentences of QADI's test set that were misclassified by the model. In order to accurately estimate the maximal accuracy for a dataset, all the samples should be checked independently by native speakers of the 18 supported Arab countries.

Acknowledgments

References

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, 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, ElMoatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The Fourth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

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 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.

Ibrahim Abu Farha and Walid Magdy. 2022. The effect of Arabic dialect familiarity on data annotation. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 399–408, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. Shami: A corpus of Levantine Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tareq Al-Moslmi, Mohammed Albared, Adel Al-Shabi, Nazlia Omar, and Salwani Abdullah. 2018. Arabic senti-lexicon: Constructing publicly available language resources for arabic sentiment analysis. *Journal of Information Science*, 44(3):345–362.

Ahmed Y Al-Obaidi and Venus W Samawi. 2016. Opinion mining: Analysis of comments written in arabic colloquial. In *Proceedings of the World Congress on Engineering and Computer Science*, volume 1.

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

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

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.

Gabriel Bernier-colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

Anis Charfi, Wajdi Zaghouani, Syed Hassan Mehdi, and Esraa Mohamed. 2019. A fine-grained annotated multi-dialectal Arabic corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 198–204, Varna, Bulgaria. INCOMA Ltd.

Ryan Cotterell and Chris Callison-Burch. 2014. A multi-dialect, multi-genre corpus of informal written Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 241–245, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

Mahmoud El-Haj, Paul Rayson, and Mariam Aboelezz. 2018. Arabic dialect identification in the context of bivalency and code-switching. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mihaela Gaman, Adrian-Gabriel Chifu, William Domingues, and Radu Tudor Ionescu. 2022. Frecdo: A large corpus for french cross-domain dialect identification.

Salima Harrat, Karima Meftouh, and Kamel Smaïli. 2017. Creating Parallel Arabic Dialect Corpus: Pitfalls to Avoid. In *18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, Budapest, Hungary.

Trang Ho. 2006-. Tatoeba: Collection of sentences and translations. Available online, Accessed: 10 September 2023.

Fei Huang. 2015. Improved Arabic dialect classification with social media data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2118–2126, Lisbon, Portugal. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

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

Karen McNeil and Miled Faiza. 2010-. Tunisian arabic corpus (tac): 895,000 words. Available online, Accessed: 10 September 2023.

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.

Hamdy Mubarak. 2018. Dial2msa: A tweets corpus for converting dialectal arabic to modern standard arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

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.

Ahmed Salama, Houda Bouamor, Behrang Mohit, and Kemal Oflazer. 2014. YouDACC: the Youtube dialectal Arabic comment corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1246–1251, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

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

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Dirk Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 1–17, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2023. Language variety identification with true labels.

A Detailed Dialect Coverage and Model Performance Report

The datasets used in the paper cover different Arabic dialects as detailed in Table A1. The **PADIC** dataset covers 4 country-level Arabic dialects from North Africa (Algeria, Tunisia), and the Levant (Syria, Palestine). On the other hand, the **QADI**, and **NADI 2023** datasets cover 18 country-level Arabic dialects.

Covering more dialects in a dataset impacts the performance of ADI models. Table A2 provides the detailed performance report of the MarBERT model fine-tuned for ADI between 18 country-level dialects, using NADI 2023's training dataset.

B The Error Analysis Survey

We created an online survey to validate the False Positives (FPs) of the MarBERT model fine-tuned on NADI 2023's training dataset. The survey aims to validate whether the errors of the model are caused by the single-label limitation of the testing dataset or are actual errors. Figure B1 shows screenshots of the Instructions, Sample examples, and the annotation interface.

Table B3 lists some examples for samples of the QADI dataset for which the model's predictions do not match the original labels, yet the annotators found these predictions to also be valid.

Dataset	Cities	Countries
PADIC	N = 5 Annaba, Algiers, Sfax, Damascus, Gaza	N = 4 Algeria, Tunisia, Syria, Palestine
MPCA	N/A	N = 5 Egypt, Syria, Jordan, Palestine, Tunisia
MADAR6	N = 5 Beirut, Cairo, Doha, Tunis, Rabat	N = 5 Lebanon, Egypt, Qatar, Tunisia, Morocco
MADAR26	N = 25 Aleppo, Damascus, Algiers, Alexandria, Aswan, Cairo, Amman, Salt, Baghdad, Basra, Mosul, Beirut, Benghazi, Tripoli, Doha, Fes, Rabat, Jeddah, Riyadh, Jerusalem, Khartoum, Muscat, Sanaa, Sfax, Tunis	N = 15 Syria, Algeria, Egypt, Jordan, Iraq, Lebanon, Libya, Qatar, Morocco, Saudi Arabia, Palestine, Sudan, Oman, Yemen, Tunisia
QADI NADI 2023	N/A	N = 18 Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Palestine, Qatar, Saudi Arabic, Sudan, Syria, Tunisia, United Arab Emirates, Yemen

Table A1: The list of labels in the different ADI datasets.

