



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

Amr Keleg

arbml Board

1 July 2024

## 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

أُسعدنا الرجل

DA

الراجل أُسعدنا

## Arabic sentences



الزلة أسعدنا

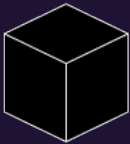
MSA

أسعدنا الرجل

DA

الراجل أسعدنا

# Arabic sentences



MSA

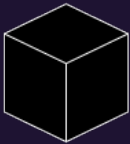
أسعدنا الرجل

DA

الراجل أسعدنا

الزلمة أسعدنا

## Arabic sentences



الزول أسعدنا

MSA

أسعدنا الرجل

DA

الراجل أسعدنا

الزلة أسعدنا

# Arabic sentences



MSA

أسعدنا الرجل

DA

الراجل أسعدنا

الزلمة 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

الزلة أسعدنا

الزول أسعدنا

## Arabic sentences



MSA

أسعدنا الرجل

Nile

الراجل أسعدنا

الزول أسعدنا

Levant

الزلة 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

الزول أسعدنا

Levant

الزلة 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., الزلّة أسعدنا)<sup>4</sup>
- 2 Generally short (e.g., وين المحطة؟)<sup>5</sup>

---

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

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



# Community's perceptions

- 1 Common for same-region dialects (e.g., الزلّة أسعدنا)<sup>4</sup>
- 2 Generally short (e.g., وين المحطة؟)<sup>5</sup>

Is it a significant limitation ?



More on this later!

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

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



# Community's perceptions

- 1 Common for same-region dialects (e.g., الزلّة أسعدنا)<sup>4</sup>
- 2 Generally short (e.g., وين المحطة؟)<sup>5</sup>

Is it a significant limitation ?



More on this later!




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

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



<sup>5</sup>Salameh, 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)

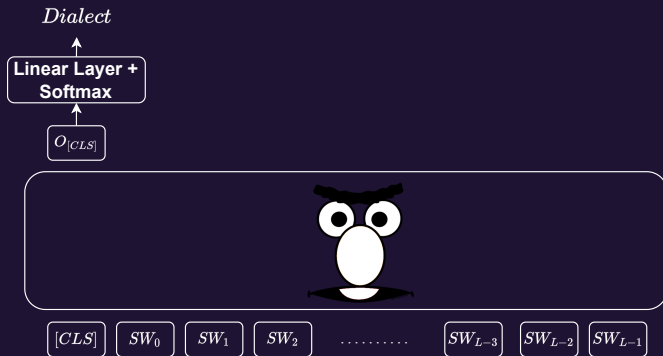
## **i** Training Dataset: NADI 2023

Target	Sentence
	وحدو الحب بيكبر و بضل
	الكحلوشه جزائريه والى مبروك علينا
...	...

## **i** Labels: 18 geolocated dialects

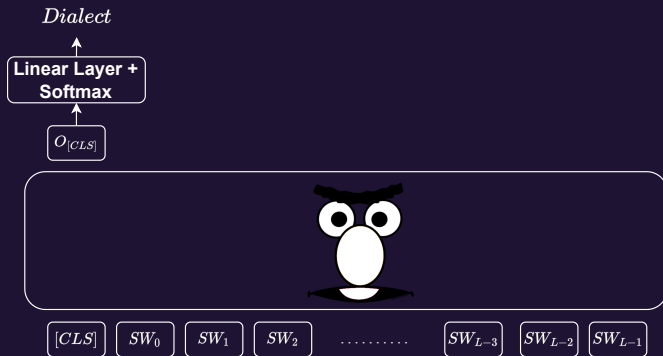
# Country-level ADI system (NADI 2023)

**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



مرتضي صوتوا ضعيف مع كامل إحترامي مايتقارنش بنسيم مجرد مقارنة

**Target:** Tunisia 🇹🇳

**Prediction:** Egypt 🇪🇬

**Verdict:** Error ❌

مرتضي صوتوا ضعيف مع كامل إحترامي مايتقارنش بنسيم مجرد مقارنة

**Target:** Tunisia 🇹🇳

**Prediction:** Egypt 🇪🇬

**Verdict:** Error ❌



Is this Sentence Valid in your dialect?

مرتضي صوتوا ضعيف مع كامل إحترامي مايتقارنش بنسيم مجرد مقارنة

**Target:** Tunisia 🇹🇳

**Prediction:** Egypt 🇪🇬

**Verdict:** Error ❌



Is this Sentence Valid in your dialect?



Yes

Not an Error

بجد الناس اللي بتنسى بسرعة بجد كييف تعملوها ! ؟ !

**Target:** Algeria 🇩🇿

**Prediction:** Egypt 🇪🇬

**Verdict:** Error ❌

بجد الناس اللي بتنسى بسرعة بجد كييف تعملوها ! ؟ !

**Target:** Algeria 🇩🇿

**Prediction:** Egypt 🇪🇬

**Verdict:** Error ❌



Is this Sentence Valid in your dialect?

يجد الناس اللي بتنسى بسرعة يجد كييف تعملوها ! ؟ !

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

---

Sentence<sub>1</sub>

Sentence<sub>2</sub>

Sentence<sub>3</sub>

...

Sentence<sub>N</sub>

# How to Build a Multilabel ADI Dataset? (1)

Sentence

---

Sentence<sub>1</sub>

Sentence<sub>2</sub>

Sentence<sub>3</sub>









...



Sentence<sub>N</sub>

Annotators















# How to Build a Multilabel ADI Dataset? (1)



Sentence					...
Sentence <sub>1</sub>					...
Sentence <sub>2</sub>					
Sentence <sub>3</sub>					
...					
Sentence <sub>N</sub>					

Annotators  





















# How to Build a Multilabel ADI Dataset? (1)

Sentence					...
Sentence <sub>1</sub>					...
Sentence <sub>2</sub>					...
Sentence <sub>3</sub>					
...					
Sentence <sub>N</sub>					

Annotators

# How to Build a Multilabel ADI Dataset? (1)

Sentence					...
Sentence <sub>1</sub>					...
Sentence <sub>2</sub>					...
Sentence <sub>3</sub>					...
...	...	...	...	...	...
Sentence <sub>N</sub>					...

