Classifying Arab Names Geographically

Hamdy Mubarak, Kareem Darwish
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
{hmubarak, kdarwish}@qf.org.qa

Abstract

Different names may be popular in different countries. Hence, person names may give a clue to a person's country of origin. Along with other features, mapping names to countries can be helpful in a variety of applications such as country tagging twitter users. This paper describes the collection of Arabic Twitter user names that are either written in Arabic or transliterated into Latin characters along with their stated geographical locations. To classify previously unseen names, we trained naive Bayes and Support Vector Machine (SVM) multi-class classifiers using primarily bag-of-words features. We are able to map Arabic user names to specific Arab countries with 79% accuracy and to specific regions (Gulf, Egypt, Levant, Maghreb, and others) with 94% accuracy. As for transliterated Arabic names, the accuracy per country and per region was 67% and 83% respectively. The approach is generic and language independent, and can be used to collect and classify names to other countries or regions, and considering language-dependent name features (like the compound names, and person titles) yields to better results.

1 Introduction

Geo-locating tweets and tweeps (Twitter users) has captured significant attention in recent years. Geographical information is important for many applications such as transliteration, social studies, directed advertisement, dialect identification, and Automatic Speech Recognition (ASR) among others. In social studies, researchers may be interested in studying the views and opinions of tweeps for specific geographical locations. Similarly, tweets can offer a tool for linguists to study different linguistic phenomena. For ASR, training language models using dialectal Arabic tweets that are associated with different regions of the Arab world was shown to reduce recognition error rate for dialectal Egyptian Arabic by 25% (Ali, et. al, 2014).

Previous work has looked at a variety of features that may geo-locate tweets and tweeps such as the dialect of tweet(s), words appearing in tweets, a tweep's social network, etc. In this work we examine the predictive power of tweep names in predicting a tweep's location or region of origin. We define geographic units at two different levels, namely: country level and region level. The country level geographic units are defined based on political boundaries regardless of the size and proximity of different geographic entities. Thus, Qatar and Bahrain as well as Lebanon and Syria are considered as different units. At the region level, we conflate nearby countries into regions. Conflation was guided by previous work on dialects, where dialects were categorized into five regional language groups, namely: Egyptian (EGY), Maghrebi (MGR), Gulf (Arabian Peninsula) (GLF), Iraqi (IRQ), and Levantine (LEV) (Zbib et al., 2012; Cotterell et al., 2014). Sometimes, the Iraqi dialect is considered to be one of the Gulf dialects (Cotterell et al., 2014). In this paper we consider Iraq as a part of the Gulf region.

Thus the goal of this work is to build a classifier that can predict a tweep's country/region of residence/origin. To build the classifier we obtained tweep names and their self-declared locations from Twitter. Many tweeps use pseudonyms, such as "white knight", and fake or irregular, such as "in phantasmagoria" or "Eastern Province". Hence, identifying fake tweep names may be necessary, and

locations need to be mapped to countries. We built multiple classifiers using either a naive Bayes or a Support Vector Machine (SVM) classifier using bag-of-words features, namely word unigrams. We also considered improvements that entailed using character n-gram features and word position weighting. For our work, we tried to collect tweets for all 22 Arab countries, but we did not find Arabic tweets from Mauritania, Somalia, Djibouti and Comoros. The contributions of this paper are:

1. We show that we can use Twitter as a source for collecting person names for different Arab countries by mapping user location to one of the Arab countries.
2. We show that we can build a classifier of Arabic names at the county level or region level with reasonable accuracy.
3. we show the characteristics of Arabic names and how they differ among different countries or regions.

The paper is organized as follows: Section 2 surveys previous work on person name classification; Section 3 describes some features of Arabic names including dialectal variation in transliteration; section 4 describes how names are collected from Twitter, cleaned and classified; section 5 shows results of name classification experiments; and Section 6 contains conclusion and future work.

2 Previous Work

The problem of classifying names at country level is not well explored. As far as we know, there are no studies for Arabic person name classification. Some work has been done on clustering and classifying person names by origin like (Fei et al., 2005), where they used the LDC bilingual person name lists to build a name clustering and classification framework. They considered that several origins may share the same pattern of transliteration and applied their technique to a name transliteration task by building letter n-gram language models for source and target languages. They clustered names into typical origin clusters (English, Chinese, German, Arabic., etc.).