Dialect	Support	Precision (P)	Recall (R)	F1-score (F1)
Algeria	170	0.63	0.42	0.51
Libya	169	0.45	0.73	0.56
Morocco	178	0.77	0.63	0.70
Tunisia	154	0.63	0.54	0.58
Bahrain	184	0.33	0.29	0.31
Iraq	178	0.69	0.62	0.65
Kuwait	190	0.38	0.43	0.40
Oman	169	0.46	0.51	0.49
Qatar	198	0.37	0.34	0.35
Saudi Arabia	199	0.40	0.44	0.42
UAE	192	0.37	0.53	0.43
Egypt	200	0.65	0.85	0.73
Sudan	188	0.91	0.68	0.78
Jordan	180	0.31	0.47	0.38
Lebanon	194	0.63	0.69	0.66
Palestine	173	0.47	0.43	0.45
Syria	194	0.56	0.31	0.40
Yemen	193	0.55	0.25	0.34
Macro avg.		0.5309	0.5085	0.5072
Weighted avg.		0.5295	0.5074	0.5058
Accuracy		0.5074		

Table A2: The evaluation metrics for the predictions of the fine-tuned MarBERT model on QADI's testing set. The model is fine-tuned on NADI 2023's training data.

Instructions

- You will be shown a set of sentences. You are asked to check if the sentences are valid/natural in your native Arabic dialect **Yes**, or not **No**.
- In case you can not decide:
  - use the **Maybe / Not sure** option.
- If you choose the **Maybe / Not sure** or the **No** options:
  - Copy the span (a set of consecutive words) that made you choose this option.
  - In case multiple spans exist, you can add all of them separated by commas.
  - Please **do not overthink** the span selection question.

Powered by Qualtrics

(a) Instructions page.

The following screenshot is an example of a judgment made by an **Egyptian Arabic speaker.**
Please check the three examples to understand how the interface works.

Example #1:

Is this sentence valid in your dialect?

إعلانه والله محروق قلبي

Yes (Y)

Maybe / Not sure (M)

No (N)

In case you select **Maybe / Not sure (M)** or **No (N)**, please copy the span that made you choose this option.

- **Span**: A set of consecutive words.
- In case multiple spans exist, copy all of them separated by commas **,**
- Please do not spend too much time identifying the spans.

Any comments you want to add?

Powered by Qualtrics

(b) First example page.

The following screenshot is an example of a judgment made by an **Egyptian Arabic speaker.**
Please check the three examples to understand how the interface works.

Example #3:

Is this sentence valid in your dialect?

جاجنو رجل ضرب هذاك بينهم مقارنه اي نقابه

Yes (Y)

Maybe / Not sure (M)

No (N)

In case you select **Maybe / Not sure (M)** or **No (N)**, please copy the span that made you choose this option.

- **Span**: A set of consecutive words.
- In case multiple spans exist, copy all of them separated by commas **,**
- Please do not spend too much time identifying the spans.

هذاك ،مكيم

Any comments you want to add?

Powered by Qualtrics

(c) Third example page.

Is this sentence valid in your dialect?

هههههههه ياي وبتصحح كمان مصدّقة نفسها ######################################
هوت بجد سايكو برضو برضه هيعبزرك مش يابنتي

○ Yes (Y)
○ Maybe / Not sure (M)
○ No (N)

In case you select **Maybe / Not sure (M)** or **No (N)**, please copy the span that made you choose this option.

- **Span**: A set of consecutive words.
- In case multiple spans exist, copy all of them separated by commas **,**
- Please do not spend too much time identifying the spans.

Any comments you want to add?

Powered by Qualtrics

(d) An annotation page.

Figure B1: Screenshots of the different pages of the annotation task described in §4.2.

Valid Label	Sentence	Original Label
Algeria	عيشك يبارك فيك و يخليك	Tunisia
	الله يرحمه ربي معك خويا و انا لله و انا اليه راجعون	Morocco
Egypt	يلعن الكورة واليوم اللي شجعت في كورة .	Palestine
	مرتضي صوتوا ضعيف مع كامل إحترامي مايتقارنش بنسيم مجرد مقارنة	Tunisia
Lebanon	حالتنا أهون من حالات كتير في الحاضر و في التاريخ . . و غيرنا كتير نجحوا .	Egypt
	هههههه مين قلك أعصابي تعبانة	Syria
Palestine	بما أنو آخر شهر يا ربي يكونو عاملين خصم عالفلافل	Lebanon
	المشكلة انه فيه ناس ماعندهم عقل عشان تعطيهم على قد عقلهم	Kuwait
Saudi Arabia	والله ماعرف عنه بس جتني الصوره على الخاص وقلت اكيد تذكرونه	Iraq
	اقرا تغريدتي بالكامل وتقرا تغريدة كساب العتيبي وتعال اسال عنها وراح اجيبك	Qatar
Sudan	هههههههه انت رجعتي في كلامك سمحتي سمحتي	Tunisia
	والله يا استاذ عوض دي عربيه	Egypt
Syria	هلق الاستعمار فرض علينا بس الاستحمار نحنا فينا نعمله او ما نعمله	Lebanon
	لابدا ناس عندهم مبدا	Iraq

Table B3: Samples of QADI for which the ADI model's predictions are also valid.