Annotators



# How to Build a Multilabel ADI Dataset? (1)

Sentence					...
Sentence <sub>1</sub>					...
Sentence <sub>2</sub>					...
Sentence <sub>3</sub>					...
...	...	...	...	...	...
Sentence <sub>N</sub>					...

Annotators  

Only need 3 annotators, but major quality issues !!

# How to Build a Multilabel ADI Dataset? (2)

## Sentence

---

Sentence<sub>1</sub>

Sentence<sub>2</sub>

Sentence<sub>3</sub>

...

Sentence<sub>N</sub>

# How to Build a Multilabel ADI Dataset? (2)

Sentence 

---

Sentence<sub>1</sub> 

Sentence<sub>2</sub> 

Sentence<sub>3</sub> 















...

Sentence<sub>N</sub> 


Annotator  





# How to Build a Multilabel ADI Dataset? (2)

Sentence		
Sentence <sub>1</sub>		
Sentence <sub>2</sub>		
Sentence <sub>3</sub>		
...	...	...
Sentence <sub>N</sub>		
Annotator	 	 





# How to Build a Multilabel ADI Dataset? (2)

Sentence					...
Sentence <sub>1</sub>					...
Sentence <sub>2</sub>					...
Sentence <sub>3</sub>					...
...	...	...	...	...	...
Sentence <sub>N</sub>					...
Annotator	 	 	 	 	...

# How to Build a Multilabel ADI Dataset? (2)

Sentence					...
Sentence <sub>1</sub>					...
Sentence <sub>2</sub>					...
Sentence <sub>3</sub>					...
...	...	...	...	...	...
Sentence <sub>N</sub>					...
Annotator	 	 	 	 	...

# How to Build a Multilabel ADI Dataset? (2)




Sentence					...
Sentence <sub>1</sub>					...
Sentence <sub>2</sub>					...
Sentence <sub>3</sub>					...
...	...	...	...	...	...
Sentence <sub>N</sub>					...
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
لمن الحياه ترسل ليك رساله	Palestine  , Sudan  , Yemen 

# Examples

Sentence	Valid in
لمن الحياه ترسل ليك رساله	Palestine  , Sudan  , Yemen 
وين يلعب هذا ما شففته	Algeria  , Palestine  , Yemen 

# Multilabel samples in NADI 2024?

Community's perceptions:

- 1 Common for same-region dialects (e.g., الزلّة أسعدنا)<sup>a</sup>
- 2 Generally short (e.g., وين المحطة؟)<sup>b</sup>

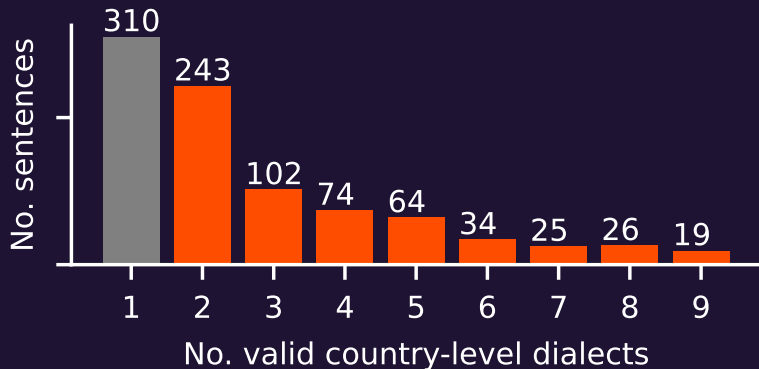
---

<sup>a</sup>Abdelali, Ahmed et al. 2021. "QADI: Arabic Dialect Identification in the Wild."

<sup>b</sup>Salameh, Mohammad, Bouamor, Houda, and Habash, Nizar. 2018. "Fine-Grained Arabic Dialect Identification."



# Multilabel samples in NADI 2024?

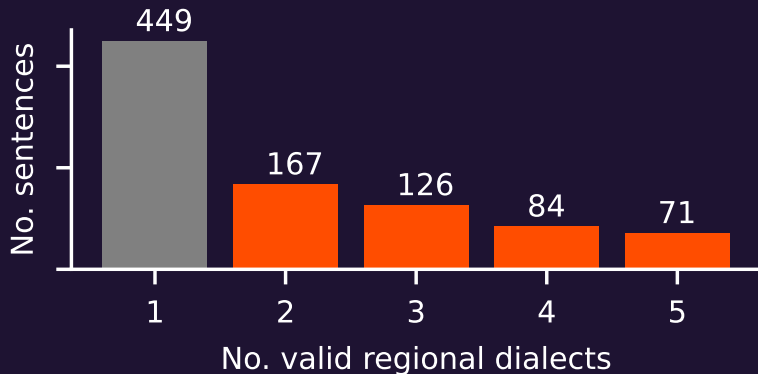


All samples but 310 are multi-dialect (country level).



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 وانا لله وانا 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

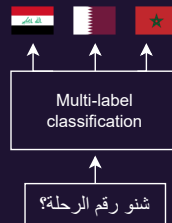
**Training datasets:**  
Single-label datasets from previous NADI subtasks

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

---

Sentence<sub>1</sub>











Sentence<sub>2</sub>

Sentence<sub>3</sub>

...

Sentence<sub>N</sub>











# Subtask 1 - Evaluation

Sentence	Target 	Prediction 
Sentence <sub>1</sub>		
Sentence <sub>2</sub>		
Sentence <sub>3</sub>		
...	...	...
Sentence <sub>N</sub>		



Binary classification task with unbalanced classes

# Subtask 1 - Evaluation











Sentence	Target 	Prediction 
Sentence <sub>1</sub>		
Sentence <sub>2</sub>		
Sentence <sub>3</sub>		
...	...	...
Sentence <sub>N</sub>		



Binary classification task with unbalanced classes



# Subtask 1 - Evaluation

Sentence	Target 	Prediction 
Sentence <sub>1</sub>		
Sentence <sub>2</sub>		
Sentence <sub>3</sub>		
...	...	...
Sentence <sub>N</sub>		



