QADI: Arabic Dialect Identification in the Wild

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan and **Kareem Darwish**
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
{aabdelali,hmubarak,ysamih,sahassan2,kdarwish}@hbku.edu.qa

Abstract

Proper dialect identification is important for a variety of Arabic NLP applications. In this paper, we present a method for rapidly constructing a tweet dataset containing a wide range of country-level Arabic dialects —covering 18 different countries in the Middle East and North Africa region. Our method relies on applying multiple filters to identify users who belong to different countries based on their account descriptions and to eliminate tweets that either write mainly in Modern Standard Arabic or mostly use vulgar language. The resultant dataset contains 540k tweets from 2,525 users who are evenly distributed across 18 Arab countries. Using intrinsic evaluation, we show that the labels of a set of randomly selected tweets are 91.5% accurate. For extrinsic evaluation, we are able to build effective country-level dialect identification on tweets with a macro-averaged F1-score of 60.6% across 18 classes.

1 Introduction

Twitter is one of the most popular social media platforms in the Middle East and North Africa (MENA) region with almost two thirds (63%) of Arab youth indicating that they look first to Facebook and Twitter for news (Radcliffe and Bruni, 2019). The popularity of Twitter in MENA is reflected by approximately 164 million active monthly users, who produce a massive volume of Arabic tweets, many of which are in Dialectal Arabic (DA). Hence, many researchers have been using Twitter as a major data source that is representative of current language usage and linguistic phenomena (Mubarak and Darwish, 2014; Samih et al., 2017; Zaghouani and Charfi, 2018a). Though Arabic is the lingua franca of most of the MENA region, different dialects of Arabic are used in different countries. While some dialects may differ significantly from each other (e.g. Egyptian dialect (EG) and Moroccan Maghrebi dialect (MA)[1]), others, particularly those in close geographic proximity, may be more difficult to tweak apart (e.g. variants of the Levantine dialect such as Syrian (SY) and Lebanese (LB)). Figure 1 highlights the dialectal variations across the Arab world. The figure shows that dialects are a continuum that often transcends geographical regions and borders. Automatically distinguishing between the different dialectal variations is valuable for many downstream applications such as machine translations (Diab et al., 2014), POS tagging (Darwish et al., 2020), geo-locating users, and author profiling (Sadat et al., 2014).

Though there has been prior work on performing Arabic Dialect Identification (ADI), much of the work was conducted on datasets with significant limitations in terms of genre (Bouamor et al., 2018; Zaidan and Callison-Burch, 2011), number of dialects (Abdul-Mageed et al., 2018), or focus (Bouamor et al., 2019; Zaghouani and Charfi, 2018a), where often the focus was on geo-locating and profiling users as opposed to dialect identification. In this work, we expand beyond these efforts by utilizing tweets from across the MENA region to build a large, non-genre specific, fine-grained, and balanced country-level dialectal Arabic dataset that we use to build effective Arabic Dialect Identification.

We rely on two main features to build the dataset. The first feature is the Twitter user profile description, where we identify users who self-declare themselves as belonging to a specific country in different forms such as showing signs of loyalty and pride (e.g. "proud Egyptian"). In the second, we use a classifier that utilizes distant supervision to accurately discriminates between MSA and di-

[1] We use ISO 3166-1 alpha-2 for country codes: https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 1–10
Kyiv, Ukraine (Virtual), April 19, 2021.

alects. In doing so, we can identify users who self-declare their identity, mostly tweet in dialectal Arabic, and only retain dialectal user tweets. Further, we use our newly constructed dataset to build models that can effectively distinguish between 18 country-level Arabic dialects. We didn't consider four Arab countries, namely Mauritania, Somalia, Djibouti, and Comoros, because we were not able to find a sufficient number of Twitter users tweeting in Arabic. This could be due to the limited use of Twitter in these countries, or that users may tweet primarily in other languages. For automated dialect identification, our models use a variety of features, such as character-level and word-level n-gram, static word embeddings, and contextual embeddings (e.g. multilingual BERT (mBERT) and AraBERT), and two classification techniques, namely Support Vector Machines (SVM) classification and fine-tuned Transformer models. The contributions of this work are:

- We introduce a method for constructing a highly accurate Arabic dialectal dataset from Twitter. This method can be completely automated such that it can be used in the future to collect fresh dialectal tweets.
- We build QADI (meaning "judge" in Arabic) dataset,[2]. It is the largest balanced non-genre specific country-level Arabic dialectal tweet dataset. The dataset contains more than 540k tweets covering 18 country-level dialects with an associated test set containing 182 tweets per country on average that was manually labeled by native speakers from 18 Arab countries.
- We provide a list of Twitter accounts from 18 Arab countries (a total of 2,525 accounts with an average of 140 accounts per country) that can be used in author profiling tasks.
- We use the new dataset to build state-of-the-art tweet-level Arabic dialect identification models using a variety of features and classifiers.

2 Related Work

Most efforts in building resources for Arabic dialect identification are limited either in terms of genre, granularity, or the size of the data. Zaidan and Callison-Burch (2011) curated the Arabic Online Commentary Dataset, a resource of more than 52M-words. They annotated over 108K sentences (41%) of the dataset with one of 5 possible dialects, namely: Maghrebi, Egyptian, Levantine, Gulf, and Iraqi. Similarly, Alshutayri and Atwell (2017), El-Haj et al. (2018), and Alsarsour et al. (2018) annotated collections of texts using the five regions/dialects. Elfardy and Diab (2013) and Darwish et al. (2014) identified whether a sentence is Modern Standard Arabic (MSA) or Egyptian.

