NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task

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

Abstract

We present the results and findings of the First Nuanced Arabic Dialect Identification Shared Task (NADI). This Shared Task includes two subtasks: country-level dialect identification (Subtask 1) and province-level sub-dialect identification (Subtask 2). The data for the shared task covers a total of 100 provinces from 21 Arab countries and are collected from the Twitter domain. As such, NADI is the first shared task to target naturally-occurring fine-grained dialectal text at the sub-country level. A total of 61 teams from 25 countries registered to participate in the tasks, thus reflecting the interest of the community in this area. We received 47 submissions for Subtask 1 from 18 teams and 9 submissions for Subtask 2 from 9 teams.

1 Introduction

The Arab world is an extensive geographical region across Africa and Asia, with a population of ∼ 400 million people whose native tongue is Arabic. Arabic could be classified into three major types: (1) Classical Arabic (CA), the language of the Qur'an and early literature, (2) Modern Standard Arabic (MSA), the medium used in education and formal and pan-Arab media, and (3) dialectal Arabic (DA), a host of geographically and politically defined variants. Modern day Arabic is also usually described as a *diglossic* language with a so-called 'High' variety that is used in formal settings (MSA), and a 'Low' variety that is the medium of everyday communication (DA). The presumably 'Low variety' is in reality a collection of variants. One axis of variation for Arabic is geography where people from various sub-regions, countries, or even provinces within the same country, may be using language differently.

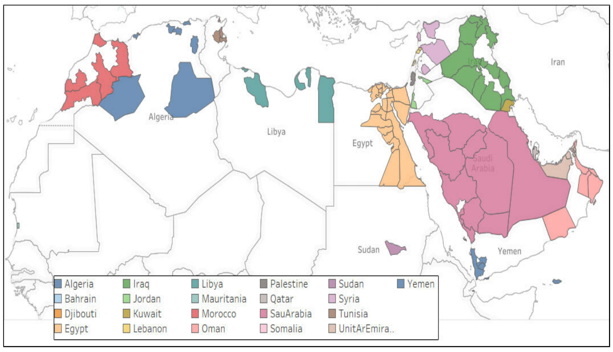

Figure 1: A map of the Arab World showing countries and provinces in the NADI dataset. Each of the 21 countries is represented in a color different from that of a neighboring country. Provinces are marked with lines inside each country.

The goal of the First Nuanced Arabic Dialect Identification (NADI) Shared Task is to provide resources and encourage efforts to investigate questions focused on dialectal variation within the collection of Arabic variants. The NADI shared task targets 21 Arab countries and a total of 100 provinces across these countries. The shared task consists of two subtasks: *country-level* dialect identification (Subtask 1) and *province-level* detection (Subtask 2). We provide participants with a new Twitter labeled dataset that we collected exclusively for the purpose of the shared task. The dataset is publicly available for research.[1] A total of 52 teams registered for the shard task, of whom 18 teams ended up submitting their systems for scoring. We then received 15 papers, of which we accepted 14.

[1]The dataset is accessible at the shared task page: http://nadi2020.arabic-nlp.net.

Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 97–110
Barcelona, Spain (Online), December 12, 2020

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

2 Related Work

As we explained in Section 1, Arabic could be viewed as comprised of 3 main types: CA, MSA, and DA. While CA and MSA have been studied and taught extensively, DA has only received more attention relatively recently (Harrell, 1962; Cowell, 1964; Badawi, 1973; Brustad, 2000; Holes, 2004).

A majority of DA computational efforts have targeted creating resources for country or regionally specific dialects (Gadalla et al., 1997; Diab et al., 2010; Al-Sabbagh and Girju, 2012; Sadat et al., 2014; Smaïli et al., 2014; Jarrar et al., 2016; Khalifa et al., 2016; Al-Twairesh et al., 2018; El-Haj, 2020). The expansion into multi-dialectal data sets and models to identify them was initially done at the regional level (Zaidan and Callison-Burch, 2011; Elfardy et al., 2014; Bouamor et al., 2014; Meftouh et al., 2015). A number of Arabic dialect identification shared tasks were organized as part of the VarDial workshop. These focused on regional varieties such as Egyptian, Gulf, Levantine, and North African based on speech broadcast transcriptions (Malmasi et al., 2016) but also acoustic features (Zampieri et al., 2017) and phonetic features (Zampieri et al., 2018) extracted from raw audio. Althobaiti (2020) presents a recent survey of computational work on Arabic dialects.

An early effort for creating finer grained parallel dialectal corpus and lexicon was done under the Multi Arabic Dialects Application and Resources (MADAR) project (Bouamor et al., 2018). The parallel data was created by commission under controlled settings to maximize its use for cross-dialectal comparisons and machine translation. Their data was also used for dialectal identification at the city level (Salameh et al., 2018; Obeid et al., 2019) of 25 Arab cities. One issue with the MADAR data in the context of identification is that it was commissioned and not naturally occurring. Concurrently, larger Twitter-based datasets covering 10-21 countries were also introduced (Mubarak and Darwish, 2014; Abdul-Mageed et al., 2018; Zaghouani and Charfi, 2018). Researchers are also starting to introduce DA datasets labeled for socio-pragmatics, e.g., (Abbes et al., 2020; Mubarak et al., 2020). The MADAR shared task (Bouamor et al., 2019) comprised two subtasks, one focusing on 21 Arab countries exploiting Twitter data manually labeled at the user level, and another on 25 Arab cities mentioned above. During the same time as NADI, Abdul-Mageed et al. (2020) describe data and models at country, province, and city levels.