Balakrishnan (Balakrishnan, 2006) extracted a list of person names from the employee database of a multinational organization covering 9 countries: US, UK, France, Germany, Canada, Japan, Italy, India, and China. Equal number of names is chosen from each country (1,000 names for each). He used pattern search for first and second names and used k-nearest neighbor and Levenshtein edit distance to measure the distance between two names. He reported a classification accuracy = 0.67 for supervised training set and 0.63 for unsupervised training set.

Fu et al. (Fu et al., 2010) mentioned that humans often identify correctly the origins of person names, and there seem to be distinctive patterns in names to distinguish origins. They constructed an ontology containing all linguistic knowledge that can directly contribute to language origin identification, and this was employed for the analysis of name structure. They reported an average performance of 87.54% using ME-based language identifier for 8 languages (Arabic, Chinese, English, French, German, Japanese, Russian, and Spanish-Portuguese).

Rao et al. (Rao et al., 2010) classified the latent user attributes including gender, age, regional origin, etc., using features like n-grams models and number of followers/followees (in a social graph information) among others.

Mahmud et al. (Mahmud et al., 2012) collected tweets using the geo-tag filter option on Twitter until they received tweets from 100 unique users from the top 100 cities in US. They used this corpus for inferring home locations of users at the level of their cities. They reported a recall of 0.7 for 100 cities.

Huang et al. (Huang et al., 2014) discussed the challenges of detecting the nationality of Twitter users using profile features and they studied the effectiveness of different features for inferring nationalities. They reported an accuracy of 83.8% for these nationality groups: Qatari, Arabs, Western, Southeast Asia, Indian, and Others. They mentioned that due to the unbalanced data distribution, the performance of less populated groups is not very high. We observe similar results in this paper.

3 Person Names in Arabic

3.1 Compound Names

Single Arabic names typically are made up of single words, but sometimes they may be composed

of 2 or 3 words. We refer here to single names with more than one word as 'compound names'. There are some words such as الله (Allh[1] – meaning "God") and الدين (Aldyn – meaning "religion") that trail other words as in عبد الله (Ebd Allh – meaning "slave of Allah") constructing the name "Abdullah" and as in صلاح الدين (SlAH Aldyn – meaning "perfection of religion") constructing the name "Salahudin" (Saladin). In some countries, father and family names are often preceded by words meaning "son of" such as بن (bn), ابن (Abn) or ولد (wld) or the word آل (|l – meaning family of). An example that combines the aforementioned variations of compound names is the name of the former king of Saudi Arabia عبد الله بن عبد العزيز آل سعود (Ebd Allh bn Ebd AlEzyz |l sEwd – "Abdullah ibn Abdelaziz Aal Saud"). A list of common words used in compound names are listed in table 1. When processing the names in our collection, we heuristically split the full names into single Arabic names, whether compound or not. As in the previous example, عبد الله بن عبد العزيز آل سعود (Ebd Allh bn Ebd AlEzyz |l sEwd), it was split into: عبد الله (Ebd Allh), بن عبد العزيز (bn Ebd AlEzyz), and آل سعود (|l sEwd). The heuristic involved always attaching the words marked in Table 1 as *pre* to the trailing words and ones that are marked as post to preceding words.

Type	Word	Example
Pre	الله، الدين، الإسلام، الرسول Allh, Aldyn, Al<slAm, Alrswl	سيف الإسلام syf Al<slAm
Post	بو،أبي،أبا،أبو،أم، بنت،ابن، بن،عبد، ولد Ebd, bn, Abn, bnt, >m, >bw, >bA, >by, bw, wld	بنت ناصر ولد محمد bnt nASr, wld mHmd

Table 1: Words that are parts of a name.

3.2 Dialectal Variations of Names

Names in Arabic are normally written without diacritics, and when they are transliterated, these hidden diacritics are shown in addition to dialectal differences in pronunciation among countries as shown in table 2. Since we are classifying names that are written in both Arabic and Latin scripts, spelling variations can perhaps be helpful in ascertaining the country/region of origin.

