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

Amr Keleg
PhD Student at the University of Edinburgh
Under the Supervision of: Walid Magdy, Sharon Goldwater

CAMeL

31 October 2024

NLP UKRI CENTRE FOR DOCTORAL TRAINING
ilcc 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: The Fifth Nuanced Arabic Dialect Identification Shared Task** (Abdul-Mageed et al., ArabicNLP-WS 2024)

2. **... nor the solution (Spoiler: Arabic Level of Dialectness)**
   - **ALDi: Quantifying the Arabic Level of Dialectness of Text** (Keleg et al., EMNLP 2023)
   - **Estimating the Level of Dialectness Predicts Inter-annotator Agreement in Multi-dialect Arabic Datasets** (Keleg et al., ACL 2024) - Outstanding Paper Award - ACL 2024

Why distinguish between varieties of Arabic?

Raw Pretraining corpus

Sentence
أمطار خفيفة على منطقة مكة المكرمة
...
طيب مافي حلقات زيادة؟ ما شبعنا والله

Why distinguish between varieties of Arabic?

Raw Pretraining corpus

Sentence
أمطار خفيفة على منطقة مكة المكرمة
...
طيب مافي حلقات زيادة؟ ما شبعنا والله

a) Representation of dialects?

Why distinguish between varieties of Arabic?

Raw Pretraining corpus

Sentence
أمطار خفيفة على منطقة مكة المكرمة
...
طيب مافي حلقات زيادة؟ ما شبعنا والله

Dataset of Tweets

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

a) Representation of dialects?

Why distinguish between varieties of Arabic?

Raw Pretraining corpus

Sentence
أمطار خفيفة على منطقة مكة المكرمة
...
طيب مافي حلقات زيادة؟ ما شبعنا والله

Dataset of Tweets

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

a) Representation of dialects?

For Annotation:

b) Routing samples.

For Modeling:

c) per-variety performance?

🤔 Can Dialect Identification help?

الزلمة أسعدنا	الراجل أسعدنا	أسعدنا الرجل
الرِجّال أسعدنا	الزول أسعدنا	

🤔 Can Dialect Identification help?

الزلمة أسعدنا	الراجل أسعدنا	أسعدنا الرجل
الرِجّال أسعدنا	الزول أسعدنا	

Goal: Automatically identify **A DIALECT** for each sentence.

Status: quite popular among the Arabic NLP community ⭐

Level of granularity #1

MSA (Fus-ha)
- shared across countries

Regional dialects

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.”

Level of granularity #2

MSA (Fus-ha)
- shared across countries

Country-level dialects
- generally targeting at least 18 labels

Investigation of country-level single-label DI

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

- Speakers from 7 countries validated 490 errors.
- Only **33**% of validated mispredictions are **true errors!**
  - i.e., 67% of them are multi-dialect samples.

- ⚠️ Inaccurate Evaluation!
- ❓ How common are these samples?

Building the First Multilabel ADI Dataset (NADI 2024)

🗞️ **NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task** (Abdul-Mageed et al., ArabicNLP 2024)

Guidelines + Shared Task Summary

Is it possible that the tweet is authored by someone who speaks one of your country's dialects?

- 1,120 sentences.
- 3 annotators from 9 different countries (total of 27)

Guidelines + Shared Task Summary

Is it possible that the tweet is authored by someone who speaks one of your country's dialects?

- 1,120 sentences.
- 3 annotators from 9 different countries (total of 27)

Sentence	Valid in
وين يلعب هذا ما شفته	Algeria🇩🇿, Palestine🇵🇸, Yemen🇾🇪

Guidelines + Shared Task Summary

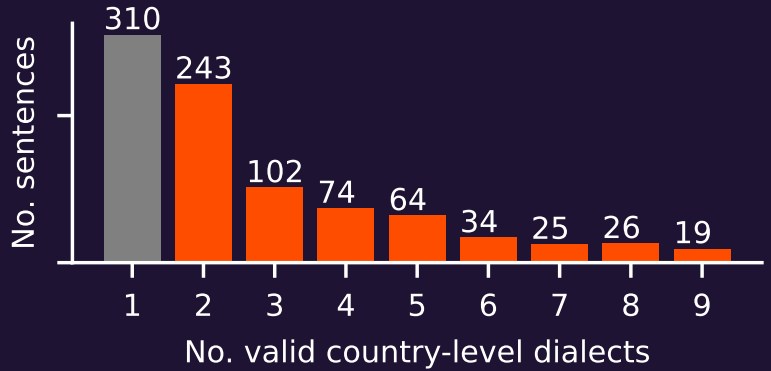
Is it possible that the tweet is authored by someone who speaks one of your country's dialects?

- 1,120 sentences.
- 3 annotators from 9 different countries (total of 27)

Sentence	Valid in
وين يلعب هذا ما شفته	Algeria🇩🇿, Palestine🇵🇸, Yemen🇾🇪

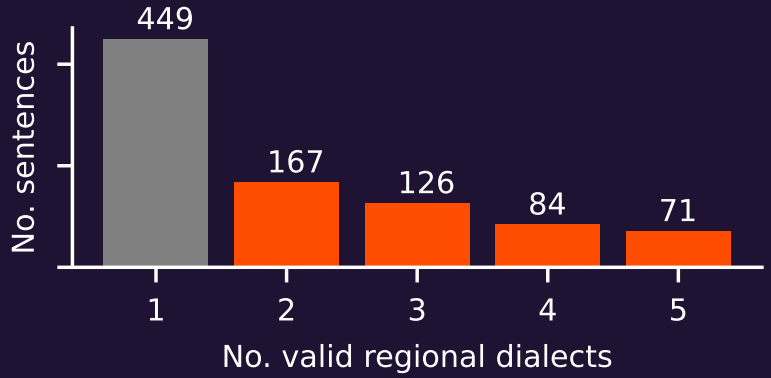
- F1-score$_{\text{macro}}$ 50.57±7.1 (Still not solved).
- Being hosted as a public leaderboard.

Multilabel samples in NADI 2024?



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

Multilabel samples in NADI 2024?



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

Dialect Identification is not the Solution

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

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

Different ways of saying: I'm happy 😀

أنا فَرِح
farih ʔna

أنا مبسوط
mbsoT ʔna

أنا مشهيص
mʃhys ʔna

Arabic Level of Dialectness (ALDi)

EMNLP 2023

Keleg, Amr, Goldwater, Sharon, and Magdy, Walid. EMNLP 2023. "ALDi: Quantifying the Arabic Level of Dialectness of Text."

Different ways of saying: I'm happy 😀

MSA root meaning to be happy

Root

فرح

frh

🗣️🇪🇬

أنا فَرِح

farih ʔna

أنا مبسوط

mbsoT ʔna

أنا مشهيص

mʃhys ʔna

Arabic Level of Dialectness (ALDi)

EMNLP 2023

Keleg, Amr, Goldwater, Sharon, and Magdy, Walid. EMNLP 2023. "ALDi: Quantifying the Arabic Level of Dialectness of Text."