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 <sub>1</sub> 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 <sub>1</sub> score (↑)
Baseline II	Random	50.14 $\pm$ 1.6	30.43 $\pm$ 8.8	50.15 $\pm$ 2.1	37.15 $\pm$ 7.2
Baseline III	Top 1	<b>73.42</b> $\pm$ 7.6	<b>76.82</b> $\pm$ 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 ( $\uparrow$ )	Precision ( $\uparrow$ )	Recall ( $\uparrow$ )	F <sub>1</sub> score ( $\uparrow$ )
1	Elyadata	67.50 $\pm$ 3.7	46.48 $\pm$ 10.1	<b>57.09<math>\pm</math>5.1</b>	<b>50.57<math>\pm</math>7.1</b>
Baseline II	Random	50.14 $\pm$ 1.6	30.43 $\pm$ 8.8	50.15 $\pm$ 2.1	<b>37.15<math>\pm</math>7.2</b>

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

# Multilabel ADI systems

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

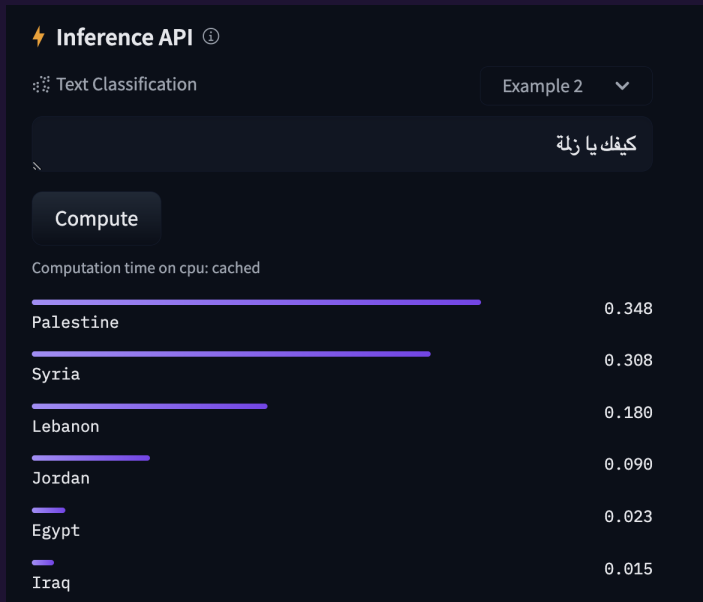
Table: Systems' performance on the test set of Subtask 1.

## Baseline I (Top 90%):

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

## Baseline I (Top 90%):

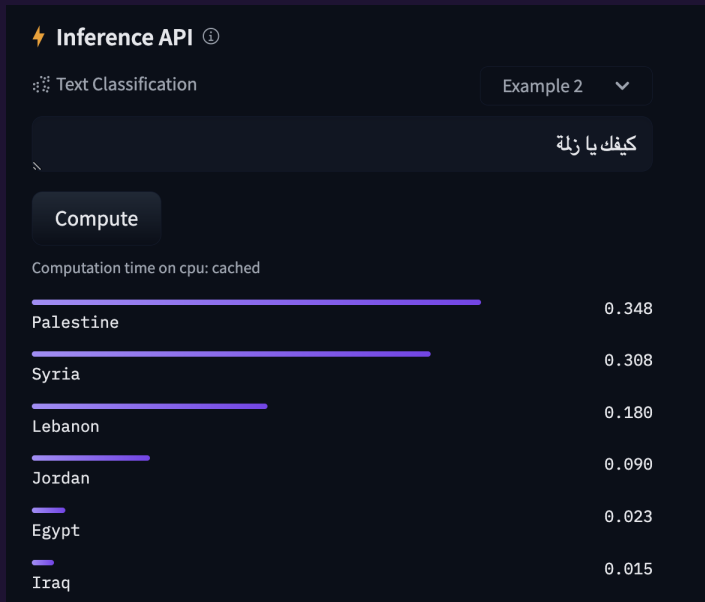
- A fine-tuned BERT-based model
- Single-label ADI



## 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>

# Arabic Level of Dialecttness (ALDi)

 **ALDi: Quantifying the Arabic Level of Dialecttness of Text**  
(Keleg et al., EMNLP 2023)

## Arabic sentences



MSA

أسعدنا الرجل

Egypt

الراجل أسعدنا

Sudan

الزول أسعدنا

Syria

الزلة أسعدنا

Palestine

الزلة أسعدنا

## Arabic sentences



الرجل بسطنا

الرجل شهيصنا

MSA

أسعدنا الرجل

Egypt

الرجل أسعدنا

Sudan

الزول أسعدنا

Syria

الزلة أسعدنا

Palestine

الزلة أسعدنا

## Arabic sentences



MSA

أسعدنا الرجل

Egypt

الراجل أسعدنا

الراجل بسطنا

الراجل شهيصنا

Sudan

الزول أسعدنا

Syria

الزلة أسعدنا

Palestine

الزلة أسعدنا

# Arabic sentences



Egypt

أسعدنا الرجل

الراجل أسعدنا

الراجل بسطنا

الراجل شهيصنا

# Arabic sentences







- **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 *Discrete* Level of Dialecttness labels!

# Dataset !?

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



Popular Dialect Identification (DI) labels.



Ignored *Discrete* Level of Dialecttness labels!



Embrace annotators disagreement!

# Sentence with two valid pronunciations

نبتدى بقى الشغل الصح فى تطوير المدارس وتوفير  
المراقبين عليها

We start with the right task of de-  
veloping schools and providing ob-  
servers over them

# Sentence with two valid pronunciations

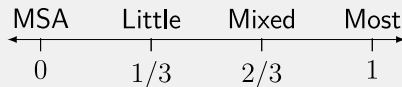
نبتدى بقي الشغل الصح فى تطوير المدارس وتوفير  
المراقبين عليها

We start with the right task of de-  
veloping schools and providing ob-  
servers over them

    
Little Little Most

## 🔄🔄🔄 From AOC to AOC-ALDi 🔄🔄🔄

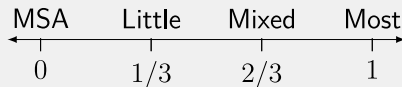
- 1 Labels into numeric values
- 2 Algebraic Mean
- 3 Regression-head on top of MarBERT



e.g.,  $\text{ALDi}(\text{MSA}, \text{MSA}, \text{Little}) = \overline{(0, 0, \frac{1}{3})} = \frac{1}{9} \approx 0.11$

