UL2C: Mapping User Locations to Countries on Arabic Twitter

Hamdy Mubarak and **Sabit Hassan**
Qatar Computing Research Institute
Doha, Qatar
{hmubarak,sahassan2}@hbku.edu.qa

Abstract

Mapping user locations to countries can be useful for many applications such as dialect identification, author profiling, recommendation systems, etc. Twitter allows users to declare their locations as free text, and these user-declared locations are often noisy and hard to decipher automatically. In this paper, we present the largest manually labeled dataset for mapping user locations on Arabic Twitter to their corresponding countries. We build effective machine learning models that can automate this mapping with significantly better efficiency compared to libraries such as geopy. We also show that our dataset is more effective than data extracted from GeoNames geographical database in this task as the latter covers only locations written in formal ways.

1 Introduction

Twitter is one of the most popular social media platforms in the Arab region. Spoken across more than 20 countries, Arabic is one of the most dominant languages on Twitter. Arabic Twittersphere presents us with an audience of diverse demographics. Identifying countries of the users can help in various NLP tasks such as dialect identification, author profiling, and recommendation systems. Identifying geolocation of Twitter users can also help in event detection (Agarwal et al., 2012) or disaster management (Earle et al., 2012; Carley et al., 2016).

Deducing a Twitter user's location from geotagged tweets is difficult because less than 1% of tweets are geotagged (Cheng et al., 2010; Hecht et al., 2011). Although Twitter provides an option to users to declare their location in their profile, this is often noisy. Users can choose to specify their location at the level of countries, regions, cities or towns. Many of these names are written in informal way and sometimes in mixed languages. Some of these names also contain emojis and special symbols. This makes it difficult to automatically infer the country of many users. This complication often prompts researchers to manually annotate user profiles for their countries. Many works therefore, (e.g., (Bouamor et al., 2019; Charfi et al., 2019)), manually annotate user profiles for their locations. Being expensive, manual annotation often limits size of datasets. This is evident in the datasets by Bouamor et al. (2019) and Charfi et al. (2019) since they both contain around 3K manually annotated users.

Related works for Arabic primarily focus on dialect identification (e.g., (Bouamor et al., 2019; Abdelali et al., 2020; Zaidan and Callison-Burch, 2011), many of which involve manual annotation of dialects for sentences. Our focus in this work significantly differs from dialect identification since our purpose is to provide a dataset that can be used to map Twitter user locations to countries, which in turn, can aid in NLP tasks such as dialect identification or event detection. To our knowledge, there has been very few works for Arabic that map noisy user locations to countries, with work by Mubarak and Darwish (2014) being one of the most notable ones and the closest to our work. Mubarak and Darwish (2014) map the most common 10K locations to Arab countries in order to build a multi-dialectal Arabic corpus. We refer to this dataset as **Top10KLoc**. Our work extends Top10KLoc by increasing unique user locations from most common 10K to random 28K user locations obtained from 160K locations that are self-declared by users[1].

The contributions of this paper are summarized below:

[1] About 2.2K locations from Top10KLoc appear in our dataset.

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 145–153
Kyiv, Ukraine (Virtual), April 19, 2021.

- We present **UL2C**; the largest dataset for mapping user locations on Arabic Twitter to countries which contains more than 28K unique locations.
- We perform analysis of the data collected, identifying key characteristics.
- We show that by using machine learning models trained on our dataset, we can achieve significantly better results compared to existing libraries (e.g. geopy package) or resources (e.g. GeoNames or Top10KLoc[2] datasets). Models trained on our dataset achieve macro-average F1 score of 88.1, which outperforms similar models trained on other datasets with at least 26% relative improvement.
- We provide a web interface that users can use to map user locations to countries.

2 Related Work

There has been a number of works that focus on identifying locations of Twitter users. Krishnamurthy et al. (2008) study geographical growth of Twitter using UTC offset data about tweets. Hecht et al. (2011) built Bayesian probability models on tweets to predict countries and state-level locations. They cover four countries (United States, the United Kingdom, Canada, and Australia) and 50 states of United States. Miura et al. (2017) use combination of text, metadata and user network representation with neurual network and utilize attention mechanism to predict geolocation. Rahimi et al. (2018) use combination of text and user network with Graph Convolutional Neural Network to predict user geolocation. Rahimi et al. (2015) note that although user network information can improve results, scaling to large datasets may become an issue. Huang and Carley (2019) build hierarchical neural networks for identifying location of Twitter users. Mahmud et al. (2014) also use hierarchical approach to predict home location of Twitter users. The authors also identify if users are traveling to improve accuracy of location detection.

