Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 198 (2022) 367–372

Procedia Computer Science

www.elsevier.com/locate/procedia

The 2nd International workshop on artificial intelligence for natural language processing (IA&NLP) November 1-4, 2021, Leuven, Belgium

Geographic Disaggregation of Textual Social Media Data: A Machine Learning-based Approach

Jihad Zahir*

LISI laboratory, Cadi Ayyad University, Marrakesh, Morocco

Abstract

This research aims to identify the geographic origin of Arabic-speaking social media users by analyzing textual data they produce and share. The paper presents an approach to infer users' region (i.e country) of origin through identification of the dialect they use in their written interactions. An Integrated Dataset for Arabic Dialect Detection (IADD) is proposed and used to train multiple classifiers which succeed in identifying the users' region and country of origin with an accuracy of 0.89 and 0.93, respectively.

Keywords: Automatic Dialect Identification; Arabic Language; Geographic Disaggregation; Machine Learning.

1. Introduction

Analyzing social media data has emerged as a form of social listening which was proved to be useful and relevant in many applications such as health [1], politics [2], marketing [3], disaster management [4], and broader domains studying users' habits and public opinion [5]. While social media data analysis provides an interesting opportunity and a new perspective to study social phenomena in a given population of web users, it might hide disparities within groups of users if data is not properly disaggregated. Breaking down data by different sets of dimensions (i.e sex, age and geographic location) reveals more details and uncovers underlying trends and insights across different groups of social media users, thus providing a more complete picture. However, information on age, sex and geographic origin is not always explicitly shared by social media users, which poses a real challenge to social media data disaggregation. One approach to address these constraints is to use Natural Language Processing (NLP) in order to identify characteristics of users by analyzing the text (i.e tweet, comment, post) they share. Papers like [6] and [7, 8] are examples of researches predicting age and sex of social media users, respectively. Arabic language has a variety of dialects cross regions, countries and even cities. This linguistic diversity presents a good opportunity to address, at least partially,

* Corresponding author.Tel.: +212-524-437-741 ; fax: +212-524-434-494.
E-mail address: j.zahir@uca.ac.ma

1877-0509
10.1016/j.procs.2021.12.255

the challenge of geographic disaggregation of social media data, as dialectal Arabic reflects geographic location of the speaker, or at least their country of origin, if mobility factors are taken into account. The main idea of this research is to infer a user's region (or country) of origin by automatically identifying the dialect they use to formulate their content. The remainder of this paper is organized as follows: Section 2 presents the overall methodology and section 3 describes how IADD, an Integrated Dataset for Arabic Dialects Detection, is created . Details about text preprocessing and vectorization and classification processes are provided in sections 4 and 5, respectively. Finally, results are presented in section 6.

2. Method

In this work, automatic geographic disaggregation of textual social media data is essentially considered as a multinomial text classification task. Three classifiers are trained to recognize the dialect in which a given text is formulated. First, a *Region-level* classifier identifies the regional dialect of text, if the identified dialect is Levantine or Maghrebi, then a country-level classifier is used to identify the country. *Country-level* classifiers identify dialects of some countries located in LEV and MGH regions. More specifically, the *LEV classifier* identifies Palestinian, Syrian, Lebanese and Jordanian dialects, while *MGH classifier* detects Algerian, Moroccan and Tunisian dialects. A description of the classifiers' targets is provided in table 1.

Table 1: Classifiers Description

Classifier	Level	Target Classes	# Classes
Region-Level Classifier	Region	EGY, IRQ, MGH, LEV, GLF, general	6
MGH Classifier	Country	MAR, TUN, DZA	3
LEV Classifier	Country	SYR, JOR, PSE, LBN	4

There is no standard classification of Arabic dialects [9, 10]. This work considers 5 main groups of Arabic regional dialects:

1. **Egyptian (EGY):** Spoken dialect in Egypt and also one of the most widely understood dialect;
2. **Gulf (GLF):** Dialects of Bahrain, Kuwait, Oman, Qatar, Kingdom of Saudi Arabia, and United Arab Emirates;
3. **Levantine (LEV):** Dialects of Lebanon (LBN), Syria (SYR), Jordan (JOR), and Palestine (PSE);
4. **Iraqi (IRQ):** Includes elements from both LEV and GLF but generally considered as a distinct class since it has distinctive features;
5. **Maghrebi (MGH):** Spoken dialects in north african countries including Algeria (DZA), Libya, Mauritania, Morocco (MAR), & Tunisia (TUN).
6. **general:** General language that might be used in any dialect.

