

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,

- _CODEC Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification (Keleg & Magdy, ArabicNLP-WS 2023)
- _CODEC NADI 2024 (Abdul-Mageed et al., To appear ArabicNLP-WS 2024)

② ... nor the solution (Spoiler: Arabic Level of Dialectness)

- _CODEC ALDi: Quantifying the Arabic Level of Dialectness of Text (Keleg et al., EMNLP 2023)
- _CODEC Estimating the Level of Dialectness 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 مايقال عنها انها رائعة
...
طيب مافي حلقات زياده؟ 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



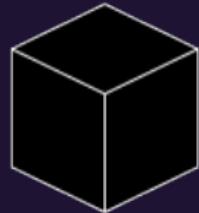
Raw Comments on YT

Comment
حلقة اقل مايقال عنها انها رائعة
...
طيب مافي حلقات زىاده؟ ما شبعنا والله

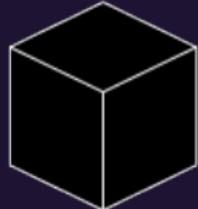
- Perform per-variety analysis:
 - a) offensive text
 - b) models' performance

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

Arabic sentences



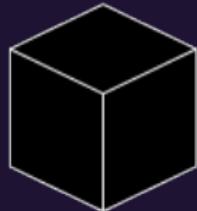
Arabic sentences



أَسْعَدْنَا الرَّجُلَ

الرَّاجِلَ أَسْعَدْنَا

Arabic sentences



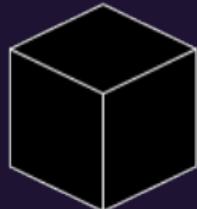
MSA

أسعدنا الرجل

DA

الرجل أسعده

Arabic sentences



الزملة أسعدنا



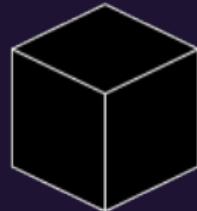
MSA

أسعدنا الرجل

DA

الرجل أسعدنا

Arabic sentences



MSA

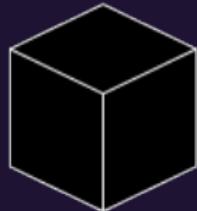
أسعدنا الرجل

DA

الرجل أسعده

الزلة أسعدها

Arabic sentences



الزول أسعدنا



MSA

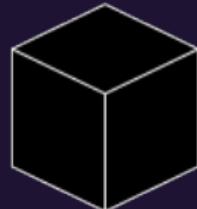
أسعدنا الرجل

DA

الراجل أسعدنا

الزلة أسعدنا

Arabic sentences



MSA

أسعدنا الرجل

DA

الرجل أسعده

الزلة أسعدها

الزول أسعدها

Proposal #1



MSA

- shared across countries



DA

- different variants

More granular scheme ?

(Proposal #2)

Regional Grouping of DA

Proposal #2



MSA

- shared across countries



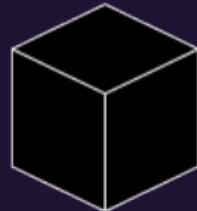
Regional dialects

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

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

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

Arabic sentences



MSA

أسعدنا الرجل

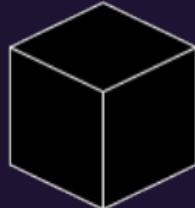
DA

الرجل أسعده

الزلة أسعدها

الزول أسعدها

Arabic sentences



MSA

أسعدنا الرجل

Nile

الراجل أسعدها

Levant

الزملة أسعدها

الزول أسعدها

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 أسعدها

Levant

الزملة أسعدها

الزول أسعدها

Arabic sentences



MSA

أسعدنا الرجل

Egypt

الراجل أسعدها

Sudan

الزول أسعدها

Syria

الزملة أسعدها

Palestine

الزملة أسعدها

Arabic sentences



ويين المحطة؟

MSA

أسعدنا الرجل

Egypt

الراجل أسعدها

Sudan

الزول أسعدها

Syria

الزملة أسعدها

Palestine

الزملة أسعدها

Arabic sentences



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



وين المحطة؟

MSA

أسعدنا الرجل

Egypt

الراجل أسعدها

Sudan

الزول أسعدها

Syria

الزلة أسعدها

Palestine

الزلة أسعدها



Community's perceptions

- ① Common for same-region dialects (e.g., أَلْزَلَةُ أَسْعَدَنَا⁴)
- ② Generally short (e.g., وَيْنِ الْمَحَطَّةُ?⁵)

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

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



Community's perceptions

- ① Common for same-region dialects (e.g., أسلمة أسعدنا⁴)
- ② Generally short (e.g., وين المحطة?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., أسلمة أسعدنا⁴)
- ② Generally short (e.g., وين المحطة؟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)



Training Dataset: NADI 2023

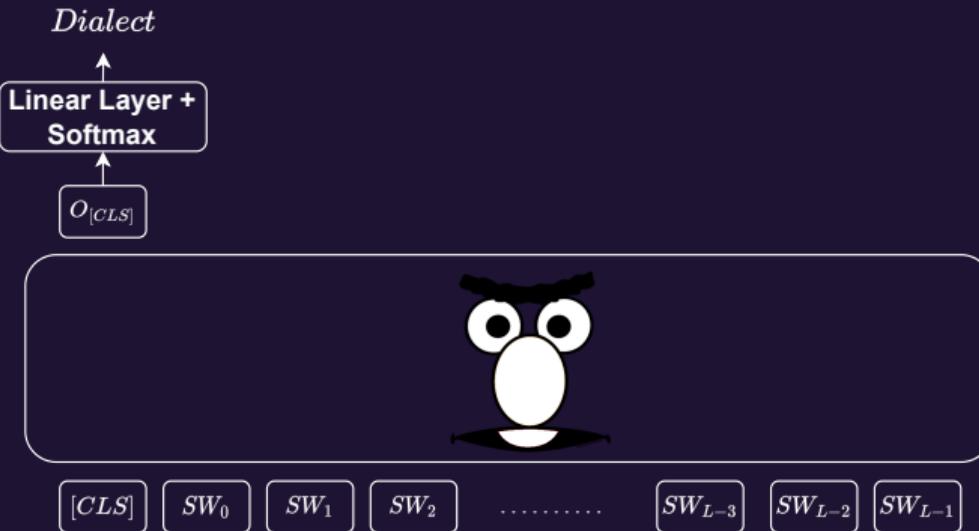
Target	Sentence
	وحدو الحب بيكبر وبضل
	الكحلوشه جزايريه والف مبروك علينا
...	...



Labels: 18 geolocated dialects

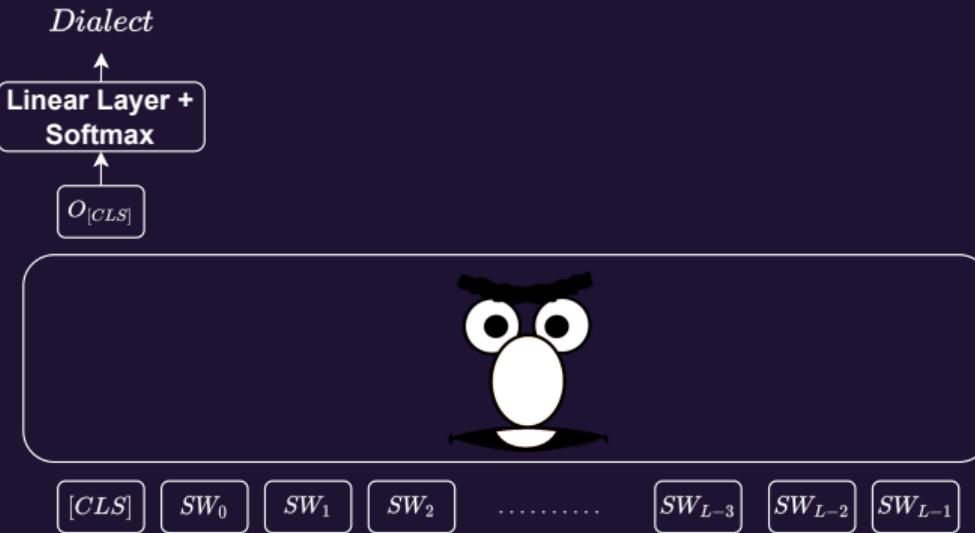
Country-level ADI system (NADI 2023)

 **Labels:** 18 dialects (country-level)



Country-level ADI system (NADI 2023)

 **Labels:** 18 dialects (country-level)



 **Accuracy:** 50.74%



Error Analysis



Annotators: 7 countries

مرتضي صوتوا ضعيف مع كامل إحترامي مايقارن بنسيم مجرد مقارنة

Target: Tunisia 

Prediction: Egypt 

Verdict: Error 

مرتضي صوتوا ضعيف مع كامل إحترامي مايقارنesh بنسيم مجرد مقارنة

Target: Tunisia 

Prediction: Egypt 

Verdict: Error 



Is this Sentence Valid in your dialect?

مرتضي صوتوا ضعيف مع كامل إحترامي مايقارنـش بنـسيـم 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 عملاوها ! ؟ !

Target: Algeria 

Prediction: Egypt 

Verdict: Error 

بجد الناس اللي بتنسى بسرعة بجد كييف عملاوها ! ؟ !

Target: Algeria 

Prediction: Egypt 

Verdict: Error 



Is this Sentence Valid in your dialect?

بجد الناس اللي بتنسى بسرعة بجد 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

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
رسالة ليك ترسل الحياة من	Palestine  , Sudan  , Yemen 

Examples

Sentence	Valid in
رسالة ليك ترسل من الحياة	Palestine  , Sudan  , Yemen 
ما شفته هذا يلعب وين	Algeria  , Palestine  , Yemen 

Multilabel samples in NADI 2024?