3.3 Religion and Gender

Names can also be indicative of other attributes such as religion and gender. For example, the names شنودة ($nwdp – "Shnouda"), عبد الحسين (Ebd AlHsyn – "Abdul Hussein"), and عمر (Emr – "Omar") are typically Coptic, Shia, and Sunni respectively. And for gender, feminine names frequently end with ة، اء، ى، ا (p, A', Y, A), such as فاطمة (FaTmp – "Fatima") and هناء (hnA' – "Hannah"). Second names, either father or family names, are mostly masculine. Though guessing a tweep's religion and gender are interesting, such is beyond the scope of this paper.

Name Variations	Phonetic Mapping
ماجد (mAjd) Ma**g**ed (EGY), Ma**j**ed (GLF)	g/j
عثمان (EvmAn) Os**h**man (EGY), O**th**man (GLF)	s/th
أشرف (A$rf) A**sh**arf (EGY), A**ch**raf (MGR)	sh/ch
فهد (fhd) Fahd (EGY), Fah**a**d (GLF)	diacritics
الوكيل (Alwkyl) **El** Wakil (EGY), **Al** Wakil (GLF)	Determiner

Table 2: Dialectal effects on Transliteration.

4 Data Collection

Twitter user profiles contain user-declared information like: Twitter account name, screen name **(user name)**, user location, description, etc. User names are normally written in Arabic or Latin characters, and user locations are written in full or abbreviated, formal or informal, etc. as shown in Figure 1.

We used the Twitter4J[2] interface to the Twitter API to collect Arabic tweets during the whole of

[1]Buckwalter transliteration is used exclusively in the paper

[2]http://twitter4j.org

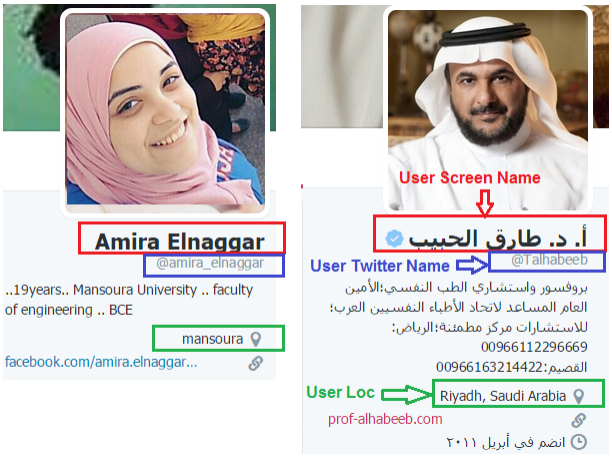

Figure 1: User Profile Information

March 2014. We searched using the query "lang:ar", which indicates any Arabic tweet. In all we collected 175 million tweets that were authored by 5.5 million unique tweeps. We used the users self-declared locations to map them to countries. We mapped the locations using the GeoNames[3] geographical database, which contains 8M place names and a database of of the most commonly used 10,000 user locations on Twitter (Mubarak et al., 2014). If the location referred to two or more different countries, as in "UK and Kuwait", it was removed. User location was successfully mapped to one of the Arab countries for 1M unique user names. After name cleaning (described later in this section), we have 170 thousand Names$_{arb}$ and 182K Names$_{trans}$ that are considered as valid names and mapped to only one country.

Per-country distributions are shown in Figure 2 and Figure 3. One of the interesting observations from these figures is that people from Saudi Arabia (SA[4]) are the majority in both cases, and they tend to write their names in Arabic, while people from Egypt (EG) tend to write their names as transliterated. We opted not to limit our collection to tweeps who have geo-tagged tweets (tweets with latitude and longitude), because geo-tagged tweets represent less than 1% of the total number of tweets[5]. We found that 0.3% of the collected tweets are geo-tagged.

Table 3 shows some examples of the collected names. We took samples of 200 random names from each set and found that 70% of the names are real and the rest are unreal person names (fake). We plan to identify fake names from real names in future.

