

Distinguishing between the Varieties of Arabic: Dialect Identification is neither Solved nor the Solution

Amr Keleg

arbml Board

1 July 2024



Institute for Language,
Cognition and Computation



THE UNIVERSITY of EDINBURGH
informatics

- 1 **Dialect Identification is not solved,**
 - 📖 **Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification** (Keleg & Magdy, ArabicNLP-WS 2023)
 - 📖 **NADI 2024** (Abdul-Mageed et al., To appear ArabicNLP-WS 2024)
- 2 **... nor the solution (Spoiler: Arabic Level of Dialecttness)**
 - 📖 **ALDi: Quantifying the Arabic Level of Dialecttness of Text** (Keleg et al., EMNLP 2023)
 - 📖 **Estimating the Level of Dialecttness Predicts Inter-annotator Agreement in Multi-dialect Arabic Datasets** (Keleg et al., To appear ACL 2024)

Why distinguish between varieties of Arabic?



Annotated Dataset of Tweets

Tweet	Label
***	🚫 OFF
...	...
الراجل بسطنا	🌟 NOT

Why distinguish between varieties of Arabic?



Annotated Dataset of Tweets

Tweet	Label
***	 OFF
...	...
الراجل بسطنا	 NOT

- Perform per-variety analysis:
 - a) offensive text
 - b) models' performance

Why distinguish between varieties of Arabic?



Annotated Dataset of Tweets

Tweet	Label
***	 OFF
...	...
الراجل بسطنا	 NOT



Raw Comments on YT

Comment
حلقة اقل ما يُقال عنها انها رائعة
...
طيب مافي حلقات زيادة؟ ما شبعنا والله

- Perform per-variety analysis:
 - a) offensive text
 - b) models' performance

Why distinguish between varieties of Arabic?



Annotated Dataset of Tweets

Tweet	Label
***	 OFF
...	...
الراجل بسطنا	 NOT



Raw Comments on YT

Comment
حلقة اقل ما يُقال عنها انها رائعة
...
طيب مافي حلقات زيادة؟ ما شبعنا والله

- Perform per-variety analysis:
 - a) offensive text
 - b) models' performance

- c) Representation of dialects?
- d) Routing samples to annotators?

Arabic sentences



Arabic sentences



أسعدنا الرجل

الراجل أسعدنا

Arabic sentences



MSA

أسعدنا الرجل

DA

الراجل أسعدنا

Arabic sentences



الزيلة أسعدنا

MSA

أسعدنا الرجل

DA

الراجل أسعدنا

Arabic sentences



MSA

أسعدنا الرجل

DA

الراجل أسعدنا

الزلمة أسعدنا

Arabic sentences



الزول أسعدنا

MSA

أسعدنا الرجل

DA

الراجل أسعدنا

الزيلة أسعدنا

Arabic sentences



MSA

أسعدنا الرجل

DA

الراجل أسعدنا

الزيلة أسعدنا

الزول أسعدنا

Proposal #1



MSA

- shared across countries



DA

- different variants

More granular scheme ?

(Proposal #2)
Regional Grouping of DA

Proposal #2



MSA

- shared across countries



Regional dialects

- Maghreb, Nile Basin, Levant,
Gulf, Gulf of Aden

Alsarsour, Israa et al. 2018. "DART: A Large Dataset of Dialectal Arabic Tweets."

Baimukan, Nurpeiis, Bouamor, Houda, and Habash, Nizar. 2022. "Hierarchical Aggregation of Dialectal Data for Arabic Dialect Identification."

Arabic sentences



MSA

أسعدنا الرجل

DA

الراجل أسعدنا

الزيلة أسعدنا

الزول أسعدنا

Arabic sentences



MSA

أسعدنا الرجل

Nile

الراجل أسعدنا

Levant

الزلة أسعدنا

الزول أسعدنا

Mutual Intelligibility Constraint

(Proposal #3)
Country-level Classification

Proposal #3



MSA

- shared across countries



Country-level dialects

- generally targeting at least 18 labels

Arabic sentences



MSA

أسعدنا الرجل

Nile

الراجل أسعدنا

Levant

الزلة أسعدنا

الزول أسعدنا

Arabic sentences



MSA

أسعدنا الرجل

Egypt

الراجل أسعدنا

Sudan

الزول أسعدنا

Syria

الزلة أسعدنا

Palestine

الزلة أسعدنا

Arabic sentences



وين المحطة؟

MSA

أسعدنا الرجل

Egypt

الراجل أسعدنا

Sudan

الزول أسعدنا

Syria

الزلة أسعدنا

Palestine

الزلة أسعدنا

Arabic sentences



Valid in: Iraq, Jordan, Lebanon, Libya, Oman, Palestine, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, Yemen



وين المحطة؟

MSA

أسعدنا الرجل

Egypt

الراجل أسعدنا

Sudan

الزول أسعدنا

Syria

الزلة أسعدنا

Palestine

الزلة أسعدنا



Community's perceptions

- 1 Common for same-region dialects (e.g.,
- 2 Generally short (e.g., @ H i)⁵

D H⁴

⁴Abdelali, Ahmed et al. 2021. "QADI: Arabic Dialect Identification in the Wild."