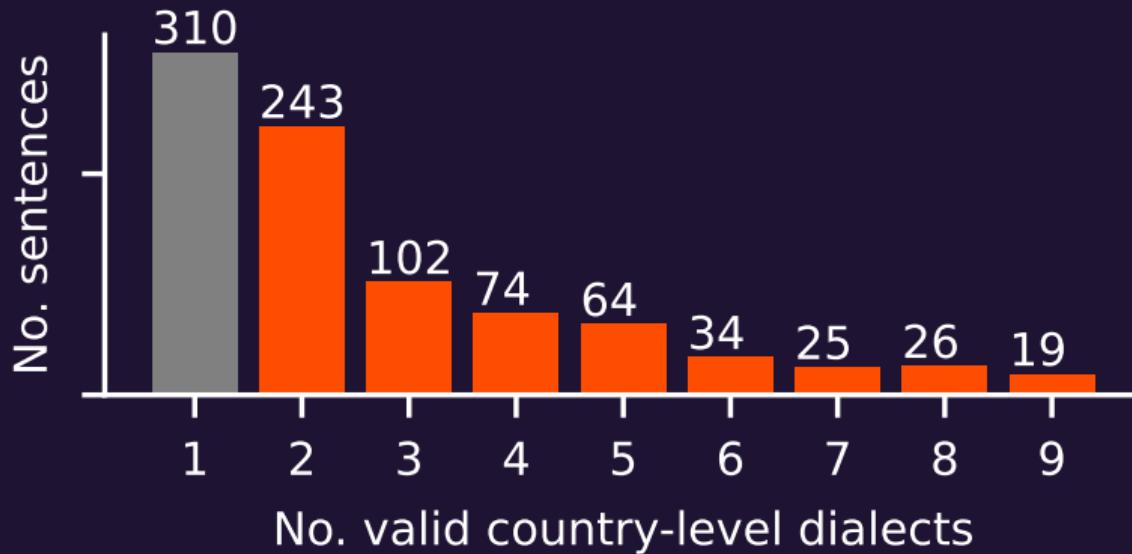
Community's perceptions:

- 1 Common for same-region dialects (e.g., أَلْزَلَةُ أَسْعَدَنَا^a)
- 2 Generally short (e.g., وَيْنِ الْمَحَطَّةُ?^b)

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

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

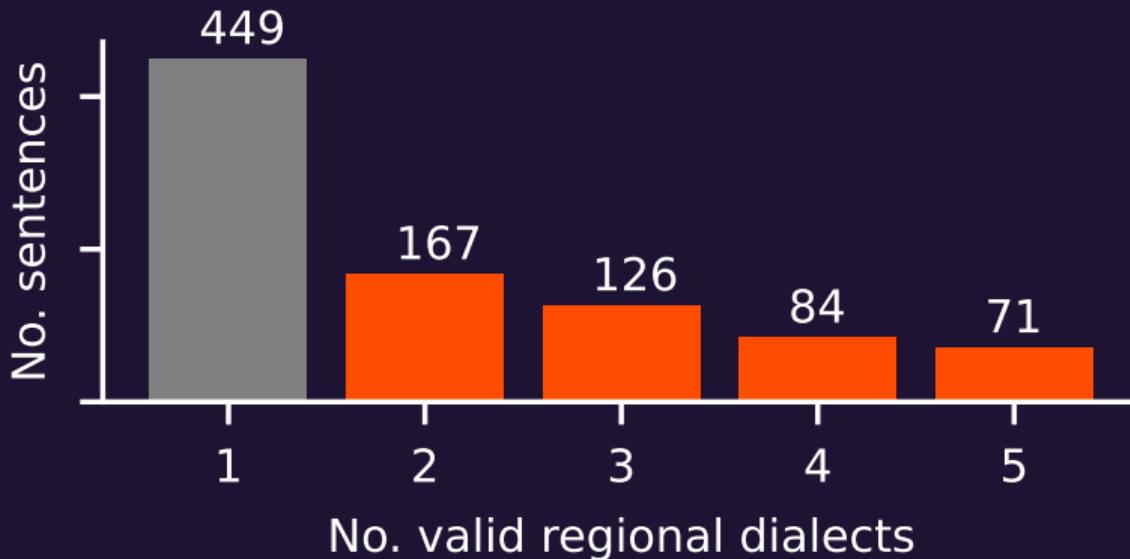
Multilabel samples in NADI 2024?



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

X 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
لله يرحمه ربى معك خربا و انا الله و انا اليه راجعون	المغرب Morocco
والله ما عرف عنه بس جتنى الصوره على الخاص و قلت اكيت تذكروننه	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
لله يرحمه ربى معك خويها و انا الله و انا اليه راجعون	المغرب Morocco
والله ما عرف عنه بس جتنى الصوره على الخاص و قلت اكيت تذكروننه	العراق Iraq
...	...

Subtask (1)

Output:
All valid dialects of the input sentence

```
graph TD; Input["شئو رقم الرحلة؟"] --> ML["Multi-label classification"]; ML --> Output["الى", "الى", "الى"]
```

The diagram illustrates the process of multi-label classification. It starts with an input sentence "شئو رقم الرحلة؟" which is processed by a "Multi-label classification" block. This block outputs three valid dialects, represented by the flags of Iraq, Morocco, and Palestine.

NADI SHARED TASK

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 _{±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 (↑)
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}

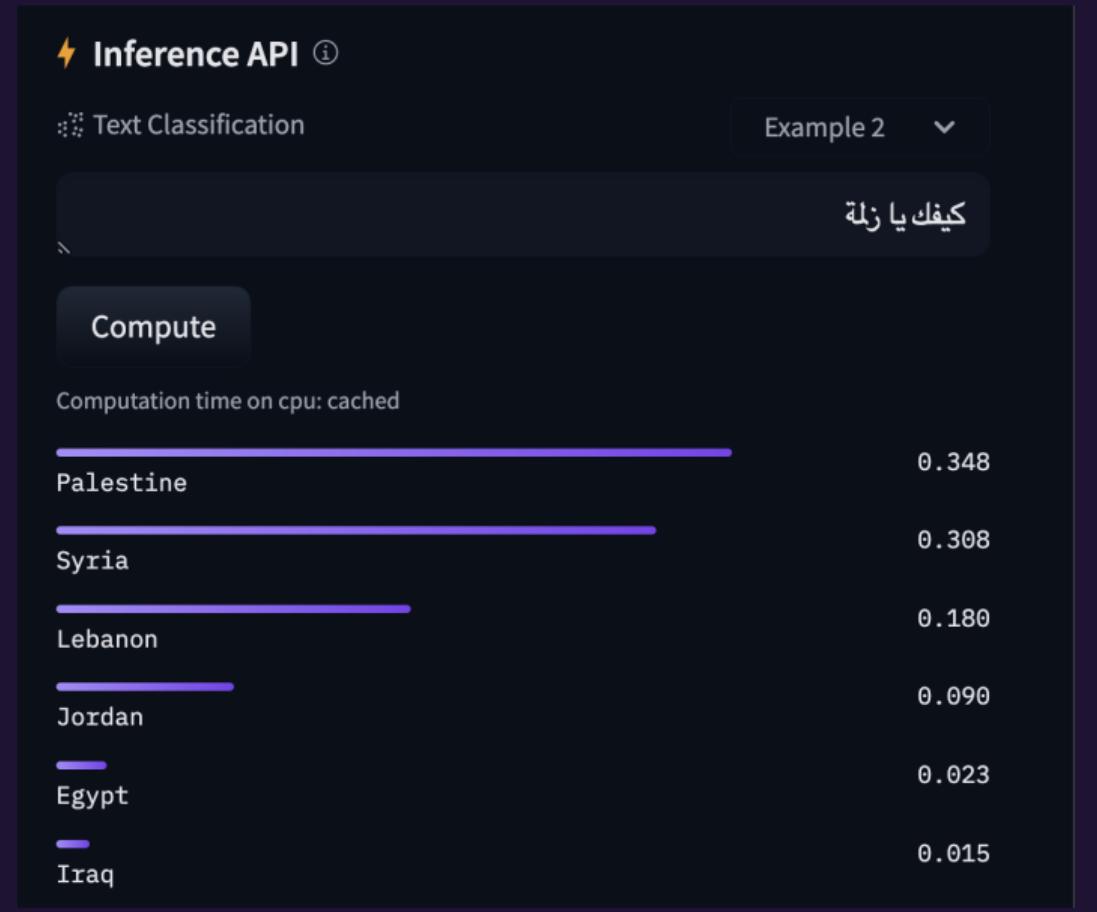
Table: Systems' performance on the test set of Subtask 1.