The NADI shared task follows these pioneering works by availing data to the (Arabic) NLP community, and encouraging work on Arabic dialects. Similar to the MADAR shared task, we include a country-level dialect identification task (Subtask 1), and a sub-country dialect identification task (Subtask 2). However, our sub-country task is a province-level identification task with a much larger label set than MADAR's city-level task, and is based on naturally occurring data. We hope that our work will be setting the stage for exploring variation in geographical regions that have not been studied before.

3 Task Description

The NADI shared task consists of two subtasks for country-level and province-level classification.

3.1 Subtask 1: Country-level Classification

The goal of Subtask 1 is to identify country-level dialects from short written sentences (tweets). NADI Subtask 1 is similar to previous works that have also taken country as their target (Mubarak and Darwish, 2014; Abdul-Mageed et al., 2018; Zaghouani and Charfi, 2018; Bouamor et al., 2019). Labeled data was provided to NADI participants with specific TRAIN and development (DEV) splits. Each of the 21 labels corresponding to the 21 countries is represented in both TRAIN and DEV. Teams could score their models through an online system on the DEV set before the deadline. Our TEST set of unlabeled tweets was released shortly before the system submission deadline. Participants were invited to submit their predictions to the online scoring system that housed the gold TEST set labels. We provide the distribution of the TRAIN, DEV, and TEST splits across countries in Table 1.

Country Name	# of Provinces	# of Tweets				
		Train	Dev	Test	Total	%
Algeria	7	1,491	359	364	2,214	7.15
Bahrain	1	210	8	20	238	0.77
Djibouti	1	210	10	51	271	0.88
Egypt	21	4,473	1,070	1,092	6,635	21.43
Iraq	12	2,556	636	624	3,816	12.33
Jordan	2	426	104	104	634	2.05
Kuwait	2	420	70	102	592	1.91
Lebanon	3	639	110	156	905	2.92
Libya	5	1,070	265	265	1,600	5.17
Mauritania	1	210	40	5	255	0.82
Morocco	5	1,070	249	260	1,579	5.10
Oman	6	1,098	249	268	1,615	5.22
Palestine	2	420	102	102	624	2.02
Qatar	2	234	104	61	399	1.29
Saudi Arabia	10	2,312	579	564	3,455	11.16
Somalia	1	210	51	51	312	1.01
Sudan	1	210	51	51	312	1.01
Syria	5	1,070	265	260	1,595	5.15
Tunisia	4	750	164	208	1,122	3.62
UAE	5	1,070	265	213	1,548	5.00
Yemen	4	851	206	179	1,236	3.99
Total	**100**	**21,000**	**4,957**	**5,000**	**30,957**	**100.00**

Table 1: Distribution of country-level dialect identification data for Subtask 1 across our data splits.

3.2 Subtask 2: Province-level Classification

The goal of Subtask 2 is to identify the specific state or province (henceforth, *province*) from a list of 100 provinces. The provinces are unequally distributed among the list of 21 countries. While efforts on city-level and country-level prediction were the topic of a previous shared task (Bouamor et al., 2019), to the best of our knowledge, the target of automatic dialect prediction at a small geographical region such as a province has not been previously investigated, thus lending novelty to this subtask. We acknowledge that this subtask has some affinity to work focused on predicting geolocation based on tweets. Nevertheless, geolocation prediction is performed at the level of users not tweets and hence is different. There are also differences between our work here and geolocation as to how the data was collected. We further explain this nuance in Section 4. The distribution of the classes across the 100 provinces in our data splits is presented in Table A1 in Appendix A.

For both Subtask 1 and Subtask 2, tweets in the TRAIN, DEV and TEST splits come from distinct sets of *users*, such that no user had their tweets in any two of the TRAIN, DEV, and TEST splits.

3.3 Restrictions and Evaluation Metrics

To ensure fair comparisons and common experimental conditions, we provided participating teams with a set of restrictions that apply to the two subtasks, and clear evaluation metrics. Our method of distributing the data as well as our evaluation setup through the CodaLab online platform also facilitated the competition management, enhanced timeliness of acquiring results upon system submission, and guaranteed ultimate transparency.[2]

We directly provided participants with the actual tweets posted to the Twitter platform, rather than tweet IDs. This enabled comparison between systems exploiting identical data. Since we shared actual tweets, we did not share tweet IDs with participants. This made it harder to collect data from the same