Despite Arabic being one of the most popular languages on Twitter, there has been very few works aimed at mapping user location in the Arab region to countries. The related field of dialect identification has received significant attention recently. Some works identify country-level dialects of Arabic tweets (e.g., (Abdelali et al., 2020) and some focus on dialects at user-level (Bouamor et al., 2019). While Abdelali et al. (2020) automatically labeled 500K+ tweets for their country-level dialects, Bouamor et al. (2019) manually labeled about 3K users for their countries of origin. Some works targeted region level classification of Arabic dialects. Zaidan and Callison-Burch (2011) collected a 52M word dataset from newspapers and annotated them for dialects of 5 Arab regions, namely, Maghrebi, Egyptian, Levantine, Gulf, and Iraqi. Alshutayri and Atwell (2017) and El-Haj et al. (2018) annotate texts from the five regions. Some other works target city level classification of Arabic dialects. Salameh et al. (2018) present a parallel corpus of dialects of 25 cities and Abdul-Mageed et al. (2018) construct a dataset representing dialects of 29 cities from 10 Arab countries. Charfi et al. (2019) presents a corpus of 5M tweets of about 3K users from 17 different Arab countries.

As discussed earlier, the closest work to ours that targets converting user locations directly to countries is by Mubarak and Darwish (2014). The authors collect 10K top user locations from 92M tweets. To map these locations to countries, the authors first use GeoNames dataset and then manually revise them. Further, the authors show that after filtering out non-dialectal tweets, the countries obtained from the user locations can be strong indicator of dialects for the remaining tweets.

3 Data Collection

We used twarc search API[3] to collect Arabic tweets. During the years 2018, 2019 and 2020, we collected 88M tweets with language filter set to Arabic (lang:ar). From unique users who authored those tweets, we randomly selected 160K unique users for which we obtained 28K unique user locations.

The user locations are information provided by Twitter users, and they can be real locations (e.g., country and city names written in formal, informal ways or nicknames), landmarks, or unreal locations, and can be written in any language. In our data collection, we found that 62% of users pro-

[2] Mubarak and Darwish (2014) didn't build an automatic classifier for mapping locations to countries.

[3] https://github.com/DocNow/twarc

vided non-empty locations. Around 60% of the non-empty locations are written in Arabic and 40% are written in other languages, mainly in English.

4 Data annotation

We used geopy[4] to map user locations to countries. geopy is a Python package that locates the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources. It includes geocoder classes for the OpenStreetMap Nominatim, Google Geocoding API (V3), and many other geocoding services. Figure 1 shows the output from geopy for an arbitrary location. In our study, we focus on mapping user locations to countries and we use ISO 3166-1 alpha-2 for country codes[5]. We observed that geopy has difficulties in identifying many locations when they are short, have special characters, or unreal locations in addition to many correct locations. Table 1 shows some examples for these errors.

We used geopy output as an initial mapping of user locations to countries then all unique user locations were revised manually by an Arabic native speaker.

In addition to mapping clear locations to countries, the annotator was asked to consider any clues in user location string that indicate belonging to a specific country. Some common examples and special annotation cases are shown in Table 2.

We randomly selected 500 unique user locations and checked annotation quality. We agreed with the manual annotations in 98% of the cases which indicates that annotation quality is very high.

User locations and their country mapping (UL2C dataset) can be downloaded from this link: https://alt.qcri.org/resources/UL2C-UserLocationsToCountries.tsv

4.1 GeoNames Dataset

GeoNames geographical database covers all countries and contains over 11M placenames whereof 4.8M populated places and 13M alternate names. Users can manually edit, correct and add new names using a user friendly interface. Dataset can be downloaded from: https://www.geonames.org/.

