NADI 2022:
The Third Nuanced Arabic Dialect Identification Shared Task

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

Abstract

We describe the findings of the third Nuanced Arabic Dialect Identification Shared Task (NADI 2022). NADI aims at advancing state-of-the-art Arabic NLP, including Arabic dialects. It does so by affording diverse datasets and modeling opportunities in a standardized context where meaningful comparisons between models and approaches are possible. NADI 2022 targeted both dialect identification (Subtask 1) and dialectal sentiment analysis (Subtask 2) at the country level. A total of 41 unique teams registered for the shared task, of whom 21 teams have participated (with 105 valid submissions). Among these, 19 teams participated in Subtask 1, and 10 participated in Subtask 2. The winning team achieved F_1=27.06 on Subtask 1 and F_1=75.16 on Subtask 2, reflecting that both subtasks remain challenging and motivating future work in this area. We describe the methods employed by the participating teams and offer an outlook for NADI.

1 Introduction

Arabic is a collection of languages and language varieties some of which are not mutually intelligible, although it is sometimes conflated as a single language. *Classical Arabic (CA)* is the variety used in old Arabic poetry and the Qur'an, the Holy Book of Islam. CA continues to be used to date, side by side with other varieties, especially in religious and literary discourses. CA is also involved in code-switching contexts with *Modern Standard Arabic (MSA)* (Abdul-Mageed et al., 2020b). In contrast, as its name suggests, MSA is a more modern variety (Badawi, 1973) of Arabic. MSA is usually employed in pan-Arab media such as AlJazeera network and in government communication across the Arab world.[1] *Dialectal Arabic (DA)* is the term used to collectively refer to

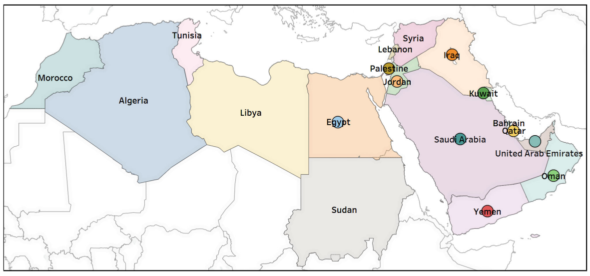

Figure 1: A map of the Arab World showing the 18 countries in the *Subtask 1* dataset and the 10 countries in the *Subtask 2* dataset. Each country is coded in a color different from neighboring countries. Subtask 2 countries are coded as circles with dark color.

Arabic dialects. DA is sometimes defined regionally into categories such as Gulf, Levantine, Nile Basin, and North African (Habash, 2010; Abdul-Mageed, 2015). More recent treatments of DA focus on more nuanced variation at the country or even sub-country levels (Bouamor et al., 2018; Abdul-Mageed et al., 2020b). Many of the works on Arabic dialects thus far have focused on dialect identification, the task of automatically detecting the source variety of a given text or speech segment.

In this paper, we introduce the findings and results of the third Nuanced Arabic Dialect Identification Shared Task (NADI 2022). NADI aims at encouraging research work on Arabic dialect processing by providing datasets and diverse modeling opportunities under a common evaluation setup. The first instance of the shared task, NADI 2020 (Abdul-Mageed et al., 2020a), focused on province-level dialects. NADI 2021 (Abdul-Mageed et al., 2021b), the second iteration of NADI, focused on distinguishing both MSA and DA according to their geographical origin at the country level. NADI 2022 extends on both editions and offers a richer context as it targets both Arabic dialect identification and *and* dialectal sentiment analysis.

NADI 2022 shared tasks proposes two subtasks:

[1]https://www.aljazeera.com/

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 85 - 97
December 8, 2022 ©2022 Association for Computational Linguistics

Subtask 1 on dialect identification, and **Subtask 2** on dialect sentiment analysis. While we invited participation in either of the two subtasks, we encouraged teams to submit systems to *both* subtasks. By offering two subtasks, our hope was to receive systems that exploit diverse machine learning and other methods and architectures such as multi-task learning systems, ensemble methods, sequence-to-sequence architectures in single models such as the text-to-text Transformer, etc. Many of the submitted systems investigated diverse approaches, thus fulfilling our objective.

A total of 41 unique teams registered for NADI 2022. Of these, 21 unique teams actually made submissions to our leaderboard (n=105 valid submissions). We received 16 papers from 15 teams, of which we accepted 15 for publication. Results from participating teams show that both dialect identification at the country level and dialectal sentiment analysis from short sequences of text remain challenging even to complex neural methods. These findings clearly motivate future work on both tasks.

The rest of the paper is organized as follows: Section 2 provides a brief overview of Arabic dialect identification and sentiment analysis. We describe the two subtasks and NADI 2022 restrictions in Section 3. Section 4 introduces shared task datasets and evaluation setup. We present participating teams and shared task results and provide a high-level description of submitted systems in Section 5. We conclude in Section 6.

2 Literature Review

2.1 Arabic Dialects

Arabic can be categorized into CA, MSA, and DA. Although CA and MSA have been studied extensively (Harrell, 1962; Cowell, 1964; Badawi, 1973; Brustad, 2000; Holes, 2004), DA is has received more attention only in recent years. One major challenge for studying DA has been the lack of resources. For this reason, most pioneering DA works focused on creating resources, usually for only a small number of regions or countries (Gadalla et al., 1997; Diab et al., 2010; Al-Sabbagh and Girju, 2012; Sadat et al., 2014; Harrat et al., 2014; Jarrar et al., 2016; Khalifa et al., 2016; Al-Twairesh et al., 2018; El-Haj, 2020). A number of works introducing multi-dialectal datasets and regional level detection models followed (Zaidan and Callison-Burch, 2011; Elfardy et al., 2014; Bouamor et al., 2014; Meftouh et al., 2015).