Name cleaning included ignoring words that are composed of single letters, special characters outside the Arabic or the Latin alphabets, entries that are single words only, and entries having stopwords. Names were normalized in the manner described by Darwish et al. (2012), which involved removing diacritics, kashidas, normalizing different forms of alef, ya and alef maqsoura, and ha and ta marbouta, and mapping letters from other languages such as Farsi that use the Arabic script to Arabic letters. Further, titles, such as Dr., and numbers were removed. We also identified compound names as described earlier. For example, the user name "Dr. Abdullah Bin Fahad AL MUTAIRI1973" will be normalized to "abdullah bin_fahad al_mutairi".

User name	Real/Unreal
طلال القحطاني (TlAl AlqHTany), Bassam Jawad	Real names
أنيقة وكفى (Anyqa wKfY), Sweet Boy	Unreal names

Table 3: Examples of user names

5 Name Classification Experiments

Given the 170K Names$_{arb}$ and 182K Names$_{trans}$ that we collected, we randomly split the set into 80/20 training and testing splits. We used word unigrams as features. We also examined giving first and last names different weights and character trigrams as a back-off for unseen words. Further, we trained two classifiers namely a Naive Bayes classifier and an SVM classifier. When using a Naive Bayes classifier and a name was not observed during training in general or for a class, we used KenLM language modeling toolkit to compute the smoothing probability of it (Heafield, 2011).

Our baseline involved tagging all test items with the tag of the majority class, which means that every tweep would assigned to SA at country level and the Gulf at region level. Table 4 shows the baseline re-

[3]http://www.geonames.org

[4]We use "ISO 3166-1 alpha-2" for country codes

[5]http://thenextweb.com/2010/01/15/twitter-geofail-023-tweets-geotagged/

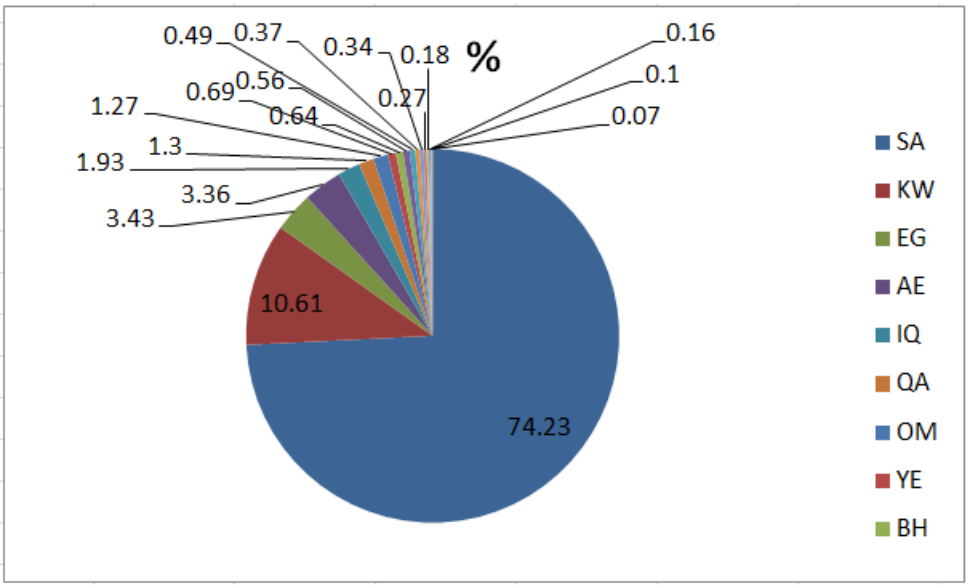

Figure 2: Country Distribution for Names$_{arb}$

sults in term of accuracy. Precision for the majority class would be identical to the overall accuracy and recall would be one. Precision and recall would be zero for all the other classes.

Name type	**Accuracy**
Names$_{arb}$ Country	74.2%
Names$_{arb}$ Region	91.4%
Names$_{trans}$ Country	44.3%
Names$_{trans}$ Region	67.4%

Table 4: Baseline Results

Table 5 and Table 6 show the results for Names$_{arb}$ per country and per region respectively using word unigrams only. Similarly, Table 7 and Table 8 show the results for Names$_{trans}$ per country and per region respectively using word unigrams only. Micro and Macro averages refer to computing metrics per test example or taking the average of per country results respectively. As can be seen, the naive Bayes classifier performed better than SVM classifier for the vast majority of countries and in overall accuracy and F-measure. Mostly the SVM classifier had higher precision with less recall.