## 🔄🔄🔄 From AOC to AOC-ALDi 🔄🔄🔄

- 1 Labels into numeric values
- 2 Algebraic Mean
- 3 Regression-head on top of MarBERT



e.g.,  $\text{ALDi}(\text{MSA}, \text{MSA}, \text{Little}) = \overline{(0, 0, \frac{1}{3})} = \frac{1}{9} \approx 0.11$



## AOC-ALDi Dataset

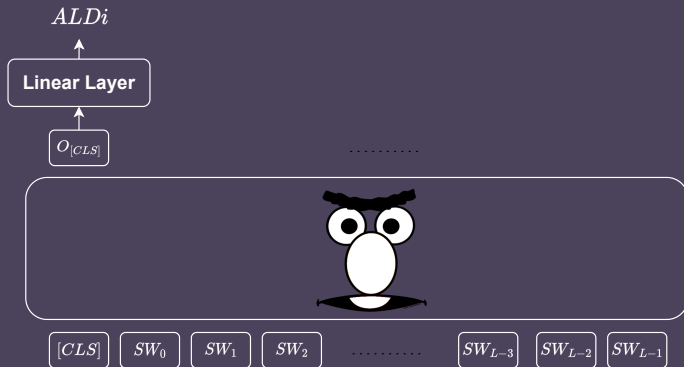
🎯 127,835 sentences (3 👤 annotations each)

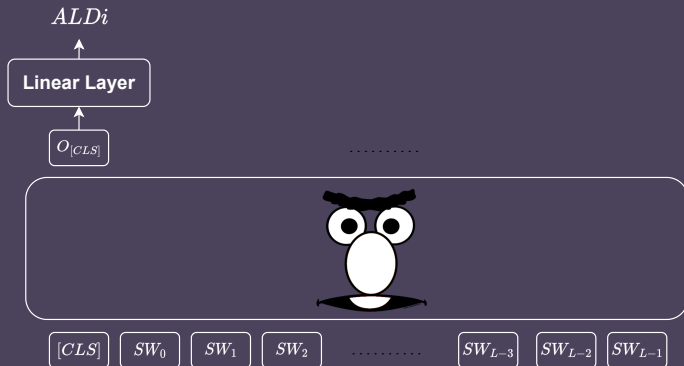
📰 Comments to news articles

🤝 Fleiss'  $\kappa = 0.44$

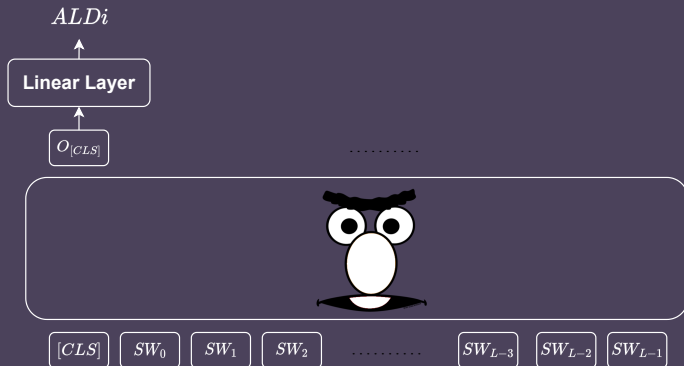
🤝 Krippendorff's  $\alpha$  (interval) = 0.63







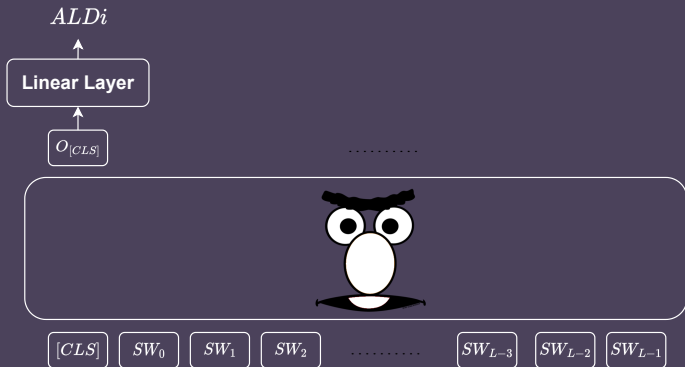
Dialect-agnostic



Dialect-agnostic



$$\text{RMSE}(\text{AOC} - \text{ALDi}_{\text{test}}) = 0.18$$



Dialect-agnostic



$$\text{RMSE}(\text{AOC} - \text{ALDi}_{\text{test}}) = 0.18$$



Demo - [huggingface.co/spaces/AMR-KELEG/ALDi](https://huggingface.co/spaces/AMR-KELEG/ALDi)

# Impact of ALDi on Inter-annotator Agreement

 **Estimating the Level of Dialectness Predicts Inter-annotator Agreement in Multi-dialect Arabic Datasets**  
(Keleg et al., To appear ACL 2024)



# Intuition

- ALDi **INCREASE**  
Mutual Intelligibility?



# Intuition

- ALDi **INCREASE**  
Mutual Intelligibility? **DECREASE**



# 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 ( $\rho$ )  $< -0.7$

# 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 X

# Thanks!


a.keleg@sms.ed.ac.uk

@Amrkeleg on X


## 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.


# References I

-  Abdelali, Ahmed et al. (Apr. 2021). "QADI: Arabic Dialect Identification in the Wild." In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Ed. by Nizar Habash et al. Kyiv, Ukraine (Virtual): Association for Computational Linguistics, pp. 1–10. URL: <https://aclanthology.org/2021.wanlp-1.1>.

# References II

-  Alsarsour, Israa et al. (May 2018). "DART: A Large Dataset of Dialectal Arabic Tweets." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari et al. Miyazaki, Japan: European Language Resources Association (ELRA). URL: <https://aclanthology.org/L18-1579>.