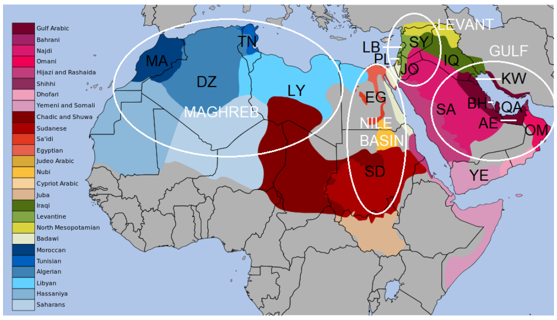


Figure 1: Geographic distribution of Arabic dialects.[3]

In recent years, there have been more efforts to cover more countries with finer granularity. Abdul-Mageed et al. (2018) constructed a dataset that covers 10 countries. Zaghouani and Charfi (2018a) built the "Arap-Tweet" dataset, which includes tweets from 15 countries. They retrieved tweets containing distinct dialectal words and expressions, and then given the users who authored these tweets, they crawled their timelines (Zaghouani and Charfi, 2018b). We approach the problem from a different angle, where we start with self-declared users rather than a pre-defined keywords, which may have limited coverage of dialectal lexical variations. The MADAR (Multi-Arabic Dialect Applications and Resources) project (Bouamor et al., 2018, 2019) produced several resources among which two corpora were used for the shared task on fine-grained dialect identification (Bouamor et al., 2019). The resource includes a lexicon and a 1,000 parallel sentences from the travel domain that were translated into local dialects of 26 Arab cities. Additionally, the project released another set of tweets by searching twitter using a set of 25 seed hashtags corresponding to the 22 states of the Arab League (e.g., #Algeria, #Egypt, #Kuwait, etc.) and relevant hashtags such as: "#ArabWorld", "#ArabLeague", and "#Arab". The approach resulted in a collection of 2,980 profiles. When inspecting the profiles, The majority of the obtained users were from

[2] QADI Dataset, is freely available for research purposes from: http://alt.qcri.org/resources/qadi/

[3] Source: https://en.wikipedia.org/wiki/Varieties_of_Arabic, country codes and regions are added

Saudi Arabia, representing 36% of the total. This was another motivation to curate a more balanced and representative dataset to use for dialect identification. Further, as we show later, this dataset is sub-optimal for tweet-level dialect identification. Abdul-Mageed et al. (2018) built a large tweet collection containing more than 200 million geo-tagged tweets that were collected over 5 years (2013-2018). The resulting collection included tweets from 29 cities from 10 Arab countries, of which 2,500 were manually annotated. The average inter-annotator agreement, Cohen's Kappa (K), was 67%, where the annotators reported not being able to distinguish between dialects from neighboring cities or countries (Abdul-Mageed et al., 2018).

Multiple approaches have been used for dialect ID that exploit a variety of features, such as character or word n-grams (Darwish et al., 2014; Zaidan and Callison-Burch, 2014; Malmasi et al., 2016; Sadat et al., 2014), and techniques such as multiple kernel learning (Ionescu and Popescu, 2016) and distributed representation of dialects (Abdul-Mageed et al., 2018; Zhang and Abdul-Mageed, 2019) to name a few. Zhang and Abdul-Mageed (2019) used semi-supervised learning using multilingual BERT for user-level dialect identification on the MADAR Shared Task. Arabic Tranformers-based approaches (Antoun et al., 2020; Safaya et al., 2020) showed competitive results in NADI (Abdul-Mageed et al., 2020) Shared Task.

3 Data Collection

It is common for users on social networks to disclose social and linguistic information about themselves in their profiles. In Twitter, the user profile provides a header and a short biography. Both fields allow users to freely describe themselves. Surveying Arabic speaking profiles, it is customary to see users declaring their patriotism and national belonging by using their county's flag or explicitly naming the city or country that they are from (e.g. "Kuwait is my home country", "I am a Libyan citizen"). To build our dataset, we obtained a collection of Arabic tweets that was crawled using the Twitter streaming API, where we set the language filter to Arabic ("lang:ar"), during the entirety of March and April, 2018. In all, the collection contains 25M tweets from which we extracted the profile information of all the users who authored these tweets. We applied three filters on user profiles and tweets as we describe in the next subsections.

Country Identification For the first stage, to identify a user's country, we filtered user profiles using a gazetteer that includes:

- All Arab country names written in either Arabic, English, or French,[4] such as المغرب (Almgrb – Morocco), Morocco, and Maroc respectively.
- The names of major cities in these countries in both Arabic and English as specified in Wikipedia,[5] such as القدس (Alqds – Jerusalem) and وهران (whrAn – Oran, Algeria).
- Arabic adjectives specifying all nationalities in both masculine and feminine forms with and without the definite article ال (Al – the) such as عراقي (ErAqy - Iraqi (m.)), عراقية (ErAqyp - Iraqi (f.)), and العراقي (AlErAqy - the Iraqi (m.)).