In further experiments, we exclusively used the naive Bayes classifier. We tried two modifications of the classifier. The first involved giving different weights to different single names in the full name, such that a person's last name would get a higher weight than his/her first name. The intuition is that different countries may have different common family names that may indicate their place of origin, family, or tribe. The weight of the word based on its position is determined using the following formula: $weight_i = \frac{1}{no\_of\_single\_names - i + 1}$ Where i ranged between 1 and number of single names in the full name. Thus the last single name would get a weight of 1 and all previous single names would get a weight of 1/2, 1/3, etc. (from end to beginning).

The second entailed using a character trigram model as a back-off for out of vocabulary words, which were not seen during training. We used KenLM to train a trigram character model using all the names in the training set (Heafield, 2011).

Table 9 and Table 10 compare the plain Bayesian classifier with using the classifier with single name weighting and character trigram back-off for Names$_{arb}$ at country and region level respectively. Table 11 and Table 12 compare the same for Names$_{trans}$. As the results show, both methods improved overall accuracy with consistent improvements in precision and improvements in recall most of the time. Using single name weighting had a greater effect on precision.

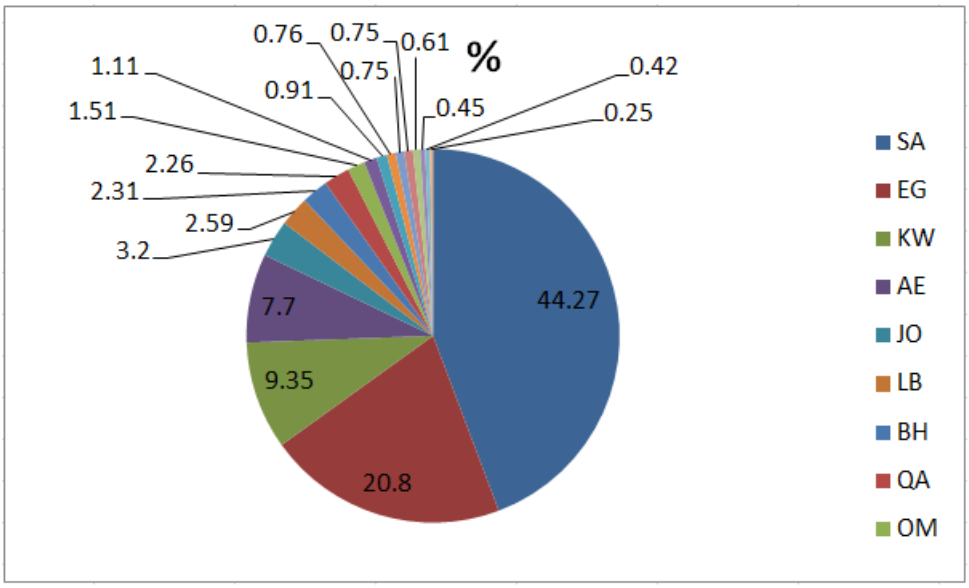

Figure 3: Country Distribution for Names$_{trans}$

6 Conclusion and Future Work

In this paper, we presented our work on classifying person names based on their country or region. To construct training data, we collected Twitter user names that authored Arabic tweets with their associated self-declared locations, which we mapped to Arab countries and regions. We experimented with Bayesian and SVM classifiers and the Bayesian classifier outperformed the SVM classifier most of the time. Adding position information and back-off to a character trigram model for names not observed during training generally improved results. Classifying user names at region level generally yielded better results than at country level.

Because majority of user names written in Arabic are from the Gulf region (93%), the classification improvement above the majority baseline was not that big, but when we applied the same approach for classifying transliterated user names, we achieved an increase of the accuracy by 52% and 20% at the country level and group level in order, and an increase in the F-measure by 135% and 46% at the country level and region level in order.

In future, we want to incorporate the user name feature in conjunction with other features in the context of geo-locating Twitter users. We need to test our engine for classifying names collected for each country from outside Twitter, think in other ways to collect user names from regions like the Maghreb, and detect more information from user profile like the gender and religion.

