

Arabic Language Processing Using AI & Machine Learning

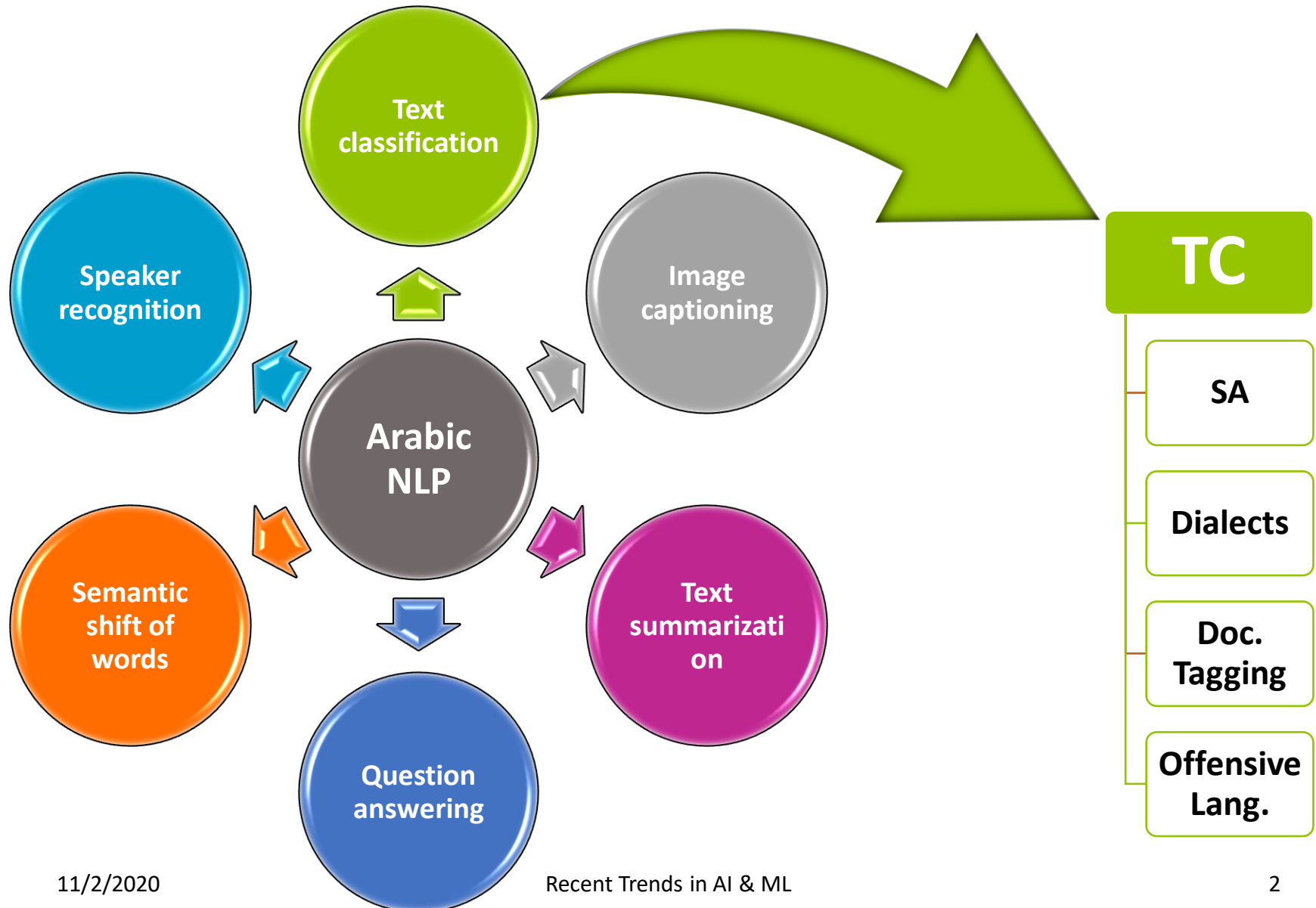
Ashraf Elnagar

Machine Learning & Arabic Computing Group

Department of Computer Science

University of Sharjah, Sharjah, UAE

MLALP RG @ UOS

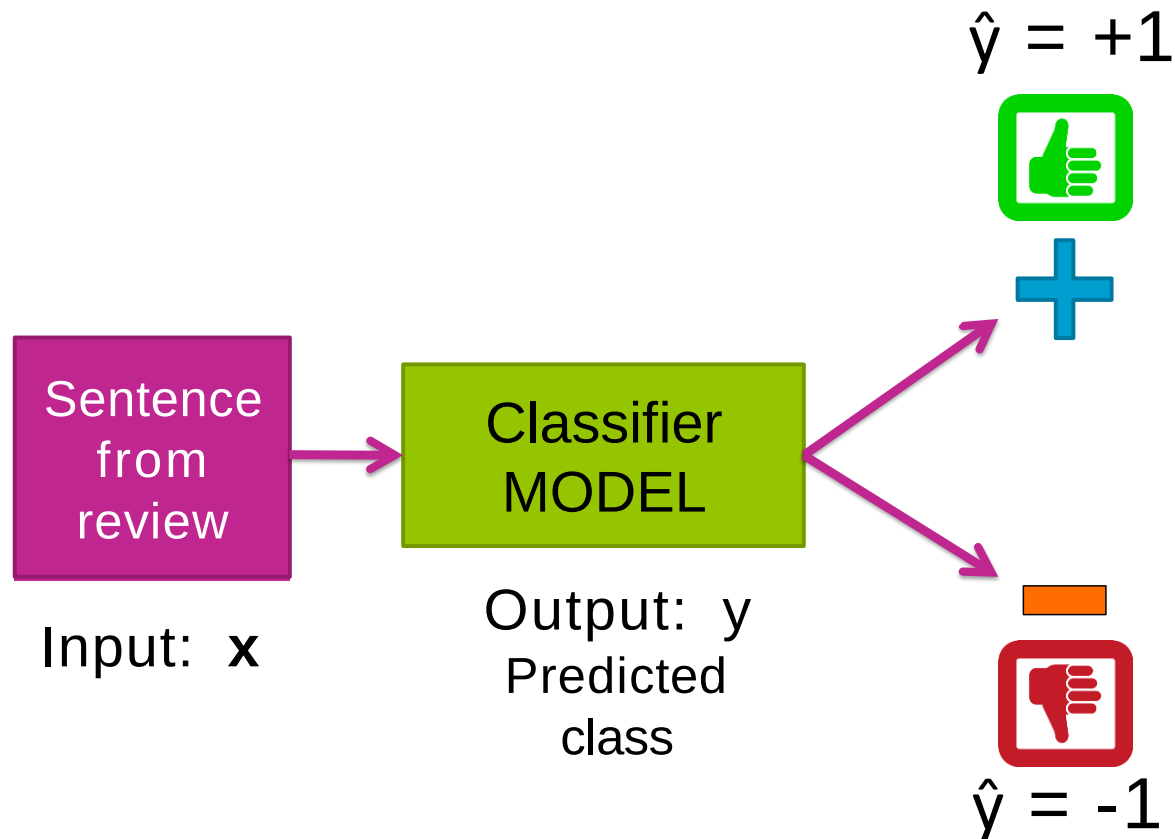


Arabic Language: Challenges & opportunities

- Rich vocabulary.
- Sentence Structures: order of words.
- Proper names: NER.
- Absence of diacritics and punctuation marks.
- New vocab due to foreign culture and technical development.
- Common spelling errors: أ، ا، آ، ة، ه، ي، ى.
- Use of colloquial dialects.
- **Lack of resources.**

Sentiment Classifier

تصنيف المشاعر



Note: multiclass classification follows same idea.

Multi-Label Classification Task

تصنيف مقالات إخبارية طبقا لموضوعاتها المتعددة



Arabic Resources

- **HARD** [373k] Hotel Arabic-Reviews Dataset
- **BRAD**: [510k] Books Reviews in Arabic Dataset
<https://github.com/elnagara/>
- **SANAD** (Akhbarona: 78,050, Alkhaleej: 45,500, Alarabiya: 71,247)
<http://dx.doi.org/10.17632/57zpx667y9.2>
- **NADiA** [485K Articles] (Masrawy, SkyNews Arabiya)
<http://dx.doi.org/10.17632/hhrb7phdyx.2>
- **Audio Clips** (15810 + 397 imitations)
<http://dx.doi.org/10.17632/3kndp5vs6b.2>





This is your published dataset

To make changes to your dataset

[Create new version](#)

NADiA: News Articles Dataset in Arabic for Multi-Label Text Categorization

876

Views

117

Downloads

Published: 02-09-2019 | **Version 2** | DOI: 10.17632/hhrb7phdyx.2Contributors: [Ridhwan Al-Debsi](#), [Ashraf Elnagar](#), [Omar Einea](#)

Description

NADiA Dataset is the largest, to the best of our knowledge, source for Arabic textual data that can be used in any NLP related task such as text classification. We chose the abbreviation NADiA as it is a common Arabic name. The data was collected by scraping 'SkyNewsArabia' and 'Masrawy' news websites using Python scripts that are fine-tuned for each website. SkyNewsArabia will be referred to as NADiA1, while the latter would be NADiA2. NADiA1 is a big dataset containing 37,445 files, while NADiA2 is a huge dataset that contains 678,563 files. However, after filtering and cleaning we reduced the numbers to 35,416 and 451,230 for NADiA 1 and 2, respectively.