Arabic Variant Identification The second filter checks if the account mainly tweets in either dialectal Arabic or MSA. Since Arabic users commonly switch between MSA and dialectal Arabic, and we were interested in strictly dialectal tweets, we sought to filter out MSA tweets. There are multiple ways to distinguish between dialectal and MSA text. One such method involves using a list of strictly dialectal words (Darwish et al., 2014). However, constructing such lists across multiple dialects can be challenging. Thus, we opted to train a text classifier using a heuristically labeled tweets. Specifically, given 50 million tweets that we collected between March and September 2018, we assumed that tweets strictly containing the MSA relative pronouns الذي، الذى، التي، التى، الذين ("Al*y, Al*Y, Alty, AltY, Al*yn" - who/that in masculine, feminine, and plural forms) were MSA, and those strictly containing the dialectal relative pronoun اللي، اللى ("Ally, AllY" – who/that) were dialectal. The major advantage of the dialectal relative pronoun اللي is that it is present in most (if not all) Arabic dialects with the same meaning but not in MSA. Table 1 shows some examples of such usage across different dialects. In doing so, we labeled 3.09M tweets as MSA and 3.17M tweets as dialectal. For these tweets, we normalized user mentions to @USER, digits to NUM, emojis to EMOJI, URLs to URL, and the aforementioned relative pronouns to RELATIVE. In doing so, we eliminated Twitter-specific features, which are not

[4] French is widely used in the Maghreb region.

[5] https://en.wikipedia.org/wiki/List_of_countries_by_largest_and_second_largest_cities

linguistic in nature, and eliminated the effect of the relative pronouns we used to construct the dataset.

Dialect	Example/Translation
Egyptian	كنتي كويسه ايه اللي حصل you were good, what happened
Levantine	في ناس منيح اللي ما في متلن it's good that no people like them
Gulf	ابي ابدل المكيف اللي في غرفتي I want to change the AC in my room
Maghrebi	اللي يحبها ما يديرلهاش المشاكل he who loves her, do not troubles her

Table 1: Examples usages of dialectal relative pronoun across dialects.

We set aside 20k MSA and dialectal tweets for testing (10k for each). We trained a fastText classifier (Joulin et al., 2016), which is a deep-learning-based classifier, using character n-grams ranging in length between 3 and 6 grams. We tested on the held-out test set, and the accuracy of distinguishing between MSA and dialectal Arabic was 98%. Using this classifier, we classified the tweets of the users. We retained users, where at least 50% of their tweets were dialectal.

Appropriateness Identification The third filter removed users who were mostly tweeting vulgar, sexually explicit, or pornographic tweets. To filter out these users, we used the obscene word list generated by Mubarak et al. (2017), which contains 288 words and 127 hashtags. We removed users if more than 50% of their tweets contained vulgar words. Removing the tweets of such users was motivated by the fact that their tweets contain strong genre specific signals, which may adversely affect the generalization of dialect identification.

Normalization Tweets often contain tokens that are specific to the Twitter platform such as hashtags and user mentions. To improve generalization of the trained models (hopefully beyond tweets), we split hashtags into their semantic constituents (Bansal et al., 2015; Declerck and Lendvai, 2015) and replaced user mentions and URLs with "@USER" and "URL" respectively.

Constructing the Dataset After applying the three aforementioned filters, we ended up with 2,525 users from 18 countries (140 users per country on average), who authored 540k tweets (30k per country on average) with a total of 8.8M words. Table 2 provides per country breakdown of the dataset.

Data Validation To assess the quality of our new data set, we resorted to manual assessment, where we manually labeled a random sample of 200 tweets from the tweets of each country. Though some expressions may be unique to a dialect of a particular country (e.g. إزيك (<zyk – how are you (Egyptian)), other expressions may be used in dialects from different countries (e.g. لا باس (lA bAs – no problem or good (Algerian (DZ), Moroccan (MA), and Tunisian (TN))). Thus, the instruction we gave to the annotators was: "Is this tweet consistent with the dialect spoken in your country?" The labeling of the tweets from each country was done by native speakers from that country. The average accuracy across countries was 91.5%. For some countries where additional annotators were available, namely Egypt, Algeria, Saudi Arabia, and Syria, we asked a second annotator to also label the tweets. For these countries, the average inter-annotator agreement using Cohen's Kappa (K) was 87%. These four countries cover the major dialect groups.

Figure 2 shows the accuracy per country for all annotators. Of the 200 tweets per country, those that were judged as correctly labeled were removed from the dataset, and we used them as a test set. In all, we had 3,303 test tweets (with 183 tweets on average for each of the 18 countries). Table 2 lists the number of test tweets per country. We are releasing the test set as a benchmark for dialect identification [6]. Additionally, the release will include the training set tweet IDs that can be hydrated in observance of Twitter's data sharing policy.

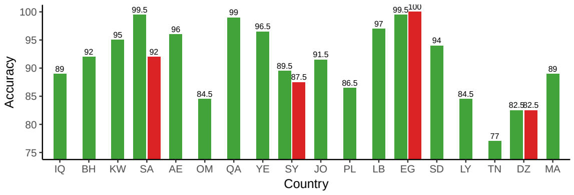


Figure 2: Annotation accuracy per country. Second annotators are colored in "Red".

The manually rejected tweets that the annotators classified as not from their dialects were mostly cases where the users interacted with or responded to users from different countries. In such cases, users tend to code-switch or adopt to other users' dialects. For example, a user identified as Tunisian tweeted انا عموما يحب الظلمة اوي (Ana EmwmA bHb AlZlmp Awy – I generally like darkness a lot).