[2] https://codalab.org/

Country	Province	Tweet
Algeria	**Bordj-Bou-Arreridj**	وانتم سيد الفاظل ملائكة اكبر مصائبنا منكم والعربان دمرتوا كل شيء جميل الله المستعان
	Jijel	ابراهيم غدوة تهلا فيه مليح
Egypt	**Asyut**	يا اقرع انت واخوك إبراهيم اللي زرع شعره واكيد الكل عارف جاب الشعر منين في جسمه وحطه في رأسه انت نسيت نفسك ولا ايه الزمالك هو اللي لمك من الشارع بعد ما صالح سليم طردك زي الكلب نسيت وانت مدرب ماحدش طلعك السما غير الزمالك ورجل شيكابالا الزمالك سيدك يا اقرع #ادعم_باسم_مرسي
	Suez	يبقي حد يوريني نفسه بقي انا قولت حلووو غلط
KSA	**Ar-Riyad**	فيه كثير أمور عني ما تعرفونها ياليت ان الأمر بهالسهولة وياليتني مرتاحه فعلا ومستقله بشكل كامل ما كان لقيتيني اشتكي ولا اقول هالكلام راح ابعد عنهم كثيير ان شاء الله واهاجر وما ارجع لهالبقعة الجغرافية ابد واترك كل شي شين وراي وابدأ حياة حقيقية من أول وجديد الله يعدي الأيام بسرعه بس
	Najran	#يسعد_مساكم حمودي يسمي عليكم ويقول رايح يشتري ثوب #العيد #عيدكم_مبارك_وعساكم_من_عواده
Morocco	**Marrakech-Tensift-Al-Haouz**	السيادة في الدنيا والسعادة في العقبى لا يوصل إليها إلا على جسر من المتاعب. #ابن_القيم
	Tanger-Tetouan	#اليابان_كولومبيا المبارة مانجا ولا فيلر
Oman	**Muscat**	بصراحة.. غريبين هدول الثورجية يلي بيجلسوا ينظروا من برا و بخبروك انو هنن تركوا شغلهم واموالهم وبيتهم وهلأ عايشين برات البلاد بسبب الظلم على قولهم.. طيب بالأول ما حدا طلب منك تترك البلاد... ثانيا.. ليش عم تحرض الناس يلي موجود بالبلد وهي يلي لازم تتبهدل وتعتقل وحضرتك برا البلاد.
	Ad-Dakhiliyah	الحين بتبدي الشماته
Sudan	**Khartoum**	نحن نبقى عليها والله شنطة زاتو ما دايرة ارح امفكو ساي
UAE	**Abu-Dhabi**	ولو بالغلط يعني . #اليوم_العالمي_للرسائل
	Ras-Al-Khaymah	عليك أغار و أكتم هالشعور أخاف غيرتي بالحيل تزعجك.

Table 2: Randomly picked examples from select provinces and corresponding countries.

user from which a tweet comes. For the two subtasks, we asked to only and exclusively use our distributed data. In other words, we provided instructions not to use any external data nor search or depend on any additional user-level information such as geolocation. In addition to our labeled TRAIN and DEV splits, we provided tweet IDs for 10M tweets and a simple script that can be used to collect the tweets. We did not provide any labels for this additional 10M tweet set, and encouraged participants to use it in developing their models in any way they deemed useful.

For both subtasks, the official metric is macro-averaged F_1 score obtained on blind test sets. We also report performance in terms of macro-averaged precision, macro-averaged recall and accuracy for systems submitted to each of the two subtasks. Each participating team was allowed to submit up to five runs for each subtask, and only the highest scoring run was kept as representing the team. Although official results are based only on a blind TEST set, we also asked participants to report their results on the DEV set in their papers. We setup two CodaLab competitions for scoring participant systems.[3,4]

[3]The CodaLab competition for Subtask 1 is accessible at: https://competitions.codalab.org/competitions/24001.

[4]The CodaLab competition for Subtask 2 is accessible at: https://competitions.codalab.org/competitions/24002.

We will keep the Codalab competition for each task live post competition, for researchers who would be interested in training models and evaluating their systems using the shared task TEST set.

4 Shared Task Datasets

We distributed a single dataset with two sets of labels, one for Subtask 1 and another for Subtask 2. In other words, the same tweet occurs in each of the two subtasks but with different subtask-specific labels. Additionally, we made available an unlabeled dataset for optional use in any of the two subtasks. We now provide more details about both the labeled and unlabeled data.

4.1 Data Collection

We used the Twitter API to crawl data from 100 provinces belonging to 21 Arab countries for 10 months (Jan. to Oct., 2019).[5] Next, we identified users who consistently and *exclusively* tweeted from a single province during the whole 10 month period. We crawled up to 3,200 tweets from each of these users.

4.2 Data Sets

Subtask 1 and Subtask 2 Data We labeled tweets from each user with the country and province from which the user posted for the whole of the 10 months period, thus exploiting user consistent posting *location* as a proxy for *dialect labels*. Note that this labeling method can still have issues as we explain in Section 4.3. We randomly sampled 30,957 tweets of length 5 words or more from the collection and split them into TRAIN (n=21,000), DEV (n=4,957), and TEST (n=5,000). Although the task is at the tweet level, we sampled the data for each of the TRAIN, DEV, and TEST from a unique set of users (i.e., users are not shared across the 3 splits). We distribute data for the two subtasks directly to participants in the form of actual tweet text (i.e., hydrated content). Tables 1 and A1 show the distribution of tweets across the data splits for both Subtask 1 and Subtask 2, respectively.

Unlabeled 10M We also crawled 10 million posts from Arabic Twitter during 2019. We call this dataset UNLABELED 10M and distribute it in the form of tweet IDs along with a script that can be used to crawl the actual tweets. We put no restrictions on using UNLABELED 10M for system development for either of the two subtasks.[6] Next, we discuss a number of nuances and issues found in the data.

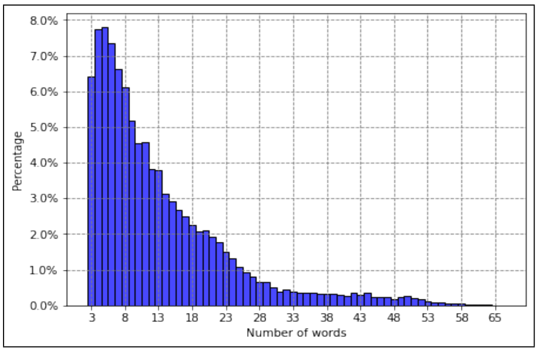

Figure 2: Distribution of tweet length in words in NADI labeled data.