Some of the earliest Arabic dialect identification shared tasks were offered as part of the VarDial workshop. These shared tasks used speech broadcast transcriptions (Malmasi et al., 2016), and later integrated acoustic features (Zampieri et al., 2017) and phonetic features (Zampieri et al., 2018) extracted from raw audio.

The Multi-Arabic Dialects Application and Resources (MADAR) project (Bouamor et al., 2018) was the first that introduced finer-grained dialectal data and a lexicon. The MADAR data was used for dialect identification at the country and city levels covering 25 cities in the Arab world (Salameh et al., 2018; Obeid et al., 2019). The MADAR data was commissioned rather than being naturally occurring, which might not be the best for dialect identification, especially when considering dialect identification in the social media context. Several larger datasets covering 10-21 countries were then introduced (Mubarak and Darwish, 2014; Abdul-Mageed et al., 2018; Zaghouani and Charfi, 2018; Abdelali et al., 2021; Issa et al., 2021; Baimukan et al., 2022). These datasets were mainly compiled from naturally-occurring posts on social media platforms such as Twitter. Some approaches for collecting dialectal data are unsupervised. A recent example is Althobaiti (2022) who describe an approach for automatically tagging Twitter posts with 15 country-level dialects and extracting relevant word lists. Some works also gather data at the fine-grained level of cities. For example, Abdul-Mageed et al. (2020b) introduced a Twitter dataset and a number of models to identify country, province, and city level variation in Arabic dialects. The NADI shared task (Abdul-Mageed et al., 2020a, 2021b) built on these efforts by providing datasets and common evaluation settings for identifying Arabic dialects. Althobaiti (2020) is a relatively recent survey of computational work on Arabic dialects.

2.2 Sentiment Analysis

Besides dialect identification, several studies investigate socio-pragmatic meaning (SM) exploiting Arabic data. SM refers to intended meaning in real-world communication and how utterances should be interpreted within the social context in which they are produced (Thomas, 2014; Zhang et al., 2022). Typical SM tasks include sentiment analysis (Abdul-Mageed et al., 2014; Abdul-Mageed, 2019), emotion recognition (Alhuzali et al., 2018), age and gender identification (Abbes et al., 2020), offensive language detection (Mubarak et al., 2020; Elmadany et al., 2020), and sarcasm detection (Abu Farha and Magdy, 2020). In NADI 2022, we focus on sentiment analysis of Arabic dialects in social media. Several studies of Arabic sentiment analysis are listed in surveys such as Elnagar et al. (2021) and Alhumoud and Wazrah (2022). Most of these studies target sentiment in MSA. Recently, there are some studies that target sentiment in Arabic dialects in social media sources such Twitter. Some of these studies create datasets (Guellil et al., 2020a; Al-Laith et al., 2021; Abo et al., 2021; Alowisheq et al., 2021; Hassan et al., 2021; Alwakid et al., 2022), focusing on one or more dialects or regions (Abdul-Mageed et al., 2020c; Fourati et al., 2020; Guellil et al., 2020b; Almuqren and Cristea, 2021; Guellil et al., 2021; Abu Farha and Magdy, 2021; Shamsi and Abdallah, 2022). Many of the previous sentiment analysis works, however, either do not distinguish dialects altogether or focus only on a few dialects such as Egyptian, Levantine, or Tunisian. This motivates us to introduce the dialectal sentiment analysis subtask as part of NADI 2022.

To the best of our knowledge, our work is the first to enable investigating sentiment analysis in 10 Arabic dialects. For our sentiment analysis subtask, we also annotate and release a novel dataset and facilitate comparisons in a standardized experimental setting.

2.3 The NADI Shared Tasks

NADI 2020 The first NADI shared task, (Abdul-Mageed et al., 2020a) was co-located with the fifth Arabic Natural Language Processing Workshop (WANLP 2020) (Zitouni et al., 2020). NADI 2020 targeted both country- and province-level dialects. It covered a total of 100 provinces from 21 Arab countries, with data collected from Twitter. It was the first shared task to target naturally occurring fine-grained dialectal text at the sub-country level.

NADI 2021 The second edition of the shared task (Abdul-Mageed et al., 2021b) was co-located with WANLP 2021 (Habash et al., 2021). It targeted the same 21 Arab countries and 100 corresponding provinces as NADI 2020, also exploiting Twitter data. NADI 2021 improved over NADI 2020 in that non-Arabic data were removed. In addition, NADI-2021 teased apart the data into MSA and DA and focused on classifying MSA and DA tweets into the countries and provinces from which they are collected. As such, NADI 2021 had four subtasks: MSA-country, DA-country, MSA-province, and DA-province.

NADI 2022 As introduced earlier, **this current edition** of NADI focuses on studying Arabic dialects at the country level as well as dialectal sentiment (i.e., sentiment analysis of data tagged with dialect labels). Our objective is that NADI 2022 can support exploring variation in social geographical regions that have not been studied before. We discuss NADI 2022 in more detail in the next section.

It is worth noting that NADI shared task datasets are starting to be used for various types of (e.g., linguistic) studies of Arabic dialects, For example, Alsudais et al. (2022) studies the effect of geographic proximity on Arabic dialects exploiting datasets from MADAR (Bouamor et al., 2018) and NADI (Abdul-Mageed et al., 2020a, 2021b).

3 Task Description

3.1 Shared Task Subtasks

The NADI 2022 shared task consists of two subtasks, both focused on dialectal Arabic at the country level. **Subtask 1** is about dialect identification and **Subtask 2** is about sentiment analysis of Arabic dialects. We now introduce each subtask.