[6] The data set can be downloaded from: http://hidden-for-blind-review/

Country	IQ	BH	KW	SA	AE	OM	QA	YE	SY	JO	PL	LB	EG	SD	LY	TN	DZ	MA
Users	142	169	160	149	172	176	139	138	139	146	145	141	150	139	149	68	130	73
Training Tweets (k)	18.4	28.3	49.9	35.4	27.8	24.8	36.7	11.6	18.3	34.1	48.6	38.4	67.8	16.3	40.9	12.9	17.6	12.8
Test tweets	178	184	190	199	192	169	198	193	194	180	173	194	200	188	169	154	170	178

Table 2: The number of users and tweets per country in our tweet corpus.

The annotator correctly tagged this as not Tunisian (TN), as it is clearly Egyptian (EG). In this example, the Tunisian user was conversing with a person from Egypt or the Levant. In another example, the tweet هسه جاي تقول أحبك؟ (hsh jAy tqwl >Hbk – just now you come to say I love you), the annotator labeled the tweet as not Yemeni (YE), mostly because of the typically Iraqi word “هسه” (hsh – just now). In this case, we found that the tweet was quoting a popular Iraqi song.

4 Corpus Statistics and Analysis

Upon constructing the dataset, we attempted to explore its characteristics. First, we extracted features that are distinctive for each dialect. To do so, we computed the so-called valence score for each word in each dialect (Conover et al., 2011). The score helps determine the distinctiveness of a given word in a specific dialect in reference to other dialects. Given $N(t, D_i)$, which is the frequency of the term t in Dialect D_i, valence is computed as follows:

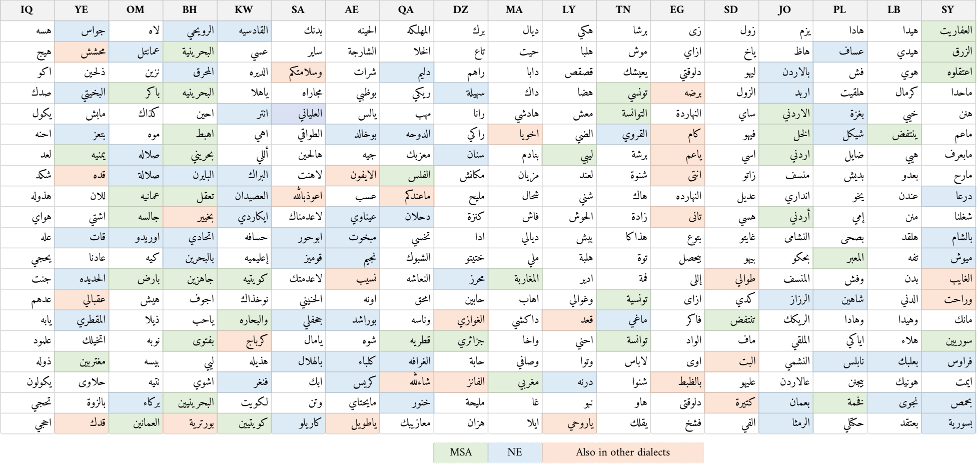
$$V(t)_i = 2\frac{\frac{N(t,D_i)}{N(D_i)}}{\sum_n \frac{N(t,D_n)}{N(D_n)}} - 1 \qquad (1)$$

Where $N(D_i)$ is the total number of occurrences of all words in dialect D_i. Figure 3 lists the words with highest valence scores per country. Though the majority of the top words were in fact distinctive dialectal words (typically function words), there were three other prominent categories of words that were not. The first was names of locations inside these countries, which implies that geographic locations in a user's Twitter timeline can be a strong features in identifying the country of the user. The second had words that appear in multiple dialects, which is expected given the overlap between dialects from different countries. The third category included MSA words. Though we intentionally excluded all tweets that were identified as MSA, the appearance of such words was expected given the large overlap between MSA and dialects and the frequent context switching between MSA and dialects in user tweets.

Next, we computed the similarity between dialects to ascertain if similarities are consistent with reports in prior literature by visualizing the similarity between different country-level dialects. For such, we constructed a list of the top 10k words with the highest valence scores across all dialects. The resulting list can be viewed as a vector of 19 valence values for each word corresponding to the valence of 18 different country-level dialects in addition to MSA. For MSA data, we used the 3.09M MSA tweets that we used earlier to train the MSA/dialect classifier. Then given the word vectors, we applied SHC bottom-up hierarchical agglomerative clustering (Li and Huang, 2009). The algorithm treats each dialect as a singleton cluster at the outset and then successively merges (or agglomerates) clusters until all clusters have been merged into a single cluster that contains all dialects. Figure 4 shows the results of hierarchical clustering. The figure reflects the similarity and the geographical proximity of various dialects. At higher levels, dialects are grouped per region, where we can identify the major dialectal groups, namely Gulf, Maghrebi, Egyptian, and Levantine. This is aligned with geographical distribution of the dialects as well as the findings of prior work (Salameh et al., 2018).