4.3 Data Issues

Location as proxy for dialect. Our method of using consistent location (i.e., posting from the same location for at least 10 months) as a proxy for assigning dialect labels is useful, but not ideal. Even though this method allows us to collect provably relevant data, as manually verified in a small random sample of users (n=30), it can be error prone since a user with a dialect of one country can be posting from a different country during this whole period of 10 months.

MSA vs. Dialect. As we explained in Section 1, Arabic is usually characterized as a diaglossic language with MSA being the 'High' variety and DA as the 'Low' variety. Arabic users also switch between these two varieties. Most relevant to our work, communication in DA over social media is not devoid of MSA even at a level as short as that of a tweet. This can vary from one dialect to another, but also

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

[6]Subtask 1, Subtask 2, and UNLABELED 10M data is available at http://nadi2020.arabic-nlp.net. More information about the data format can be found in the accompanying README file.

depending on a range of other factors including the user educational background, career, and the actual goal of the post itself. To illustrate, based on our intuition and occasional observation, users with training in language sciences (education), those in careers such as media or higher education (job), or those trying to reach out to especially older generations or project religious or cultural authority (goal or pragmatic function) will likely use more MSA. Due to the co-existence of MSA and DA in the same tweet, we opted for using all the data we collected from the users for the competition. Alternatively, we could have identified MSA tweets either manually or automatically and removed these. We did not take that step in order to keep the task more challenging since a classifier would need to learn about patterns of MSA-DA mixing to perform well. A model would also need to acquire skills enabling it to tease apart tweets that may be overly or exclusively MSA. To explore the extent of MSA in the dataset, we run an in-house neural MSA-DA model ($acc = 89.1\%$, $F_1 = 88.6\%$) on it. The model predicts the percentages of MSA data as follows: 49.5% for TRAIN, 46.6%for DEV, and 49.7% for TEST. This distribution needs to be couched with caution, however, since dialectal features can be subtle and mixed with MSA to varying degrees. Upon manual inspection of a random sample of tweets the model labeled as MSA, we identify examples that are fully and verifiably MSA such as #1 below. In addition, we observe sequences that carry dialectal features. For example #2 and #3 below have dialectal words (highlighted in orange):

1. الحمد لله رب العالمين على نعمه التي لا تعد ولا تحصى ولم يغفل عن رزق كل خلقه.
2. البيض الفاسد يتدرب على بعضه يا صاحبي . قاتلهم الله
3. ما أبي اخذك منهم ، أبيك انت تبديني عليهم

We also observe that the more confident the MSA-DA model is (based on softmax value), the more likely its decision is correct. This suggests we can use a thresholding approach to filter out MSA tweets, should we desire to reduce MSA in the data. We leave further investigation of this issue to the future.

Non-Arabic Text. Despite efforts to exclusively keep Arabic content, our dataset had a small percentage (2.52%) of Farsi. While collecting the data, we only kept tweets assigned an Arabic language tag by the Twitter API. However, the API is error prone and hence some non-Arabic was not filtered out. To circumvent this, we only kept tweets that have at least three word written in Arabic script after running an internal normalizer that removes diacritics and reduced repetitions of consecutive characters of > 2 to only 2, replaced URLs and usernames with the generic strings URL and $@USR$. Even after this step, some Farsi leaked to our data. The reason is that Farsi is written in the same script as Arabic, with only a few differences. Aliwy et al. (2020) manually inspected the NADI TRAIN set and provided a distribution of Farsi tweets over the different countries. We share this distribution in Table 3.

5 Shared Task Teams & Results

5.1 Our Baseline Systems

We have two baseline classifier, Baseline I and Baseline II. **Baseline I** is based on the majority class in the TRAIN data for each subtask. It scores at $accuracy = 21.84\%$ and $F_1 = 1.71\%$ for Subtask 1 and $accuracy = 1.92\%$ and $F_1 = 0.04\%$ for Subtask 2. For **Baseline II**, we initially train two classifiers for these two sub-tasks individually. For each task, we fine-tune on Google's pre-trained multilingual BERT-Base (mBERT).[7] We set the maximum length of sequences in our model to 50 tokens, and employ batch training with a batch size of 8 for this model. We run the network for 20 epochs and save the model at the end of each epoch, choosing the model that performs highest on DEV as our best model. For country-level identification (Subtask 1), our best result is acquired with 16 epochs. Our best result is obtained with 20 epochs on province-level task (Subtask 2). Our mBERT model obtains $accuracy = 32.38\%$ and $F_1 = 13.32\%$ on country-level classification and $accuracy = 3.32\%$ and $F_1 = 2.13\%$ for province-level classification.

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

Country	# tweet	# Farsi	% Farsi
Algeria	1,491	5	0.34
Egypt	4,473	2	0.04
Iraq	2,556	382	14.95
Morocco	1,070	1	0.09
Oman	1,098	26	2.37
Saudi Arabia	2,312	6	0.26
Syria	1,070	3	0.28
Tunisia	750	2	0.27
UAE	1,070	7	0.65
Yemen	851	70	8.23
Total	21,000	504	2.40

Table 3: Distribution of tweets manually labeled as Farsi in TRAIN. We only provide countries with Farsi tweets, and remove the rest of 21 countries in NADI.