⁵Salameh, Mohammad, Bouamor, Houda, and Habash, Nizar. 2018. "Fine-Grained Arabic Dialect Identification."



Community's perceptions

- 1 Common for same-region dialects (e.g., D H⁴)
- 2 Generally short (e.g., @ H i)⁵

Is it a significant limitation ?



More on this later!

⁴Abdelali, Ahmed et al. 2021. "QADI: Arabic Dialect Identification in the Wild."

⁵Salameh, Mohammad, Bouamor, Houda, and Habash, Nizar. 2018. "Fine-Grained Arabic Dialect Identification."



Community's perceptions

- 1 Common for same-region dialects (e.g., D H⁴)
- 2 Generally short (e.g., @ H i)⁵

Is it a significant limitation ?



More on this later!




Multi-dialect samples were ignored. We can do better now!

⁴Abdelali, Ahmed et al. 2021. "QADI: Arabic Dialect Identification in the Wild."



⁵Salameh, Mohammad, Bouamor, Houda, and Habash, Nizar. 2018. "Fine-Grained Arabic Dialect Identification."

Impact of Single-label Modeling on Error Analysis

 **Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification** (Keleg & Magdy, ArabicNLP-WS 2023)

Country-level ADI system (NADI 2023)

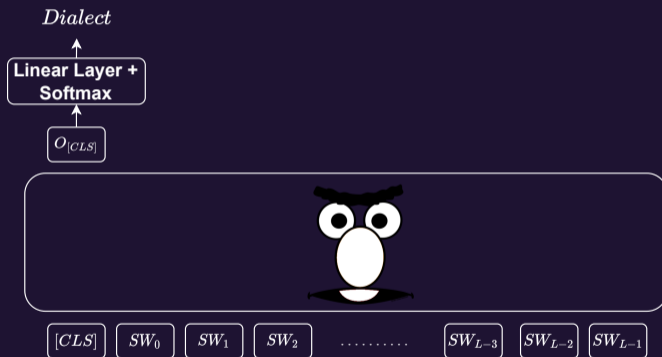
i Training Dataset: NADI 2023

Target	Sentence				
		i	H	i	
	di	H	H		H
...					...

i Labels: 18 geolocated dialects

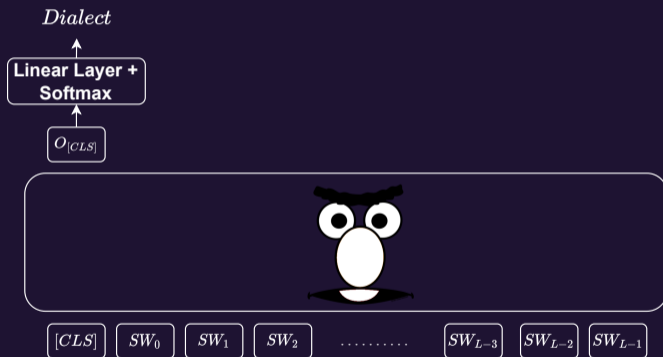
Country-level ADI system (NADI 2023)

i Labels: 18 dialects (country-level)



Country-level ADI system (NADI 2023)

i Labels: 18 dialects (country-level)



Accuracy: 50.74%



Error Analysis



Annotators: 7 countries

R P

R H F

H

Target: Tunisia 

Prediction: Egypt 

Verdict: Error 

R P

R H F

H

Target: Tunisia 🇹🇳

Prediction: Egypt 🇪🇬

Verdict: Error ❌



Is this Sentence Valid in your dialect?

R P

R H F

H

Target: Tunisia 🇹🇳

Prediction: Egypt 🇪🇬

Verdict: Error ❌



Is this Sentence Valid in your dialect?



Yes

Not an Error

~@~

HT H

Target: Algeria 

Prediction: Egypt 

Verdict: Error 

~@~

HT H

Target: Algeria 

Prediction: Egypt 

Verdict: Error 



Is this Sentence Valid in your dialect?