Baseline I (Top 90%):

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

Baseline I (Top 90%):

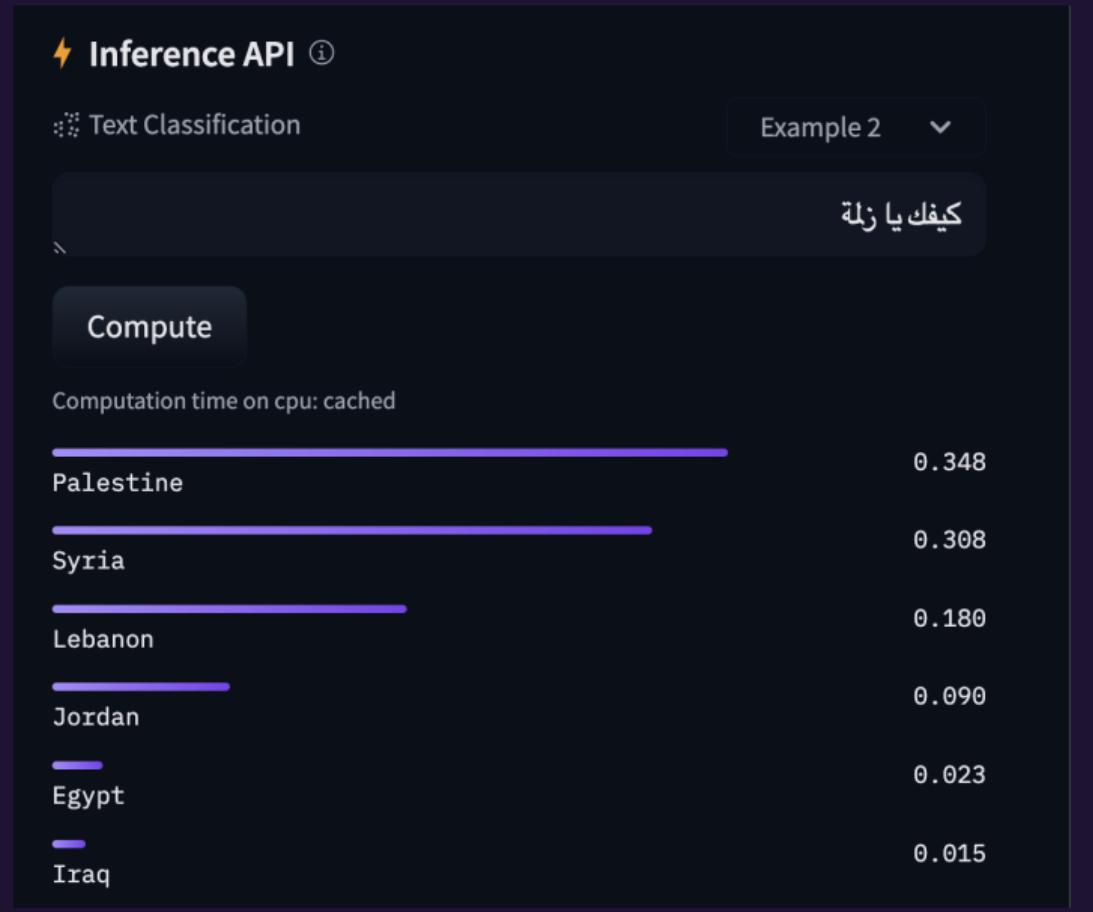
- A fine-tuned BERT-based model
- Single-label ADI



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>

Arabic Level of Dialectness (ALDi)



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

Arabic sentences



MSA

أسعدنا الرجل

Egypt

الراجل أسعدها

Sudan

الزول أسعدها

Syria

الزلة أسعدها

Palestine

الزلة أسعدها

Arabic sentences



الراجل بسطنا

الراجل شهيضنا

MSA

أسعدنا الرجل

Egypt

الراجل أسعدهنا

Sudan

الزول أسعدهنا

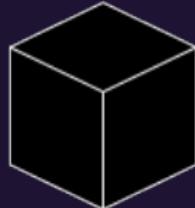
Syria

الزلة أسعدهنا

Palestine

الزلة أسعدهنا

Arabic sentences



MSA

أسعدنا الرجل

Egypt

الراجل أسعدهنا

Sudan

الزول أسعدهنا

Syria

الزلة أسعدهنا

Palestine

الزلة أسعدهنا

Arabic sentences



Egypt

أسعدنا الرجل

الرجل أسعدهنا

الرجل بسطنا

الرجل شهيضنا

Arabic sentences



Egypt

أسعدنا الرجل

الراجل أسعدنا

الراجل بسطنا

الراجل شهيضنا

0 (MSA)

Arabic Level of Dialectness
(ALDi)

1 (High Divergence from MSA)



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



- **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.
- trash bin icon Ignored *Discrete* Level of Dialectness labels!

Zaidan, Omar F. and Callison-Burch, Chris. 2011. "The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content."

Dataset !?



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

- ⭐ Popular Dialect Identification (DI) labels.
- 🗑 Ignored *Discrete* Level of Dialectness labels!
- 🤔 Embrace annotators disagreement!

Zaidan, Omar F. and Callison-Burch, Chris. 2011. "The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content."

Sentence with two valid pronunciations

نبدأ بـالشغل الصحي في تطوير المدارس وتوفير

المراقبين عليها

We start with the right task of developing schools and providing observers over them

Sentence with two valid pronunciations

نبتدى بقى الشغل الصح في تطوير المدارس وتوفير

المراقبين عليها

We start with the right task of developing schools and providing observers 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., $\text{ALDi}(\text{MSA}, \text{MSA}, \text{Little}) = \overline{(0, 0, \frac{1}{3})} = \frac{1}{9} \approx 0.11$



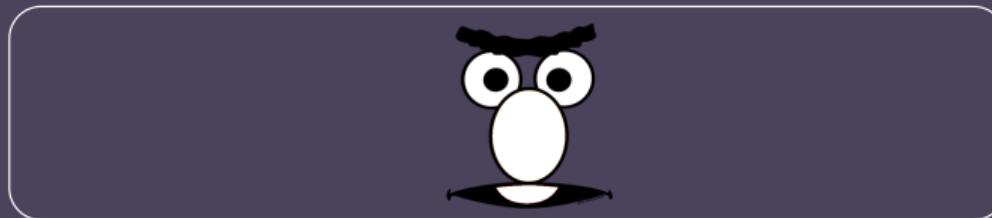
AOC-ALDi Dataset

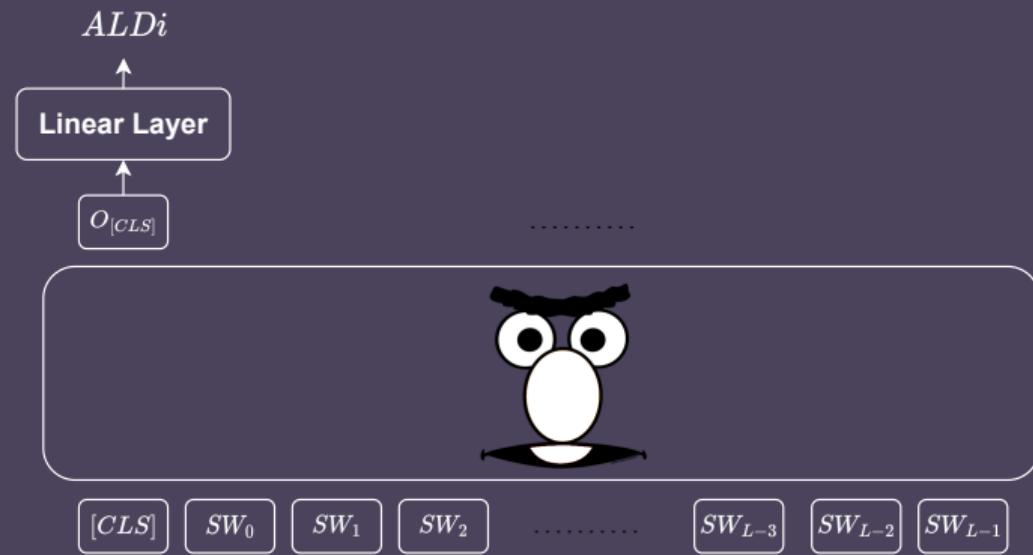
- ⌚ 127,835 sentences (3 🧑 annotations each)
- 📰 Comments to news articles
- 🤝 Fleiss' $\kappa = 0.44$
- 🤝 Krippendorff's α (interval) = 0.63