Figure 2 shows some information about Damascus, the capital of Syria, and its alternate names written in tens of languages as obtained from GeoNames. We extracted Arabic and English names of all places with population of 10K or more[6]. The figure shows also an example of the excluded locations that we anticipate users will not use to describe their locations. We ended up with having a list of 66K English location names (ASCII name field) and a shorter list of 13K Arabic names for some of them.

In the experiments section, we will examine the efficiency of using GeoNames to identify countries of Twitter users.

5 Data Analysis

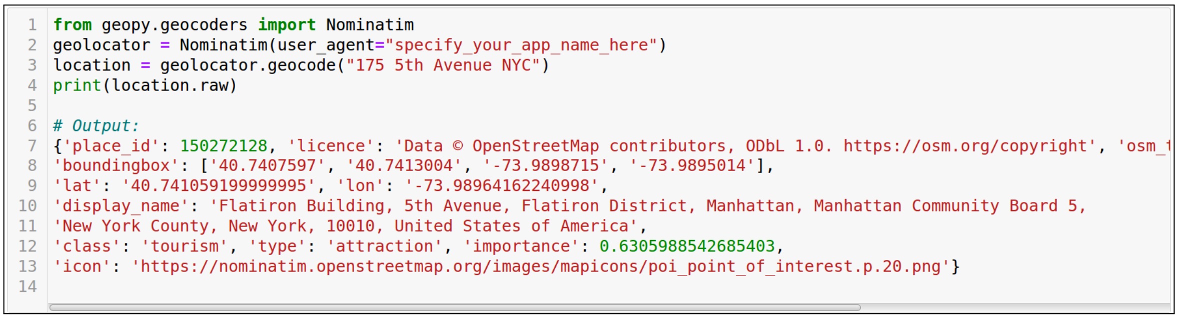
By assigning countries to unique user locations, we could map locations of ≈ 90K users to countries which represent 56% of the 160K users in our dataset. Many of user locations were either empty (38%) or cannot be mapped to a specific country (6%). Distribution of user countries is shown in Figure 3. Although there are 22 Arab countries, in our collection we didn't find locations from two countries, namely Djibouti and Comoros. We observe that users from Saudi Arabia (SA) represent more than half of Arab Twitter users. Around 70% of Twitter users are from Gulf region (countries: SA, KW, OM, AE, QA and BH), 4% of users are from Levant region (JO, PS, LB and SY), 3% of users are from Maghreb region (DZ, LY, MA and TN), and users from other regions (EG, YE and IQ) are in between. It's worth mentioning that country distribution obtained from UL2C and Top10KLoc datasets are very similar which indicates that most probably any random big collection of tweets will have similar country distribution.

From the geographical map of all Twitter users shown in Figure 4, we can see that Arabic tweets come from almost all world countries. The top 5 countries outside the Arab World where Arabic tweets come from are: US (United States), GB (United Kingdom), TR (Turkey), DE (Germany) and FR (France) in order. This can give an estimation about countries that Arabs live in

[4] https://pypi.org/project/geopy/

[5] https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

[6] We want to exclude river names, mountain names, lakes, etc.



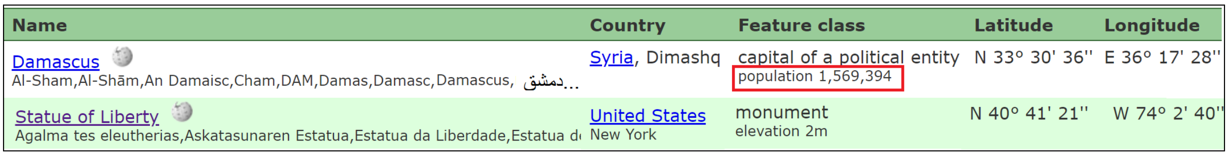
```python
from geopy.geocoders import Nominatim
geolocator = Nominatim(user_agent="specify_your_app_name_here")
location = geolocator.geocode("175 5th Avenue NYC")
print(location.raw)

# Output:
{'place_id': 150272128, 'licence': 'Data © OpenStreetMap contributors, ODbL 1.0. https://osm.org/copyright', 'osm_
'boundingbox': ['40.7407597', '40.7413004', '-73.9898715', '-73.9895014'],
'lat': '40.741059199999995', 'lon': '-73.98964162240998',
'display_name': 'Flatiron Building, 5th Avenue, Flatiron District, Manhattan, Manhattan Community Board 5,
'New York County, New York, 10010, United States of America',
'class': 'tourism', 'type': 'attraction', 'importance': 0.6305988542685403,
'icon': 'https://nominatim.openstreetmap.org/images/mapicons/poi_point_of_interest.p.20.png'}
```