~ @ ~

HT H

Target: Algeria 

Prediction: Egypt 

Verdict: Error 



Is this Sentence Valid in your dialect?

No

An error


Limitations of single-label ADI

Only **33%** of validated mispredictions are **true errors!**

i.e., 67% of them are multi-dialect samples.

 Inaccurate Evaluation

Hindering progress? 🤔

 How common are these samples in the whole dataset?

Building a Multilabel ADI Dataset (NADI 2024)

 **NADI 2024** (To appear, ArabicNLP 2024)

How to Build a Multilabel ADI Dataset? (1)

Sentence

Sentence₁

Sentence₂

Sentence₃

...

Sentence_N

How to Build a Multilabel ADI Dataset? (1)

Sentence

Sentence₁

Sentence₂

Sentence₃









...

Sentence_N

Annotators















How to Build a Multilabel ADI Dataset? (1)

Sentence					...
Sentence ₁					...
Sentence ₂					
Sentence ₃					
...					
Sentence _N					

Annotators























How to Build a Multilabel ADI Dataset? (1)

Sentence					...
Sentence ₁					...
Sentence ₂					...
Sentence ₃					
...					
Sentence _N					

Annotators























How to Build a Multilabel ADI Dataset? (1)

Sentence					...
Sentence ₁					...
Sentence ₂					...
Sentence ₃					...
...
Sentence _N					...

Annotators



How to Build a Multilabel ADI Dataset? (1)

Sentence					...
Sentence ₁					...
Sentence ₂					...
Sentence ₃					...
...
Sentence _N					...

Annotators



Only need 3 annotators, but major quality issues !!

How to Build a Multilabel ADI Dataset? (2)

Sentence

Sentence₁

Sentence₂

Sentence₃

...

Sentence_N

How to Build a Multilabel ADI Dataset? (2)

Sentence 

Sentence₁ 

Sentence₂ 















Sentence₃ 

...

Sentence_N 

Annotator  

How to Build a Multilabel ADI Dataset? (2)

Sentence		
Sentence ₁		
Sentence ₂		
Sentence ₃		
...
Sentence _N		
Annotator	 	 

How to Build a Multilabel ADI Dataset? (2)

Sentence					...
Sentence ₁					...
Sentence ₂					...
Sentence ₃					...
...
Sentence _N					...
Annotator	 	 	 	 	...

How to Build a Multilabel ADI Dataset? (2)

Sentence					...
Sentence ₁					...
Sentence ₂					...
Sentence ₃					...
...
Sentence _N					...
Annotator	 	 	 	 	...

How to Build a Multilabel ADI Dataset? (2)

Sentence					...
Sentence ₁					...
Sentence ₂					...
Sentence ₃					...
...
Sentence _N					...
Annotator	 	 	 	 	...

Need 3 annotators PER COUNTRY LABEL 😊

NADI 2024's evaluation sets

3 Annotators from 9 different countries (total of 27)

Maghreb (Morocco, Algeria, Tunisia)

Nile (Sudan, Egypt)

Levant (Palestine, Syria)

Gulf (Iraq)

Gulf of Aden (Yemen)

1,120 sentences (120 in dev. set + 1,000 in test set)

To be made available on request!

Examples

Sentence	Valid in
R h H	Palestine 🇵🇸, Sudan 🇸🇩, Yemen 🇲🇪

Examples

Sentence		Valid in	
R	h H	Palestine 🇵🇸, Sudan 🇸🇩, Yemen 🇲🇪	
	H	i	Algeria 🇩🇿, Palestine 🇵🇸, Yemen 🇲🇪

Multilabel samples in NADI 2024?

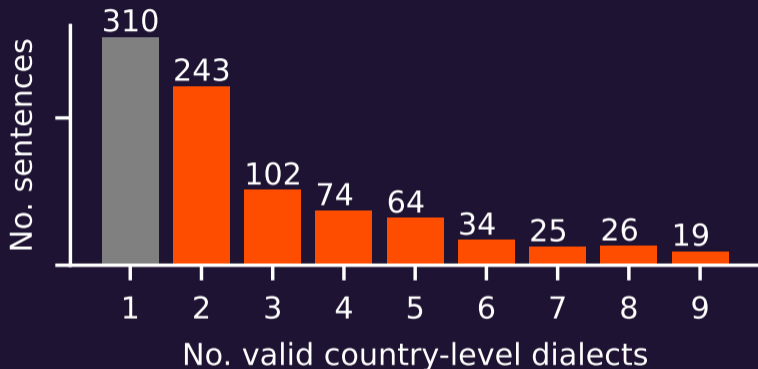
Community's perceptions:

- 1 Common for same-region dialects (e.g., D H)
- 2 Generally short (e.g., @ H i)^V

^UAbdelali, Ahmed et al. 2021. "QADI: Arabic Dialect Identification in the Wild."