ALDi

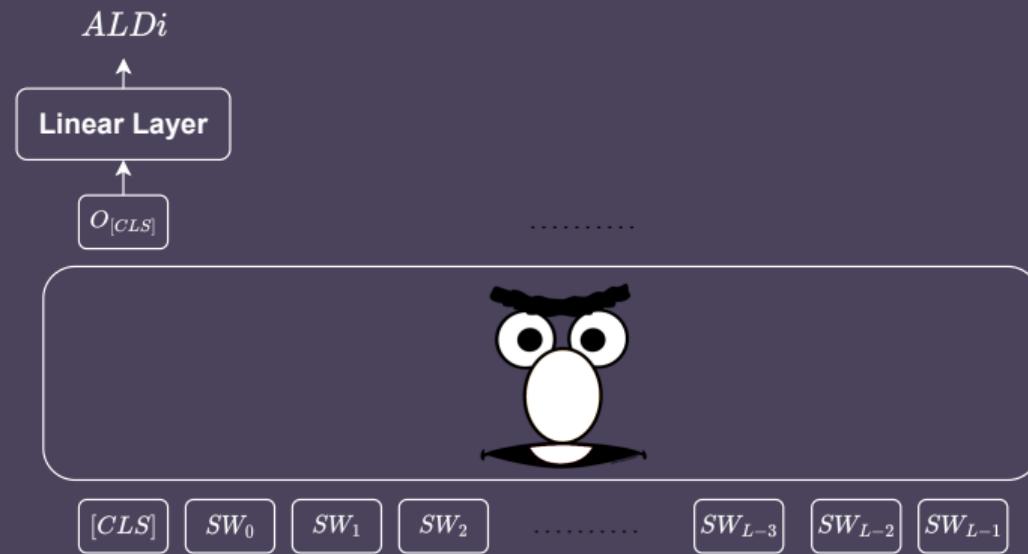
Linear Layer

$O_{[CLS]}$





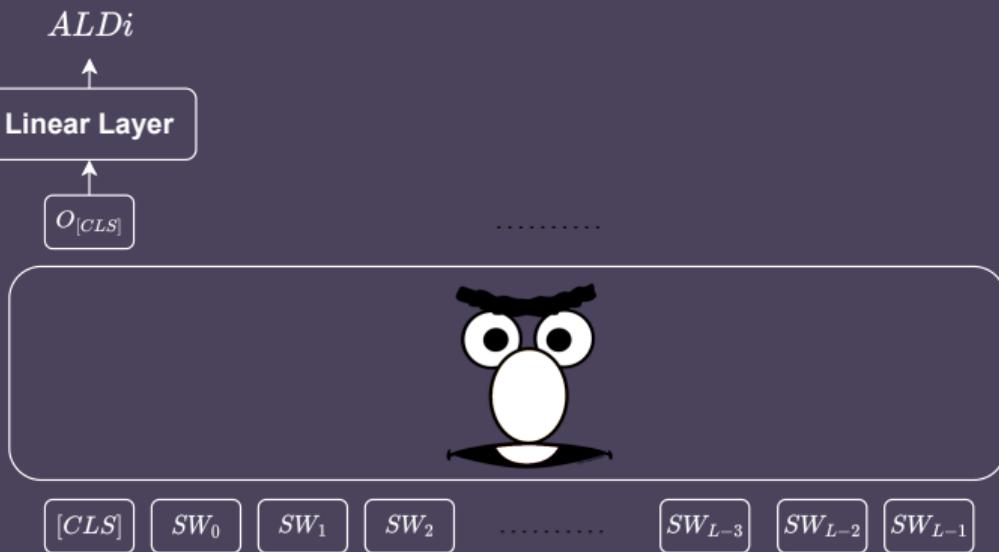
Dialect-agnostic



Dialect-agnostic



$RMSE(AOC - ALDi_{test}) = 0.18$



- Dialect-agnostic**
- $RMSE(AOC - ALDi_{test}) = 0.18$
- Demo - 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 (ρ) < -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



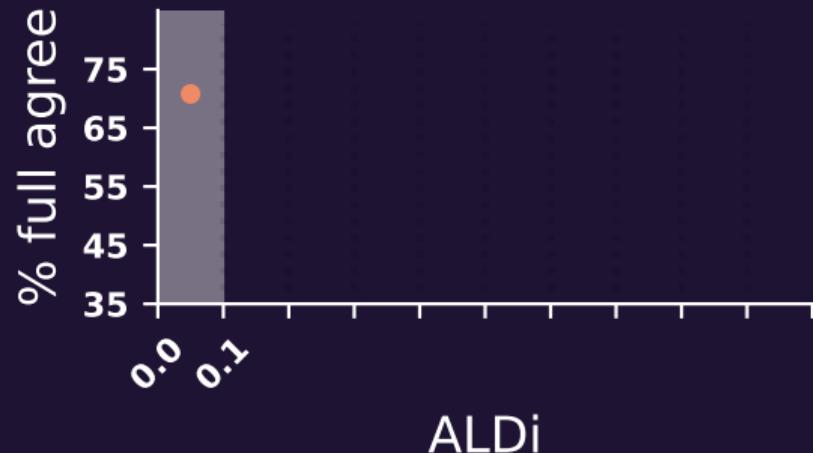
Methodology:

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



Labels:

Positive/ Neutral/ Negative



Sentiment Analysis Dataset



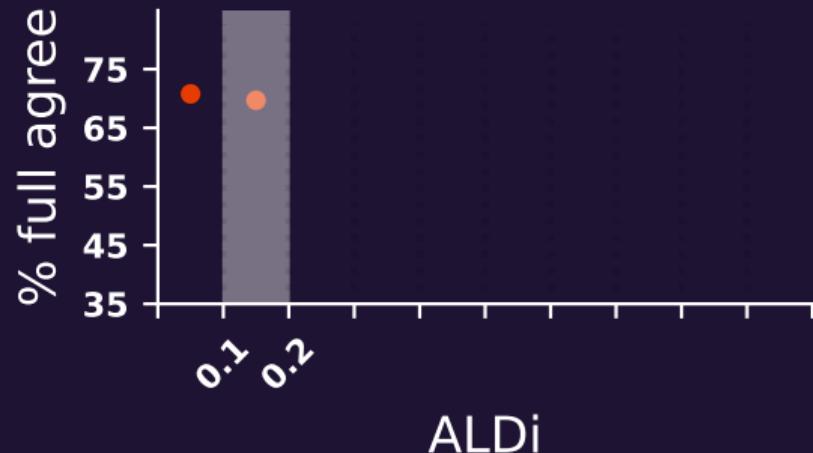
Methodology:

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



Labels:

Positive/ Neutral/ Negative



Sentiment Analysis Dataset



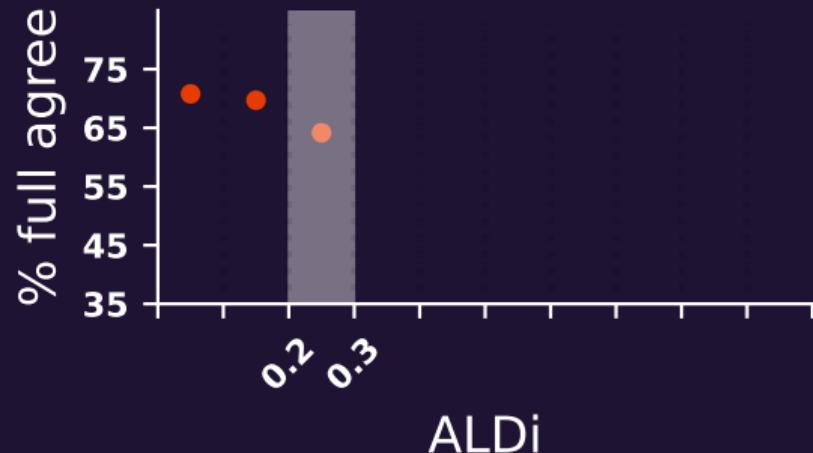
Methodology:

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



Labels:

Positive/ Neutral/ Negative



Sentiment Analysis Dataset



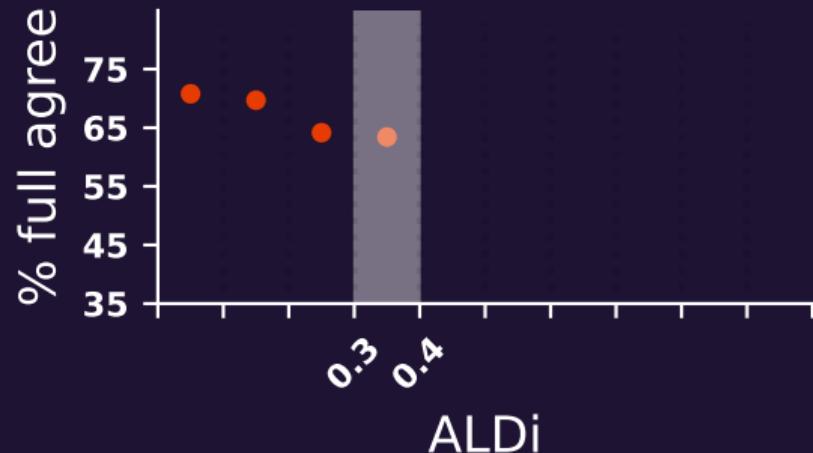
Methodology:

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



Labels:

Positive/ Neutral/ Negative



Sentiment Analysis Dataset



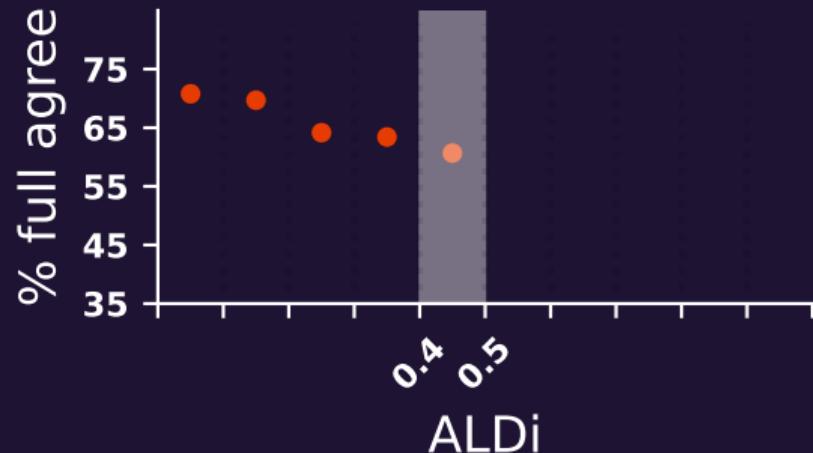
Methodology:

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



Labels:

Positive/ Neutral/ Negative



Sentiment Analysis Dataset



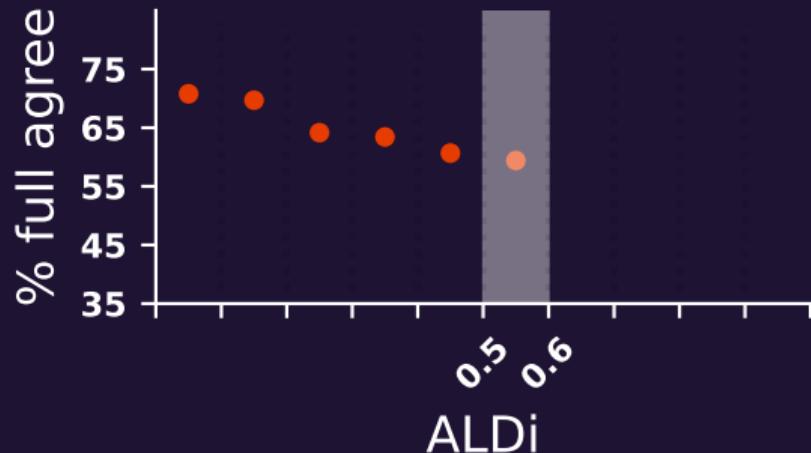
Methodology:

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



Labels:

Positive/ Neutral/ Negative



Sentiment Analysis Dataset



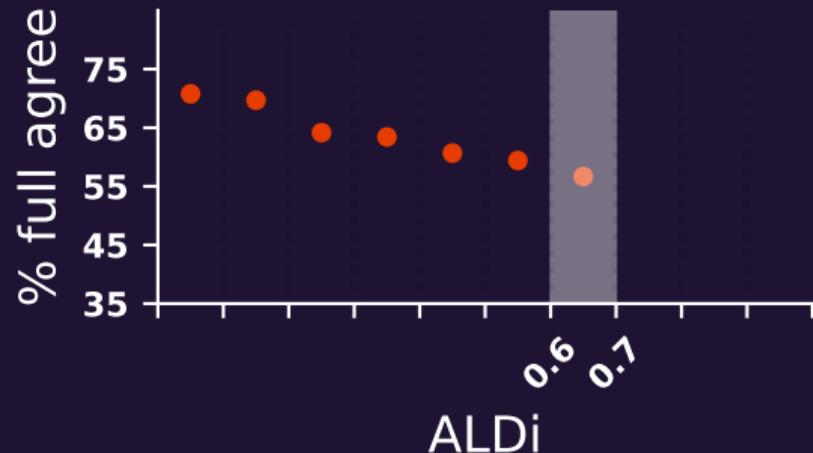
Methodology:

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



Labels:

Positive/ Neutral/ Negative



Sentiment Analysis Dataset



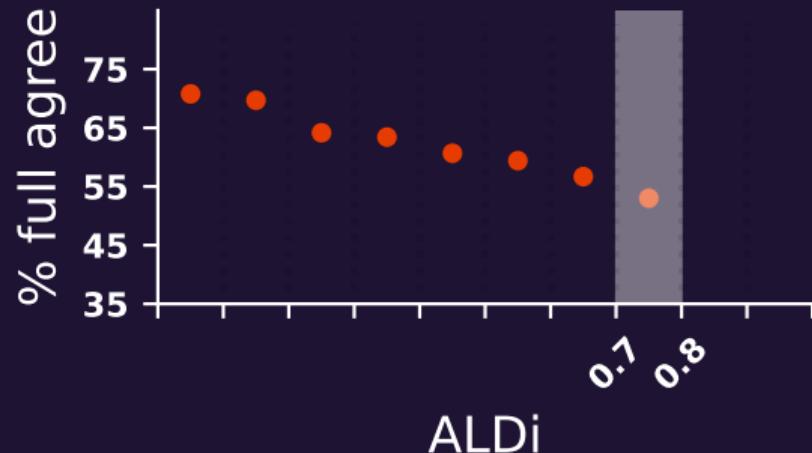
Methodology:

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



Labels:

Positive/ Neutral/ Negative



Sentiment Analysis Dataset



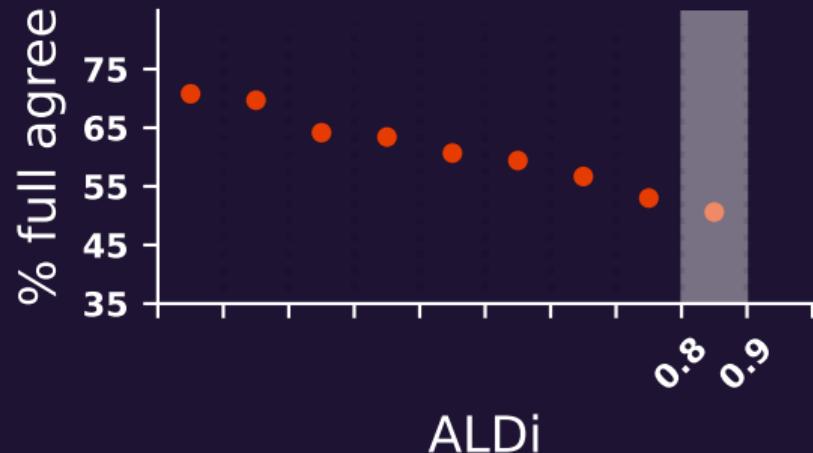
Methodology:

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



Labels:

Positive/ Neutral/ Negative



Sentiment Analysis Dataset



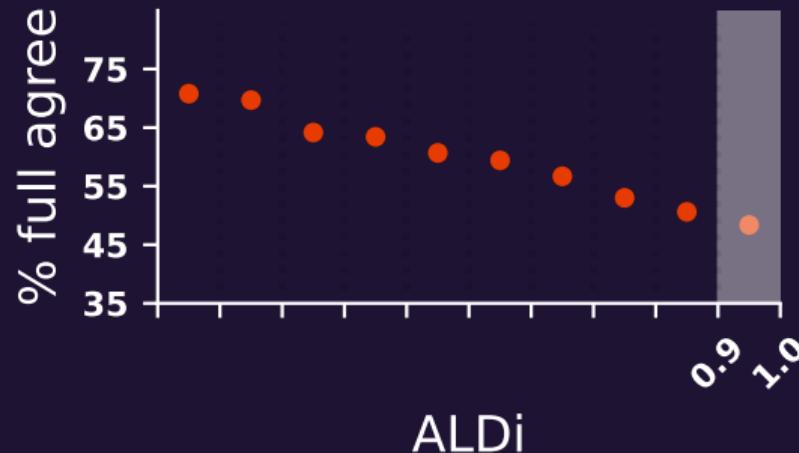
Methodology:

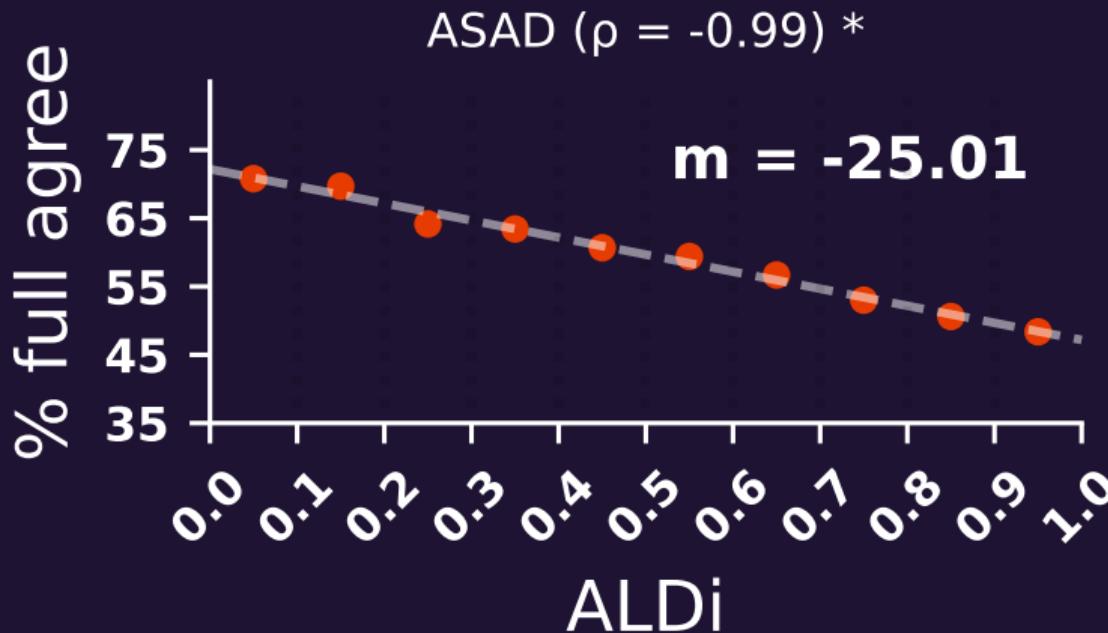
- ① Estimate ALDi of samples.
- ② Bin samples.
- ③ Estimate % samples Full Agreement.



Labels:

Positive/ Neutral/ Negative





IAA - Dialect Identification Dataset

 **Labels (Macro-regional):**

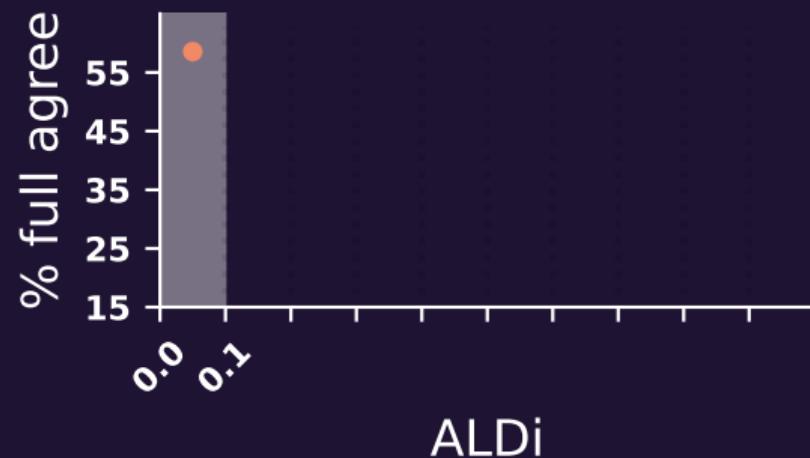
MSA, Maghreb, Egypt, Levant, Gulf



IAA - Dialect Identification Dataset

i Labels (Macro-regional):