Different ways of saying: I'm happy 😀

MSA root meaning	Root	🗣️🇪🇬
to be happy	فرح	أنا فَرِح
	frh	farih ʔna
extend - cheer	بسط	أنا مبسوط
	bsT	mbsoT ʔna
N/A	شهيص	أنا مشهيص
	ʃhys	mʃhys ʔna

Arabic Level of Dialectness (ALDi)

EMNLP 2023

Keleg, Amr, Goldwater, Sharon, and Magdy, Walid. EMNLP 2023. "ALDi: Quantifying the Arabic Level of Dialectness of Text."

- **ALDi**: Divergence from Standard Arabic (MSA).
- Continuous score in [0, 1].
- Sentence-like level

Keleg, Amr, Goldwater, Sharon, and Magdy, Walid. EMNLP 2023. "ALDi: Quantifying the Arabic Level of Dialectness of Text."

Sentence-ALDi model

Sentence-ALDi model

$\text{RMSE}(\text{test set}) = 0.18$

Sentence-ALDi model

🔬 $\mathrm{RMSE}(\text{test set}) = 0.18$

🌍 Dialect-agnostic

Applications of ALDi

1. Studying Intraspeaker Variation (Presidential Speeches)

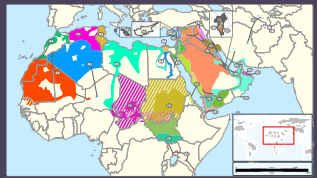
2) Annotating Multi-Dialect Arabic Datasets

Common Practice: 🎲 randomly assign to Arabic speakers

☢☢☢ **Annotator's dialect ≠ Sample's dialect** ☢☢☢

📎 More strict annotating Hate Speech 🤬
(Bergman and Diab, 2022)

📎 Less accurate identifying Sarcasm 🤪
(Abu Farha and Magdy, 2022)



Bergman, A. and Diab, Mona. ACL (findings) 2022. "Towards Responsible Natural Language Annotation for the Varieties of Arabic."

Abu Farha, Ibrahim and Magdy, Walid. WANLP 2022. "The Effect of Arabic Dialect Familiarity on Data Annotation."

Annotation Codebook (v1.0) 📚

🏷️ **Step 1:** Identify the dialect of each sample

🔀 **Step 2:** Route the sample to speakers of its dialect

Mubarak, Hamdy and Darwish, Kareem. 2016. “Demographic surveys of Arab annotators on CrowdFlower.”

Annotation Codebook (v1.0) 📚

🏷️ **Step 1:** Identify the dialect of each sample

🔀 **Step 2:** Route the sample to speakers of its dialect

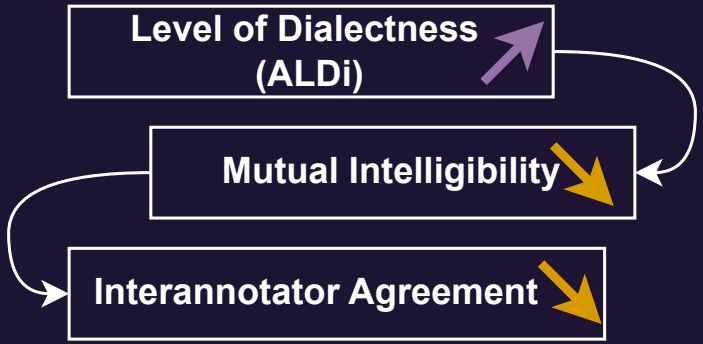
Hard to crowdsource speakers of some dialects (i.e., Limited resource 💎💎)

(Mubarak and Darwish, 2016)

Mubarak, Hamdy and Darwish, Kareem. 2016. “Demographic surveys of Arab annotators on CrowdFlower.”

🤔 Should some dialectal samples be priotrized?

Intuition


Level of Dialectness (ALDi)
Mutual Intelligibility
Interannotator Agreement

🧪 Analysis

📄 15 public datasets covering 6 Tasks:
Hatespeech, Sentiment Analysis, Dialect Identifcation, ...

(1) sentence-level classification datasets
(2) multi-dialect samples
(3) **samples randomly assigned to annotators**

(4) **individual annotator labels**

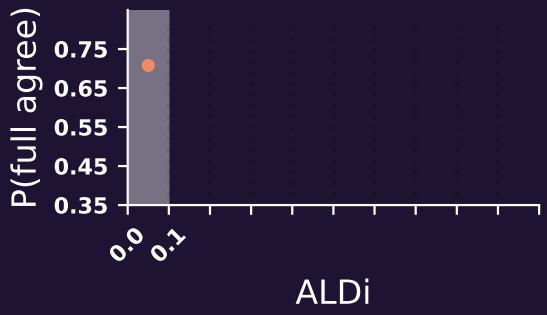
🧪 Methodology:

1. Estimate ALDi of samples.
2. Bin samples.
3. $\mathrm{P}_{\mathrm{bin}}(\text{Full Agreement})$

$$\mathrm{P}_{\mathrm{bin}}(\text{Full Agreement}) \approx \frac{\mathrm{N}_{(\mathrm{bin})\ \text{Full Agreement}}}{\mathrm{N}_{(\mathrm{bin})\ \text{Total Samples}}}$$

🧪 Methodology:

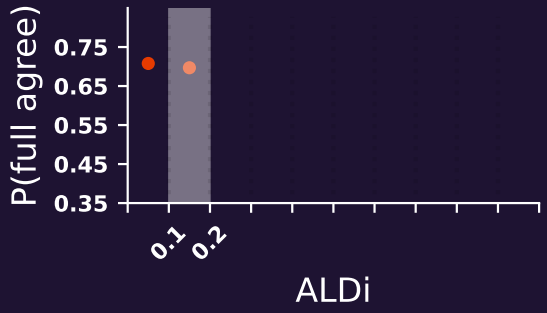
1. Estimate ALDi of samples.
2. Bin samples.
3. $\mathrm{P}_{\mathrm{bin}}(\text{Full Agreement})$



$$\mathrm{P}_{\mathrm{bin}}(\text{Full Agreement}) \approx \frac{\mathrm{N}_{(\mathrm{bin})\ \text{Full Agreement}}}{\mathrm{N}_{(\mathrm{bin})\ \text{Total Samples}}}$$

🧪 Methodology:

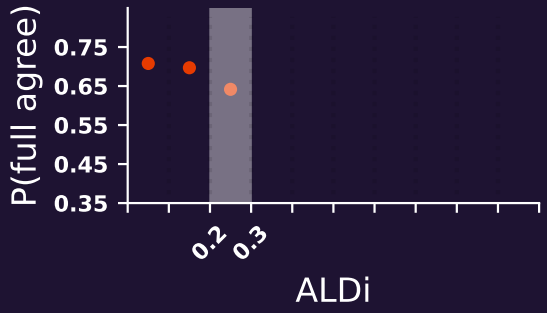
1. Estimate ALDi of samples.
2. Bin samples.
3. $\mathrm{P}_{\mathrm{bin}}(\text{Full Agreement})$



$$\mathrm{P}_{\mathrm{bin}}(\text{Full Agreement}) \approx \frac{\mathrm{N}_{(\mathrm{bin})\ \text{Full Agreement}}}{\mathrm{N}_{(\mathrm{bin})\ \text{Total Samples}}}$$