Figure 1: geopy example

Location	Translation	geopy Country	Correct Country	Case
ولاية صحار سلطنة عمان	Sohar, Sultanate of Oman	-	OM	Unknown
Al Taif - Ksa	Ta'if city - KSA	SD	SA	Transliteration
الوطن العربي	the Arab World	JO	Unknown	Unspecified location
7th sky		DE	Unknown	Unreal location
cario	Cairo (capital of Egypt)	SA	EG	Letter case
Q8	Abbreviation for Kuwait	IT	KW	Short text, abbrev.
(11 . !)		FR	Unknown	Punctuation, numbers

Table 1: geopy sample errors



Name	Country	Feature class	Latitude	Longitude
Damascus Al-Sham,Al-Shām,An Damaisc,Cham,DAM,Damas,Damasc,Damascus, دمشق...	Syria, Dimashq	capital of a political entity population 1,569,394	N 33° 30' 36''	E 36° 17' 28''
Statue of Liberty Agalma tes eleutherias,Askatasunaren Estatua,Estatua da Liberdade,Estatua de	United States New York	monument elevation 2m	N 40° 41' 21''	W 74° 2' 40''

Figure 2: GeoNames examples: First place is included and second place is excluded from our dataset

outside Arab countries. We notice also that tweets from US are more than those from individual Maghreb countries like DZ and TN.

Figure 5 shows the most common words used in user locations for 4 countries, namely SA (Saudi Arabia), EG (Egypt), SY (Syria) and DZ (Algeria) which represent major regions in the Arab World (Gulf, Levant, Nile Basin and Maghreb regions respectively). We can see country and major city names are written in bigger font in different languages. For example, while majority of names are written in Arabic in SA (Gulf region), they are written in Arabic, English and French in DZ (Maghreb region). Arabic and English names are widely used in EG and SY. This gives an indication about the popularity of language usage across different regions in the Arab World.

6 Experiments and Results

In this section, we compare effectiveness of mapping user locations to countries with classifiers trained on different datasets, namely, UL2C, Top10KLoc, GeoNames and combination of these datasets. In our experiments, we merge all countries that are outside Arab region to ”UNK/OTHER” class, yielding a total of 21 classes (20 Arab countries + Unknown/Other).

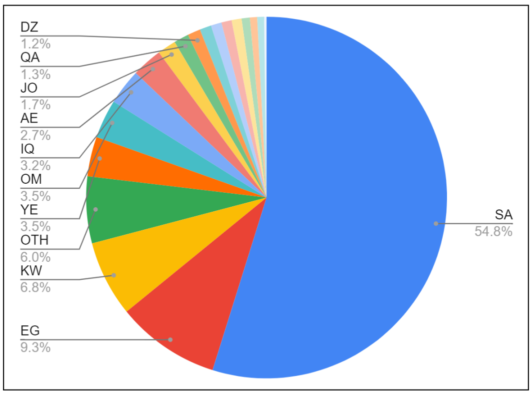


Figure 3: Country distribution of Twitter users

6.1 Preprocessing text

In order to reduce noise in user-declared locations, we perform the following preprocessing steps:

Location	Translation	Country	Case
مدينة رسول الله	City of Prophet Mohamed	SA	Common knowledge in Arabic culture
xx@outlook.kw		KW	Website/email domain name
كعبة المضيوم	Kaaba (Destination) of the oppressed	QA	Nickname or informal common name
WhatsApp 00964xx		IQ	Country calling code
لبناني وأفتخر	Lebanese and proud	LB	Nationality
Alhilal_FC	Al Hilal Saudi Football Club	SA	Club fans
دبي - لندن	Dubai-London	AE	Multiple countries (1st is selected)