^VSalameh, Mohammad, Bouamor, Houda, and Habash, Nizar. 2018. "Fine-Grained Arabic Dialect Identification."

Multilabel samples in NADI 2024?

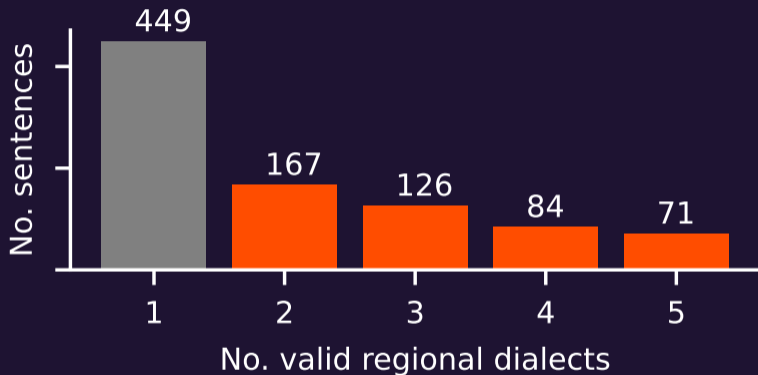


All samples but 310 are multi-dialect (country level).



Not just the short ones!

Multilabel samples in NADI 2024?





50% of samples are valid in multiple regions.
Not just because of within-region similarities!



Multi-dialect samples are much more common than expected!

NADI 2024 - Subtask 1 - Multilabel ADI

Training datasets:
Single-label datasets from previous NADI subtasks

يلعن الكورة واليوم اللي شجعت في كورة	فلسطين Palestine	
الله يرحمه ربي معك خويا و انا لله و انا اليه راجعون	المغرب Morocco	
والله ما عرف عنه بس جتني الصورة على الخاص وقلت اكيد تذكرونه	العراق Iraq	
...	...	

Subtask (1)



NADI 2024 - Subtask 1 - Multilabel ADI

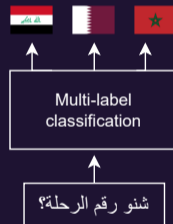
Training datasets:
Single-label datasets from previous NADI subtasks

يلعن الكورة واليوم اللي شجعت في كورة	فلسطين Palestine	
الله يرحمه ربي معك خويا و انا لله و انا اليه راجعون	المغرب Morocco	
والله ما عرف عنه بس جتني الصورة على الخاص وقلت اكيد تذكرونه	العراق Iraq	
...	...	

Subtask (1)



Output:
All valid dialects of the input sentence



Subtask 1 - Evaluation

Sentence

Sentence₁











Sentence₂

Sentence₃

...

Sentence_N











Subtask 1 - Evaluation

Sentence	Target 	Prediction 
Sentence ₁		
Sentence ₂		
Sentence ₃		
...
Sentence _N		



Binary classification task with unbalanced classes











Subtask 1 - Evaluation

Sentence	Target 	Prediction 
Sentence ₁		
Sentence ₂		
Sentence ₃		
...
Sentence _N		



Binary classification task with unbalanced classes

Subtask 1 - Evaluation

Sentence	Target 	Prediction 
Sentence ₁		
Sentence ₂		
Sentence ₃		
...
Sentence _N		



Binary classification task with unbalanced classes

Compute country-level metrics

Take their macro-average

Multilabel ADI systems

Rank	System	Macro-average			
		Accuracy (%)	Precision (%)	Recall (%)	F ₁ score (%)
Baseline III	Top 1	73.42 \pm 7.6	76.82 \pm 10.6	17.77 \pm 10.8	27.30 \pm 12.6

Table: Systems' performance on the test set of Subtask 1.

Multilabel ADI systems

Rank	System	Macro-average			
		Accuracy (%)	Precision (%)	Recall (%)	F ₁ score (%)
Baseline II	Random	50.14 _{±1.6}	30.43 _{±8.8}	50.15 _{±2.1}	37.15 _{±7.2}
Baseline III	Top 1	73.42 _{±7.6}	76.82 _{±10.6}	17.77 _{±10.8}	27.30 _{±12.6}

Table: Systems' performance on the test set of Subtask 1.