🧪 Methodology:

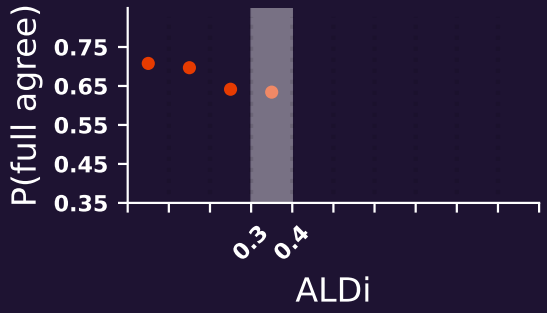
1. Estimate ALDi of samples.
2. Bin samples.
3. $\mathrm{P_{bin}}(\text{Full Agreement})$



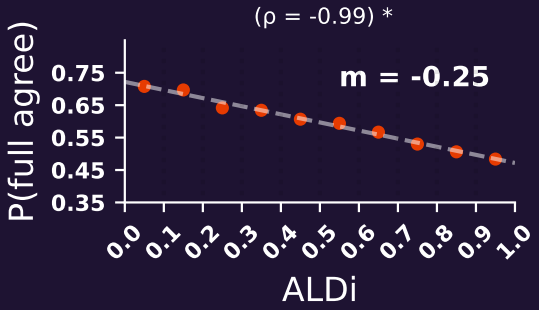
$$\mathrm{P_{bin}}(\text{Full Agreement}) \approx \frac{\mathrm{N_{(bin)\ Full\ Agreement}}}{\mathrm{N_{(bin)\ Total\ Samples}}}$$

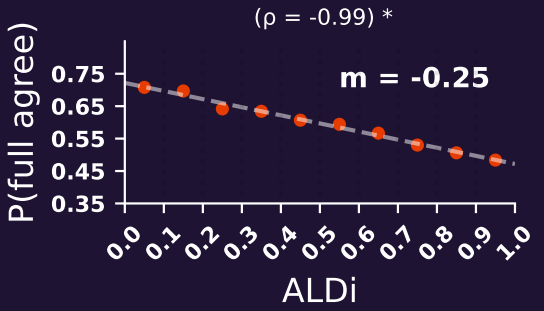
🧪 Methodology:

1. Estimate ALDi of samples.
2. Bin samples.
3. $\mathrm{P_{bin}}(\text{Full Agreement})$



$$\mathrm{P_{bin}}(\text{Full Agreement}) \approx \frac{\mathrm{N_{(bin)\ Full\ Agreement}}}{\mathrm{N_{(bin)\ Total\ Samples}}}$$


(ρ = -0.99) *
m = -0.25
P(full agree)
0.75
0.65
0.55
0.45
0.35
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
ALDi



Finding (1) - For 8 of 12 non Dialect Identification datasets

ALDi ↗ **Interannotator Agreement** ↘

with significant strong negative $\rho < -0.7$

Annotation Codebook (v1.1) 📙

🚨 Prioritize routing high-ALDi samples to speakers of the samples' respective dialects, (Finding 1)
🕵️ for which Dialect Identification is more accurate. (Finding 2)

Collecting ALDi Annotations

1) AOC Dataset

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

- 📰 Comments to news articles
- 🎯 127,835 sentences (3👤annotations each)
- 🌟 Popular Dialect Identification (DI) labels.

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

1) AOC Dataset

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

- 📰 Comments to news articles
- 🎯 127,835 sentences (3👤annotations each)
- 🌟 Popular Dialect Identification (DI) labels.
- 🗑️ 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."

Tell us how much dialect (عامية) is in the sentence.

Tell us how much dialect (عامية) is in the sentence.

🤝 Fleiss' κ = 0.44

Tell us how much dialect (عامية) is in the sentence.

🤝 Fleiss' κ = 0.44

🤔 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 developing schools and providing observers over them

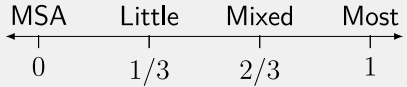
Sentence with two valid pronunciations

نبتدى بقى الشغل الصح فى تطوير المدارس وتوفير المراقبين عليها

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

♻♻♻ From AOC to AOC-ALDi ♻♻♻

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



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

🤝 Krippendorff's α (interval) = 0.63
🎯 Sentence-ALDi's RMSE(AOC-ALDi$_{\text{test}}$) = 0.18

Reflections on the AOC ratings

⚠ Annotators randomly assigned to sentences.
⚠ Underspecified guidelines.

2) NADI 2024 Dataset

IFF an annotator labels the tweet as written in one of their country-level dialects.

Please evaluate the Level of Dialectness of each tweet as:

L0 Sound MSA
L1 Formal Colloquial or Colloquial-influenced MSA
L2 Natural/Ordinary Colloquial
L3 Informal (or Vulgar) Colloquial

Note: The levels and their descriptions were provided in Arabic.

Country	N valid	Krip. α
Algeria	333	0.66
Morocco	230	0.74
Tunisia	189	0.75
Egypt	353	0.82
Sudan	393	0.66
Palestine	375	0.68
Syria	475	0.79
Iraq	271	0.73
Yemen	454	0.50

✅ Improved alpha scores than AOC-ALDi.

Reflections 🤔

ALDi: Divergence from Modern Standard Arabic (MSA).

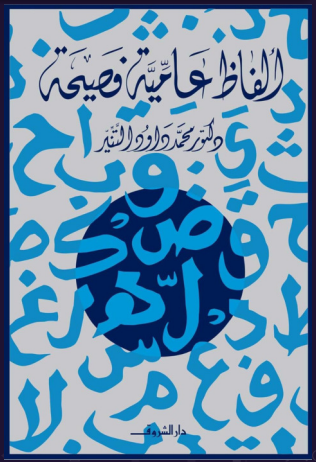
1. Differences in ALDi Ratings across Countries
   e.g., بسمتك يا زين تسوي الف بسمه
   (O' Zain, your smile is worth thousand smiles)
   - 🇩🇿 L1 - Formal Colloquial
   - 🇾🇪 L2 - Natural/Ordinary Colloquial

Country-level ALDi scores VS Single-aggregated ALDi score?

Reflections 🤔

ALDi: Divergence from Standard Arabic.

2. Perception of Standard Arabic might not be linguistically accurate.
   - Two words for Wine (خمر, خمرة)
   - Both are grammatical in Standard Arabic.
   - خمرة is perceived as Dialectal in Egypt.



References:

📖 ألفاظ عامية فصيحة (Mohamed Dawoud El-tanir, 2017)

🗞️ مفرداث عامية عربية فصيحة أهملت العربية الفصحى الحديثة معظمها (Ben-Zarrouk Hussein, 2021)

Thanks!