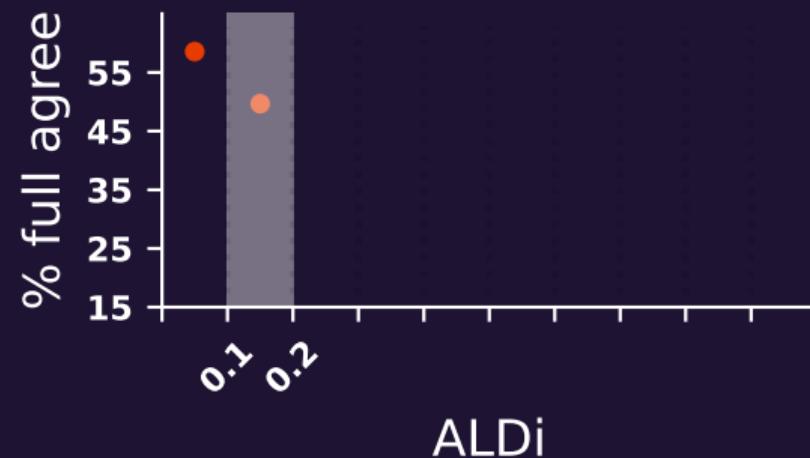
MSA, Egypt, Gulf, Levant, Maghreb



IAA - Dialect Identification Dataset

i Labels (Macro-regional):

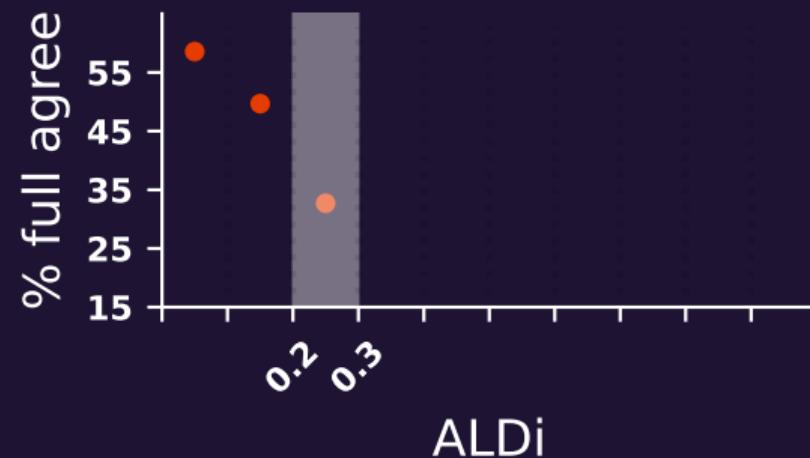
MSA, Egypt, Gulf, Levant, Maghreb



IAA - Dialect Identification Dataset

i Labels (Macro-regional):

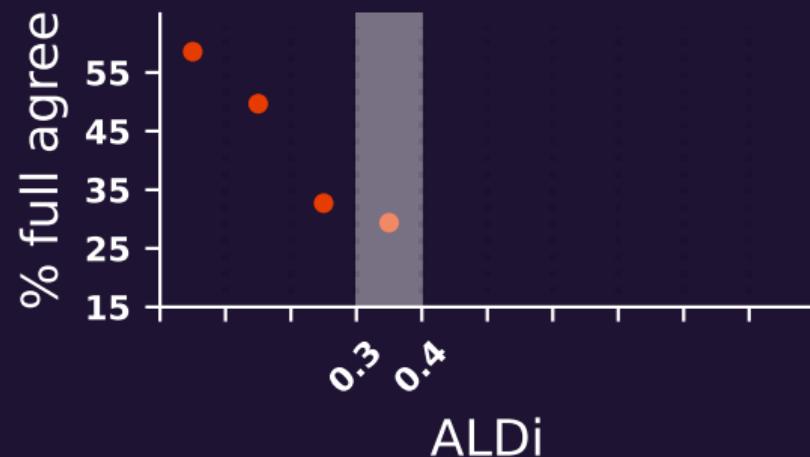
MSA, Egypt, Gulf, Levant, Maghreb



IAA - Dialect Identification Dataset

i Labels (Macro-regional):

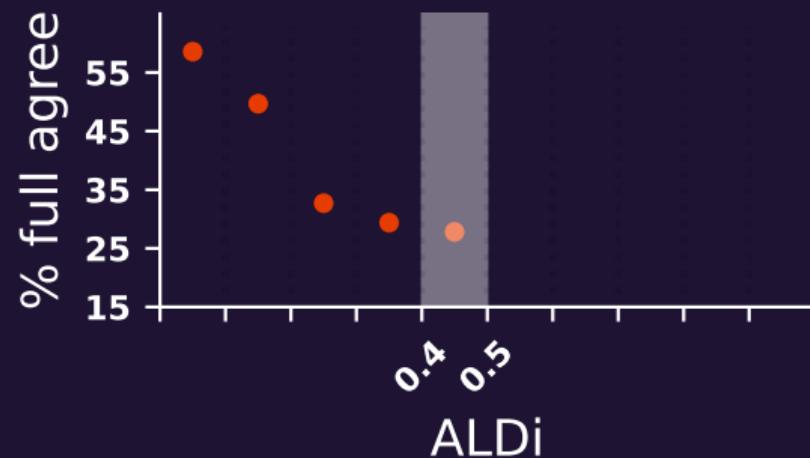
MSA, Egypt, Gulf, Levant, Maghreb



IAA - Dialect Identification Dataset

i Labels (Macro-regional):

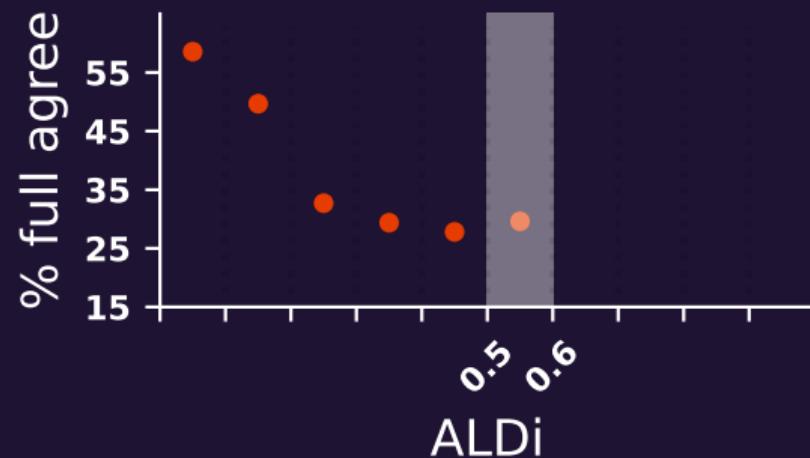
MSA, Egypt, Gulf, Levant, Maghreb



IAA - Dialect Identification Dataset

i Labels (Macro-regional):

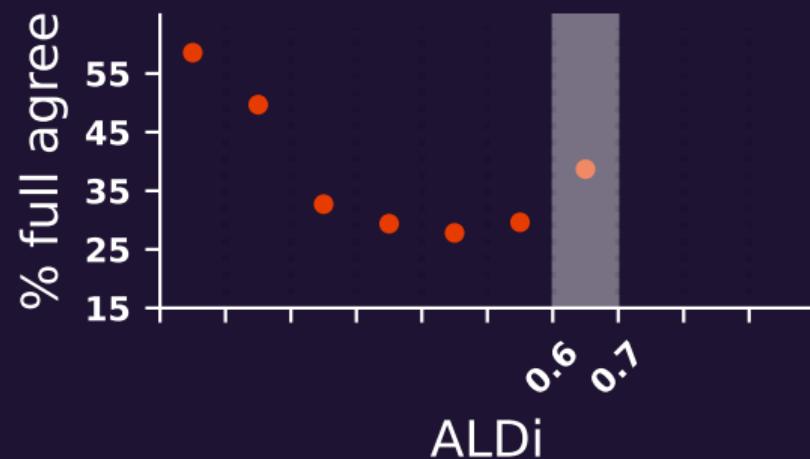
MSA, Egypt, Gulf, Levant, Maghreb



IAA - Dialect Identification Dataset

i Labels (Macro-regional):

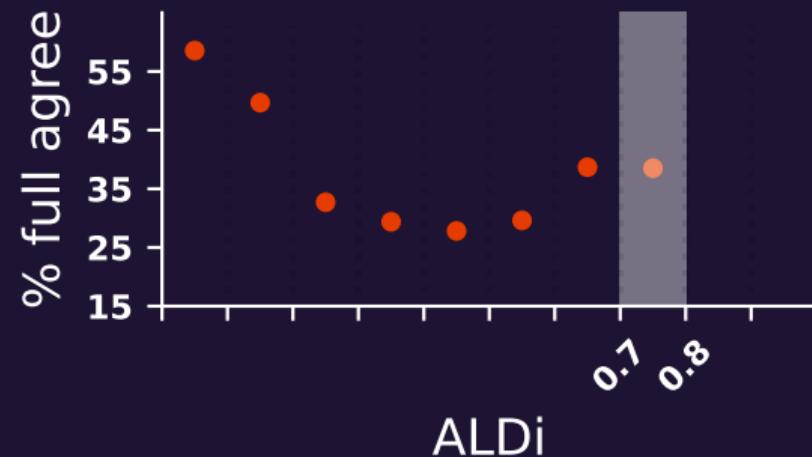
MSA, Egypt, Gulf, Levant, Maghreb



IAA - Dialect Identification Dataset

i Labels (Macro-regional):

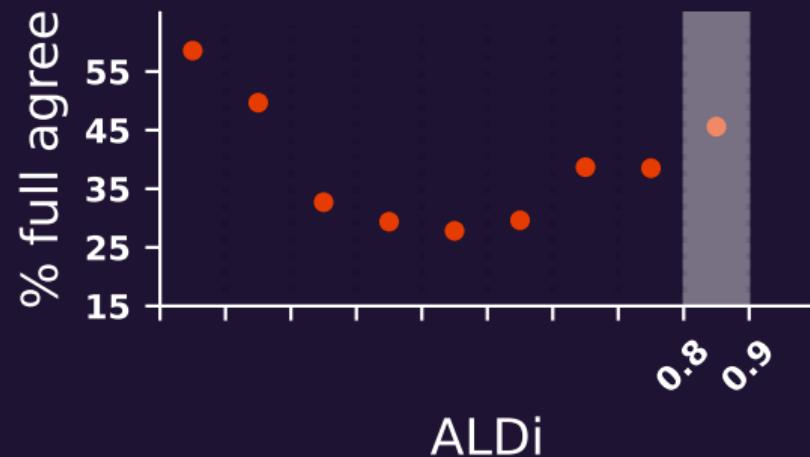
MSA, Egypt, Gulf, Levant, Maghreb



IAA - Dialect Identification Dataset

i Labels (Macro-regional):