Table 2: Annotation clues and special cases

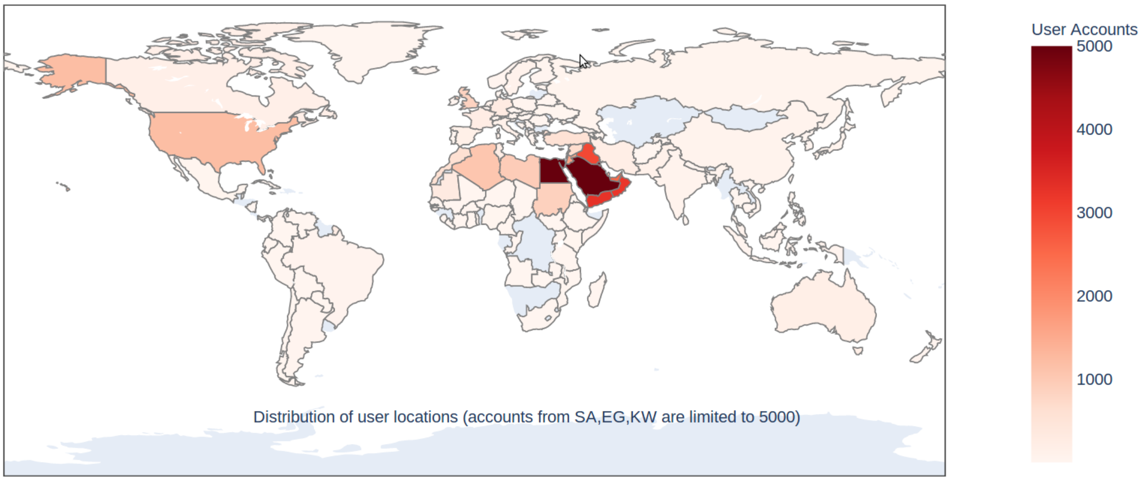

User Accounts
5000
4000
3000
2000
1000
Distribution of user locations (accounts from SA,EG,KW are limited to 5000)

Figure 4: Country map of Twitter users

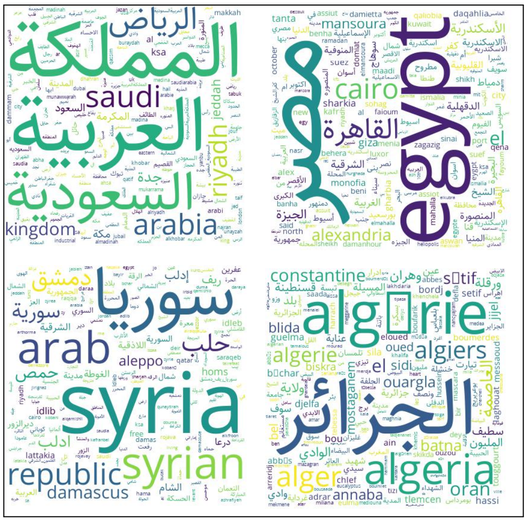


Figure 5: Top locations in 4 countries: SA, EG, SY and DZ in order

- Remove all Arabic diacritics
- Remove all punctuation marks and emojis
- Convert all Latin letters to lowercase
- Normalize different shapes of Arabic Hamza, final Alif Maqsoura and Taa Marbouta letters to plain Alif, Yaa and Haa letters respectively
- Convert English and Arabic decorated letters (e.g. some Farsi letters) to original letters using the mapping list shared by (Mubarak and Abdelali, 2016).For example, we map “α, β” to “a, B” and “پ ، چ ، گ” to “ب ، ج ، ک” in order.

In summary, a location like “$\alpha\beta$HA1 :) أَبْهَا*, ” is converted to “abha ابها” (city in SA) after decorated letters mapping and applying other normalization steps.

6.2 Features

Word n-gram Since our input text is name of a location and therefore, typically short, we limit range of word n-gram to [1-2]. We use term frequency-inverse term document frequency (tf-idf) for weighting the n-gram vectors.

Character n-gram We experimented with character n-grams ranging from [2-3] to [2-5], beyond which, we did not see any further improvement. Similar to word n-gram, we used tf-idf weighting.