NADiA1 consists of the following categories (24, displayed in English for easy referencing):

News, North Africa, Levant, Middle East, The Americas, Research, Finance & Economy, War & Terrorism, Gulf, Europe, Political Figures, Iran, Technology, Russia, Sports, Tennis, Football, English League, Arabian Sports, Spanish League, Health, East Asia, Environment, Other Countries

NADiA2 consists of the following categories (28, displayed in English for easy referencing):

Politics, Middle East, Asia, Africa, United States, Europe, Other Countries, Leaders, Sports, Arabian Sports, Football Clubs, Spanish League, Egyptian League, Finance, Arts, Cinema & TV, Fashion, Health, Pregnancy & Delivery, Cancer, Obesity, Social Media, Technology, Religion, Islamic, Fatawa, Worship, Prophet Biography

Latest version

Version 2

Published:

02-09-2019

DOI:

10.17632/hhrb7phdyx.2

Cite this dataset

Al-Debsi, Ridhwan; Elnagar, Ashraf; Einea, Omar (2019), "NADiA: News Articles Dataset in Arabic for Multi-Label Text Categorization", Mendeley Data, V2, doi: 10.17632/hhrb7phdyx.2

<http://dx.doi.org/10.17632/hhrb7phdyx.2>

[Cite this dataset](#)

Previous versions

Version 1

28-03-2019

Word Embedding Models

سَخِيفَ

books	ركيك	مرذول	ساذج	تافه	بذيء	سحف	سَخِيفَ_العقل
wiki	مضحك	غير_منطقي	ساذج	مخادع	أحمق	تافه	بذيء
news	ساذج	غير_معقول	مستفز	مخادع	مثير_للسخرية	مقزز	خاطئ
msa_dialects	مستفز	ساذج	تافه	مبتذل	سمج	ركيك	مضحك

قَبْلَهُ

books	قبلته	وقبله	قبله_ترضاها	القبله	كعبة	للقبله	قبلتهم
wiki	مقصدا	محجا	مزارا	مقصد	معلما_سياحيا	موئلا	قبله_للسياح
news	وقبله	وجهة_رئيسية	مقصدا	منارة	مقصدا_مهما	الوجهة_المفضلة	مكانا_مفضلا
msa_dialects	بقبله	بوسة	فقبله	قبالات	قبله_حنينة	طبع_قبله	قبله_خفيفة

قَمَرٌ

books	قمر	بدر_الدجى	كبدر	قمره	البدر	كشمس	هلالها
wiki	وقمر	زحل	والقمر	القمر	ترايتون	إنسيلاتوس	قمره
news	وقمر	القمر	قمرى	قمرين	مسبار	والقمر	كوكب
msa_dialects	غدير	شروق	إسراء	اكرام	بانا	سيندي	ريم

BRAD: Goodreads.com
HARD: booking.com

15	41	0	3	2	6	—		1	7	12	25	الدوري الإسباني	12–2011	برشلونة
11	46	0	2	1	9	—		2	6	8	29		13–2012	
21	54	0	2	0	9	—		2	9	19	34		14–2013	
47	141	0	7	3	24	—		5	22	39	88	المجموع		
25	52	0	1	4	9	1	1	4	6	16	35	الدوري الإنجليزي	15–2014	أرسنال
17	41	0	0	3	7	0	1	1	3	13	30		16–2015	
30	51	—		3	8	0	0	3	5	24	38		17–2016	
8	21	0	0	1	1	0	1	0	0	7	19		18–2017	
80	165	0	1	11	25	1	3	8	14	60	122	المجموع		
1	3	—		0	0	—		0	1	1	2	الدوري الإنجليزي	18–2017	مانشستر يونايتد
174	550	0	8	18	85	1	3	14	45	141	409	الإجمالي		

Actual Tags:

- 'Sports',
- 'Football',
- 'English League'

Predicted tags:

- ✓ ('Sports', 1.0),
- ✓ ('Football', 0.99999976),
- ✗ ('Spanish League', 0.99999607),
- ✓ ('Europe', 0.67192304),
- ('The Americas', 0.00059251)