Multilabel ADI systems

Rank	System	Macro-average			
		Accuracy (%)	Precision (%)	Recall (%)	F ₁ score (%)
1	Elyadata	67.50 _{±3.7}	46.48 _{±10.1}	57.09_{±5.1}	50.57_{±7.1}
	Baseline II Random	50.14 _{±1.6}	30.43 _{±8.8}	50.15 _{±2.1}	37.15_{±7.2}

Table: Systems' performance on the test set of Subtask 1.

Multilabel ADI systems

Rank	System	Macro-average			
		Accuracy (%)	Precision (%)	Recall (%)	F ₁ score (%)
1	Elyadata	67.50 _{±3.7}	46.48 _{±10.1}	57.09 _{±5.1}	50.57 _{±7.1}
Baseline I	Top 90%	73.40 _{±6.1}	60.67 _{±14.5}	39.22 _{±14.6}	45.09 _{±11.3}

Table: Systems' performance on the test set of Subtask 1.

Baseline I (Top 90%):

- A fine-tuned BERT-based model
- Single-label ADI

Baseline I (Top 90%):

- A fine-tuned BERT-based model
- Single-label ADI

Baseline I (Top 90%):

- A fine-tuned BERT-based model
- Single-label ADI

Predictions:

Palestine, Syria,
Lebanon, Jordan

Interim Summary

>70% of samples are multi-dialect

Multilabel ADI is not solved (yet 🤔)

Could be you: <https://codalab.lisn.upsaclay.fr/competitions/18130>

\$UDELF /HYHO RI 'LDOHF

\$/ 'L 4XDQWLI\LQJ WKH \$UDELF /HYHO RI
.HOHJ HW DO (01/3

'LVWLQJXLVKLQJ EHWZHHQ WKH 9DULHWLHV RI \$UDELF 'LDOHFW -GHQWL4

'LVWLQJXLVKLQJ EHWZHHQ WKH 9DULHWLHV RI \$UDELF 'LDOHFW -GHQWL4

'LVWLQJXLVKLQJ EHWZHHQ WKH 9DULHWLHV RI \$UDELF 'LDOHFW -GHQWL4

\$ / ' L

- **Definition:** Divergence from Standard Language.

ALDi

- **Definition:** Divergence from Standard Language.
- **Operationalization:** Score in $[0; 1]$ on sentence-like level.

Dataset

Arabic Online Commentary Dataset (Zaidan et. al, 2011)

Popular Dialect Identification (DI) labels.

Ignored 8]gWYH Level of Dialectness labels!

' D W D V H W

\$ U D E L F 2 Q O L Q H & R P P H Q W D U \ ' D W D V H W
3 R S X O D U ' L D O H F W - G H Q W L 4 F D W L R Q ' -
- J Q R U L H G F U / H W H O R I ' L D O H F W Q H V V O D E H
(P E U D F H D Q Q R W D W R U V G L V D J U H H P H Q V

=DLGDQ 2PDU) DQG &DOOLVRQ %XUFK &KULV
'LDOHFWDO &RQWHQW

7KH \$UDELF 2QOLQH &RPPHQWDU \ 'DWDVHW



'LVWLQJXLVKLQJ EHWZHHQ WKH 9DULHWWLHV RI \$UDELF 'LDOHFW -GHQWL4

6HQWHQFH ZLWK WZR YDOLG SURQXQFLDW

6HQWHQFH ZLWK WZR YDOLG SURQXQFLDW

AOC-ALDi Dataset

127,835 sentences (3  annotations each)

Comments to news articles

Fleiss' $\kappa = 0.44$

Krippendorff's κ (interval) = 0.63



'LDOHFW DJQRVWLF

'LDOHFW DJQRVWLF
RMSE(AOC ALD_i_{test}) = 0:18

'LDOHFW DJQRVWLF

RMSE(AOC $ALD_{i_{test}}$) = 0:18

'HPR KXJJLQJIDFH FR VSDFHV \$05 .(/(

- P S D F W R I \$ / ' L R Q - Q W H
\$ J U H H P H Q W

(V W L P D W L Q J W K H / H Y H O R I ' L D O H F W Q H V
- Q W H U D Q Q R W D W R U \$ J U H H P H Q W L Q 0 X O W
. H O H J H W D O 7 R D S S H D U \$ & /

- Q W X L W L R Q

\$ / ' L - 1 & 5 (\$ 6 (

0 X W X D O - Q W H O O L J L E L O L W \ "

- Q W X L W L R Q

\$ / ' L - 1 & 5 (\$ 6 (

0 X W X D O - Q W H O O L \$ 6 (E L O L W \ "



Intuition

ALDi **INCREASE**

Mutual Intelligibility? **DECREASE**

Mutual Intelligibility **DECREASE**

Inter-annotator Agreement?