Data	Model	Features	Acc	P	R	F1
-	geopy (baseline)	-	68.1	78.2	65.3	69.2
UL2C	SVM	Word [1-2]	90.1	93.1	81.8	86.7
		Char [2-3]	90.2	93.7	80.6	86.0
		Char [2-4]	91.1	**94.4**	83.1	88.0
		Char [2-5]	**91.2**	94.1	**83.5**	**88.1**
GeoNames	SVM	Word [1-2]	59.9	58.6	44.5	46.6
		Char [2-3]	39.7	56.9	35.0	38.9
		Char [2-4]	47.0	60.4	47.4	48.8
		Char [2-5]	48.7	60.7	49.2	50.1
Top10KLoc	SVM	Word [1-2]	83.2	85.0	62.1	70.5
		Char [2-3]	79.4	85.2	52.7	63.2
		Char [2-4]	80.6	85.2	55.0	65.2
		Char [2-5]	81.2	85.4	56.6	66.5
UL2C + GeoNames	SVM	Word [1-2]	90.2	90.1	84.3	86.7
		Char [2-3]	90.3	90.5	83.2	86.1
		Char [2-4]	91.5	90.2	86.7	87.9
		Char [2-5]	91.6	90.7	86.7	88.2
UL2C + Top10KLoc	SVM	Word [1-2]	90.0	93.7	81.2	86.6
		Char [2-3]	89.4	94.5	78.7	85.2
		Char [2-4]	90.4	**95.0**	81.5	87.3
		Char [2-5]	90.6	**95.0**	81.8	87.5

Table 3: Evaluation of different datasets

6.3 Classification Models

geopy baseline The geopy library acts as our baseline. We call library with Twitter user locations and extract countries they are mapped to by the library.

Support Vector Machines (SVMs) SVMs have traditionally been used for many classification tasks. Even in recent Arabic text classification tasks such as offensive language identification (Hassan et al., 2020b,a), spam detection (Mubarak et al., 2020), adult content detection (Mubarak et al., 2021), and dialect identification (Abdelali et al., 2020), SVMs have shown promising results. We used LibSVM implementation with default parameters by scikit-learn[7] for training.

6.4 Experiment results

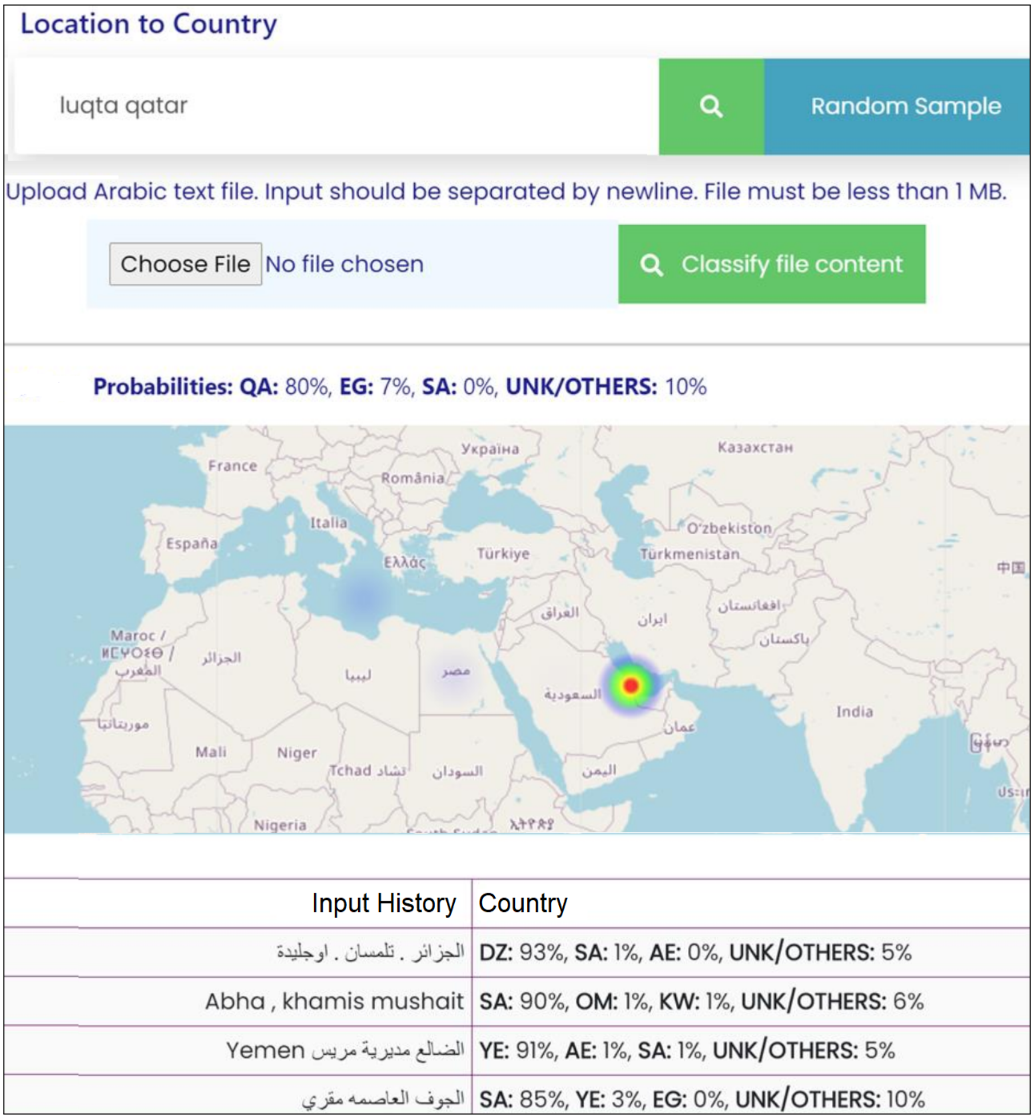
geopy baseline Serving as our baseline model, geopy achieves F1 score of 69.2 when evaluated on UL2C dataset.