5 Experimental Setup

Given our new dataset, we conducted a battery of experiments on the dataset to build effective country-level Arabic dialect identification. We experimented with several tweet representation and classification models. For tweet representations, we used: surface features, namely words and character n-grams, static embeddings, and deep contextual embeddings, namely AraBERT and mBERT. For classification, we used two different classifiers,



IQ	YE	OM	BH	KW	SA	AE	QA	DZ	MA	LY	TN	EG	SD	JO	PL	LB	SY
هسه	جواس	لاه	الرويحي	القادسيه	بدنك	الحينه	المهلكه	برك	ديال	هكي	برشا	زى	زول	يزم	هادا	هيدا	العفاريت
هيج	محشش	عمانتل	البحرينية	عسي	ساير	الشارجة	الخلا	تاع	حيت	هلبا	موش	ازاي	ياخ	هاظ	عساف	هيدي	الزرق
اكو	ذلحين	نزين	المحرق	الديره	وسلامتكم	شرات	دليم	راهم	دابا	قصقص	يعيشك	دلوقتي	ليهو	بالاردن	فش	هوي	اعتقلوه
صدك	البخيتي	باكر	البحرينيه	ياهلا	مجاراه	بوظبي	ريكي	سهيلة	داك	هضا	تونسي	برضه	الزول	اربد	هلقيت	كرمال	ماحدا
يكول	مابش	كذاك	احين	انتر	العلياني	يالس	مهب	رانا	هادشي	معش	التوانسة	النهاردة	ساي	الاردني	بغزة	خيي	هنن
احنه	بتعز	موه	اهبط	اهي	الطواقي	بوخالد	الدوحه	راكي	اخويا	الضي	القروي	كام	فيهو	انخل	شيكل	ينتفض	ماعم
لعد	يمنيه	صلاله	بحريني	ألّي	هالحين	جيه	معزبك	سنان	بنادم	ليبي	برشة	ياعم	اسي	اردني	ضايل	هيي	مابعرف
شكد	قده	صلالة	البايرن	البراك	لاهنت	الايفون	الفلس	مكانش	مزيان	لعند	شنوة	انتى	زاتو	منسف	بديش	بعدو	مارح
هذوله	للان	عمانيه	تعقل	العصيدان	اعوذبالله	عسب	ماعندكم	مليح	شحال	شني	هاك	النهارده	عديل	انداري	يخو	عندن	درعا
هواي	اشتي	جالسه	بخيير	ايكاردي	لاعدمناك	عيناوي	دحلان	كنزة	فاش	الحوش	زادة	تانى	هسي	أردني	إمي	منن	شغلنا
عله	قات	اوريدو	اتحادي	حسافه	ابوحور	مبخوت	تخسي	ادا	ديالي	بيش	هذاكا	بتوع	غايتو	النشامى	بصحى	هلقد	بالشام
يحجي	عادنا	كيه	بالبحرين	إعليميه	قوميز	نجيم	الشبوك	ختيتو	ملي	هلبة	توة	بيحصل	بيهو	بحكو	المعبر	تفه	ميوش
جنت	الحديده	بارض	جاهزين	كويتيه	لاعدمتك	نسيب	النعاشه	محرز	المغاربة	ادير	فمة	إللى	طوالي	المنسف	وفش	بدن	الغايب
عدهم	عقبالي	هيش	اجوف	نوخذاك	الحنيني	اونه	امحق	حابين	اهاب	وغوالي	تونسية	ازاى	كدي	الرزاز	شاهين	الدني	وراحت
يابه	المقطري	ذيلا	ياحب	والبحاره	جحفلي	بوراشد	وناسه	الغوازي	داكشي	قعد	ماغي	فاكر	تنتفض	الريكك	وهادا	وهيدا	مانك
علمود	اتخيلك	نوبه	بفتوى	كرباج	يامال	شوه	قطريه	جزائري	واخا	احني	توانسة	الواد	ماف	الملقي	اياكي	هلاء	سوريين
ذوله	مغتربين	بيسه	لبي	هذيله	بالهلال	كلباء	الغرافه	حابة	وصافي	وتوا	لاباس	اوى	البت	النشمي	نابلس	بعلبك	فراوس
يكولون	حلاوى	نتيه	اشوي	فنغر	ابك	كريس	شاءالله	الفانز	مغربي	درنه	شنوا	بالظبط	عليهو	عالاردن	بيجنن	هونيك	ايمت
تحجي	بالزوة	بركاء	البحرينيين	لكويت	وتن	مايحتاي	خنور	مليحة	غا	نبو	هاو	دلوقتى	كتيرة	بعمان	نخمة	نجوى	بحمص
اجي	قدك	العمانين	بورترية	كويتيين	كاريلو	ياطويل	معازيبك	هزان	ايلا	ياروحي	يقلك	فشخ	الفي	الرمثا	حكّلي	بعتقد	بسورية

MSA | NE | Also in other dialects

Figure 3: Highest valence words for each country.

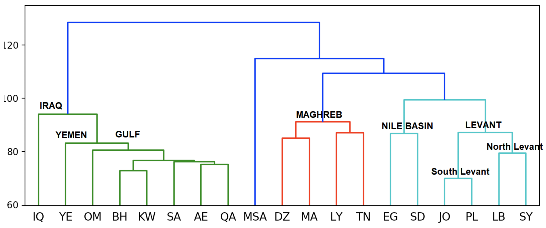


Figure 4: Clustering of Arabic Dialects using valence scores on top 10k words.

namely an SVM classifier and a fine-tuned Transformer model. For comparison, we conducted the same experiments on MADAR dataset. In the following subsections, we present tweet representations and classification models.