Intuition

ALDi **INCREASE**

Mutual Intelligibility? **DECREASE**

Mutual Intelligibility **DECREASE**

Inter-annotator Agreement? **DECREASE**

Finding

For 8 datasets across 5 different tasks:

ALDi **INCREASE** IAA **DECREASE**

Pearson Correlation Coefficient () < -0.7

Updated Recommendation

- 1 Prioritize high-ALDi samples to native speakers.
- 2 For high-ALDi samples, dialects identified with higher accuracy.

Thanks!

a.keleg@sms.ed.ac.uk


@Amrkeleg on L

Thanks!

a.keleg@sms.ed.ac.uk

@Amrkeleg on L

Summary

- 1 >70% of NADI 2024's test set are multi-dialect.
- 2 Multilabel DI is not solved (yet).
- 3 High-ALDi samples have less mutual intelligibility.
 Important for accurate annotation.

5 H I H U H Q F H V

\$ E G H O D O L \$ K P H G H W D O 4 \$ ' - \$ \$ U D E L F ' L D O
- G H Q W L 4 F D W L R Q - Q Q W K H H G L Q J V R I W K H
1 D W X U D O / D Q J X D J H 3 U R F (H G V E \ Q J } B U N D
H W D O . \ L Y 8 N U D L Q H 9 L U W X D O \$ V V R F
/ L Q J X L V W L F V S S 8 5 /

<https://aclanthology.org/2021.wanlp-1.1>


5 H I H U H Q F H V

\$ O V D U V R X U - V U D D H W \$ 5 7 \$ 0 / D \ U J H ' D W D '
' L D O H F W D O \$ U D E L F 7 Z H H W V
" Q W H U Q D W L R Q D O & R Q I H U H Q F H R Q / D Q J '
(Y D O X D W L R Q / 5 G & E \ 1 L F R O H W W D & D O J R
0 L \ D J D N L - D S D Q (X U R S H D Q / D Q J X D J H 5
(/ 5 \$ 8 5 <https://aclanthology.org/L18-1579>


5 H I H U H Q F H V

% D L P X N D Q 1 X U S H L L V + R X G D % R X D P R U
+ L H U D U F K L F D O \$ J J U H J D W L R Q R I ' L
' L D O H F W - G H Q W L 4 F D W L R Q
/ D Q J X D J H 5 H V R X U F H V D Q G (Y D (C X E W L R Q
1 L F R O H W W D & D O J R O D U L H W D O 0 D U V H L
/ D Q J X D J H 5 H V R X U F H V \$ V V R F L D W L R Q S S
<https://aclanthology.org/2022.lrec-1.489>


References IV

-  Bergman, A. and Mona Diab (May ACL (findings) 2022).
“Towards Responsible Natural Language Annotation for the Varieties of Arabic.” In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 57-66. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 364–371. DOI: 10.18653/v1/2022.findings-acl.31. URL: <https://aclanthology.org/2022.findings-acl.31>.

References V

-  Salameh, Mohammad, Houda Bouamor, and Nizar Habash (Aug. 2018). “Fine-Grained Arabic Dialect Identification.” In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, August 1-7, 2018, Brussels, Belgium. Association for Computational Linguistics, pp. 1332–1344. URL: <https://aclanthology.org/C18-1113>.

References VI

-  Zaidan, Omar F. and Chris Callison-Burch (June 2011). "The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content." In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 37–41. URL: <https://aclanthology.org/P11-2007>.

Analyzed Datasets

Searching catalog of **public datasets** (Masader)

Searching catalog of **public datasets** (Masader)

- 1 **Language:** Mixture of MSA and DA.

Searching catalog of **public datasets** (Masader)

- 1 **Language:** Mixture of MSA and DA.

151 datasets

Searching catalog of **public datasets** (Masader)

- 2 **Tasks Setup:** Sentence-level classification.
- 3 **Samples Variation:** multiple variants of DA.
- 4 **Annotators:**
 - speakers of different variants of DA
 - randomly assigned to the samples.

Searching catalog of **public datasets** (Masader)

- 2 **Tasks Setup:** Sentence-level classification.
- 3 **Samples Variation:** multiple variants of DA.
- 4 **Annotators:**
 - speakers of different variants of DA
 - randomly assigned to the samples.

28 datasets

Searching catalog of **public datasets** (Masader)

- 5 **Released Labels:** Individual annotator labels.

 Searching catalog of **public datasets** (Masader)

5 **Released Labels:** Individual annotator labels.