@Amrkeleg
a.keleg@sms.ed.ac.uk

Thanks!

𝕏 @Amrkeleg
a.keleg@sms.ed.ac.uk

Summary

1. Multi-label setup is more realistic for ADI.
   - ⚠️ Not solved yet.
2. Arabic sentences have different levels of dialectness (ALDi) i.e., not just MSA or a dialect.
   - 🤖 Automatically estimated by Sentence-ALDi model.
   - 🌱 aids Content Moderation, and Analysis of different styles.
     - Demo on 🤗: `huggingface.co/spaces/AMR-KELEG/ALDi`

References I

- Abu Farha, Ibrahim and Walid Magdy (Dec. WANLP 2022). "The Effect of Arabic Dialect Familiarity on Data Annotation." In: *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*. Ed. by Houda Bouamor et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 399–408. DOI: 10.18653/v1/2022.wanlp-1.39. URL: https://aclanthology.org/2022.wanlp-1.39.

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

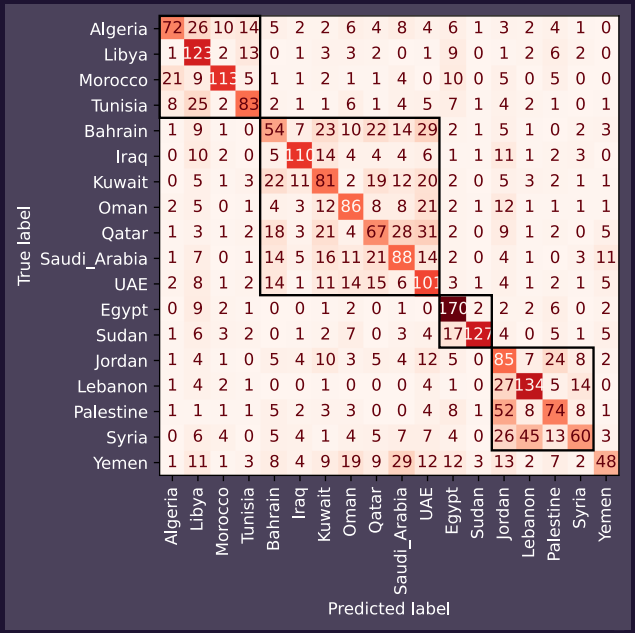
- Bergman, A. and Mona Diab (May ACL (findings) 2022). "Towards Responsible Natural Language Annotation for the Varieties of Arabic." In: *Findings of the Association for Computational Linguistics: ACL 2022*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 364–371. DOI: 10.18653/v1/2022.findings-acl.31. URL: https://aclanthology.org/2022.findings-acl.31.

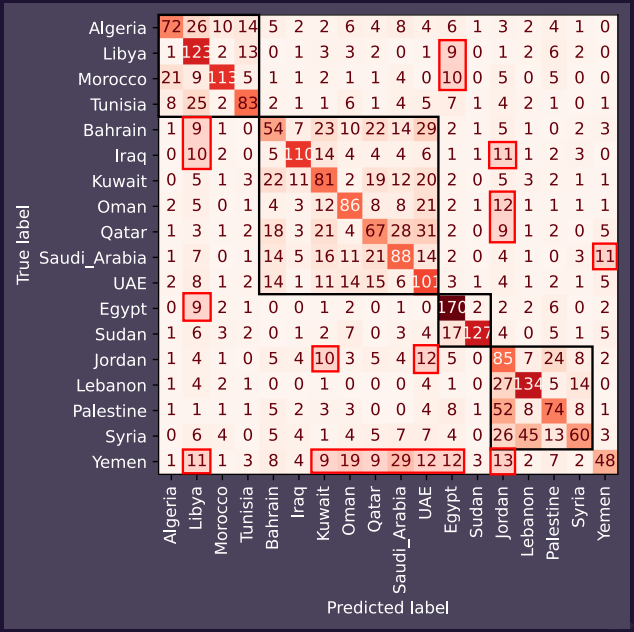
References V

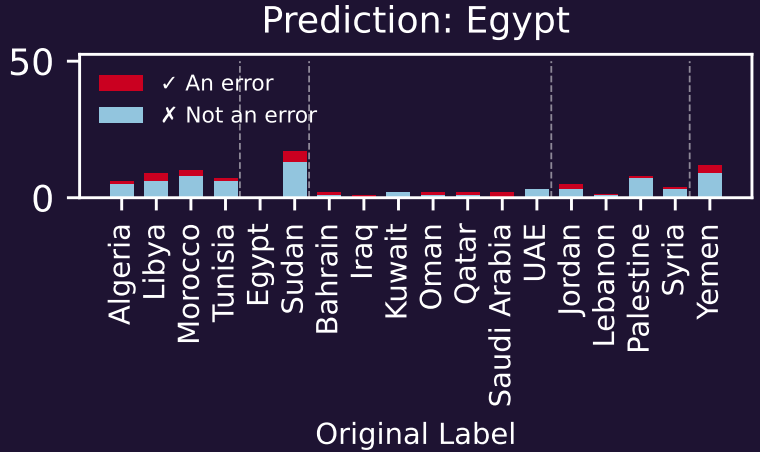
- Keleg, Amr, Sharon Goldwater, and Walid Magdy (Dec. EMNLP 2023). “ALDi: Quantifying the Arabic Level of Dialectness of Text.” In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 10597–10611. DOI: 10.18653/v1/2023.emnlp-main.655. URL: https://aclanthology.org/2023.emnlp-main.655.

References VI

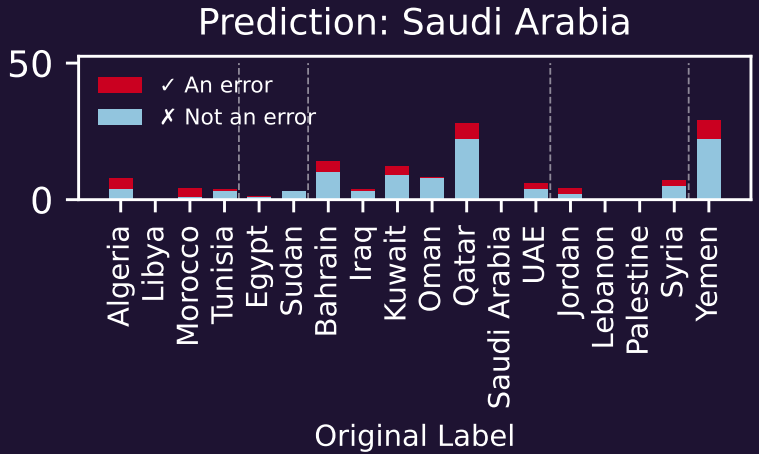
- Mubarak, Hamdy and Kareem Darwish (2016). “Demographic surveys of Arab annotators on CrowdFlower.” In: *Proceedings of ACM WebSci16 Workshop “Weaving Relations of Trust in Crowd Work: Transparency and Reputation across Platforms*.
- Zaidan, Omar F. and Chris Callison-Burch (June 2011). “The Arabic 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.