5.1 Representations

Surface Features: We used two different surface-level features, namely word and character n-grams. Specifically, we represented tweets using: i) character n-grams, where we used 2 to 6-grams (C{2-6}); ii) word n-grams, where we used unigrams (W{1}) and unigrams to 6-grams (W{1-6}); and iii) a combination of word and character n-grams. For our dataset and MADAR , we normalized URLs, numbers, and user mentions to URL, NUM, and MENTION respectively. We used tf-idf weighting for character and word n-grams.

Static Embeddings: We used **Mazajak** word-level skip-gram embeddings (Abu Farha and Magdy, 2019) that were trained on 250M Arabic tweets with 300-dimensional vectors.

Deep Contextualized Embeddings: We also experimented with two pre-trained contextualized embeddings with fine-tuning for down-stream tasks, namely BERT$_{\text{base-multilingual}}$ (mBERT) and AraBERT (Antoun et al., 2020). Recently, deep contextualized language models such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), UMLFIT (Howard and Ruder, 2018), and OpenAI GPT (Radford et al., 2018), to name but a few, have achieved ground-breaking results in many NLP classification and language understanding tasks.

Both mBERT and AraBERT are pre-trained on identical architectures, namely an encoder with 12 Transformer blocks, hidden size of 768, and 12 self-attention heads. However, they differ in one major way. While mBERT is pre-trained on Wikipedia text for 104 languages,AraBERT is trained on a large Arabic news corpus containing 8.5M articles composed of roughly 2.5B tokens. For consistency with mBERT, we used AraBERT with BP. Following Devlin et al. (2019), the classification consists of introducing a dense layer over the final hidden state h corresponding to first token of the sequence, [CLS], adding a softmax activation on top of BERT to predict the probability of the l label: $p(l|h) = softmax(Wh)$, where W is the task-specific weight matrix. We set the learning rate to 2e-5, batch size to 8, max sequence length to 128, and the number fine-tuning epochs to 6. During fine-tuning, all mBERT or AraBERT parameters together with W are optimized end-to-end to maximize the log-probability of the correct labels.

5.2 Classification Models

For classification, we used an SVM classifier and fine-tuned mBERT and AraBERT. We utilized the SVM classifier when using surface features and static pre-trained Mazajak embeddings. We used the Scikit Learn libsvm implementations of the SVM classifier with a linear kernel. When using

	Training Set	
Classifier	QADI	MADAR
MultiLangBERT	58.9	25.3
AraBERT	**60.6**	29.0
Mazajak	39.8	24.6
SVM$_{C\{2-6\}}$	57.3	25.6
SVM$_{W\{1\}}$	50.8	23.7
SVM$_{W\{1-6\}}$	51.8	20.6
SVM$_{C\{2-6\},W\{1-6\}}$	57.6	26.4

Table 3: Classification results for QADI and MADAR sets using the various models

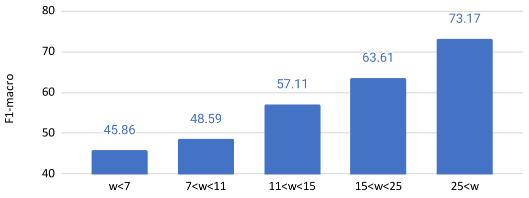


Figure 5: Macro-averaged F1-score given tweet length using AraBERT.

contextualized embeddings, we fine-tuned mBERT or AraBERT by adding a fully-connected dense layer followed by a softmax classifier, minimizing the binary cross-entropy loss function for the training data. We used the PyTorch[7] implementation by HuggingFace[8].

5.3 Experiments and Results

As stated earlier, we ran a number of country-level dialect ID experiments on our new dataset and on MADAR dataset for comparison. The details of the training and test splits for the dataset as as follows:
QADI Dataset: Table 2 provides the statistics of the training and test parts of QADI dataset. Given that manual verification was done at tweet-level, all the experiments on QADI dataset were done at tweet level. In all, the dataset contains 540k training tweets and 3,303 test tweets.
MADAR Dataset: MADAR task 2 dataset was designed for user-level classification, where each user is assigned a country label. The dataset is split into train/dev/test splits that contain 2,180, 300, and 500 users respectively, with approximately 100 sample tweets per users. For our experiments, we merged the training and development splits. Since we were performing tweet-level classification, we assigned the user label to all their tweets, and proceeded to perform tweet-level training and testing. We normalized tweets in the same manner applied on QADI dataset .