References

Ahmed Ali, Hamdy Mubarak, Stephan Vogel. 2014. Advances in Dialectal Arabic Speech Recognition: A Study Using Twitter to Improve Egyptian ASR. International Workshop on Spoken Language Translation (IWSLT 2014).

Balakrishnan, Raju. 2006. Country wise classification of human names. Proceedings of the 5th WSEAS Int. Conf. on Artificiall Intelligence, Knowledge Engineering and Data Bases, Madrid, 2006.

Ryan Cotterell and Chris Callison-Burch. 2014. A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. LREC-2014, pages 241–245.

Kareem Darwish, Walid Magdy, and Ahmed Mourad. 2012. Language processing for arabic microblog retrieval. Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012.

Huang, Fei, Stephan Vogel, and Alex Waibel. 2005. Clustering and classifying person names by origin. Proceedings of the National Conference on Artificial Intelligence. Vol. 20. No. 3. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

Fu, Yu, Feiyu Xu, and Hans Uszkoreit. 2010. Determin-

	NB			SVM		
Country	P	R	F	P	R	F
EG	0.40	0.50	0.45	0.48	0.14	0.22
DZ	0.18	0.16	0.17	0.71	0.07	0.13
SD	0.13	0.07	0.09	0.50	0.03	0.06
IQ	0.51	0.33	0.40	0.69	0.12	0.20
MA	0.20	0.07	0.10	0.00	0.00	0.00
SA	0.84	0.91	0.88	0.77	0.99	0.87
YE	0.35	0.17	0.22	0.50	0.02	0.03
SY	0.21	0.11	0.14	0.62	0.07	0.12
TN	0.04	0.03	0.04	0.50	0.03	0.06
AE	0.52	0.38	0.44	0.71	0.11	0.19
JO	0.21	0.11	0.14	0.62	0.03	0.06
LY	0.28	0.11	0.16	0.90	0.06	0.11
PL	0.18	0.10	0.13	0.83	0.03	0.06
LB	0.03	0.09	0.05	0.71	0.08	0.14
OM	0.50	0.29	0.37	0.85	0.07	0.14
KW	0.49	0.34	0.40	0.58	0.10	0.17
QA	0.50	0.23	0.32	0.63	0.06	0.12
BH	0.28	0.15	0.20	0.70	0.08	0.15
Macro Avg	0.33	0.23	0.26	0.63	0.12	0.16
Micro Avg	0.74	0.77	0.75	0.73	0.76	0.69
Accuracy	0.77			0.76		

Table 5: Names$_{arb}$ Results per country

	NB			SVM		
Country	P	R	F	P	R	F
EGY	0.40	0.50	0.45	0.48	0.14	0.22
GLF	0.96	0.96	0.96	0.94	0.99	0.97
LEV	0.19	0.14	0.16	0.70	0.05	0.09
MGR	0.24	0.13	0.17	0.65	0.05	0.09
OTHER	0.29	0.14	0.19	0.50	0.02	0.04
Macro Avg	0.42	0.38	0.39	0.66	0.25	0.28
Micro Avg	0.92	0.92	0.92	0.92	0.94	0.91
Accuracy	0.92			0.94		

Table 6: Names$_{arb}$ Results per region

ing the Origin and Structure of Person Names. LREC. 2010.

Heafield, Kenneth. 2011. KenLM: Faster and Smaller Language Model Queries. Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation. pp 187–197

Huang, Wenyi, Ingmar Weber, and Sarah Vieweg. 2014. Inferring nationalities of Twitter users and studying inter-national linking. Proceedings of the 25th ACM conference on Hypertext and social media. ACM, 2014.

Mahmud, Jalal, Jeffrey Nichols, and Clemens Drews. 2012. Where Is This Tweet From? Inferring Home Locations of Twitter Users. ICWSM. 2012.

Hamdy Mubarak, Kareem Darwish. 2014. Using Twitter to collect a multi-dialectal corpus of Arabic. ANLP 2014.