True label
Predicted label
Algeria 72 26 10 14 5 2 2 6 4 8 4 6 1 3 2 4 1 0
Libya 1 123 2 13 0 1 3 3 2 0 1 9 0 1 2 6 2 0
Morocco 21 9 113 5 1 1 2 1 1 4 0 10 0 5 0 5 0 0
Tunisia 8 25 2 83 2 1 1 6 1 4 5 7 1 4 2 1 0 1
Bahrain 1 9 1 0 54 7 23 10 22 14 29 2 1 5 1 0 2 3
Iraq 0 10 2 0 5 110 14 4 4 4 6 1 1 11 1 2 3 0
Kuwait 0 5 1 3 22 11 81 2 19 12 20 2 0 5 3 2 1 1
Oman 2 5 0 1 4 3 12 86 8 8 21 2 1 12 1 1 1 1
Qatar 1 3 1 2 18 3 21 4 67 28 31 2 0 9 1 2 0 5
Saudi_Arabia 1 7 0 1 14 5 16 11 21 88 14 2 0 4 1 0 3 11
UAE 2 8 1 2 14 1 11 14 15 6 101 3 1 4 1 2 1 5
Egypt 0 9 2 1 0 0 1 2 0 1 0 170 2 2 2 6 0 2
Sudan 1 6 3 2 0 1 2 7 0 3 4 17 127 4 0 5 1 5
Jordan 1 4 1 0 5 4 10 3 5 4 12 5 0 85 7 24 8 2
Lebanon 1 4 2 1 0 0 1 0 0 0 0 4 1 0 27 134 5 14 0
Palestine 1 1 1 1 5 2 3 3 0 0 4 8 1 52 8 74 8 1
Syria 0 6 4 0 5 4 1 4 5 7 7 4 0 26 45 13 60 3
Yemen 1 11 1 3 8 4 9 19 9 29 12 12 3 13 2 7 2 48
Algeria
Libya
Morocco
Tunisia
Bahrain
Iraq
Kuwait
Oman
Qatar
Saudi_Arabia
UAE
Egypt
Sudan
Jordan
Lebanon
Palestine
Syria
Yemen

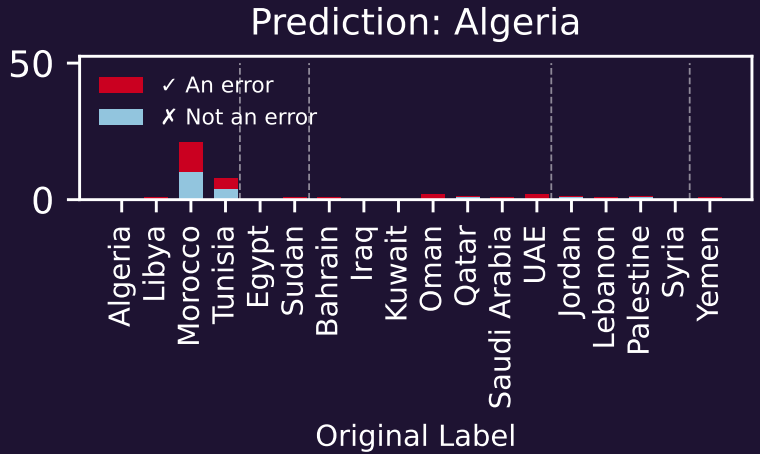

True label
Predicted label
Algeria 72 26 10 14 5 2 2 6 4 8 4 6 1 3 2 4 1 0
Libya 1 123 2 13 0 1 3 3 2 0 1 9 0 1 2 6 2 0
Morocco 21 9 113 5 1 1 2 1 1 4 0 10 0 5 0 5 0 0
Tunisia 8 25 2 83 2 1 1 6 1 4 5 7 1 4 2 1 0 1
Bahrain 1 9 1 0 54 7 23 10 22 14 29 2 1 5 1 0 2 3
Iraq 0 10 2 0 5 110 14 4 4 4 6 1 1 11 1 2 3 0
Kuwait 0 5 1 3 22 11 81 2 19 12 20 2 0 5 3 2 1 1
Oman 2 5 0 1 4 3 12 86 8 8 21 2 1 12 1 1 1 1
Qatar 1 3 1 2 18 3 21 4 67 28 31 2 0 9 1 2 0 5
Saudi_Arabia 1 7 0 1 14 5 16 11 21 88 14 2 0 4 1 0 3 11
UAE 2 8 1 2 14 1 11 14 15 6 101 3 1 4 1 2 1 5
Egypt 0 9 2 1 0 0 1 2 0 1 0 170 2 2 2 6 0 2
Sudan 1 6 3 2 0 1 2 7 0 3 4 17 127 4 0 5 1 5
Jordan 1 4 1 0 5 4 10 3 5 4 12 5 0 85 7 24 8 2
Lebanon 1 4 2 1 0 0 1 0 0 0 4 1 0 27 134 5 14 0
Palestine 1 1 1 1 5 2 3 3 0 0 4 8 1 52 8 74 8 1
Syria 0 6 4 0 5 4 1 4 5 7 7 4 0 26 45 13 60 3
Yemen 1 11 1 3 8 4 9 19 9 29 12 12 3 13 2 7 2 48
Algeria
Libya
Morocco
Tunisia
Bahrain
Iraq
Kuwait
Oman
Qatar
Saudi_Arabia
UAE
Egypt
Sudan
Jordan
Lebanon
Palestine
Syria
Yemen