# References III

-  Baimukan, Nurpeiis, Houda Bouamor, and Nizar Habash (June 2022). "Hierarchical Aggregation of Dialectal Data for Arabic Dialect Identification." In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 4586–4596. URL: <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: *Findings of the Association for Computational Linguistics: ACL 2022*. 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 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: 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 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 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



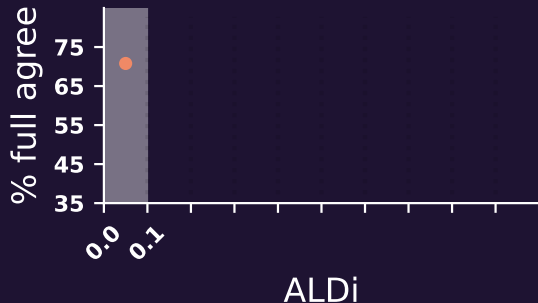
## Methodology:

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



## Labels:

Positive/ Neutral/ Negative



# Sentiment Analysis Dataset



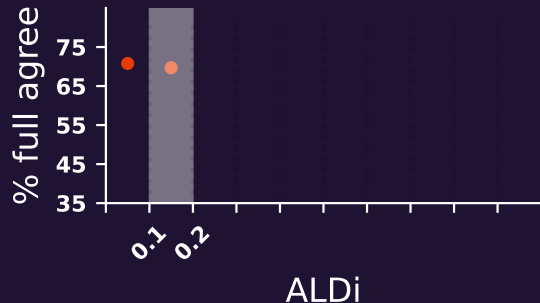
## Methodology:

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



## Labels:

Positive/ Neutral/ Negative



# Sentiment Analysis Dataset



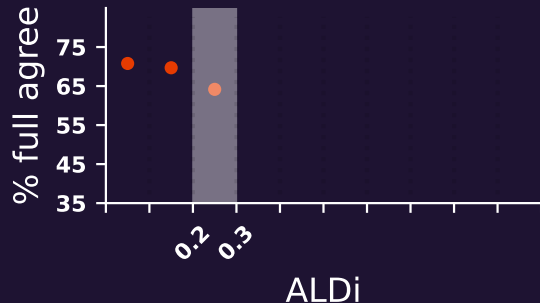
## Methodology:

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



## Labels:

Positive/ Neutral/ Negative



# Sentiment Analysis Dataset



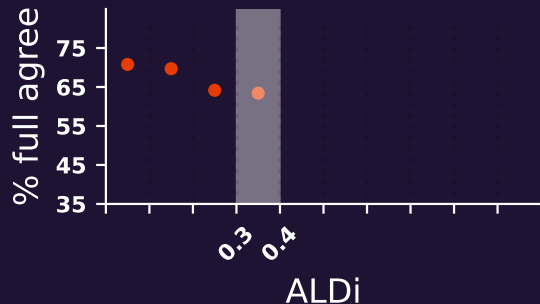
## Methodology:

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



## Labels:

Positive/ Neutral/ Negative





# Sentiment Analysis Dataset



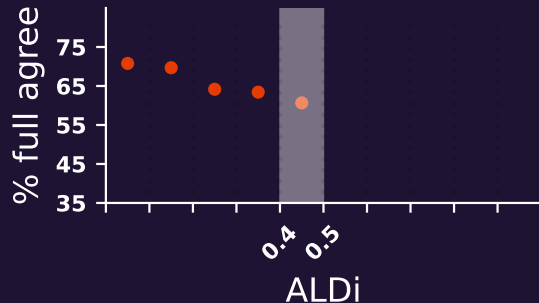
## Methodology:

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



## Labels:

Positive/ Neutral/ Negative



# Sentiment Analysis Dataset



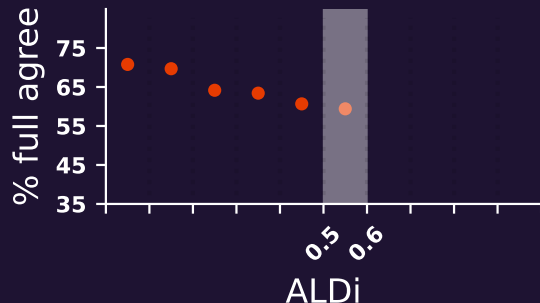
## Methodology:

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



## Labels:

Positive/ Neutral/ Negative



# Sentiment Analysis Dataset



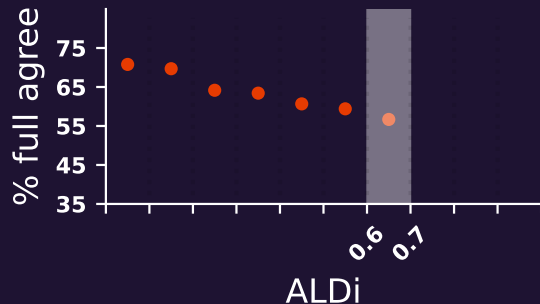
## Methodology:

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



## Labels:

Positive/ Neutral/ Negative



# Sentiment Analysis Dataset



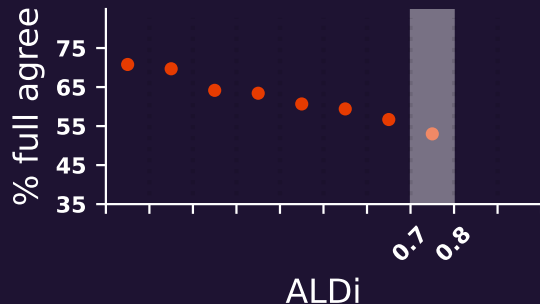
## Methodology:

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



## Labels:

Positive/ Neutral/ Negative



# Sentiment Analysis Dataset



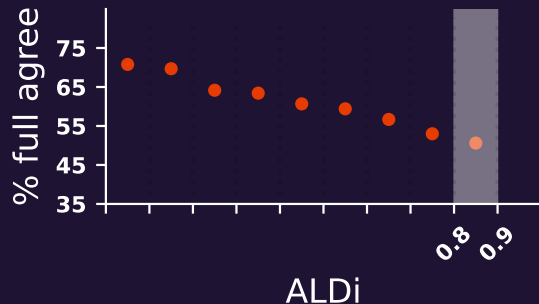
## Methodology:

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



## Labels:

Positive/ Neutral/ Negative



# Sentiment Analysis Dataset



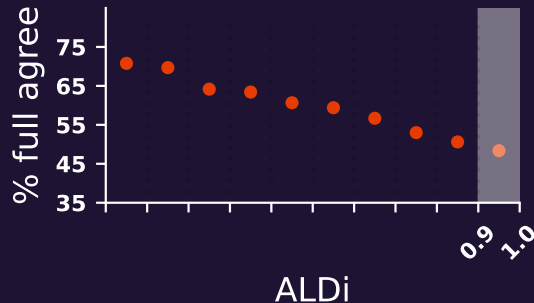
## Methodology:

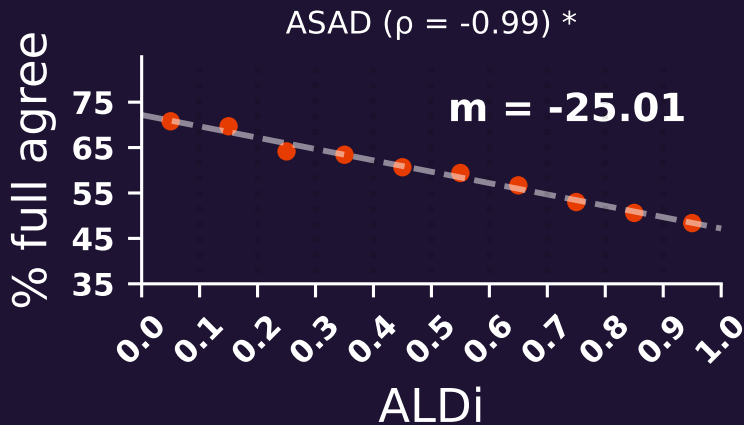
- 1 Estimate ALDi of samples.
- 2 Bin samples.
- 3 Estimate % samples Full Agreement.



## Labels:

Positive/ Neutral/ Negative





# IAA - Dialect Identification Dataset

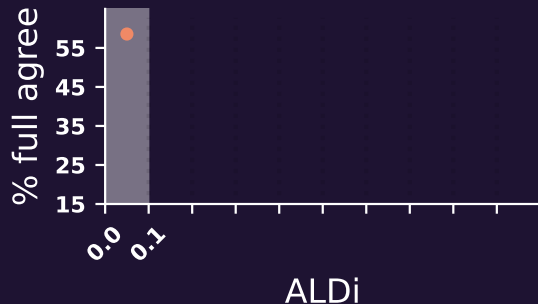
**i Labels (Macro-regional):**  
MSA, Maghreb, Egypt, Levant, Gulf