GeoNames We trained several classifiers on GeoNames dataset and evaluated on our UL2C dataset. The classifiers yielded very poor results initially. We noticed that many of the locations were outside Arab region which likely caused the poor performance (omitted here for conciseness). To address this problem, we rebalanced the dataset to have equal number of locations from within and outside the Arab region. This yielded a total of 8,632 unique instances (4,316 from the Arab region and 4,316 from outside the Arab region). The results are summarized in Table 3. We can see that even after rebalancing the data, with maximum F1 score of 50.1, the classifiers trained on GeoNames are outperformed by all others.

Top10KLoc We trained similar set of classifiers on Top10KLoc dataset and evaluated on our dataset. The models were seen to outperform geopy by a small margin and GeoNames by a large margin with maximum F1 score of 70.5 (see Table 3).

UL2C (Our dataset) We performed 5-fold cross-validation with same set of classifiers on the dataset presented in this paper. The results were seen to improve by a significant margin since the best results were obtained by SVM trained with character [2-5] n-grams, a relative improvement of 26% in F1 score from previous best (70.5) when trained on Top10KLoc.

[7] http://scikit-learn.org/



Probabilities: QA: 80%, **EG:** 7%, **SA:** 0%, **UNK/OTHERS:** 10%

Input History	Country
الجزائر . تلمسان . اوجليدة	**DZ:** 93%, **SA:** 1%, **AE:** 0%, **UNK/OTHERS:** 5%
Abha , khamis mushait	**SA:** 90%, **OM:** 1%, **KW:** 1%, **UNK/OTHERS:** 6%
Yemen الضالع مديرية مريس	**YE:** 91%, **AE:** 1%, **SA:** 1%, **UNK/OTHERS:** 5%
الجوف العاصمه مقري	**SA:** 85%, **YE:** 3%, **EG:** 0%, **UNK/OTHERS:** 10%

Figure 6: Online interface for mapping user locations to countries (https://asad.qcri.org)

UL2C + GeoNames In this set of experiments, we modified the cross-validation setting by adding GeoNames dataset with each of the 5 folds of our dataset during training. We can see from Table 3 that adding our dataset to GeoNames dataset offsets the lower performance when using GeoNames alone and with F1 score of 88.2, improves the results from using UL2C dataset alone by a small margin.

UL2C + Top10KLoc Lastly, we modified the cross-validation setting by adding Top10KLoc to each of the folds during training. From Table 3, we see a similar trend where the lower performance of using Top10KLoc only is offset by use of UL2C. However, there is no significant improvement when additional data is used compared to when using only UL2C.

7 Interface

We build an interface for users to map user locations to countries. The web interface is added to Arabic Social media Analytics and unDerstanding (ASAD) (Hassan et al., 2021) at `https://asad.qcri.org`. Figure 6 shows sample outputs from the website.