3. Building IADD dataset

Integrated Arabic Dialect Dataset (IADD) is created in two steps: 1) Data sources identification and 2) data preparation and insertion.

3.1. Data sources identification

IADD is created from the combination of subsets of five corpora: DART, SHAMI, TSAC, PADIC and AOC. Each corpus supports a different set of dialects, as shown in table 2. The Dialectal ARabic Tweets dataset (DART) [14] has about 25,000 tweets that are annotated via crowdsourcing while the SHAMI dataset [11] consists of 117,805 sentences and covers levantine dialects spoken in Palestine, Jordan, Lebanon and Syria. TSAC [12] is a Tunisian dialect corpus of 17,000 comments collected mainly from Tunisian Facebook pages. Parallel Arabic Dialect Corpus

(PADIC) [15] is made of sentences transcribed from recordings or translated from MSA. Finally, the Arabic Online Commentary (AOC) dataset [13] is based on reader commentary from the online versions of three Arabic newspapers, and it consists of 1.4M comments.

Table 2: Description of corpora composing IADD

Corpus	**Source**	**Supported Dialects**	
		Regional Level	**Country Level**
DART [14]	Twitter	Egyptian, Maghrebi, Levantine, Gulf& Iraqi	Egypt, Iraq
SHAMI [11]	Twitter	Levantine	Palestine, Jordan, Lebanon, Syria
TSAC [12]	Facebook users comments	Maghrebi	SyriaTunisia
PADIC [15]	Manual transcription from recordings of conversations, movies or shows	Levantine and Maghrebi	Syria, Palestine,Algeria, Morocco
AOC [13]	Readers' comments in websites of Arabic newspapers	Egyptian, Maghrebi, Levantine, Gulf & Iraqi	Egypt, Iraq

Table 3: Detailed overview of IADD

Region	**Country**	**#Sentences**
Maghrebi (MGH)	Algeria	14,426
	Morocco	7,213
	Tunisia	11,998
Total		33,996(25%)
Levantine (LEV)	Palestine	17,855
	Jordan	7,017
	Syria	44,972
	Lebanon	10,829
Total		87,573 (≈ 64%)
Egypt (EGY)	Egypt	4,837(3.6%)
Iraq (IRQ)	Iraq	216 (< 1%)
Gulf (GLF)	− − −	7,195 (≈ 5%)
general	− − −	2,500 (≈ 2%)
Total		136,317 (100%)

3.2. Data preparation and insertion

Sentences from SHAMI and TSAC are directly inserted in IADD. *Region* is set to "LEV" for SHAMI data and to "MGH" for TSAC data. Regarding DART, besides the five groups of regional dialects (EGY, IRQ, GLF, LEV, MGH), it contains also an additional group named "Other". The items corresponding to the "Other" category are removed and are therefore not added in IADD. Sentences in PADIC are initially classified into 6 categories of dialects: *ALGIERS, ANNABA, MODERN-STANDARD-ARABIC, SYRIAN, PALESTINIAN* and *MOROCCAN*.

- *ALGIERS* and *ANNABA* are two cities in Algeria. These tags are used to distinguish sentences written in Annaba dialect from those written in Algiers dialect.
- *MODERN-STANDARD-ARABIC* tag is associated to sentences written in MSA;
- *SYRIAN, PALESTINIAN* and *MOROCCAN* are dialects corresponding to Syria, Palestine and Morocco, respectively.

Texts in AOC dataset have 3 annotations given by 3 different reviewers. Annotators judged each text and assigned, according to their judgment, one of the following labels: "*notsure*", "*junk*", "*levantine*", "*egyptian*", "*gulf*", "*'iraqi'*", "*maghrebi*", "*general*" and "*'msa'*". Only texts with at least two identical annotations are considered. From these, texts annotated as "*msa*", "*'junk'*" or "*notsure*" are discarded, as sentences with the "*msa*" tag are in modern standard language and the two other tags are associated with noisy and ambiguous sentences, respectively.