Team	Affiliation	Tasks
Mawdoo3 AI (Talafha et al., 2020)	Mawdoo3 AI, Jordan	1
BERT_NGRAMS (El Mekki et al., 2020)	Mohammed VI Polytechnic University, Morocco	1,2
ArabicProcessors (Gaanoun and Benelallam, 2020)	Institut National de Statistique et d'Economie, Morocco	1,2
Tri-directional (Beltagy et al., 2020)	Faculty of Engineering, Alexandria University, Egypt	1
MMZ (Mansour et al., 2020)	Faculty of Engineering, Alexandria University, Egypt	1
QMUL Team (Aloraini et al., 2020)	Queen Mary University of London, United Kingdom	1
Code Lyoko (Tahssin et al., 2020)	Faculty of Engineering, Alexandria University, Egypt	1
TRY_NLP (Balaji and Bharathi, 2020)	SSN College of Engineering, India	1,2
Sorbonne (Ghoul and Lejeune, 2020)	Sorbonne Université, France	1
Speech Translation (Lichouri and Abbas, 2020)	CRSTDLA Research Center, Algeria	1
LTG-ST (Touileb, 2020)	University of Oslo, Norway	1
Alexa (Bni Younes et al., 2020)	Jordan University of Science and Technology, Jordan	1
Alpha (AlShenaifi and Azmi, 2020)	King Saud University, Saudi Arabia	1
IRAQ (Aliwy et al., 2020)	University of Kufa, Iraq	1

Table 4: List of the 14 teams that participated in Subtasks 1 and 2 *and* submitted description papers.

5.2 Participating Teams

We received a total of 61 unique team registrations, among which 7 teams registered to participate in Subtask 1 only, 1 team registered to participate in Subtask 2 only, and 53 teams registered to participate in both subtasks. After evaluation phase, we received 47 submissions for Subtask 1 from 18 teams and 9 submissions for subtask 2 from 4 teams. Of participating teams, a total of 15 teams submitted description papers all of which except one were accepted for publication. Table 5.2 lists the 14 teams whose papers were accepted.

5.3 Shared Task Results

Table 5 presents the best TEST results for all 18 teams who submitted systems for Subtask 1, regardless of whether they have submitted a paper. Based on the official metric, $macro-F_1$, Mawdoo3-AI obtained the best performance with 26.78% F_1 score. Table 6 presents the best TEST results of each of the 4 teams who submitted systems to Subtask 2. Team BERT-NGRAMS achieved the best F_1 score that is 6.39%. [8]

5.4 General Description of Submitted Systems

In Table 7, we provide a high-level description of the systems submitted to each subtask. For each team, we list the overall number of submissions per subtask, their overall best score, the features employed, the methods adopted/developed, and whether they have used the 10M unlabeled tweet dataset we provided

[8] The full sets of results for Subtask 1 and Subtask 2 are in Tables A2 and A3, respectively, in Appendix A.

Team	F1	Accuracy	Precision	Recall
Mawdoo3 AI	**26.78 (1)**	**42.86 (1)**	**32.52 (1)**	**25.19 (1)**
BERT_NGRAMS	25.99 (2)	39.66 (2)	30.26 (2)	24.85 (2)
Arabic Processors	23.26 (3)	38.34 (3)	27.17 (4)	22.43 (5)
Tri-directional	23.09 (4)	37.70 (5)	26.40 (5)	23.04 (4)
MMZ	22.58 (5)	38.28 (4)	24.28 (8)	23.36 (3)
QMUL Team	20.77 (6)	34.32 (11)	21.62 (13)	21.09 (6)
Code Lyoko	20.34 (7)	36.26 (8)	27.83 (3)	20.56 (8)
TRY_NLP	20.04 (8)	33.66 (15)	20.07 (14)	21.07 (7)
Sorbonne	18.80 (9)	36.54 (7)	24.87 (7)	18.05 (12)
Iktishaf	18.63 (10)	33.98 (14)	20.21 (15)	18.76 (9)
Speech Translation	18.27 (11)	36.68 (6)	23.75 (10)	18.06 (11)
LTG-ST	17.71 (12)	36.22 (9)	24.93 (6)	17.21 (13)
Alexa	17.29 (13)	34.16 (12)	22.09 (12)	16.81 (15)
NAYEL	16.84 (14)	30.98 (18)	17.88 (16)	18.20 (10)
DNLP	16.50 (15)	31.28 (17)	17.84 (17)	17.04 (14)
NLPRL	15.77 (16)	35.06 (10)	23.96 (9)	15.92 (16)
Alpha	15.10 (17)	34.02 (13)	22.34 (11)	14.71 (17)
Our Baseline II	13.32	32.38	14.57	14.69
IRAQ	12.45 (18)	31.60 (16)	16.39 (18)	12.67 (18)
Our Baseline I	1.71	21.84	1.04	4.76

Table 5: Results for Subtask 1. The numbers in parentheses are the ranks. The table is sorted on the $macro - F1$ score, the official metric. Some teams did not submit description papers.

Team	F1	Accuracy	Precision	Recall
BERT_NGRAMS	**6.39 (1)**	6.50 (2)	**7.84 (1)**	6.54 (2)
Arabic Processors	5.75 (2)	**6.80 (1)**	6.78 (2)	**6.74 (1)**
NAYEL	4.99 (3)	5.22 (3)	5.52 (3)	5.17 (3)
TRY_NLP	4.03 (4)	4.86 (4)	3.74 (4)	4.68 (4)
Our Baseline II	2.13	3.32	4.04	3.22
Our Baseline I	0.03	1.92	0.02	1.00

Table 6: Results for Subtask 2. The numbers in parentheses are the ranks. Table is sorted on the $macro - F1$ score, the official metric. Team NAYEL does not have a description paper.

to all teams. As can be seen from the table, the majority of the top teams have (1) used Transformers, (2) exploited the unlabeled data for further pre-training, and/or (3) have used self-training to enhance their models. The rest of participating teams have either used a type of neural networks other than Transformers or resorted to linear machine learning models, usually with some form of ensembling.