Rao, Delip, David Yarowsky, Abhishek Shreevats, Manaswi Gupta. 2010. Classifying latent user attributes in twitter. Proceedings of the 2nd international workshop on Search and mining user-generated contents. ACM, 2010.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, Chris Callison-Burch. 2012. Machine translation of Arabic dialects. NAACL-2012, pages 49–59.

	NB			SVM		
EG	0.68	0.78	0.73	0.68	0.43	0.53
DZ	0.18	0.19	0.18	0.33	0.08	0.14
SD	0.22	0.14	0.17	0.26	0.05	0.08
IQ	0.15	0.12	0.14	0.25	0.04	0.08
MA	0.38	0.27	0.32	0.50	0.23	0.32
SA	0.71	0.80	0.75	0.54	0.94	0.68
YE	0.00	0.00	0.00	0.50	0.01	0.02
SY	0.22	0.10	0.14	0.18	0.02	0.03
TN	0.29	0.32	0.30	0.46	0.16	0.24
AE	0.43	0.33	0.37	0.50	0.16	0.24
JO	0.38	0.24	0.29	0.39	0.07	0.13
LY	0.03	0.10	0.05	0.33	0.01	0.03
PL	0.13	0.08	0.10	0.22	0.01	0.02
LB	0.48	0.42	0.45	0.54	0.31	0.39
OM	0.61	0.40	0.48	0.69	0.13	0.22
KW	0.54	0.40	0.46	0.52	0.18	0.27
QA	0.44	0.23	0.31	0.75	0.06	0.11
BH	0.45	0.28	0.34	0.46	0.09	0.15
Macro Avg	0.35	0.29	0.31	0.45	0.17	0.20
Micro Avg	0.60	0.62	0.61	0.55	0.55	0.49
Accuracy	0.62			0.55		

Table 7: Names$_{trans}$ Results per country

	NB			SVM		
EGY	0.68	0.78	0.73	0.68	0.43	0.53
GLF	0.88	0.86	0.87	0.77	0.94	0.85
LEV	0.51	0.35	0.42	0.59	0.17	0.27
MGR	0.25	0.38	0.30	0.57	0.22	0.31
OTHER	0.22	0.10	0.14	0.15	0.04	0.06
Macro Avg	0.51	0.50	0.49	0.55	0.36	0.40
Micro Avg	0.79	0.79	0.79	0.73	0.75	0.72
Accuracy	0.79			0.75		

Table 8: Names$_{arb}$ Results per region

		P	R	F	Acc
NB	Macro	0.33	0.23	0.26	0.77
	Micro	0.74	0.77	0.75	
Pos weight	Macro	0.51	0.20	0.26	0.79
	Micro	0.75	0.79	0.75	
Char n-gram	Macro	0.37	0.26	0.30	0.79
	Micro	0.76	0.79	0.77	

Table 9: Names$_{arb}$ Results by country for plain Naive Bayes, position weighting, and char n-gram back-off

		P	R	F	Acc
NB	Macro	0.42	0.38	0.39	0.92
	Micro	0.92	0.92	0.92	
Pos weight	Macro	0.59	0.34	0.39	0.94
	Micro	0.93	0.94	0.93	
Char n-gram	Macro	0.47	0.38	0.41	0.93
	Micro	0.93	0.93	0.93	

Table 10: Names$_{arb}$ Results by country for plain Naive Bayes, position weighting, and char n-gram back-off

		P	R	F	Acc
NB	Macro	0.35	0.29	0.31	0.62
	Micro	0.60	0.62	0.61	
Pos weight	Macro	0.48	0.27	0.32	0.65
	Micro	0.62	0.65	0.61	
Char n-gram	Macro	0.45	0.30	0.35	0.66
	Micro	0.63	0.66	0.63	

Table 11: Names$_{trans}$ Results by country for plain Naive Bayes, position weighting, and char n-gram back-off

		P	R	F	Acc
NB	Macro	0.51	0.50	0.49	0.79
	Micro	0.79	0.79	0.79	
Pos weight	Macro	0.64	0.46	0.50	0.81
	Micro	0.80	0.81	0.80	
Char n-gram	Macro	0.61	0.50	0.53	0.82
	Micro	0.81	0.82	0.81	

Table 12: Names$_{trans}$ Results by country for plain Naive Bayes, position weighting, and char n-gram back-off