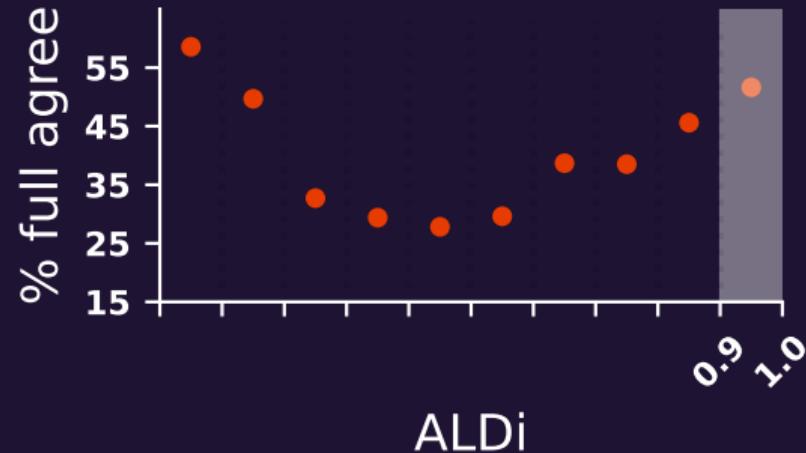
MSA, Egypt, Gulf, Levant, Maghreb

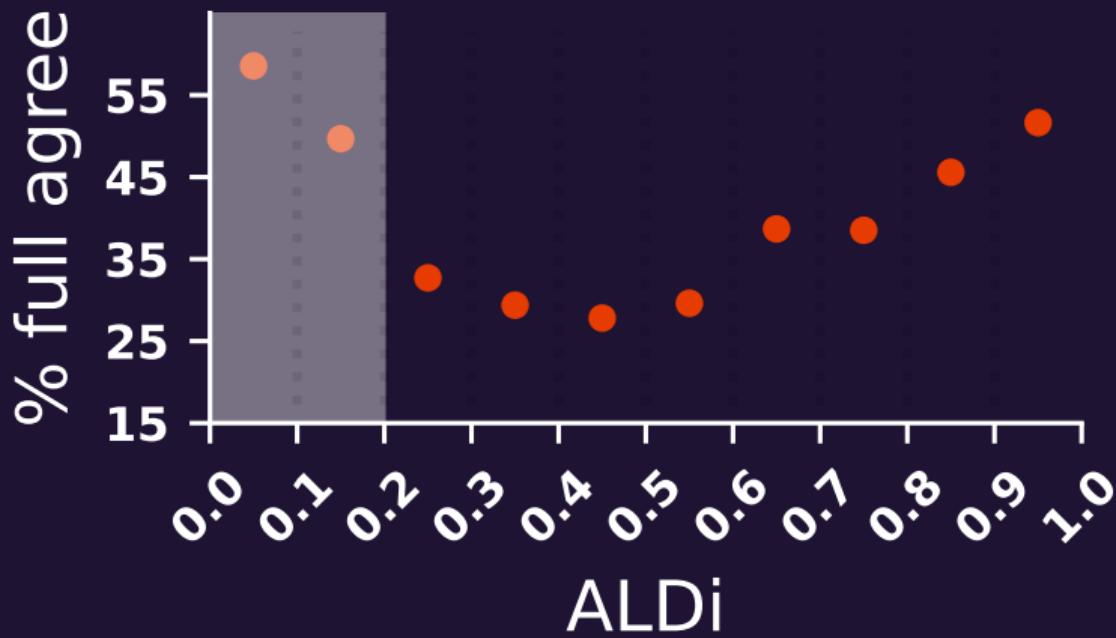


IAA - Dialect Identification Dataset

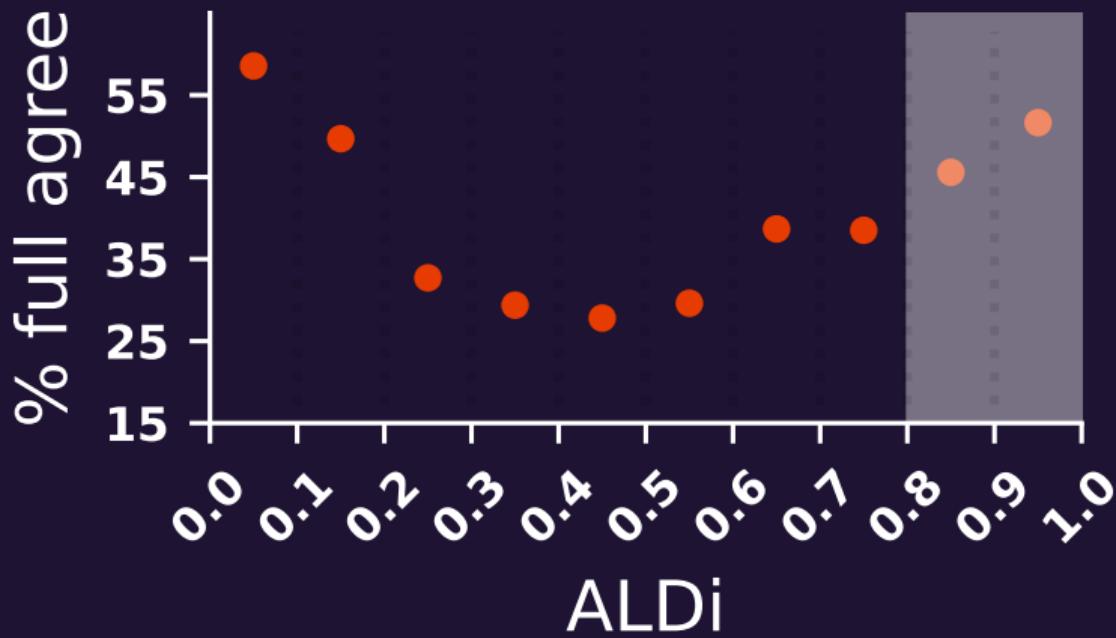
i Labels (Macro-regional):