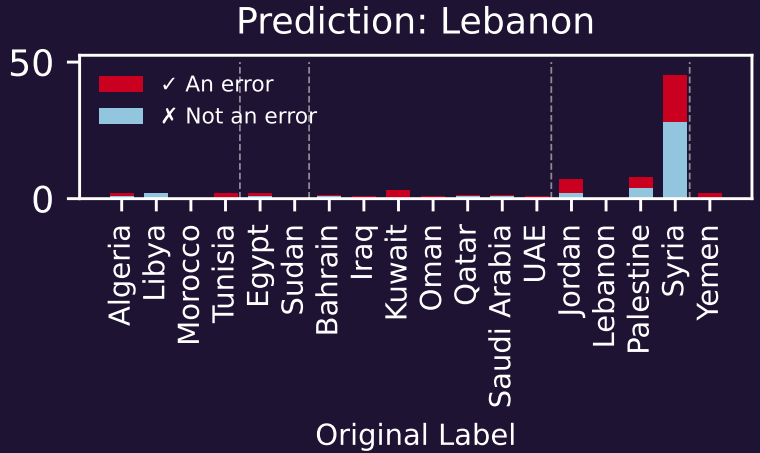
% of True Errors ≈ 26%



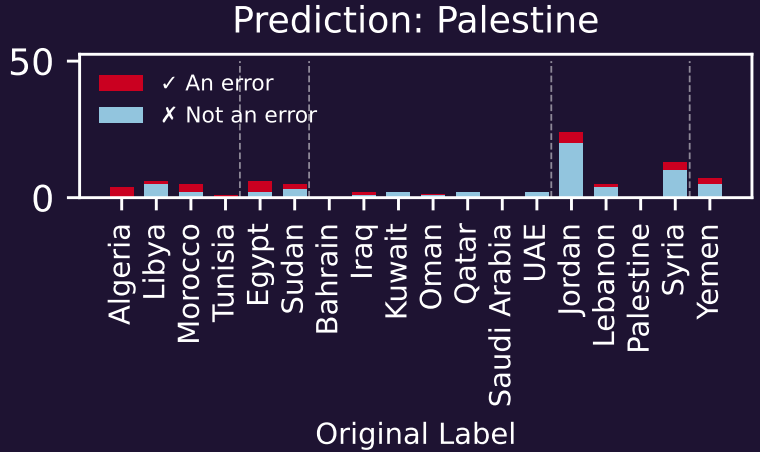
% of True Errors ≈ 26.5%



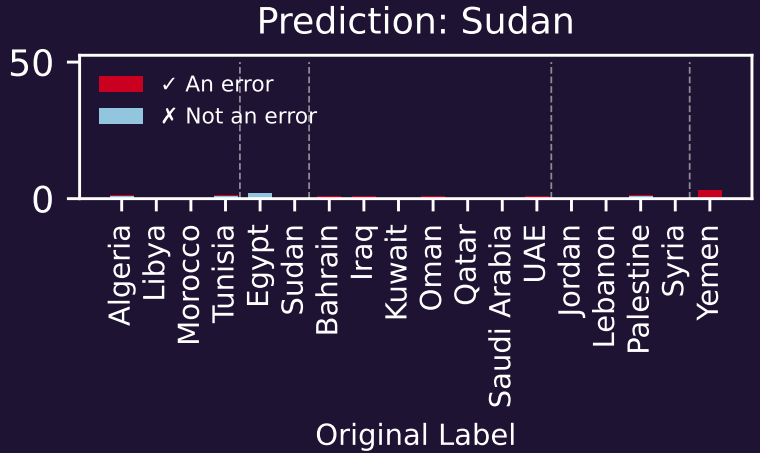
% of True Errors ≈ 59.5%



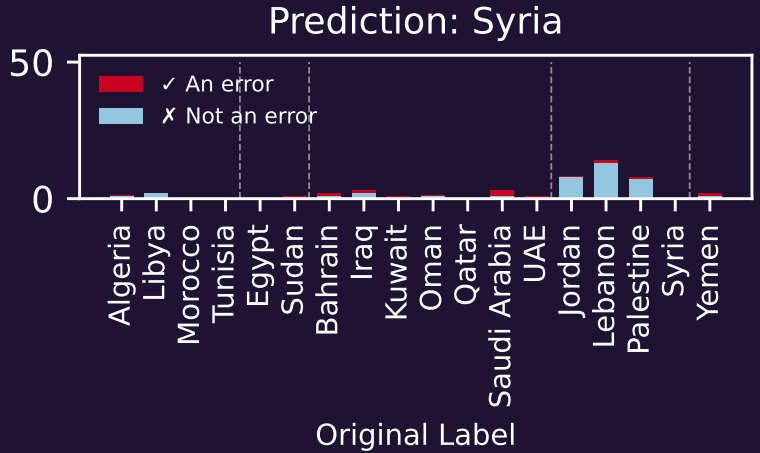
% of True Errors ≈ 48.1%



% of True Errors ≈ 30.6%



% of True Errors ≈ 58.3%



% of True Errors ≈ 21.28%

Samples from the NADI Dataset

Sentence	Valid in
لمن الحياه ترسل ليك رساله	Palestine🇵🇸, Sudan🇸🇩, Yemen🇾🇪

Samples from the NADI Dataset

Sentence	Valid in
لمن الحياه ترسل ليك رساله	Palestine🇵🇸, Sudan🇸🇩, Yemen🇾🇪
وين يلعب هذا ما شفته	Algeria🇩🇿, Palestine🇵🇸, Saudi🇸🇦, Yemen🇾🇪

Mulitlabel 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.

Mulitlabel ADI systems

Rank	System	Macro-average			
		Accuracy (↑)	**Precision (↑)**	**Recall (↑)**	**F_1 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	$\mathbf{73.42_{\pm 7.6}}$	$\mathbf{76.82_{\pm 10.6}}$	$17.77_{\pm 10.8}$	$27.30_{\pm 12.6}$

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

Mulitlabel ADI systems

Rank	System	Macro-average			
		Accuracy (↑)	**Precision (↑)**	**Recall (↑)**	**F_1 score (↑)**
1	**Elyadata**	$67.50_{\pm 3.7}$	$46.48_{\pm 10.1}$	$\mathbf{57.09_{\pm 5.1}}$	$\mathbf{50.57_{\pm 7.1}}$
Baseline II	Random	$50.14_{\pm 1.6}$	$30.43_{\pm 8.8}$	$50.15_{\pm 2.1}$	$37.15_{\pm 7.2}$

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

Mulitlabel ADI systems

Rank	System	Macro-average			
		Accuracy (↑)	**Precision (↑)**	**Recall (↑)**	**F_1 score (↑)**
1	**Elyadata**	$67.50_{\pm 3.7}$	$46.48_{\pm 10.1}$	$\mathbf{57.09_{\pm 5.1}}$	$\mathbf{50.57_{\pm 7.1}}$
Baseline I	Top 90%	$73.40_{\pm 6.1}$	$60.67_{\pm 14.5}$	$39.22_{\pm 14.6}$	$45.09_{\pm 11.3}$

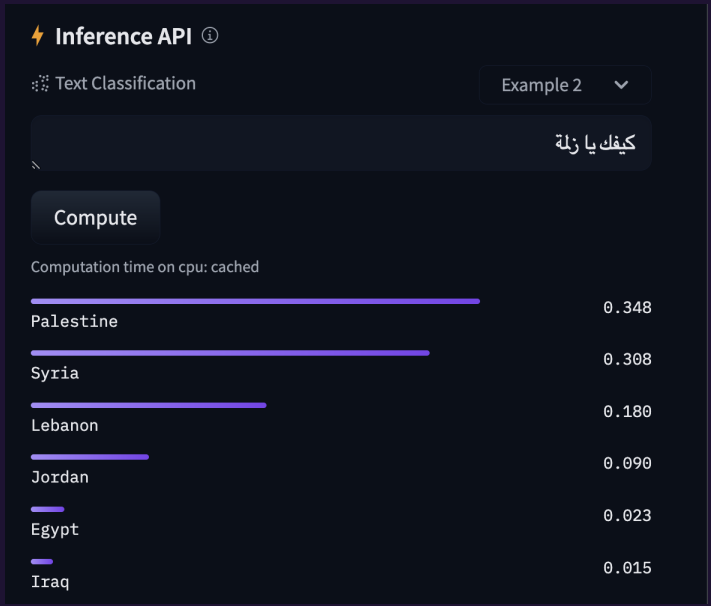
Table: Systems' performance on the test set of Subtask 1.