3.3. IADD: Overview

At the end, IADD is stored in a JSON-like object with six keys: 1) ***idSentence*** to identify sentences/ text; 2) ***Sentence***; 3) ***Region*** that stores the corresponding dialectal region (MGH, LEV, EGY, IRQ, GLF or general); 4) ***Country*** which specifies the corresponding country, if available (MAR, TUN, DZ, EGY, IRQ, SYR, JOR, PSE, LBN); 5) ***City*** contains the corresponding city, if available; and 6) ***DataSource*** indicates the source of the data (PADIC, DART, AOC, SHAMI or TSAC). An overview of IADD, describing the number and percentage of sentences by region and country, and the vocabulary size is provided in table 3.

4. Text preprocessing and vectorization

First, items contained in IADD are cleaned by removing mentions, URLs and emojis and next, stop words from classical Arabic and words composed of less than two characters, punctuation, repeating letters and diacritics are also removed. It's worth noting that some sentences, especially those originating from TSAC, contain words written in French and in Arabizi, which represents Arabic text written using Latin characters. These words are left in their original format and no transformation is applied to them, except for the repeating letters removal. After preprocessing, sentences are transformed into vectors using the bag-of-words approach. It consists of generating a vocabulary from all the texts in the dataset, then for each text, a vector is created associating each word from the vocabulary with a value representing the frequency of the particular word in the text.

5. Classification

Multinomial Naive Bayes (MNB), a probabilistic learning method, is used to generate the classifiers. This choice is motivated by three factors. Firstly, results of MNB are easily interpretable, secondly this classifier is fast in both training and prediction and finally, it has very few tunable parameters. In this work, scikit-learn [1] implementation of MNB is utilized for experimentations. Relevant subsets of IADD are split into training and test sets to train and validate the classifiers. Sizes of training and test sets are reported in table 4(a).

6. Results and Discussion

6.1. Geographic disaggregation using IADD

Accuracy and micro-average precision are considered to assess the performance of the classifiers trained on IADD. As shown in table 4, accuracy of the Region-Level classifier is 0.89 while micro-average precision is 0.95. Figure1 presents the fifteen most predictive words for each dialect class for the regional classifier. Table 4(b) shows the accuracy and precision of the Country-Level classifiers. The MGH classifier achieves better performance compared to the LEV classifier, which can be explained by the increased similarity between Levantine dialects. In fact, Syrian, Jordanian, Palestinian and Lebanese dialects have many words in common as observed in figure 2. Confusion matrices for Country-Level classifiers are presented in table 5.

[1] A machine learning library for the Python programming language: https://scikit-learn.org/stable/

Table 4: Experimentation settings and classifiers accuracy

(a) Training and test sets size

Classifier	Training set size	Test set size
Region-Level Classifier	122, 685	13, 632
MGH Classifier	30, 273	3, 364
LEV Classifier	72, 605	8, 068

(b) Classifiers accuracy and precision

Classifier	Accuracy	Micro Average Precision
Region-Level Classifier	0.8856	0.9486
MGH Classifier	0.9348	0.9869
LEV Classifier	0.8228	0.9116

Table 5: Confusion Matrices

(a) Levantine (LEV) Dialects

		Predicted			
		Jordan	Lebanon	Palestine	Syria
Actual	Jordan	**290**	19	38	36
	Lebanon	19	**407**	42	64
	Palestine	23	18	**440**	43
	Syria	13	8	39	**492**

(b) Maghrebi (MGH) Dialects

		Predicted		
		Algeria	Morocco	Tunisia
Actual	Algeria	**543**	11	5
	Morocco	42	**305**	5
	Tunisia	24	7	**500**

EGY	GLF	IRQ	LEV	MGH	general
احنا	اقول	بيج	هاد	برافو	وعقبال
كده	طيب	شتريد	وانا	barcha	العالم
الزمالك	شاء	اني	انتي	حاجه	لازم
الشعب	ليش	خطيه	امي	كيما	ابو
يعني	الهلال	بيه	ليش	ماسط	يستاهل
واله	انت	ليش	يعني	بصح	طيب
الناس	عشان	يمك	بدي	قالت	الناس
الاهلي	وين	وانا	واله	باش	شاء
الناس	الناس	غردليشعر	كنت	واله	وين
انت	عشان	حجي	هيك	الي	مبروك
علي	النصر	اله	كتير	ربي	يعني
ايه	الهلال	الي	انو	تاع	يعني
اله	انا	هيج	اله	واش	واله
مصر	واله	كلشي	الي	اله	الي
الي	اله	كلش	انا	bravo	اله

Fig. 1: The fifteen most discriminative terms per regional dialect

7. Conclusion

This paper presented an approach to identify the geographic origin of Arab web users from text they share in social media. Geographic disaggregation was modeled as a 2-levels multinomial classification task. With the new integrated corpus and Naive Multinomial Bayes, it was possible to generate classifiers that identify dialect in Arabic text and therefore infer the country of origin of the author. In future future work, other methods and datasets will be explored to further improve the accuracy of the region-level classifier and to support dialect identification for Gulf countries.