15 datasets

6 Tasks including: Sentiment Analysis, Sarcasm Detection, Dialect Identification.

Methodology

- 1 Estimate ALDi of samples.
- 2 Bin samples.
- 3 Estimate % samples Full Agreement.

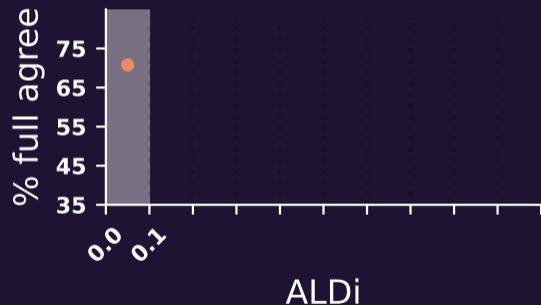
Sentiment Analysis Dataset

Methodology:

- 1 Estimate ALDi of samples.
- 2 Bin samples.
- 3 Estimate % samples Full Agreement.

i Labels:

Positive/ Neutral/ Negative



6HQWLPHQW \$QDO\VLV 'DWDV

0HWKRGRORJ\ /DEHOV

- (VWLPDWH \$/'LRI V3DPVLOWLWH 1HXWUDO
- %LQVDPSOHV
- (VWLPDWH VDP SOHV)XOO
\$JUHPHQW

6HQWLPHQW \$QDO\VLV 'DWDV

0HWKRGRORJ\ /DEHOV

- (VWLPDWH \$/'LRI V3DPVLOWLWH 1HXWUDO
- %LQVDPSOHV
- (VWLPDWH VDP SOHV)XOO
\$JUHPHQW

6HQWLPHQW \$QDO\VLV 'DWDV

0HWKRGRORJ\ /DEHOV

- (VWLPDWH \$/'LRI V3DPVSLWLVH 1HXWUDO
- %LQVDPSOHV
- (VWLPDWH VDPSONHV)XOO
\$JUHPHQW

6HQWLPHQW \$QDO\VLV 'DWDV

0HWKRGRORJ\ /DEHOV

- (VWLPDWH \$/'LRI V3DPVSLWLVH 1HXWUDO
- %LQVDPSOHV
- (VWLPDWH VDPSONHV)XOO
\$JUHPHQW

6HQWLPHQW \$QDO\VLV 'DWDV

0HWKRGRORJ\ /DEHOV

- (VWLPDWH \$/'LRI V3DPVLOWLWH 1HXWUDO
- %LQVDPSOHV
- (VWLPDWH VDPSONHV)XOO
\$JUHPHQW

6HQWLPHQW \$QDO\VLV 'DWDV

0HWKRGRORJ\ /DEHOV

- (VWLPDWH \$/'LRI V3DPVSLWLVH 1HXWUDO
- %LQVDPSOHV
- (VWLPDWH VDPSONHV)XOO
\$JUHPHQW

6HQWLPHQW \$QDO\VLV 'DWDV

0HWKRGRORJ\ /DEHOV

- (VWLPDWH \$/'LRI V3DPVSLWLVH 1HXWUDO
- %LQVDPSOHV
- (VWLPDWH VDPSONHV)XOO
\$JUHPHQW

6HQWLPHQW \$QDO\VLV 'DWDV

0HWKRGRORJ\ /DEHOV

- (VWLPDWH \$/'LRI V3DPVLOWLWYH 1HXWUDO
- %LQVDPSOHV
- (VWLPDWH VDPSONHV)XOO
\$JUHPHQW

6HQWLPHQW \$QDO\VLV 'DWDV

0HWKRGRORJ\ /DEHOV

- (VWLPDWH \$/'LRI V3DPVSLWLVH 1HXWUDO
- %LQVDPSOHV
- (VWLPDWH VDP SOHV)XOO
\$JUHPHQW

'LVWLQJXLVKLQJ EHWZHHQ WKH 9DULHWLHV RI \$UDELF 'LDOHFW -GHQWL4

-\$ \$

'LDOHFW -GHQWL4FDWLI

/DEHOV 0DFUR UHJLRQDO

06\$ 0DJKU HJAS WHYD QXOI

-\$ \$

'LDOHFW -GHQWL4FDWLI

/DEHOV 0DFUR UHJLRQDO

06\$ (J\SW *XOI /HYDQW 0DJKUHE

-\$ \$

'LDOHFW -GHQWL4FDWLI

/DEHOV 0DFUR UHJLRQDO

06\$ (J\SW *XOI /HYDQW 0DJKUHE

-\$ \$

'LDOHFW -GHQWL4FDWLI