6 Conclusion and Future Work

We presented an overview of the NADI 2020 shared task. We described the dataset and identified areas of improvement especially related to its collection. We also provided a high-level description of participating teams. The number of submissions to the shared task reflects an interest in the community and calls for further work in the area of Arabic dialect detection, but also more generally Arabic dialect processing.

In the future, we plan to host a second iteration of the NADI shared task that will use new datasets and pursue a number of novel questions inspired by the issues discovered in this year's task. For example, in addition to DA classification, we will propose *MSA regional use classification* as a subtask. Since MSA is shared across the Arab world, we hypothesize this will be a more challenging task than DA classification. We will also encourage teams to experiment with studying the interaction between MSA and DA in

Team	# Submissions	F_1	Features					Methods					Uses Unlabelled 10M		
			N-grams	TF-IDF	Word Embed.	Topic Models	Sampling	Classical ML	Neural Nets	Transformer	Ensemble	Hierarchical	Pre-Training	Data Augment.	Self-Training
SUBTASK 1															
Mawdoo3 AI	3	26.78	✓	✓	✓		✓	✓	✓	✓	✓		✓		
BERT_NGRAMS	4	25.99	✓	✓				✓		✓	✓			✓	
Arabic Processors	3	23.26	✓	✓				✓		✓	✓				✓
Tri-directional	1	23.09					✓	✓		✓		✓	✓		
MMZ	2	22.58	✓	✓				✓		✓	✓		✓		
QMUL Team	2	20.77	✓	✓	✓	✓			✓				✓		
Code Lyoko	2	20.34	✓	✓	✓			✓	✓	✓			✓		
TRY_NLP	3	20.04	✓	✓	✓				✓	✓					✓
Sorbonne	3	18.80	✓	✓	✓			✓	✓		✓		✓		
Speech Translation	3	18.27	✓	✓			✓	✓			✓				
LTG-ST	2	17.71	✓	✓				✓			✓				
Alexa	3	17.29	✓	✓				✓			✓				
Alpha	5	15.10	✓	✓				✓							
IRAQ	3	12.45	✓	✓				✓			✓				
SUBTASK 2															
BERT_NGRAMS	3	6.39	✓	✓				✓	✓	✓	✓	✓		✓	
Arabic Processors	1	5.75	✓	✓				✓	✓	✓					✓
TRY_NLP	2	4.03	✓	✓	✓				✓	✓					✓

Table 7: Summary of approaches used by participating teams in Subtasks 1 and 2. Classical ML refers to any non-neural machine learning methods such as naive Bayes and support vector machines. The term "neural nets" refers to any model based on neural networks (e.g., RNN, CNN) except Transformer models. Transformer refers to neural networks based on a Transformer architecture such as BERT. The table is sorted by official metric , $macro - F_1$. We only list teams that submitted a description paper.

novel ways. For example, questions as to the utility of using DA data to improve MSA regional use classification systems and vice versa can be investigated exploiting various machine learning methods.

Acknowledgments

We gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), the Social Sciences Research Council of Canada (SSHRC), and Compute Canada (www.computecanada.ca). We thank AbdelRahim Elmadany for assisting with dataset preparation, setting up the Codalab for the shared task, and providing the map in Figure 2.

References

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

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of Arabic dialects. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020. Toward micro-dialect identification in diaglossic and code-switched environments. *arXiv preprint arXiv:2010.04900*.

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

Nora Al-Twairesh, Rawan Al-Matham, Nora Madi, Nada Almugren, Al-Hanouf Al-Aljmi, Shahad Alshalan, Raghad Alshalan, Nafla Alrumayyan, Shams Al-Manea, Sumayah Bawazeer, Nourah Al-Mutlaq, Nada Al-manea, Waad Bin Huwaymil, Dalal Alqusair, Reem Alotaibi, Suha Al-Senaydi, and Abeer Alfutamani. 2018. SUAR: Towards building a corpus for the Saudi dialect. In *Proceedings of the International Conference on Arabic Computational Linguistics (ACLing)*.

Ahmed Aliwy, Hawraa Taher, and Zena AboAltaheen. 2020. Arabic Dialects Identification for All Arabic countries. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

Abdulrahman Aloraini, Ayman Alhelbawy, and Massimo Poesio. 2020. The QMUL/HRBDT contribution to the NADI Arabic Dialect Identification Shared Task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

Nouf AlShenaifi and Aqil Azmi. 2020. Faheem at NADI shared task: Identifying the dialect of Arabic tweet. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

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

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

Nitin Balaji and B. Bharathi. 2020. Semi-supervised Fine-grained Approach for Arabic Dialect Detection . In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

Ahmad Beltagy, Abdelrahman Abouelenin, and Omar ElSherief. 2020. Arabic Dialect Identification Using BERT-Based Domain Adaptation. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

Mutaz Bni Younes, Nour Al-Khdour, and Mohammad AL-Smadi. 2020. Team Alexa at NADI Shared Task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

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

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

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

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

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

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. COLABA: Arabic dialect annotation and processing. In *LREC workshop on Semitic language processing*, pages 66–74.