Subtask 1 (Dialect Identification) The goal of Subtask 1 is to identify the specific country-level dialect of a given Arabic tweet. For this subtask, we reuse the training, development, and test datasets of 18 countries from NADI 2021 (Abdul-Mageed et al., 2021b). In addition to the test set of NADI 2021, we introduce a *new* test set manually annotated with k country-level dialects, where $k = 10$ but is kept unknown to teams. We ask participants to submit system runs on these two test sets.

Subtask 2 (Dialectal Sentiment Analysis) The goal of Subtask 2 is to identify the sentiment of a given tweet written in Arabic. Tweets are collected from 10 different countries during the year of 2018 and involve both MSA and DA. The data are manually labeled with sentiment tags from the set {*positive*, *negative*, *neutral*}. More information about our data splits and evaluation settings for both Subtask 1 and Subtask 2 is given in Section 4.

Figure 1 shows the countries covered in NADI 2022 for both subtasks.

3.2 Shared Task Restrictions

We follow the same general approach to managing the shared task we adopted in NADI 2020 and NADI 2021. This includes providing participating teams with a set of restrictions that apply to all subtasks, and clear evaluation metrics. The purpose of our restrictions is to ensure fair comparisons and common experimental conditions. In addition, similar to NADI 2020 and 2021, our data release strategy and our evaluation setup through the CodaLab online platform facilitated competition management, enhanced timeliness of acquiring results upon system submission, and guaranteed ultimate transparency. Once a team registered in the shared task, we directly provided the registering member with the data via a private download link. We provided the data in the form of the actual tweets posted to the Twitter platform, rather than tweet IDs. This guaranteed comparison between systems exploiting identical data.

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

4 Shared Task Datasets and Evaluation

TWT-10 We collected $\sim 10K$ tweets covering 10 Arab countries (*Egypt*, *Iraq*, *Jordan*, *KSA*, *Kuwait*, *Oman*, *Palestine*, *Qatar*, *UAE*, and *Yemen*) via the Twitter API.[2] The tweets were collected during the year of 2018. We asked a total of three college-educated Arabic native speakers to annotate these tweets with three types of information: (1) *dialectness* (MSA vs. DA), (2) 10-*way country-level dialects*, and (3) *three-way sentiment labels* (i.e., $\{positive, negative, neutral\}$). For each of the 10 countries, 500 tweets were labeled by two different annotators. We calculated the inter-annotator agreement using Cohen's Kappa . We obtained a Kappa (K) of 0.85 for the sentiment labeling task and K of 0.41 for the 10-way dialect identification one. Table 1 also presents the distribution of dialect and sentiment classes. It also shows that MSA comprises 50.86% of TWT-10 (while DA is 49.14%). Table 2 shows tweet examples with sentiment labels randomly selected from a number of countries representing different regions in our annotated dataset.

Country	Dialect		Sentiment			Total
	MSA	DA	Pos	Neg	Neut	
Egypt	137	363	176	187	137	500
Iraq	314	186	230	219	51	500
Jordan	257	243	169	253	78	500
KSA	300	200	194	152	154	500
Kuwait	170	330	203	227	70	500
Oman	340	160	166	179	155	500
Palestine	248	252	159	169	172	500
Qatar	181	319	288	194	18	500
UAE	270	230	232	112	156	500
Yemen	326	174	118	198	184	500
Total	**2,543**	**2,457**	**1,935**	**1,890**	**1,175**	**5,000**

Table 1: The TWT-10 dataset class distributions.

Subtask 1 (Dialect Identification) We use the dataset of Subtask 1.2 of NADI 2021 (i.e., country-level DA) (Abdul-Mageed et al., 2021b). This dataset was collected using tweets covering 21 Arab countries during a period of 10 months (Jan. to Oct.) during the year of 2019. It was heuristically labelled exploiting the users' geo-location feature and mobility patterns and automatically cleaned to exclude non-Arabic and MSA tweets. For the purpose of this shared task, we keep the same training, development, and test splits as NADI 2021 but we exclude data from Djibouti, Somalia, and Mauritania since these are poorly represented in the dataset. We call the resulting dataset **TWT-GEO**. TWT-GEO includes 18 country-level dialects, split into **Train** ($\sim 20K$ tweets), **Dev** ($\sim 5K$ tweets), and **Test-A** ($\sim 4.8K$ tweets). We refer to the test set of TWT-GEO as Test-A since we use an additional test split for evaluation, **Test-B**. Test-B contains 1.5K dialect tweets randomly sampled from the TWT-10 dataset described earlier. Table 3 presents the class distributions in Subtask 1 Train, Dev, and Test splits (Test-A and Test-B).

[2] https://developer.twitter.com/en/docs/twitter-api

Country	Example	Sentiment
Egypt	محدش عنده فكة وجع !!	Negative
	مرتضى : معروف و كوفي ملهمش مليم عندي و من يمتلك عرضا يغور	Neutral
	كوباية شاي خيالية لدرجة أني عايز آكل الكوباية الفاضية تسلم أيدي اصلن	Positive
Jordan	وييننننك يا زووجي توخدني عبارريس	Negative
	منرجع منحكي: التعداد السكاني والمساكن، بدها تعرف الدولة وين ساكن	Neutral
	شفتو حق جلسات وناسة وش سوا هههههههاااااااي	Positive
KSA	لما تبي تحاكي احد بس ماتدري شتقول	Negative
	مادري ليش آذآ شفت احد نآيم احسه مسكين !!	Neutral
	الفجر يقيم خمس وعشر عشر ساعات يمديك توصل الشرقية تغطّ رجولك بالبحر وتتعشا وترجع	Positive
Kuwait	كالعاده الخشوف طاحسين بالتايم والسحوت يرضعون.	Negative
	احدٍ اذا مديتله طيبك .. يغيب ماتملي عيونه كبار الفعايل واحدٍ يموت إن كان سويت به طيب يخاف مايقدر يرد الجمايل	Neutral
	اي والله أغليه و نفسي فيه مفتونه.	Positive
UAE	ذا كنت تحبني صج ومن كل قلبك تعشقني ، حسسني بهالشي خلني أحس اني اهمك صج مابي كم كلمه حلوه وانتا طول يومك مو معاي ومشغول بغيري	Negative
	اقولك تدري منو سوالي فولو	Neutral
	يبيله احلى بوسه على الصبح و لحس الخبق	Positive

Table 2: Randomly picked dialectal tweets from select countries in our annotated data for Subtask 2.