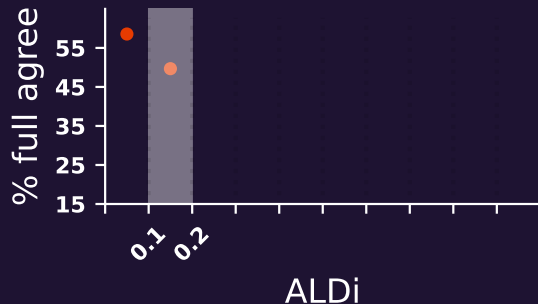
# IAA - Dialect Identification Dataset

**i Labels (Macro-regional):**  
MSA, Egypt, Gulf, Levant, Maghreb



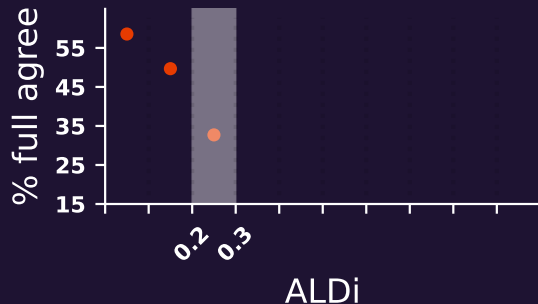
# IAA - Dialect Identification Dataset

**i Labels (Macro-regional):**  
MSA, Egypt, Gulf, Levant, Maghreb



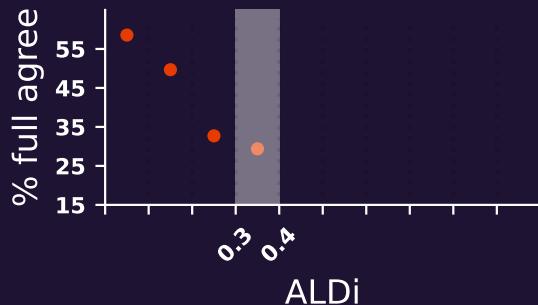
# IAA - Dialect Identification Dataset

**i Labels (Macro-regional):**  
MSA, Egypt, Gulf, Levant, Maghreb



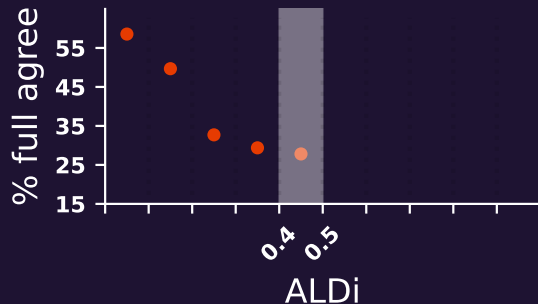
# IAA - Dialect Identification Dataset

**i Labels (Macro-regional):**  
MSA, Egypt, Gulf, Levant, Maghreb



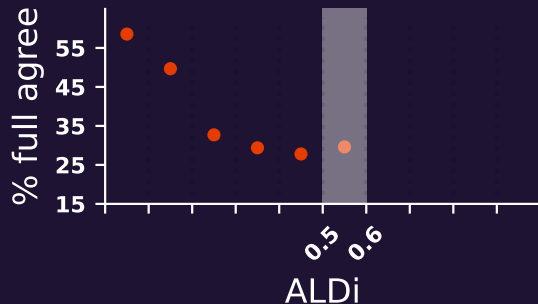
# IAA - Dialect Identification Dataset

**i Labels (Macro-regional):**  
MSA, Egypt, Gulf, Levant, Maghreb



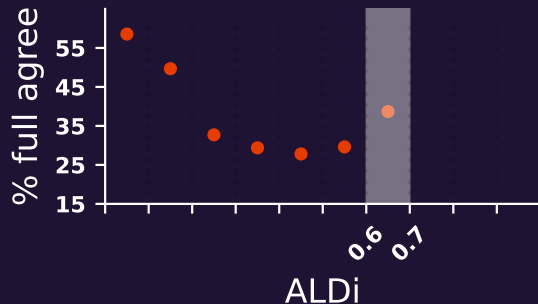
# IAA - Dialect Identification Dataset

**i Labels (Macro-regional):**  
MSA, Egypt, Gulf, Levant, Maghreb



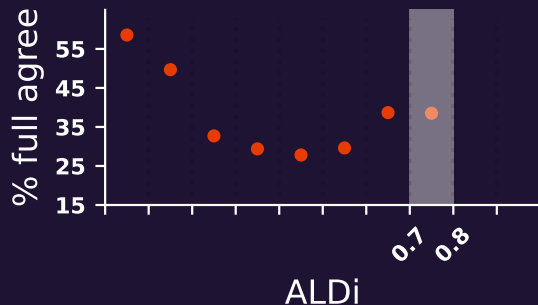
# IAA - Dialect Identification Dataset

**i Labels (Macro-regional):**  
MSA, Egypt, Gulf, Levant, Maghreb



# IAA - Dialect Identification Dataset

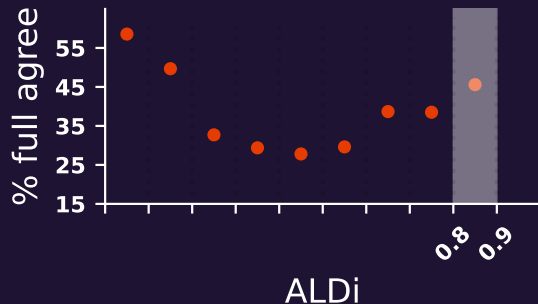
**i Labels (Macro-regional):**  
MSA, Egypt, Gulf, Levant, Maghreb





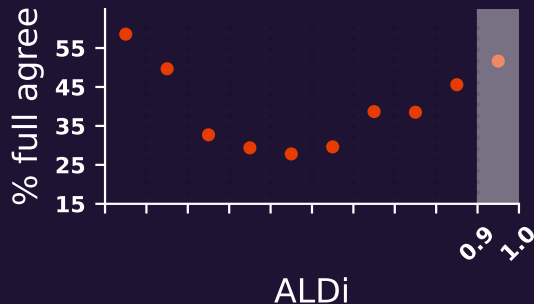
# IAA - Dialect Identification Dataset

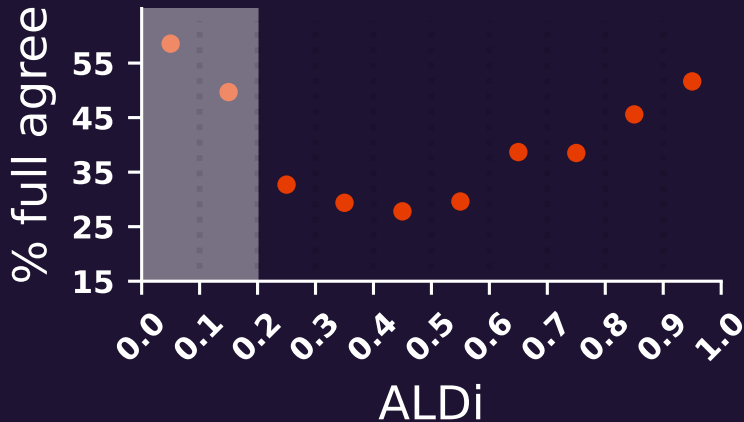
**i Labels (Macro-regional):**  
MSA, Egypt, Gulf, Levant, Maghreb