7.1 Design

The user can type user location to be mapped to countries. The user can also test random samples from UL2C dataset to see their mapping. This allows the user to easily understand the functionalities of the interface. The user is then shown probabilities of the location belonging to different countries. To help the users visualize distribution of possible countries related to the location, we display a heatmap of the probabilities. We also allow the user to upload a file consisting user locations. This allows users to map many user locations at the same time. We impose a restriction on file size in order to limit abuse of our system. The user is then able to download a file containing predictions and probabilities of the user locations belonging to different countries.

7.2 Implementation

We use Bootstrap[8] for our responsive frontend design. We use Flask[9], a python-based web development framework, for backend development and javascript for communication between backend and frontend. To visualize the heatmap, we use Leaflet.js[10] and Heatmap.js[11] with OpenStreetMap[12] map server.

8 Conclusion

In this paper, we have presented a large manually annotated and publicly available dataset of Twitter user locations from the Arab region, mapped to their respective countries. We analyzed different characteristics of our data such as country distribution and top locations. We built machine learning models that can use the data to map user locations to countries more effectively compared to existing resources such as geopy Python package or GeoNames geographical database. Lastly, we provide a web interface to access this service easily.

References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2020. Arabic dialect identification in the wild. *ArXiv*, abs/2005.06557.

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 Languages Resources Association (ELRA).

Puneet Agarwal, R. Vaithiyanathan, Saurabh Sharma, and G. Shroff. 2012. Catching the long-tail: Extracting local news events from twitter. In *ICWSM*.

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).

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.

Kathleen M. Carley, Momin Malik, Peter M. Landwehr, Jürgen Pfeffer, and Michael Kowalchuck. 2016. Crowd sourcing disaster management: The complex nature of twitter usage in padang indonesia. *Safety Science*, 90:48 – 61. Building Community Resilience to Global Hazards: A Sociotechnical Approach.

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

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *CIKM '10*.

Paul Earle, Daniel Bowden, and Michelle Guy. 2012. Twitter earthquake detection: Earthquake monitoring in a social world. *Annals of geophysics = Annali di geofisica*, 54.

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).

Sabit Hassan, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. 2021. Asad: Arabic social media analytics and understanding. In *Proceedings of the Software Demonstrations of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

Sabit Hassan, Younes Samih, Hamdy Mubarak, and Ahmed Abdelali. 2020a. ALT at SemEval-2020 task 12: Arabic and English offensive language identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1891–1897, Barcelona (online). International Committee for Computational Linguistics.

Sabit Hassan, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, Ammar Rashed, and Shammur Absar Chowdhury. 2020b. ALT submission for OSACT shared task on 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 61–65,

[8] https://getbootstrap.com/

[9] https://flask.palletsprojects.com/en/1.1.x/

[10] https://leafletjs.com/

[11] https://www.patrick-wied.at/static/heatmapjs/

[12] https://www.openstreetmap.org/

Marseille, France. European Language Resource Association.

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from justin bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 237–246, New York, NY, USA. Association for Computing Machinery.

Binxuan Huang and Kathleen Carley. 2019. A hierarchical location prediction neural network for twitter user geolocation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4732–4742, Hong Kong, China. Association for Computational Linguistics.

Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. 2008. A few chirps about twitter. pages 19–24.

Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. 2014. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology*, 5(3):1–21.

Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2017. Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1260–1272, Vancouver, Canada. Association for Computational Linguistics.

Hamdy Mubarak and Ahmed Abdelali. 2016. Arabic to english person name transliteration using twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 351–355.

Hamdy Mubarak, Ahmed Abdelali, Sabit Hassan, and Kareem Darwish. 2020. Spam detection on arabic twitter. In *Social Informatics*, pages 237–251, Cham. Springer International Publishing.

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, Sabit Hassan, and Ahmed Abdelali. 2021. Adult content detection on arabic twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2015. Twitter user geolocation using a unified text and network prediction model. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 630–636, Beijing, China. Association for Computational Linguistics.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2018. Semi-supervised user geolocation via graph convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2009–2019, Melbourne, Australia. Association for Computational Linguistics.

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.

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.