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

Abdellah El Mekki, Ahmed Alami, Hamza Alami, Ahmed Khoumsi, and Ismail Berrada. 2020. Weighted combination of BERT and N-GRAM features for Nuanced Arabic Dialect Identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

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

Kamel Gaanoun and Imade Benelallam. 2020. Arabic dialect identification: An Arabic-BERT model with data augmentation and ensembling strategy. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

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

Dhaou Ghoul and Gael Lejeune. 2020. Comparison between Voting Classifier and Deep Learning methods for Arabic Dialect Identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

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

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

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

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

Mohamed Lichouri and Mourad Abbas. 2020. Simple vs Oversampling-based Classification Methods for Fine Grained Arabic Dialect Identification in Twitter. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

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.

Moataz Mansour, Moustafa Tohamy, Zeyad Ezzat, and Marwan Torki. 2020. Arabic Dialect Identification Using BERT Fine-Tuning. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

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

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

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

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

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

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

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

Rawan Tahssin, Youssef Kishk, and Marwan Torki. 2020. Identifying Nuanced Dialect for Arabic Tweets with Deep Learning and Reverse Translation Corpus Extension System. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

Bashar Talafha, Mohamed Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein T. AL-NATSHEH. 2020. Multi-dialect Arabic BERT for Country-level Dialect Identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

Samia Touileb. 2020. LTG-ST at NADI Shared Task 1: Arabic Dialect Identification using a Stacking Classifier. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

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

Omar F Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, Organization = Association for Computational Linguistics*, pages 37–41.

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

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

Appendices

A Appendix

Province Name	# of Tweets			Province Name	# of Tweets		
	train	dev	test		train	dev	test
Damascus City	214	53	52	Minya	213	53	52
Ariana	212	52	52	BBordj Bou Arreridj	213	53	52
Asyut	213	53	52	Hawalli	210	51	51
Marrakech-Tensift-Al Haouz	214	53	52	Al Butnan	214	53	53
Bouira	213	53	52	Abu Dhabi	214	53	52
Ash Sharqiyah	32	10	8	Kairouan	114	8	52
Khenchela	213	53	52	Banaadir	210	51	51
As-Sulaymaniyah	213	53	52	Ar Riyad	213	54	52
Ad Dakhiliyah	213	53	52	Baghdad	213	53	52
Fujairah	214	53	52	Djibouti	210	10	51
An-Najaf	213	53	52	Musandam	213	27	52
Oriental	214	37	52	Muscat	213	53	52
Ibb	213	50	52	Doha	24	52	10
Al Quassim	213	53	52	Khartoum	210	51	51
Qena	213	53	52	Aswan	213	53	52
Sousse	212	52	52	North Sinai	213	53	52
As-Suwayda	214	53	52	Ash Sharqiyah	395	97	96
Ha'il	213	54	52	Beni Suef	213	53	52
Jizan	213	53	52	Tabuk	213	53	52
Jijel	213	53	52	Tripoli	214	53	53
Mahdia	212	52	52	Béchar	213	41	52
Ismailia	213	53	52	Najran	213	53	52
Meknes-Tafilalet	214	53	52	West Bank	210	51	51
Wasit	213	53	52	Alexandria	213	53	52
Gaza Strip	210	51	51	Dhofar	213	53	52
Kafr el-Sheikh	213	53	52	Capital	210	8	20
Nouakchott	210	40	5	Misrata	214	53	53
Gharbia	213	53	52	Aqaba	213	52	52
Al Anbar	213	53	52	Cairo	213	10	52
Arbil	213	53	52	North Lebanon	213	52	52
Akkar	213	6	52	South Lebanon	213	52	52
Makkah	213	54	52	Faiyum	213	53	52
Hims	214	53	52	Souss-Massa-Draa	214	53	52
Benghazi	214	53	53	Beheira	213	53	52
Lattakia	214	53	52	Al Jabal al Akhdar	214	53	53
Port Said	213	53	52	Ouargla	213	53	52
Oran	213	53	52	Monufia	213	53	52
Aden	213	52	52	Sohag	213	53	52
Al Madinah	213	54	52	Al Batnah	214	53	52
Red Sea	213	53	52	Dubai	214	53	52
Karbala	213	53	52	Maysan	213	53	52
Zarqa	213	52	52	Ninawa	213	53	52
Basra	213	53	52	Al Hudaydah	213	52	52
Suez	213	53	52	Jahra	210	19	51
Dakahlia	213	53	52	Al-Muthannia	213	53	52
South Sinai	213	53	52	Ras Al Khaymah	214	53	52
Umm Al Qaywayn	214	53	5	Dihok	213	53	52
Aleppo	214	53	52	Ar Rayyan	210	52	51
Tanger-Tetouan	214	53	52	Luxor	213	53	52
Asir	213	54	52	Dhamar	212	52	23

Table A1: Distribution of the NADI data over provinces, by country, across our TRAIN, DEV, and TEST splits.