Country	TRAIN	DEV	TEST-A	TEST-B
Algeria	1,809	430	379	—
Bahrain	215	52	50	—
Egypt	4,283	1,041	1,025	219
Iraq	2,729	664	648	117
Jordan	429	104	101	144
KSA	2,140	520	501	116
Kuwait	429	105	103	202
Lebanon	644	157	119	—
Libya	1,286	314	309	—
Morocco	8,58	207	210	—
Oman	1,501	355	360	91
Palestine	428	104	99	160
Qatar	215	52	51	190
Sudan	215	53	53	—
Syria	1,287	278	279	—
Tunisia	859	173	211	—
UAE	642	157	157	136
Yemen	429	105	103	99

Table 3: Distribution of classes for Subtask 1 data.

Subtask 2 (Sentiment Analysis) For this subtask, we use the manually annotated 5,000 tweets (including both MSA and dialects) in TWT-10. We randomly split the tweets into **Train** (1,500 tweets), **Dev** (500 tweets), and **Test** (3,000 tweets). We intentionally provide a small training dataset to encourage various approaches (e.g., *few-shot* learning). Figure 2 shows the distribution of sentiment classes across the data splits.

Unlabeled Dataset We provide participants with a total of 10M unlabeled Arabic tweets in the form of tweet IDs. We refer to this collection as **UNLABELED-10M**. We collected these tweets in 2019. In UNLABELED-10M, Arabic was identified using Twitter language tag (`ar`). We included in our data package released to participants a simple script to collect these tweets. Participants were free to use UNLABELED-10M for any of the two subtasks.[3]

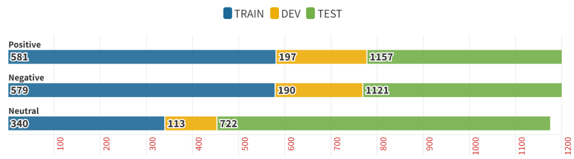

Figure 2: Subtask 2 class distributions across data splits.

Evaluation Metrics The official evaluation metric for **Subtask 1** is Macro-Averaged F_1-score. We evaluate on Test-A and Test-B separately, and use the average score between these two test sets as the final score of Subtask 1. For **Subtask 2**, F_{NP}-score is the official metric, where we use the average of the F_1 scores of the *positive* and *negative* classes only while neglecting the neutral class. These metrics are 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

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

Team	Affiliation	Tasks
259 (Qaddoumi, 2022)	New York University, USA	1
Ahmed and Khalil (El-Shangiti and Mrini, 2022)	Independent Researcher, Morocco	1, 2
ANLP-RG (Fsih et al., 2022)	Faculty of Economics and Management of Sfax, Tunisia	2
BFCAI (Sobhy et al., 2022)	Benha University, Egypt	1
BhamNLP	King Abdulaziz University, KSA and Uni. of Birmingahm, UK	2
Elyadata	ELYADATA, Tunisia	1
Giyaseddin (Bayrak and Issifu, 2022)	Marmara University, Turkey	1, 2
GOF (Jamal et al., 2022)	University of Windsor, Canada	1
iCompass (Messaoudi et al., 2022)	iCompass, Tunisia	1
ISL-AAST	Arab academy for science and technology, Egypt	1, 2
MTU_FIZ (Shammary et al., 2022)	Munster Technological University, Ireland	1
NLP_DI (Kanjirangat et al., 2022)	Dalle Molle Institute for AI, Switzerland	1
Oscar_Garibo	Valencian International University, Spain	1, 2
Pythoneers (Attieh and Hassan, 2022)	Aalto University, Finland	1, 2
rematchka (Abdel-Salam, 2022)	Cairo University, Egypt	1, 2
RUTeam	Reichman University, Israel	1, 2
SQU (AAlAbdulsalam, 2022)	Sultan Qaboos University, Oman	1
SUKI (Jauhiainen et al., 2022)	University of Helsinki, Finland	1
UniManc (Khered et al., 2022)	The University of Manchester, UK	1, 2
XY (AlShenaifi and Azmi, 2022)	Kind Saud University, KSA	1
zTeam	British University in Dubai, UAE	1

Table 4: List of teams that participated in either one or the two of subtasks. Teams with accepted papers are cited.

up to five runs for each test set of a given subtask, and only the highest scoring run was kept for each team. Although official results are based only on a blind test set, we also asked participants to report their results on the Dev sets in their papers. We set up two CodaLab competitions for scoring participant systems.[4] We plan to keep the Codalab competition for each subtask live post competition for researchers who would be interested in training models and evaluating their systems using the shared task blind test sets. For this reason, we will not release labels for the test sets of any of the subtasks.

5 Shared Task Teams & Results

5.1 Participating Teams

We received a total of 41 unique team registrations. After the testing phase, we received a total of 105 valid submissions from 21 unique teams. The breakdown across the subtasks is as follows: 42 submission for Test-A of Subtask 1 from 19 teams, 41 submissions for Test-B of Subtask 1 from 19 teams, 22 submissions for Subtask 2 from 10 teams. Table 4 lists the 21 teams. A total of 15 teams submitted a total of 16 description papers from which we accepted 15 papers for publication. Accepted papers are given in Table 4.