Jordan	Lebanon	Palestine	Syria	Algeria	Morocco	Tunisia
حدا	الناس	لازم	لان	وين	مزيان	bravo
ليش	انك	انو	متل	وعلاه	علاش	اله
اني	بلد	ليش	هاد	اني	بصح	ماسط
علي	متل	انت	ليش	درك	اله	ربي
يعني	هيدا	ايش	انتي	باش	هاد	barcha
حالي	بلي	انتي	وانا	حاجه	ديال	maset
كنت	انت	عشان	هيك	لالا	قلت	mala
هاي	انا	اشي	امي	ايه	ليك	الي
هيك	هيك	يعني	بدي	برك	غادي	برافو
اله	كتير	كتير	كنت	بصح	بزاف	malem
مشان	حدا	واله	اله	واله	فاش	تونس
اشي	لبنان	هيك	كتير	كيما	بلي	rabi
انو	الي	انا	انو	قالت	واش	معلم
الي	اله	اله	الي	واش	شنو	ala
انا	انو	الي	انا	تاع	غدي	واله

Fig. 2: The fifteen most discriminative terms per country dialect

References

[1] M. Sharma, K. Yadav, N. Yadav, K. C. Ferdinand, Zika virus pandemic: analysis of facebook as a social media health information platform, American journal of infection control 45 (3) (2017) 301–302.

[2] N. Anstead, B. O'Loughlin, Social media analysis and public opinion: The 2010 uk general election, Journal of Computer-Mediated Communication 20 (2) (2014) 204–220.

[3] W. He, S. Zha, L. Li, Social media competitive analysis and text mining: A case study in the pizza industry, International Journal of Information Management 33 (3) (2013) 464–472.

[4] J. B. Houston, J. Hawthorne, M. F. Perreault, E. H. Park, M. Goldstein Hode, M. R. Halliwell, S. E. Turner McGowen, R. Davis, S. Vaid, J. A. McElderry, et al., Social media and disasters: a functional framework for social media use in disaster planning, response, and research, Disasters 39 (1) (2015) 1–22.

[5] S. Abbar, Y. Mejova, I. Weber, You tweet what you eat: Studying food consumption through twitter, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, 2015, pp. 3197–3206.

[6] L. Sloan, J. Morgan, P. Burnap, M. Williams, Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data, PloS one 10 (3) (2015) e0115545.

[7] E. AlSukhni, Q. Alequr, Investigating the use of machine learning algorithms in detecting gender of the arabic tweet author, International Journal of Advanced Computer Science and Applications 7 (7) (2016) 319–328.

[8] J. Zahir, Y. M. Oukaja, H. Mousannif, Author gender identification from arabic youtube comments, in: 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), IEEE, 2019, pp. 00–00.

[9] M. Abdul-Mageed, H. Alhuzali, M. Elaraby, 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), 2018.

[10] O. F. Zaidan, C. Callison-Burch, Arabic dialect identification, Computational Linguistics 40 (1) (2014) 171–202.

[11] C. Qwaider, M. Saad, S. Chatzikyriakidis, S. Dobnik, Shami: A corpus of levantine arabic dialects, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), 2018.

[12] S. Medhaffar, F. Bougares, Y. Estève, L. Hadrich-Belguith, Sentiment analysis of tunisian dialects: Linguistic ressources and experiments, in: Proceedings of the third Arabic natural language processing workshop, 2017, pp. 55–61.

[13] O. F. Zaidan, C. Callison-Burch, The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, Association for Computational Linguistics, 2011, pp. 37–41.

[14] I. Alsarsour, E. Mohamed, R. Suwaileh, T. Elsayed, Dart: A large dataset of dialectal arabic tweets, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), 2018.

[15] K. Meftouh, S. Harrat, S. Jamoussi, M. Abbas, K. Smaili, Machine translation experiments on padic: A parallel arabic dialect corpus, 2015.