Baseline I (Top 90%):

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

Baseline I (Top 90%):

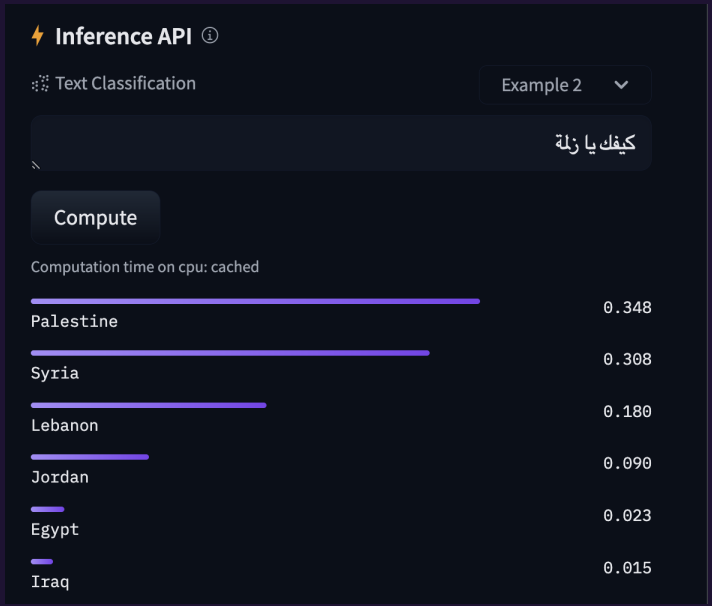
- A fine-tuned BERT-based model
- Single-label ADI

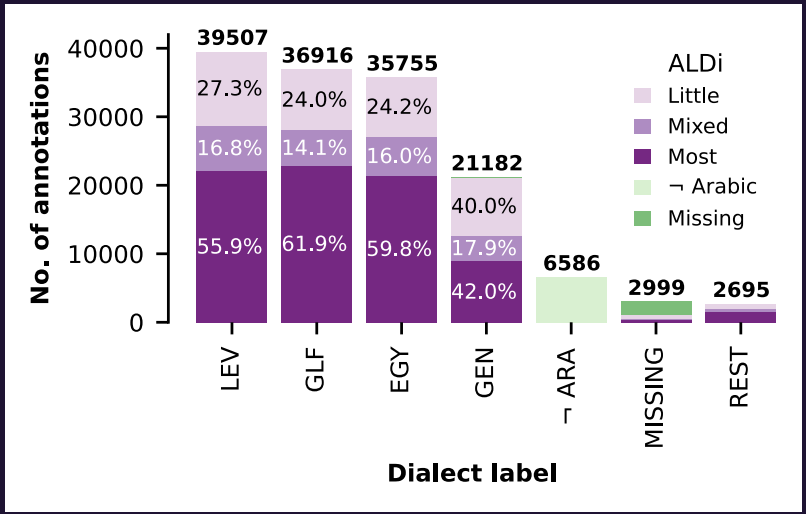


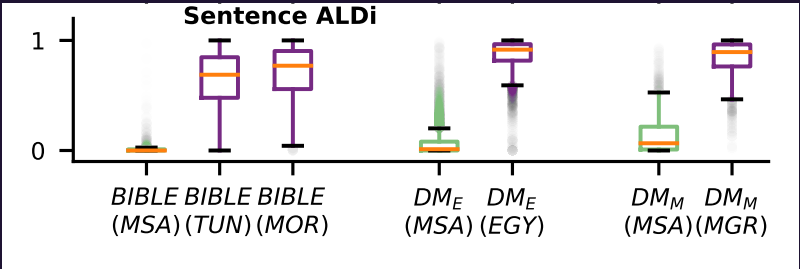
Baseline I (Top 90%):

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

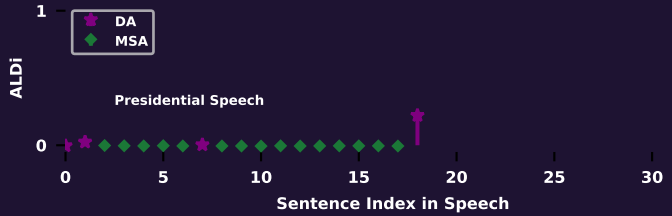
Predictions: Palestine, Syria, Lebanon, Jordan



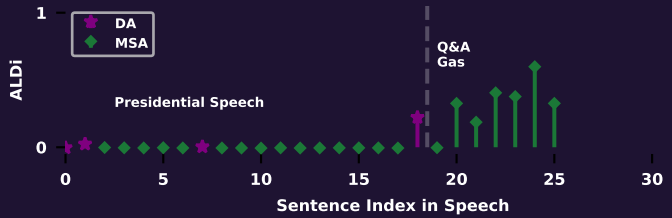

No. of annotations
40000
30000
20000
10000
0
39507
27.3%
16.8%
55.9%
36916
24.0%
14.1%
61.9%
35755
24.2%
16.0%
59.8%
21182
40.0%
17.9%
42.0%
6586
2999
2695
LEV
GLF
EGY
GEN
¬ ARA
MISSING
REST
Dialect label
ALDi
Little
Mixed
Most
¬ Arabic
Missing


Sentence ALDi
1
0
BIBLE (MSA)
BIBLE (TUN)
BIBLE (MOR)
DM_E (MSA)
DM_E (EGY)
DM_M (MSA)
DM_M (MGR)