[4]The different CodaLab competitions are available at the following links: Subtask 1; Subtask 2.

5.2 Baselines

We provide three baselines for each of the two subtasks. **Baseline-I** is based on the majority class in the Train data for each subtask. For Subtask 1, Baseline-I performs at F_1=1.97 on Test-A and F_1=2.59 on Test-B, hence it obtains an average F_1 of 2.28. For Subtask 2, Baseline-I performs at F_{NP}=27.83. **Baseline-mBERT**, **Baseline-XLMR**, and **Baseline-MARBERT** are fine-tuned multilingual BERT-Base model (mBERT) (Devlin et al., 2019), cross-lingual RoBERTa (XLMR) (Conneau and Lample, 2019), and MARBERT (Abdul-Mageed et al., 2021a), respectively. More specifically, we take checkpoints for these models from Hugginface Library (Wolf et al., 2020) and fine-tune each of them for 20 epochs with a learning rate of $2e$-5 and batch size of 32. The maximum length of input sequence is set to 64 tokens. We evaluate each model at the end of each epoch and choose the best model based on performance on the respective Dev set. We then report performance of the best model on test sets. Baseline-MARBERT is our strongest baseline: it obtains F_1=31.39 on Test-A of Subtask 1, F_1=16.94 on Test-B of Subtask 1, average F_1=24.17 over Test-A and Test-B, and F_{NP}=72.36 on Subtask 2.

5.3 Shared Task Results

Table 5 presents the leaderboard of Subtask 1 and is sorted by the main metric of Subtask 1, i.e., average macro-F_1 score. As Tables 6 and 7 show, for each

Team	Avg. Macro-F_1
1 rematchka	27.06
2 UniManc	26.86
3 GOF	26.44
4 mtu_fiz	25.50
5 iCompass	25.32
6 ISL-AAST	24.59
7 Ahmed_and_Khalil	24.35
Baseline-MARBERT	24.17
8 Pythoneers	24.12
9 Giyaseddin	22.42
10 SQU	22.42
11 Elyadata	22.41
12 NLP_DI	21.28
13 RUTeam	17.28
14 259	16.89
15 zTeam	16.12
16 XY	15.80
Baseline-mBERT	15.70
17 BFCAI	15.48
18 SUKI	15.11
Baseline-XLMR	14.68
19 Oscar_Garibo	14.45
Baseline-I	2.28

Table 5: Results for Subtask 1 (Country-Level DA).

Team	Macro-F_1	Acc	Rec	Prec
1 rematchka	36.48	53.05	35.22	41.89
2 GOF	35.68	52.10	34.91	39.18
3 UniManc	34.78	52.33	34.74	38.74
4 iCompass	33.70	51.91	33.71	35.86
5 mtu_fiz	33.32	51.18	32.42	38.87
6 Pythoneers	32.63	48.91	31.77	36.77
7 ISL_AAST	32.24	50.27	32.07	37.53
8 Ahmed_and_Khalil	31.54	50.34	32.04	34.00
Baseline-MARBERT	31.39	47.77	31.01	35.53
9 Giyaseddin	30.55	47.65	30.04	34.18
10 SQU	30.01	46.85	29.75	34.57
11 Elyadata	29.35	45.84	28.60	31.27
12 NLP_DI	26.12	42.08	25.75	28.29
13 RUTeam	23.20	36.61	22.84	24.00
14 XY	22.36	39.85	21.33	30.52
15 259	21.93	34.11	22.69	22.32
16 zTeam	21.76	39.43	20.77	27.25
17 BFCAI	21.25	38.63	20.47	25.25
Baseline-mBERT	20.88	35.22	20.67	21.82
18 Oscar_Garibo	20.50	36.80	20.06	22.15
Baseline-XLMR	19.74	36.22	19.83	21.00
19 SUKI	19.63	29.23	20.85	21.95
Baseline-I	1.97	21.54	5.55	1.20

Table 6: Results on Test-A of Subtask 1.

Team	Macro-F_1	Acc	Rec	Prec
1 UniManc	18.95	36.84	20.48	25.82
2 mtu_fiz	17.67	33.92	18.79	25.03
3 rematchka	17.64	36.50	19.62	23.59
4 GOF	17.19	34.60	18.56	22.12
5 Ahmed_and_Khalil	17.15	34.67	19.47	23.39
6 ISL-AAST	16.95	35.07	18.40	22.47
7 iCompass	16.94	34.94	19.52	19.01
Baseline-MARBERT	16.94	34.06	18.82	23.19
8 NLP_DI	16.44	27.68	18.49	20.28
9 Pythoneers	15.61	29.51	15.90	19.51
10 Elyadata	15.46	29.85	16.34	20.25
11 SQU	14.84	30.12	16.80	21.32
12 Giyaseddin	14.30	29.92	15.59	21.95
13 259	11.85	22.25	11.43	14.21
14 RUTeam	11.35	22.80	11.86	14.60
15 SUKI	10.58	20.56	10.11	12.98
Baseline-mBERT	10.53	22.05	11.42	14.06
16 zTeam	10.47	25.71	13.23	16.29
17 BFCAI	9.71	23.13	11.99	14.54
Baseline-XLMR	9.62	21.91	11.33	14.05
18 XY	9.25	23.74	11.73	17.57
19 Oscar_Garibo	8.40	19.40	9.80	11.74
Baseline-I	2.59	14.86	10.00	1.49

Table 7: Results on Test-B of Subtask 1.