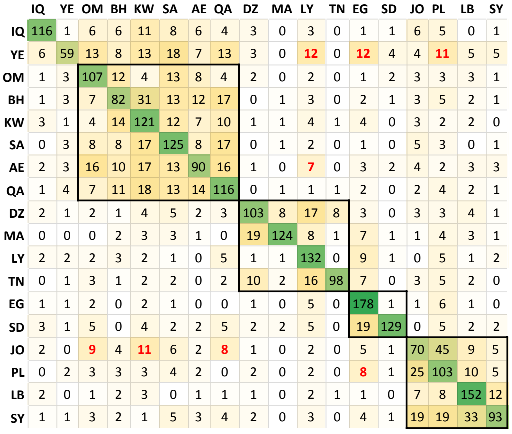
Results Table 3 reports on the macro-averaged F1-score results of training and testing using QADI and MADAR datasets. As QADI results show, using contextual embeddings yielded the best results with AraBERT results edging mBERT results. Using an SVM classifier that is trained using either character n-grams only (C{2-6}) or a combination of character and word n-grams (C{2-6},W{1-6}) was slightly lower than using contextual embeddings. Using an SVM classifier is computationally more efficient than using contextual embeddings. Further, character n-grams performed better than using word n-grams, with the combination of both character and word n-grams performing slightly better than using character n-grams alone. Using Mazajak embeddings led to significantly lower results. Further, when inspecting the best classification results (AraBERT), we noted that the length of the tweets impacted the classification results. The longer a tweet, the more accurate the prediction was. Figure 5 shows the accuracy of the classifier for various tweet lengths. This is expected given that longer tweets potentially contain more clues for the classifier. Training using MADAR led to significantly lower results compared to training using QADI . This likely stems from a mismatch between the problem at hand (tweet-level dialect ID) and the purpose for which MADAR was constructed (user-level dialect/country ID). Further, belonging to a country does not guarantee that a user will always tweet in the dialect of that country. Often users from different countries use MSA (or even other languages). We speculated that many of the tweets in the MADAR data are actually MSA, because the tweets were collected without taking into account whether they were actually dialectal or not. To test this hypothesis, we used our aforementioned MSA/dialectal classifier. When we classified the MADAR tweets, the classifier tagged 29% of the tweets as dialectal and the rest as MSA (71%), confirming our hypothesis. Since the vast majority of the tweets were MSA, training on the MADAR dataset led to significantly lower tweet-level dialect classification results. Since QADI filters out MSA tweets, it doesn't have the same issue.

Error Analysis We inspected tweets from the QADI test set that were misclassified by AraBERT (our best system). Generally, the most prominent reason for incorrect classification could be attributed to the fluidity of geolinguistic distinctions between Arabic dialects. To some degree,

[7] https://pytorch.org/

[8] https://github.com/huggingface/transformers

geographical proximity is associated with dialectal closeness, making it difficult for classifiers to distinguish between the dialects at hand. Note that these dialects share a plethora of linguistic features to warrant their subsumability under the same dialect. As shown in Figure 6, the dialects from the Gulf region (OM, BH, KW, SA, AE, and QA) show the largest confusion due to their similarity. For example, the tweet, دخلت البرنامج سويت حساب... والرقم ما يدخل وكلمتكم كم مره شوفو لي حل (I logged into the program, created an account, and the number is not accepted ... Find me a solution), could be plausibly attributed to any of the Gulf dialects.



	IQ	YE	OM	BH	KW	SA	AE	QA	DZ	MA	LY	TN	EG	SD	JO	PL	LB	SY
IQ	116	1	6	6	11	8	6	4	3	0	3	0	1	1	6	5	0	1
YE	6	59	13	8	13	18	7	13	3	0	12	0	12	4	4	11	5	5
OM	1	3	107	12	4	13	8	4	2	0	2	0	1	2	3	3	3	1
BH	1	3	7	82	31	13	12	17	0	1	3	0	2	1	3	5	2	1
KW	3	1	4	14	121	12	7	10	1	1	4	1	4	0	3	2	2	0
SA	0	3	8	8	17	125	8	17	0	1	2	0	1	0	5	3	0	1
AE	2	3	16	10	17	13	90	16	1	0	7	0	3	2	4	2	3	3
QA	1	4	7	11	18	13	14	116	0	1	1	1	2	0	2	4	2	1
DZ	2	1	2	1	4	5	2	3	103	8	17	8	3	0	3	3	4	1
MA	0	0	0	2	3	3	1	0	19	124	8	1	7	1	1	4	3	1
LY	2	2	2	3	2	1	0	5	1	1	132	0	9	1	0	5	1	2
TN	0	1	3	1	2	2	0	2	10	2	16	98	7	0	3	5	2	0
EG	1	1	2	0	2	1	0	0	1	0	5	0	178	1	1	6	1	0
SD	3	1	5	0	4	2	2	5	2	2	5	0	19	129	0	5	2	2
JO	2	0	9	4	11	6	2	8	1	0	2	0	5	1	70	45	9	5
PL	0	2	3	3	3	4	2	0	2	0	2	0	8	1	25	103	10	5
LB	2	0	1	2	3	0	1	1	1	0	2	1	1	0	7	8	152	12
SY	1	1	3	2	1	5	3	4	2	0	3	0	4	1	19	19	33	93

Figure 6: Confusion matrix for the test set. The bulk of the mis-classification happens within the region (marked with thick border). Outliers are marked in red where the classification is beyond the region.

Similarly, the second greatest confusion is among dialects from the Levant region (JO, PL, LB, and SY), where we found a considerable amount of mix-up between LB and SY. The tweet, لما الانسان بفكر حالو الوحيد يلي بيفهم . . تأكد انه حمار (When the human thinks that he is the only one who understands ... be sure that he is a donkey), can be equally valid for both dialects. Similar to the results observed for both the Gulf and Levant regions, the Maghrebi dialects (MA, DZ, LY, TN) exhibit a similar pattern. MA and DZ account for considerable confusion. For instance, the tweet الله يبارك فيك خويااا (God bless you, brother!!), could be used in both dialects. As for the Nile Basin dialects, Egyptian (EG) and Sudanese (SD) could also be confused with one another. The tweet, التويته دي معدلة فوتوشوب (This tweet is modified in Photoshop), is equally valid in both dialects. This is normal since SD is similar to central and southern Sa'idi Egyptian Arabic.[9] Interestingly, we found that about 2% of the misclassified tweets were outliers that were classified outside of their region (highlighted in red in Figure 6). The main reasons for incorrect classification, beyond the region, is due to the fact that many of them contain quotes from popular songs and poems or in few cases they have MSA words. As this YE tweet, وإنت عارف مكانتك ومتأكد منها من غير مايعيشك شعور مرات إنت العمر ومرات ما أعرفك ... (And you know your status, and you are certain about it without making you feel...), misclassified as LY. This tweet despite being manually labelled as YE, it could fit in either country.