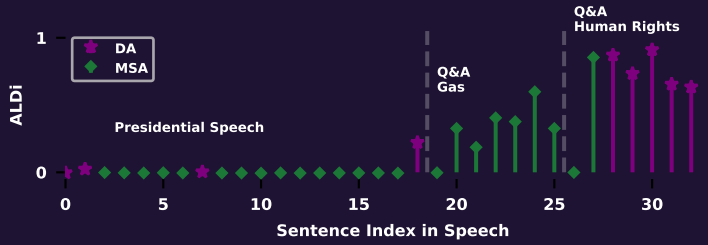
1) Studying Intraspeaker Variation



1) Studying Intraspeaker Variation

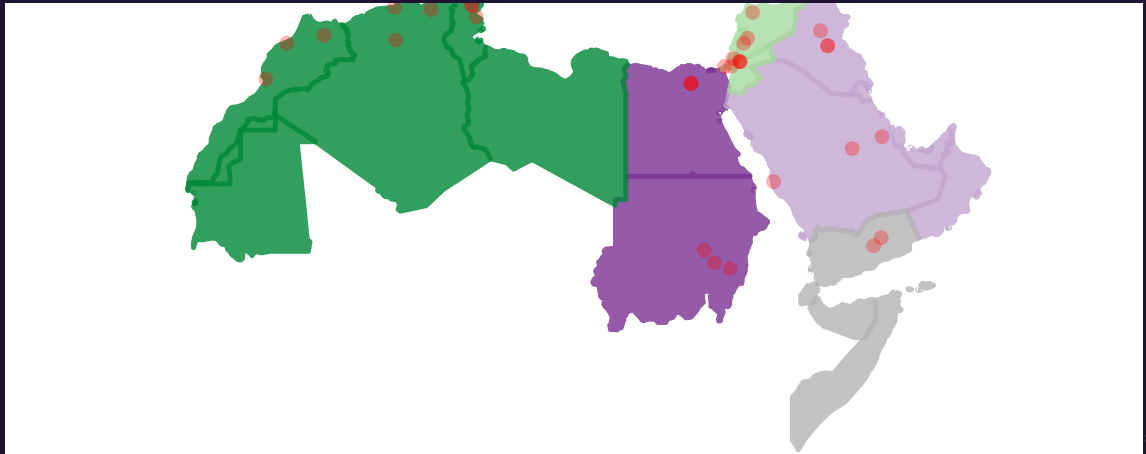


1) Studying Intraspeaker Variation



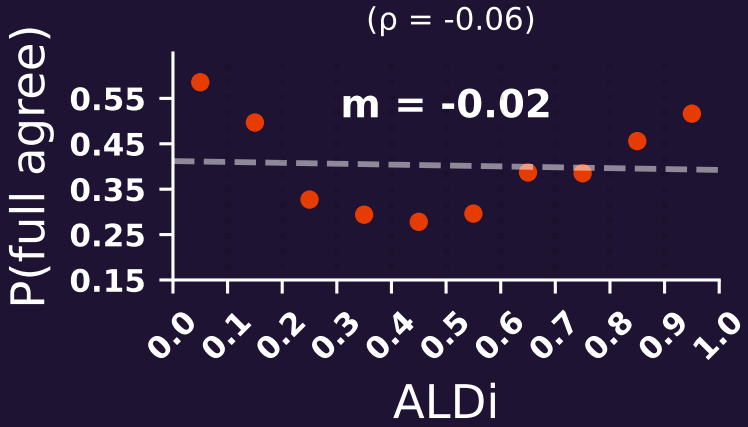
Finding (2) - Dialect Identification Datasets

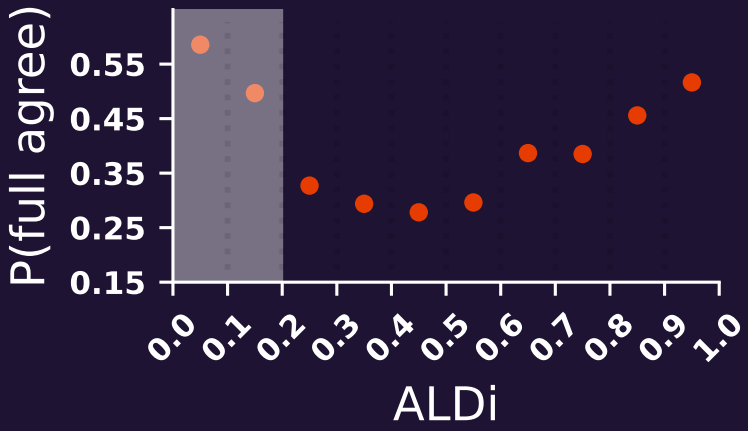
ℹ️ Labels (Macro-regional):
Modern Standard Arabic (MSA), Maghreb, Egypt, Levant, Gulf



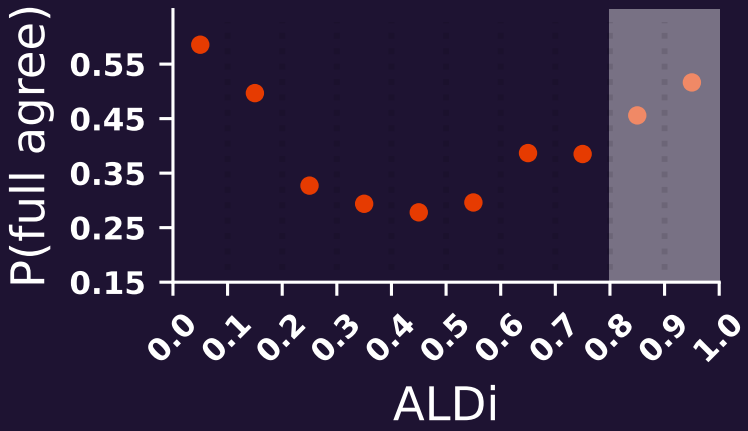
Finding (2) - Dialect Identification Datasets

ℹ️ Labels (Macro-regional):
Modern Standard Arabic (MSA), Maghreb, Egypt, Levant, Gulf

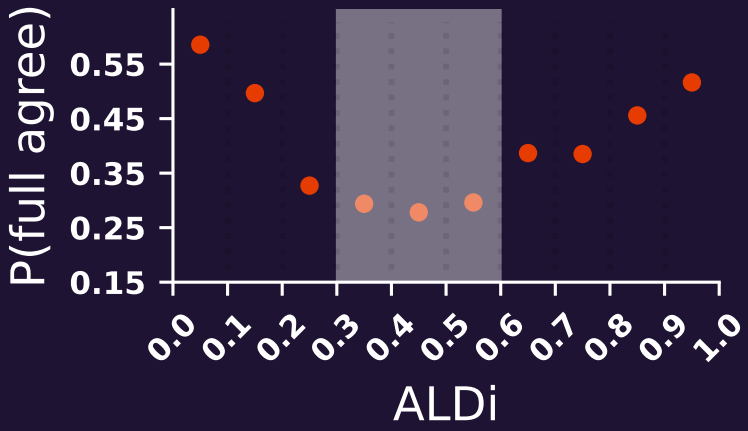



P(full agree)
0.55
0.45
0.35
0.25
0.15
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
ALDi

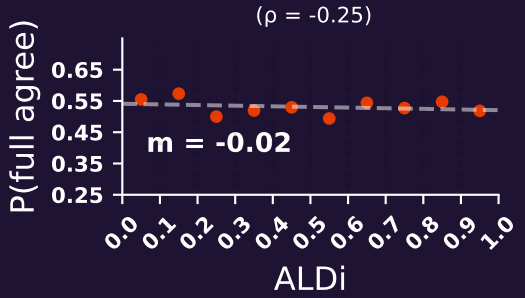
📎 **MSA samples**

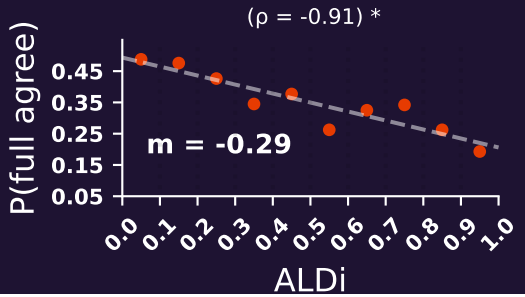


📎 **DA samples with multiple distinctive cues of a dialect**



📎 **Less cues - harder to determine the dialect.**


(ρ = -0.25)
P(full agree)
0.65
0.55
0.45
0.35
0.25
m = -0.02
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
ALDi


(ρ = -0.91) *
P(full agree)
0.45
0.35
0.25
0.15
0.05
m = -0.29
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
ALDi