Team	F_1-PN	Acc	Rec	Prec
1 rematchka	75.16	69.70	66.22	67.57
2 UniManc	73.54	67.70	63.92	65.27
3 BhamNLP	73.46	67.33	62.83	65.24
4 Pythoneers	73.40	68.23	65.87	66.08
Baseline-MARBERT	72.36	66.66	63.92	64.50
5 Ahmed_and_Khalil	71.46	66.03	63.73	63.84
6 Giyaseddin	71.43	65.80	62.20	63.51
7 ISL_AAST	70.55	64.97	61.41	62.58
8 ANLP-RG	67.31	61.90	59.67	59.69
Baseline-XLMR	63.24	57.30	55.53	55.66
9 RUTeam	61.07	56.17	53.58	53.90
Baseline-mBERT	55.84	50.13	49.00	49.47
10 Oscar_Garibo	46.43	43.00	41.92	42.00
Baseline-I	27.83	38.57	33.33	12.86

Table 8: Results for Subtask 2 (Sentiment Analysis).

team, we take their best score of Test-A and Test-B and then calculate the average macro-F_1 score over the best scores of these two test sets (i.e., Test-A and Test-B). `Team rematchka` (Abdel-Salam, 2022) obtained the best performance on Subtask 1 with 27.06 average macro-F_1. We can observe that seven teams outperform our strongest baseline, Baseline-MARBERT. `Team rematchka` also achieved the best F_1 of 36.48 on Test-A of Subtask 1. `Team UniManc` (Khered et al., 2022) acquired the best F_1 of 18.95 on Test-B of Subtask 1. Results show that dialect identification based on text input is challenging. We note that there is a sizable discrepancy between test results on Test-A and Test-B: Test-B results are much lower. We believe the reason is that Test-B is derived from a different distribution (e.g., different collection time) as compared to training data of Subtask 1.

Table 8 shows the leaderboard of Subtask 2 and is sorted by the main metric of Subtask 2, F_{NP} score. Again, `Team rematchka` achieved the best F_{NP} score of 75.16. We observe that four and then eight teams outperformed our Baseline-MARBERT and Baseline-XLMR, respectively.

Team Name	# submit	Main Metric	Features					Techniques									Use unlabeled
			N-gram	TF-IDF	Linguistic	Word embeds	Sampling	Classical ML	Neural nets	Transformer	Ensemble	Adapter	Multitask	Prompting	Distillation	Data Aug.	Pre-training
Subtask 1																	
rematchka	6	27.06							✓	✓	✓			✓			
UniManc	6	26.86					✓			✓							
GOF	4	26.44								✓							
mtu_fiz	8	25.50							✓	✓		✓				✓	
iCompass	2	25.32								✓							
ISL_AAST	5	24.59								✓		✓				✓	
Ahmed_and_Khalil	2	24.35								✓							
Pythoneers	4	24.12								✓	✓		✓		✓		
Giyaseddin	3	22.42								✓							
SQU	4	22.42	✓	✓		✓		✓	✓	✓							✓
NLP_DI	9	21.28	✓						✓	✓							✓
RUTeam	2	17.28			✓					✓							
259	2	16.89	✓					✓		✓							
zTeam	2	16.12				✓		✓	✓	✓							
XY	10	15.80						✓	✓	✓	✓						
BFCAI	6	15.48		✓		✓		✓	✓								
SUKI	2	15.11	✓					✓									
Subtask 2																	
rematchka	4	75.16			✓				✓	✓	✓			✓			
UniManc	3	73.54								✓							✓
BhamNLP	3	73.46	✓			✓			✓	✓							
Pythoneers	1	73.40								✓	✓		✓		✓		
Ahmed_and_Khalil	1	71.46								✓							
Giyaseddin	1	71.43								✓							
ISL_AAST	3	70.55								✓		✓				✓	
ANLP-RG	3	67.31								✓							
RUTeam	1	61.07			✓					✓							

Table 9: Summary of approaches used by participating teams who also submitted system descriptions. Teams are sorted by their performance on official metric, the average $Macro\text{-}F_1$ score over Test-A and Test-B for Subtask 1 and $F1_{NP}$ score over the positive and negative classes for Subtask 2. Classical machine learning (ML) refers to any non-neural machine learning methods such as naive Bayes and support vector machines. The term "neural nets" refers to any model based on neural networks (e.g., FFNN, RNN, and CNN) except Transformer models. Transformer refers to neural networks based on a Transformer architecture such as BERT. **Data Aug.**: Data Augmentation.

5.4 General Description of Submitted Systems

In Table 9, we provide a high-level summary of the submitted systems. For each team, we list their best score with the the main metric of each subtask and the number of their submissions. As shown in this table, most teams used Transformer-based pre-trained language models, including mBERT (Devlin et al., 2019), ArabBERT (Antoun et al., 2020), MARBERT (Abdul-Mageed et al., 2021a).

The top team of Subtasks 1 and 2, i.e., `rematchka`, exploited MARBERT, AraBERT, and AraGPT2 (Antoun et al., 2021) with different prompting techniques and added linguistic features to their models.

The team placing first on Test-B of Subtask 1, i.e., `UniManc`, used MARBERT and enhanced the model on under-represented classes by introducing a sampling strategy.

Teams `mtu_fiz` (Shammary et al., 2022) and `ISL_AAST` used adapter modules to fine-tune MARBERT and applied data augmentation techniques.

`Team UniManc` found that further pre-training MARBERT on the 10M unlabelled tweets we released does not benefit Subtask 1 but improves performance on Subtask 2.

Six teams also utilized classical machine learning methods (e.g., SVM and Naive Bayes) to develop their systems.