6 Conclusion

In this paper, we presented a method for building a country-level dialectal tweet corpus. The construction of the corpus relied on a cascade of filters, where user accounts were filtered on keywords indicating country, and tweets were filtered to remove users who predominantly tweet in MSA or vulgar language. We built a large corpus containing 540k tweets from 2,525 Twitter accounts that cover 18 Arab countries. Based on a manual inspection of a random sample of tweets from the corpus, the estimated accuracy of country-level dialectal tags was 91.5%. We also showed that the resultant corpus can be effective in training a country-level dialect classifier for tweets that achieves a macro-averaged F1-score of 60.6% across 18 different classes. We compared to training on a publicly available dataset, namely MADAR dataset, and MADAR results were significantly lower.

Based on our error analysis, we discovered that a large source of errors was due to the naturally occurring overlap between dialects from neighboring countries and to code switching between different dialects. While overlap between dialects is potentially an intractable problem, detecting code switching between dialects is a future direction that can further help filter training data and identify tweets that may include multiple dialects simultaneously. For future work, we plan to investigate code switching and examine the efficacy of extending our dataset to perform user-level geotagging. Though identifying a user's country may depend on multiple signals, accurate dialect identification is likely a strong signal that can aid classification.

[9] https://en.wikipedia.org/wiki/Egyptian_Arabic

References

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, Houda Bouamor, and Nizar Habash. 2020. 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.

Ibrahim Abu Farha and Walid Magdy. 2019. Mazajak: An online Arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 192–198, Florence, Italy. Association for Computational Linguistics.

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

Areej Alshutayri and Erik Atwell. 2017. Exploring twitter as a source of an arabic dialect corpus. *International Journal of Computational Linguistics (IJCL)*, 8(2):37–44. This is an open access article under the terms of the Creative Commons Attribution License (CC-BY).

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*, pages 9–15.

Piyush Bansal, Romil Bansal, and Vasudeva Varma. 2015. Towards deep semantic analysis of hashtags. In *European conference on information retrieval*, pages 453–464. Springer.

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, Florence, Italy. Association for Computational Linguistics.

Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. *ICWSM*, 133:89–96.

Kareem Darwish, Mohammed Attia, Hamdy Mubarak, Younes Samih, Ahmed Abdelali, Lluís Màrquez, Mohamed Eldesouki, and Laura Kallmeyer. 2020. Effective multi dialectal arabic pos tagging. *Natural Language Engineering*, 1(1):18.

Kareem Darwish, Hassan Sajjad, and Hamdy Mubarak. 2014. Verifiably Effective Arabic Dialect Identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar.

Thierry Declerck and Piroska Lendvai. 2015. Processing and normalizing hashtags. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 104–109.

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 T Diab, Mohamed Al-Badrashiny, Maryam Aminian, Mohammed Attia, Heba Elfardy, Nizar Habash, Abdelati Hawwari, Wael Salloum, Pradeep Dasigi, and Ramy Eskander. 2014. Tharwa: A Large Scale Dialectal Arabic-Standard Arabic-English Lexicon. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 3782–3789, Reykjavik, Iceland.

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

Heba Elfardy and Mona Diab. 2013. Sentence Level Dialect Identification in Arabic. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics.

Radu Tudor Ionescu and Marius Popescu. 2016. UnibucKernel: An approach for Arabic dialect identification based on multiple string kernels. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 135–144, Osaka, Japan. The COLING 2016 Organizing Committee.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

X. Li and J. Huang. 2009. Shc: A spectral algorithm for hierarchical clustering. In *2009 International Conference on Multimedia Information Networking and Security*, volume 2, pages 197–200.

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.

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.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.

Damian Radcliffe and Payton Bruni. 2019. *State of Social Media Middle East: 2018*. University of Oregon Libraries.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic Identification of Arabic Dialects in Social Media. In *Proceedings of the Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 22–27, Dublin, Ireland.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media.

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.

Younes Samih, Mohamed Eldesouki, Mohammed Attia, Kareem Darwish, Ahmed Abdelali, Hamdy Mubarak, and Laura Kallmeyer. 2017. Learning from relatives: Unified dialectal Arabic segmentation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 432–441, Vancouver, Canada. Association for Computational Linguistics.

Wajdi Zaghouani and Anis Charfi. 2018a. Araptweet: A large multi-dialect twitter corpus for gender, age and language variety identification. *CoRR*, abs/1808.07674.

Wajdi Zaghouani and Anis Charfi. 2018b. Guidelines and annotation framework for arabic author profiling.

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 Conference of the Association for Computational Linguistics (ACL)*, pages 37–41.

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

Chiyu Zhang and Muhammad Abdul-Mageed. 2019. No army, no navy: Bert semi-supervised learning of arabic dialects. pages 279–284.