Team Name	F1	Accuracy	Precision	Recall
Mawdoo3 AI	**26.78 (1)**	42.86 (2)	**32.52 (1)**	25.19 (2)
Mawdoo3 AI	26.77 (2)	42.56 (3)	31.51 (4)	**25.45 (1)**
Mawdoo3 AI	26.47 (3)	**43.18 (1)**	31.59 (3)	25.12 (3)
BERT_NGRAMS	25.99 (4)	39.66 (5)	30.26 (6)	24.85 (4)
BERT_NGRAMS	25.99 (4)	39.66 (5)	30.26 (6)	24.85 (4)
BERT_NGRAMS	25.02 (5)	38.92 (6)	30.92 (5)	23.81 (5)
BERT_NGRAMS	23.83 (6)	40.88 (4)	32.50 (2)	23.36 (7)
Arabic Processors	23.26 (7)	38.34 (8)	27.17 (9)	22.43 (9)
Tri-directional	23.09 (8)	37.70 (10)	26.40 (11)	23.04 (8)
Arabic Processors	23.03 (9)	38.42 (7)	27.40 (8)	22.40 (10)
MMZ	22.58 (10)	38.28 (9)	24.28 (15)	23.36 (6)
MMZ	22.58 (10)	38.28 (9)	24.28 (15)	23.36 (6)
Arabic Processors	22.52 (11)	38.28 (9)	26.70 (10)	22.12 (11)
QMUL team	20.77 (12)	34.32 (22)	21.62 (28)	21.09 (12)
Code Lyoko	20.34 (13)	36.26 (13)	27.83 (7)	20.56 (15)
TRY_NLP	20.04 (14)	33.66 (27)	20.70 (29)	21.07 (13)
TRY_NLP	20.01 (15)	33.58 (28)	20.66 (30)	21.03 (14)
TRY_NLP	19.84 (16)	34.80 (20)	20.54 (31)	20.17 (16)
QMUL team	19.45 (17)	33.74 (26)	20.40 (33)	19.84 (17)
Sorbonne	18.80 (18)	36.54 (12)	24.87 (14)	18.05 (21)
Iktishaf	18.63 (19)	33.98 (25)	20.21 (34)	18.76 (18)
Speech Translation	18.27 (20)	36.68 (11)	23.75 (20)	18.06 (20)
Speech Translation	17.90 (21)	35.68 (16)	22.40 (23)	17.64 (23)
Iktishaf	17.84 (22)	33.48 (29)	19.07 (36)	17.98 (22)
Sorbonne	17.77 (23)	35.44 (18)	23.79 (19)	17.15 (26)
LTG-ST	17.71 (24)	36.22 (15)	24.93 (13)	17.21 (25)
Speech Translation	17.69 (25)	36.24 (14)	22.17 (25)	17.41 (24)
Alexa	17.29 (26)	34.16 (23)	22.09 (26)	16.81 (29)
Alexa	17.20 (27)	35.64 (17)	23.53 (21)	16.86 (28)
NAYEL	16.84 (28)	30.98 (41)	17.88 (38)	18.20 (19)
LTG-ST	16.81 (29)	34.78 (21)	23.90 (18)	16.46 (31)
DNLP	16.50 (30)	31.28 (39)	17.84 (39)	17.04 (27)
DNLP	16.27 (31)	31.24 (40)	17.66 (40)	16.77 (30)
Sorbonne	16.06 (32)	31.90 (36)	22.00 (27)	15.90 (34)
NAYEL	15.81 (33)	32.22 (35)	17.91 (37)	16.01 (32)
NLPRL	15.77 (34)	35.06 (19)	23.96 (17)	15.92 (33)
Alpha	15.10 (35)	34.02 (24)	22.34 (24)	14.71 (39)
Alexa	15.09 (36)	34.78 (21)	23.00 (22)	15.68 (35)
Alpha	14.91 (37)	32.80 (32)	25.43 (12)	14.30 (42)
Alpha	14.72 (38)	33.00 (30)	20.44 (32)	14.61 (40)
Alpha	14.61 (39)	32.24 (34)	17.30 (41)	14.81 (37)
NAYEL	14.37 (40)	29.42 (42)	15.32 (45)	14.90 (36)
Alpha	14.27 (41)	32.84 (31)	24.15 (16)	14.33 (41)
Sorbonne	14.21 (42)	32.38 (33)	19.13 (35)	14.25 (43)
Code Lyoko	13.57 (43)	31.70 (37)	15.32 (44)	14.74 (38)
IRAQ	12.45 (44)	31.60 (38)	16.39 (42)	12.67 (44)
IRAQ	12.20 (45)	31.28 (39)	15.76 (43)	12.45 (45)

Table A2: Full results for Subtask 1. The numbers in parentheses are the ranks. The table is sorted on the $macro\ F_1$ score, the official metric.

Team Name	F1	Accuracy	Precision	Recall
BERT_NGRAMS	**6.39 (1)**	6.50 (2)	7.84 (2)	6.54 (2)
BERT_NGRAMS	6.08 (2)	6.16 (3)	7.78 (3)	6.03 (3)
Arabic Processors	5.75 (3)	**6.80 (1)**	6.78 (4)	**6.74 (1)**
BERT_NGRAMS	5.42 (4)	5.32 (4)	**8.00 (1)**	5.24 (4)
NAYEL	4.99 (5)	5.22 (5)	5.52 (5)	5.17 (5)
NAYEL	4.28 (6)	4.48 (8)	4.34 (6)	4.69 (6)
TRY_NLP	4.03 (7)	4.86 (6)	3.74 (9)	4.68 (7)
TRY_NLP	3.94 (8)	4.54 (7)	3.86 (7)	4.45 (8)
NAYEL	3.60 (9)	3.84 (9)	3.83 (8)	3.85 (9)

Table A3: Full results for Subtask 2. The numbers in parentheses are the ranks. The table is sorted on the $macro\ F_1$ score, the official metric.