6 Conclusion and Future Work

We presented the findings and results of the third Nuanced Arabic Dialect Identification shared task, NADI 2022. The shared task has two subtasks: Subtask 1 on country-level dialect identification (including 18 countries) and Subtask 2 on dialectal sentiment analysis (including 10 countries). NADI continues to be an attractive shared task, as reflected by the wide participation: 41 registered teams, 21 submitting teams scoring 105 valid models, and 15 published papers. Results obtained by the various teams show that both dialect identification and dialectal sentiment analysis of short text sequences remain challenging tasks. This motivates further work on Arabic dialects, and so we plan to run future iterations of NADI. Our experience from NADI 2022 shows that inclusion of additional subtasks, along with dialect identification, provides a rich context for modeling. Hence, we intend to continue adding at least one subtask (e.g., sentiment analysis covering more countries, emotion detection) to our main focus of dialect identification. We will also consider adding a data contribution track to NADI. In that track, teams may collect and label new datasets for public release.

Acknowledgements

MAM acknowledges support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), and Digital Research Alliance of Canada,[5] and UBC Advanced Research Computing-Sockeye.[6]

References

Abdulrahman AAlAbdulsalam. 2022. SQU-CS @ NADI 2022: Dialectal Arabic Identification using One-vs-One Classification with TF-IDF Weights Computed on Character n-grams. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. Association for Computational Linguistics.

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, Marseille, France. European Language Resources Association.

Reem Abdel-Salam. 2022. Dialect & sentiment identification in nuanced Arabic tweets using an ensemble of prompt-based, fine-tuned and multitask bert-based models. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. Association for Computational Linguistics.

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

Muhammad Abdul-Mageed. 2015. *Subjectivity and sentiment analysis of Arabic as a morophologically-rich language*. Ph.D. thesis, Indiana University.

Muhammad Abdul-Mageed. 2019. Modeling Arabic subjectivity and sentiment in lexical space. *Information Processing & Management*, 56(2):291–307.

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

Muhammad Abdul-Mageed, Mona T. Diab, and Sandra Kübler. 2014. SAMAR: subjectivity and sentiment analysis for arabic social media. *Comput. Speech Lang.*, 28(1):20–37.

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, 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.

[5]https://alliancecan.ca

[6]https://arc.ubc.ca/ubc-arc-sockeye

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.

Muhammad Abdul-Mageed, Chiyu Zhang, Azadeh Hashemi, and El Moatez Billah Nagoudi. 2020c. AraNet: A deep learning toolkit for Arabic social media. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 16–23, Marseille, France. European Language Resource Association.

Mohamed Elhag Mohamed Abo, Norisma Idris, Rohana Mahmud, Atika Qazi, Ibrahim Abaker Targio Hashem, Jaafar Zubairu Maitama, Usman Naseem, Shah Khalid Khan, and Shuiqing Yang. 2021. A multi-criteria approach for Arabic dialect sentiment analysis for online reviews: Exploiting optimal machine learning algorithm selection. *Sustainability*, 13(18):10018.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France. European Language Resource Association.

Ibrahim Abu Farha and Walid Magdy. 2021. Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 21–31, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Ali Al-Laith, Muhammad Shahbaz, Hind F Alaskar, and Asim Rehmat. 2021. Arasencorpus: A semi-supervised approach for sentiment annotation of a large Arabic text corpus. *Applied Sciences*, 11(5):2434.

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'12)*, pages 2882–2889, Istanbul, Turkey. European Language Resources Association (ELRA).

Nora Al-Twairesh, Rawan N. Al-Matham, Nora Madi, Nada Almugren, Al-Hanouf Al-Aljmi, Shahad Alshalan, Raghad Alshalan, Nafla Alrumayyan, Shams Al-Manea, Sumayah Bawazeer, Nourah Al-Mutlaq, Nada Almanea, Waad Bin Huwaymil, Dalal Alqusair, Reem Alotaibi, Suha Al-Senaydi, and Abeer Alfutamani. 2018. SUAR: towards building a corpus for the saudi dialect. In *Fourth International Conference On Arabic Computational Linguistics, ACLING 2018, November 17-19, 2018, Dubai, United Arab Emirates*, volume 142 of *Procedia Computer Science*, pages 72–82. Elsevier.

Sarah Omar Alhumoud and Asma Ali Al Wazrah. 2022. Arabic sentiment analysis using recurrent neural networks: a review. *Artif. Intell. Rev.*, 55(1):707–748.

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

Latifah Almuqren and Alexandra I. Cristea. 2021. Aracust: a saudi telecom tweets corpus for sentiment analysis. *PeerJ Comput. Sci.*, 7:e510.

Areeb Alowisheq, Nora Al-Twairesh, Mawaheb Altuwaijri, Afnan AlMoammar, Alhanouf Alsuwailem, Tarfa Albuhairi, Wejdan Alahaideb, and Sarah Alhumoud. 2021. MARSA: multi-domain Arabic resources for sentiment analysis. *IEEE Access*, 9:142718–142728.

Nouf AlShenaifi and Aqil Azmi. 2022. Arabic dialect identification using machine learning and transformer-based models: Submission to the NADI 2022 Shared Task. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. Association for Computational Linguistics.

Abdulkareem Alsudais, Wafa Alotaibi, and Faye Alomary. 2022. Similarities between Arabic dialects: Investigating geographical proximity. *Information Processing & Management*, 59(1):102770.

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

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

Ghadah Alwakid, Taha Osman, Mahmoud El Haj, Saad Alanazi, Mamoona Humayun, and Najm Us Sama. 2022. MULDASA: Multifactor lexical sentiment analysis of social-media content in nonstandard Arabic social media. *Applied Sciences*, 12(8):3806.