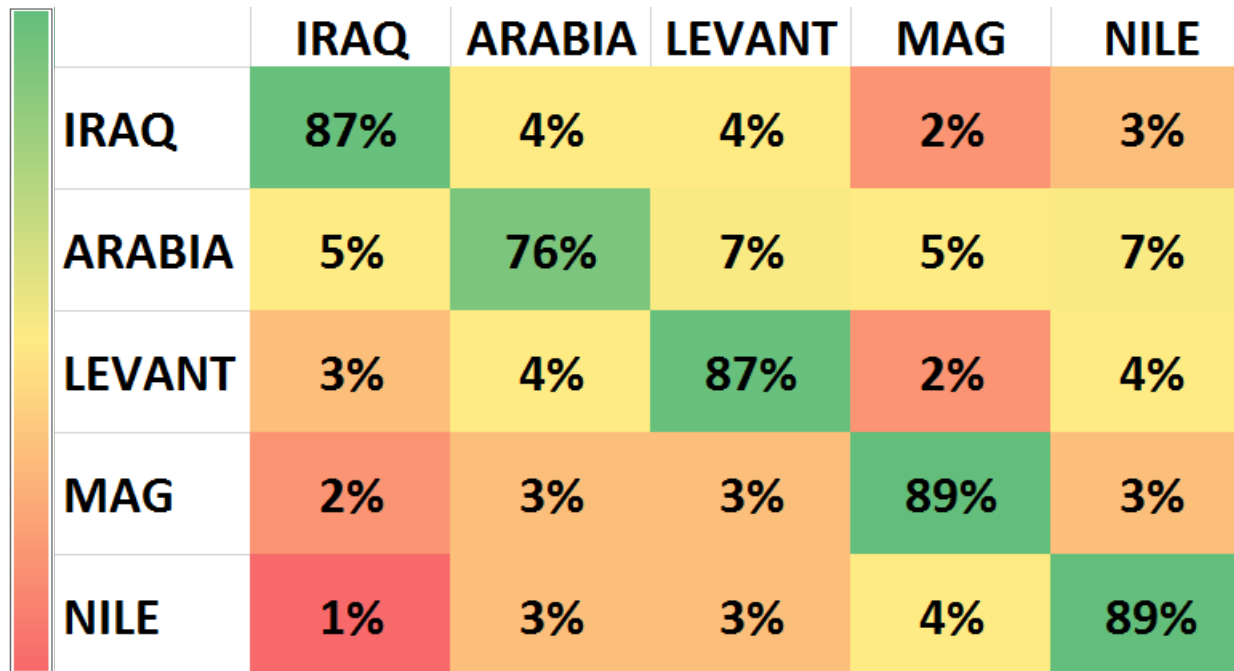
وقد يؤدي اعتراف سانشير لتغريمه وربما عقوبة سجن مع وقف التنفيذ.

So is it
'Spanish League'
or
'English League' ?

وسبق لنجم برشلونة الأرجنتيني ليونيل ميسي ووالده أن حكما بالسجن 21 شهرا يوليو 2016 بسبب التهرب من الضرائب خلال الفترة الممتدة من 2007 حتى 09

Text Classification: Arabic Dialects

Example



	IRAQ	ARABIA	LEVANT	MAG	NILE
IRAQ	87%	4%	4%	2%	3%
ARABIA	5%	76%	7%	5%	7%
LEVANT	3%	4%	87%	2%	4%
MAG	2%	3%	3%	89%	3%
NILE	1%	3%	3%	4%	89%

Text Classification: Arabic Dialects

Example

Dialect identification: **which city?**

Enter your sentence: كنت فين لحد دلوقتي

Enter your sentence: باجر بطرش البريد

Enter your sentence: هكا خفت بزاف عليها

Enter your sentence: ما زبطت معنا

Enter your sentence: شكو بيها ؟

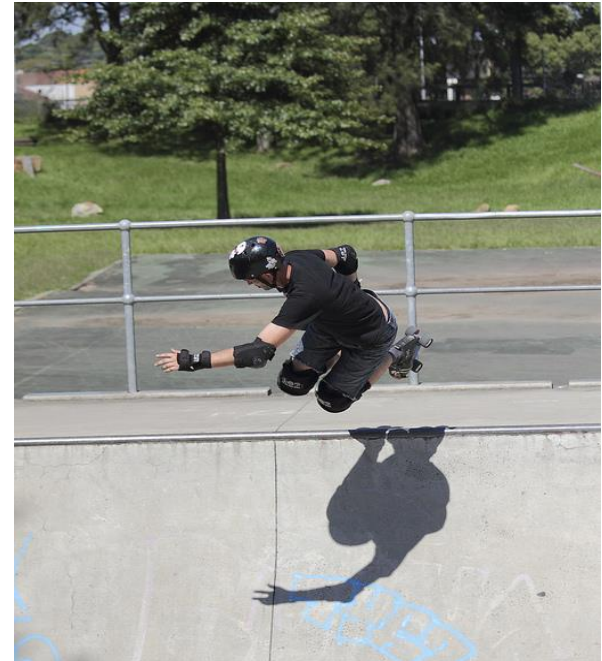
Image Captioning Example



قطعة تجلس فوق جهاز كمبيوتر محمول



دراجة نارية تمر في منطقة ريفية



شخص يقفز في الهواء مع لوح تزلج



Thank you
ashraf@sharjah.ac.ae