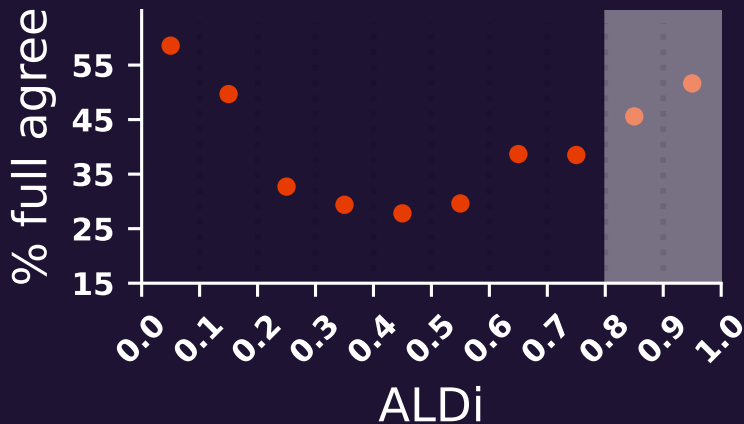
# IAA - Dialect Identification Dataset

**i Labels (Macro-regional):**  
MSA, Egypt, Gulf, Levant, Maghreb

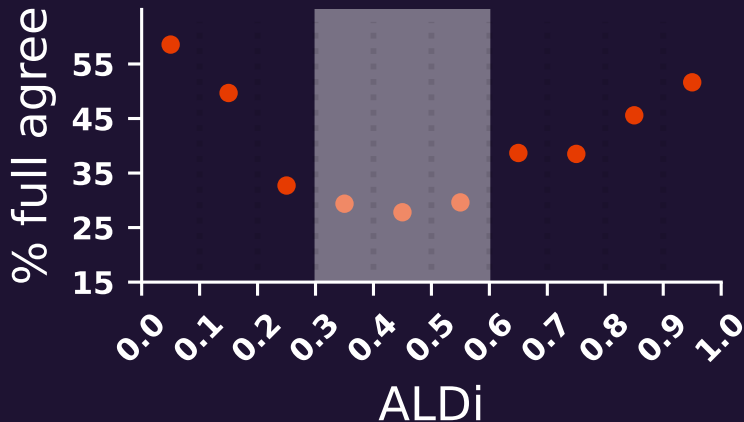




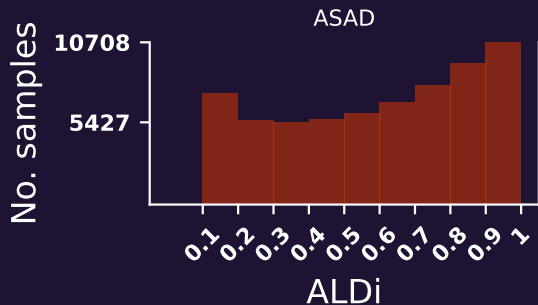
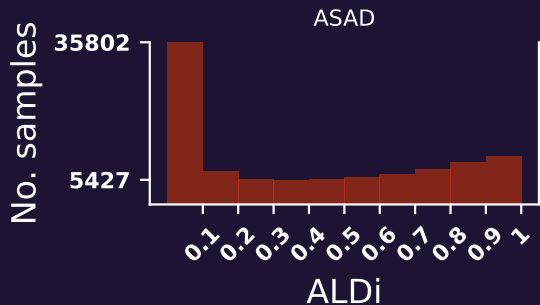
**Easily identifiable MSA samples**



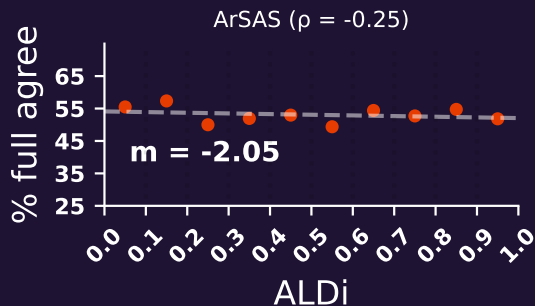
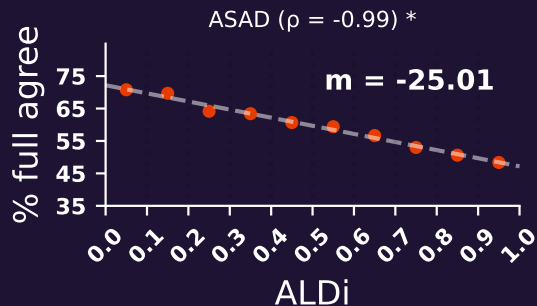
 **DA samples with multiple distinctive cues of a dialect**

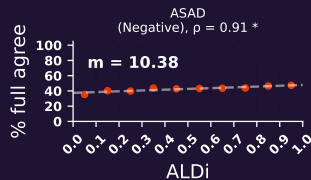
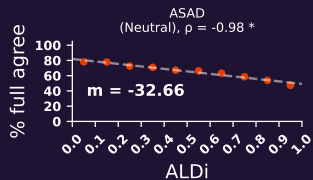
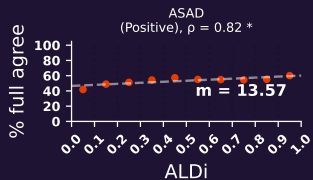
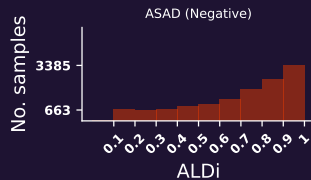
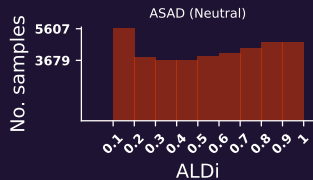
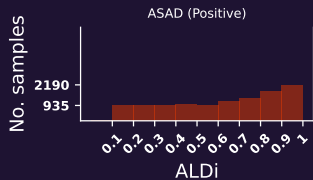


- 📎 1) Hard to determine the dialect?  
OR 📎 2) Valid in multiple dialects?

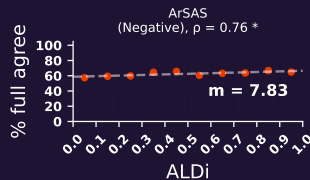
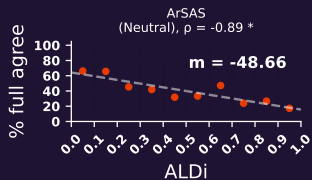
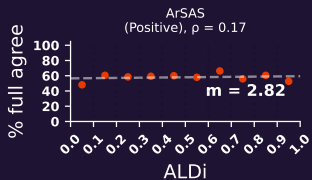
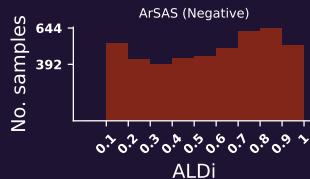
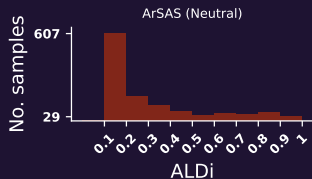
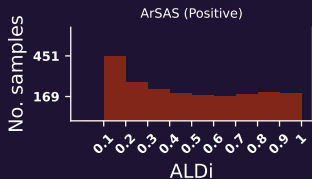


# Two Sentiment Analysis Datasets









# AOC's annotation guidelines

- Tell us how much dialect (عامية) is in the sentence.
- Dialect Level (كمية اللهجة العامية):
  - No dialect (فصحى فقط)
  - A bit of dialect (القليل من العامية)
  - Mixed (خليط من الفصحى والعامية)
  - Mostly dialect (معظمها عامية)
  - Not Arabic (لغة أخرى أو رموز)

# Discarded Samples

Type	Sentence	Source	Level of Dialect-ness
Symbols	؟؟؟؟؟	Cmnt (Y7)	↯ Arabic (x13), Missing (x2)
	*****	Cmnt (Ri)	↯ Arabic (x3)
English	gloves to protect the baby from infection ! تلبس	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://elbeet-elmuslim.ace.st/forum.htm	Cmnt (Y7)	↯ Arabic (x3)
	Ahmad.altamimi@alghad.jo	Cntrl (Gh)	↯ Arabic (x3)
Presence of HTML	&#9608;&#9608;&#9608;&#9608;&#9608;5000 &#8730;DONE	Cmnt (Y7)	↯ Arabic (x3)
	<a href="EditorOpinions.asp?EditorID=404">أشرف ببيع</a>	Cntrl (Y7)	↯ Arabic (x3)