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

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraGPT2: Pre-trained transformer for Arabic language generation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 196–207, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Joseph Attieh and Fadi Abdulfattah Mohammed Hassan. 2022. Arabic Dialect Identification and Sentiment Classification using Transformer-based Models. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. Association for Computational Linguistics.

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

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, LREC 2022, Marseille, France, 20-25 June 2022*, pages 4586–4596. European Language Resources Association.

Gıyaseddin Bayrak and Abdul Majeed Issifu. 2022. Domain-Adapted BERT-based models for Nuanced Arabic Dialect Identification and Tweet Sentiment Analysis. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. 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).

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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. European Language Resources Association.

Ahmed Oumar El-Shangiti and Khalil Mrini. 2022. Ahmed and Khalil at NADI 2022: Transfer Learning and Addressing Class Imbalance for Arabic Dialect Identification and Sentiment Analysis. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. Association for Computational Linguistics.

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

AbdelRahim Elmadany, Chiyu Zhang, Muhammad Abdul-Mageed, and Azadeh Hashemi. 2020. Leveraging affective bidirectional transformers for offensive language detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 102–108.

Ashraf Elnagar, Sane Yagi, Ali Bou Nassif, Ismail Shahin, and Said A. Salloum. 2021. Sentiment analysis in dialectal Arabic: A systematic review. In *Advanced Machine Learning Technologies and Applications - Proceedings of AMLTA 2021, Cairo, Egypt, March 22-24, 2021*, volume 1339 of *Advances in Intelligent Systems and Computing*, pages 407–417. Springer.

Chayma Fourati, Abir Messaoudi, and Hatem Haddad. 2020. Tunizi: a tunisian arabizi sentiment analysis dataset. *arXiv preprint arXiv:2004.14303*.

Emna Fsih, Saméh Kchaou, Rahma Boujelbane, and Lamia Hadrich Belguith. 2022. Benchmarking Transfer Learning Approaches for Sentiment Analysis of Arabic Dialect. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. Association for Computational Linguistics.

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.

Imane Guellil, Ahsan Adeel, Faiçal Azouaou, Fodil Benali, Ala-Eddine Hachani, Kia Dashtipour, Mandar

Gogate, Cosimo Ieracitano, Reza Kashani, and Amir Hussain. 2021. A semi-supervised approach for sentiment analysis of arab(ic+izi) messages: Application to the algerian dialect. *SN Comput. Sci.*, 2(2):118.

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

Imane Guellil, Marcelo Mendoza, and Faiçal Azouaou. 2020b. Arabic dialect sentiment analysis with ZERO effort. \\ case study: Algerian dialect. *Inteligencia Artif.*, 23(65):124–135.

Nizar Habash, Houda Bouamor, Hazem Hajj, Walid Magdy, Wajdi Zaghouani, Fethi Bougares, Nadi Tomeh, Ibrahim Abu Farha, and Samia Touileb, editors. 2021. *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics, Kyiv, Ukraine (Virtual).

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

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

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

Sabit Hassan, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. 2021. ASAD: Arabic social media analytics and unDerstanding. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 113–118, Online. Association for Computational Linguistics.

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

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

Salma Jamal, Aly M. Kassem, Omar Mohamed, and Ali Ashraf. 2022. On The Arabic Dialect. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. Association for Computational Linguistics.

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

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022. Optimizing Naive Bayes for Arabic Dialect Identification. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. Association for Computational Linguistics.

Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2022. NLP_DI at NADI Shared Task Subtask-1: Sub-word Level Convolutional Neural Models and Pre-trained Binary Classifiers for Dialect Identification. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. Association for Computational Linguistics.

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

Abdullah Khered, Ingy Abdelhalim, and Riza Batista-Navarro. 2022. Building an Ensemble of Transformer Models for Arabic Dialect Classification and Sentiment Analysis. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. Association for Computational Linguistics.

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

Karima Meftouh, 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.

Abir Messaoudi, Chayma Fourati, Hatem Haddad, and Moez Ben HajHmida. 2022. iCompass Working Notes for the Nuanced Arabic Dialect Identification Shared task. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. Association for Computational Linguistics.

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.

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, Marseille, France. European Language Resource Association.

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

Abdelrahim Qaddoumi. 2022. Arabic Sentiment Ensemble NADI Shared Task 2. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. Association for Computational Linguistics.

Fatiha Sadat, Farzindar Kazemi, and Atefeh Farzindar. 2014. Automatic identification of Arabic language varieties and dialects in social media. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 22–27, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Fouad Shammary, Yiyi Chen, Zsolt T. Kardkovács, Haithem Afli, and Mehwish Alam. 2022. TF-IDF or Transformers for Arabic Dialect Identification? ITFLOWS participation in the NADI 2022 Shared Task. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. Association for Computational Linguistics.

Arwa A. Al Shamsi and Sherief Abdallah. 2022. Sentiment analysis of Emirati dialect. *Big Data Cogn. Comput.*, 6(2):57.

Mahmoud Sobhy, Ahmed H. Abu El Atta, Ahmed A. El-Sawy, and Hamada Nayel. 2022. Word Representation Models for Arabic Dialect Identification. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP 2022)*. Association for Computational Linguistics.

Jenny A Thomas. 2014. *Meaning in interaction: An introduction to pragmatics*. Routledge.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

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

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.

Chiyu Zhang, Muhammad Abdul-Mageed, and El Moatez Billah Nagoudi. 2022. Decay no more: A persistent twitter dataset for learning social meaning. *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media*.

Imed Zitouni, Muhammad Abdul-Mageed, Houda Bouamor, Fethi Bougares, Mahmoud El-Haj, Nadi Tomeh, and Wajdi Zaghouani, editors. 2020. *Proceedings of the Fifth Arabic Natural Language Processing Workshop*. Association for Computational Linguistics, Barcelona, Spain (Online).