/DEHOV 0DFUR UHJLRQDO

06\$ (J\SW *XOI /HYDQW 0DJKUHE

-\$ \$

'LDOHFW -GHQWL4FDWLI

/DEHOV 0DFUR UHJLRQDO

06\$ (J\SW *XOI /HYDQW 0DJKUHE

-\$ \$

'LDOHFW -GHQWL4FDWLI

/DEHOV 0DFUR UHJLRQDO

06\$ (J\SW *XOI /HYDQW 0DJKUHE

-\$ \$

'LDOHFW -GHQWL4FDWLI

/DEHOV 0DFUR UHJLRQDO

06\$ (J\SW *XOI /HYDQW 0DJKUHE

-\$ \$

'LDOHFW -GHQWL4FDWLI

/DEHOV 0DFUR UHJLRQDO

06\$ (J\SW *XOI /HYDQW 0DJKUHE

-\$ \$

'LDOHFW -GHQWL4FDWLI

/DEHOV 0DFUR UHJLRQDO

06\$ (J\SW *XOI /HYDQW 0DJKUHE

-\$ \$

'LDOHFW -GHQWL4FDWLI

/DEHOV 0DFUR UHJLRQDO

06\$ (J\SW *XOI /HYDQW 0DJKUHE

-\$ \$

'LDOHFW -GHQWL4FDWLI

/DEHOV 0DFUR UHJLRQDO

06\$ (J\SW *XOI /HYDQW 0DJKUHE

(DVLO\ LGHQWL4DEOH 06\$ VDPS

'\$ VDPSOHV ZLWK PXOWLSOH GLVWLQFV

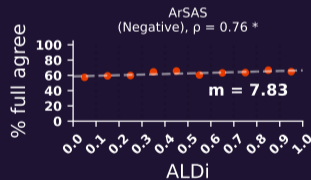
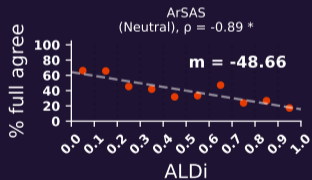
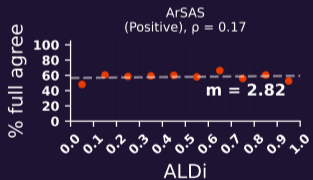
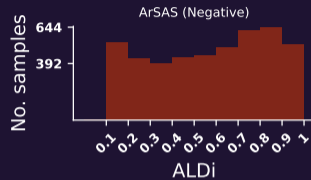
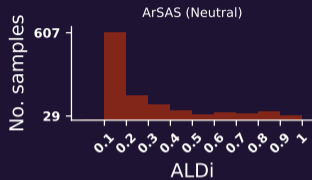
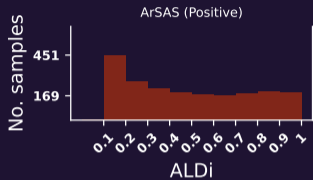
25

+ DUG WR GHWHUPLQH WKH GLD
9DOLG LQ PXOWLSOH GLDOH

'LVWLQJXLVKLQJ EHWZHHQ WKH 9DULHWLHV RI \$UDELF 'LDOHFW -GHQWL4

7ZR 6HQWLP HQW \$QDO\VLV 'D

'LVWLQJXLVKLQJ EHWZHHQ WKH 9DULHWLHV RI \$UDELF 'LDOHFW -GHQWL4



AOC's annotation guidelines

- Tell us how much dialect () is in the sentence.
- Dialect Level (H H):
 - No dialect ()
 - A bit of dialect (H H)
 - Mixed (H H)
 - Mostly dialect ()
 - Not Arabic (S Ri Dj D)

Discarded Samples

Type	Sentence	Source	Level of Dialectness
Symbols	@@@@@@	Cmnt (Y7)	↯ Arabic (x13), Missing (x2)
	*****	Cmnt (Ri)	↯ Arabic (x3)
English	gloves to protect the baby from infection !	Cmnt (Ri)	↯ Arabic (x2), MSA (x1)
	very nice...	Cmnt (Ri)	↯ Arabic (x3)
Arabizi	ya zamalek ya 7arameyaaaa	Cmnt (Y7)	↯ Arabic (x2), Most (x1)
URLs and Emails	http://el beet-el musl i m. ace. st/forum. htm	Cmnt (Y7)	↯ Arabic (x3)
	Ahmad. al tami mi @al ghad. j o	Cntrl (Gh)	↯ Arabic (x3)
Presence of HTML	█████ 5000 √DONE	Cmnt (Y7)	↯ Arabic (x3)
	 b D'P	Cntrl (Y7)	↯ Arabic (x3)