MSA, Egypt, Gulf, Levant, Maghreb

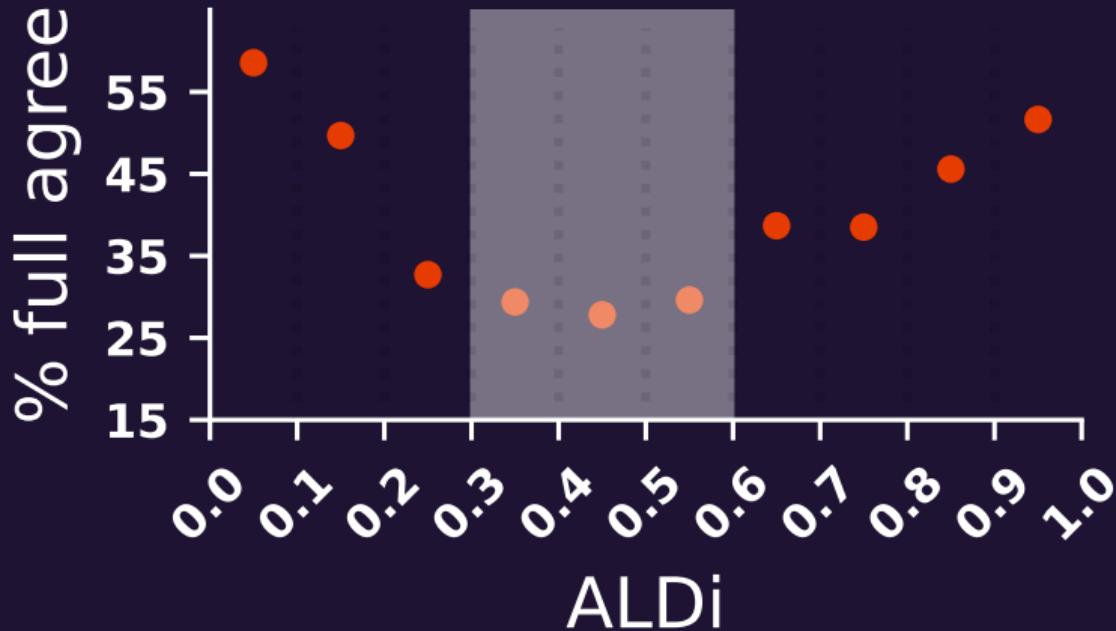




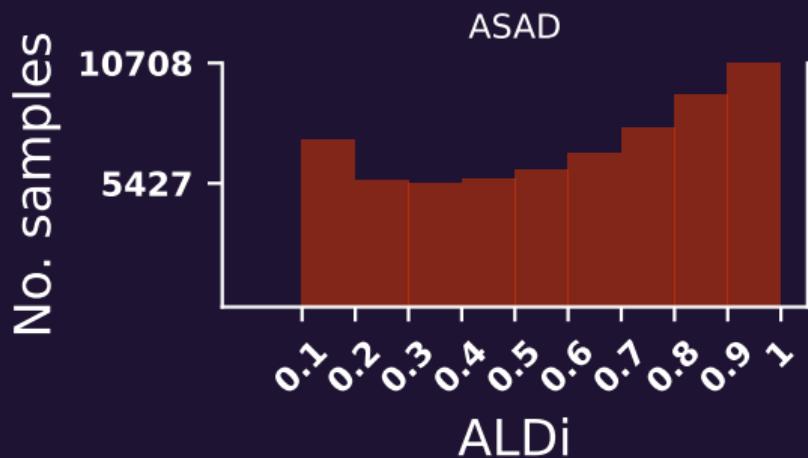
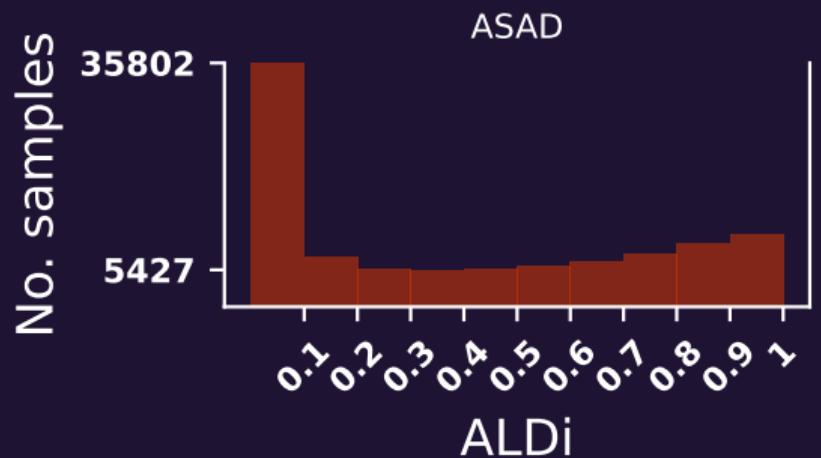
Easily identifiable MSA samples



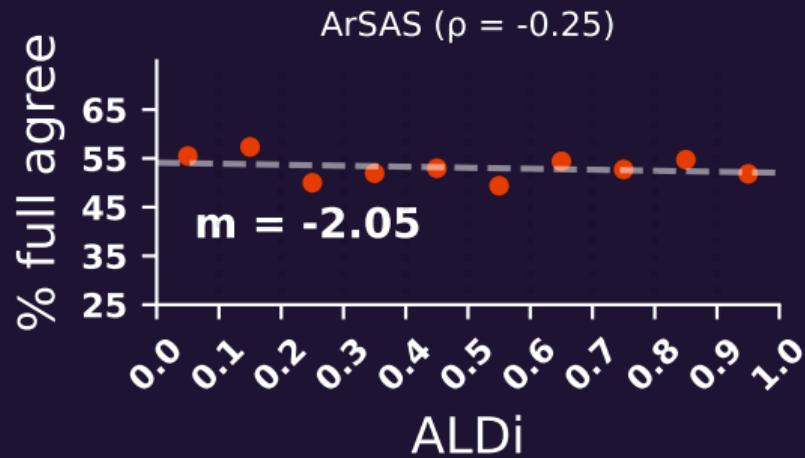
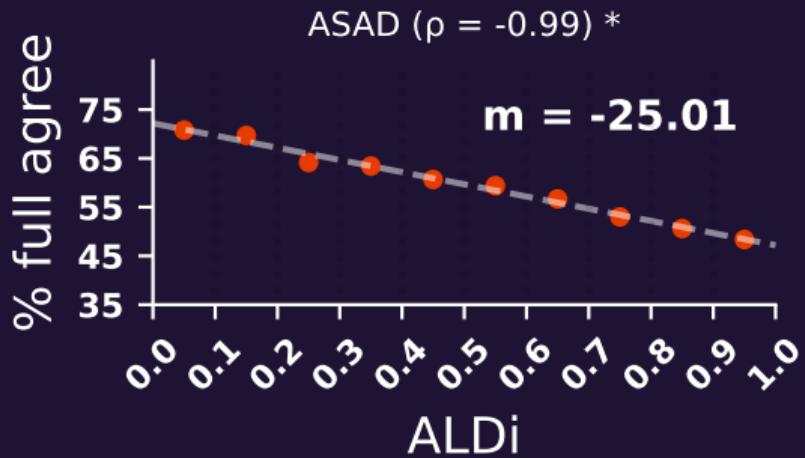
DA samples with multiple distinctive cues of a dialect

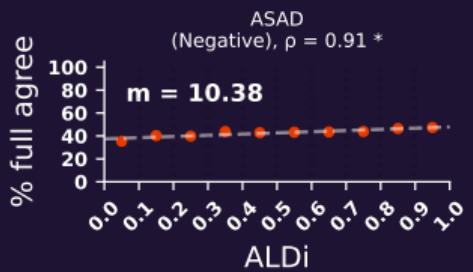
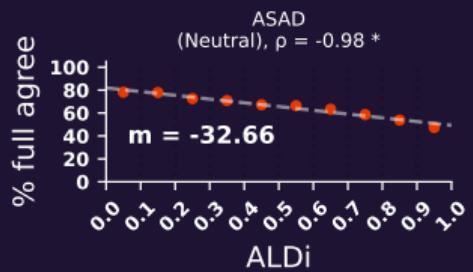
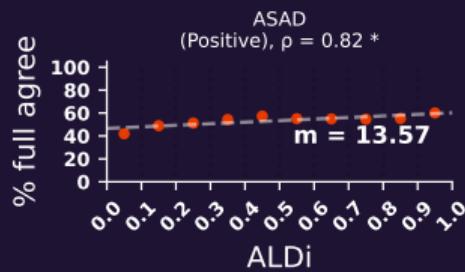
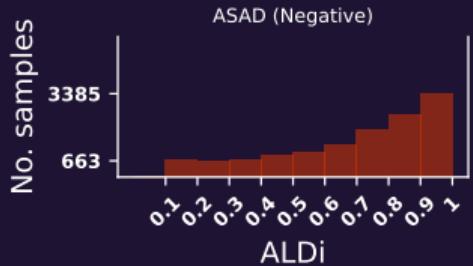
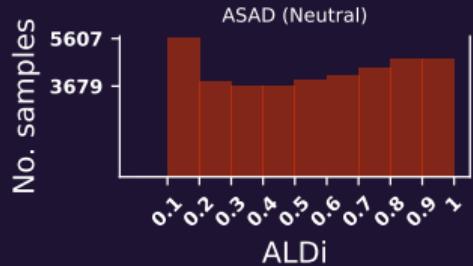
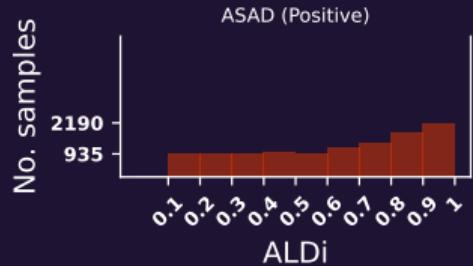


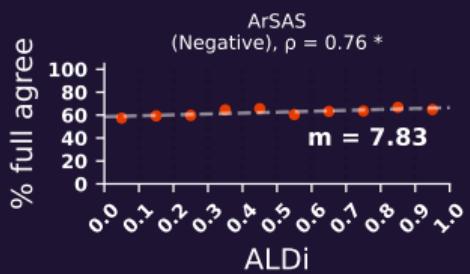
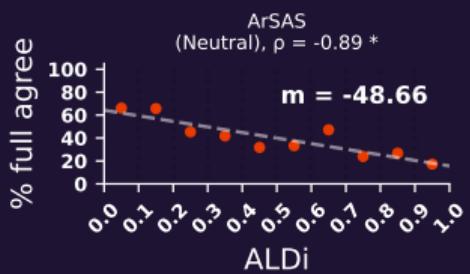
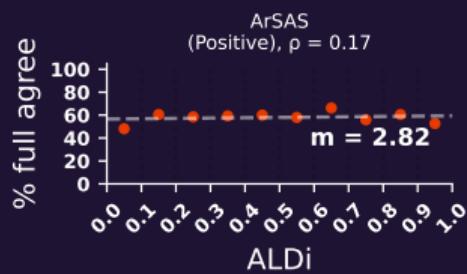
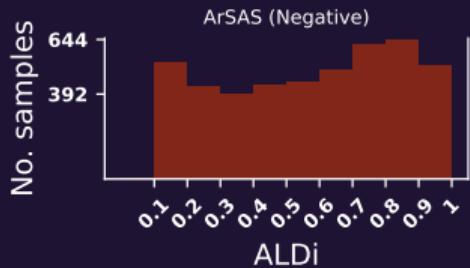
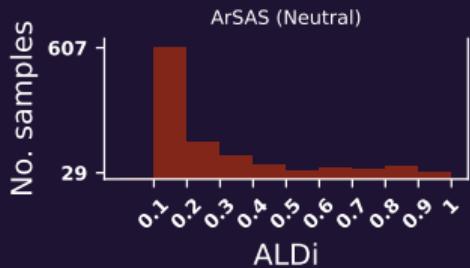
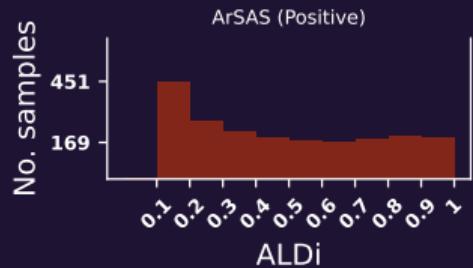
- 📎 1) Hard to determine the dialect?
- OR
- 📎 2) Valid in multiple dialects?



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 (فصحي فقط)
 - A bit of dialect (القليل من العامية)
 - Mixed (خلط من الفصحي 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 !	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	█████ 5000 √DONE	Cmnt (Y7)	¬ Arabic (x3)
	د. أشرف بلعج	Cntrl (Y7)	¬ Arabic (x3)