

How Well Do Tweets Represent Sub-Dialects of Egyptian Arabic?

Mai Mohamed Eida¹, Mayar Nassar², Jonathan Dunn¹

University of Illinois at Urbana-Champaign¹, Ain Shams University²

Egyptian Arabic Sub-Dialects

Egyptian Arabic Sub-Dialects can be grouped into five major groups (Behnstedt and Woidich, 1985). These include the most prestigious Egyptian Dialect: **Cairene Egyptian Arabic (CEA)**, and the most stigmatized Egyptian Dialect: **Sa'idi Egyptian Arabic (SEA)**.

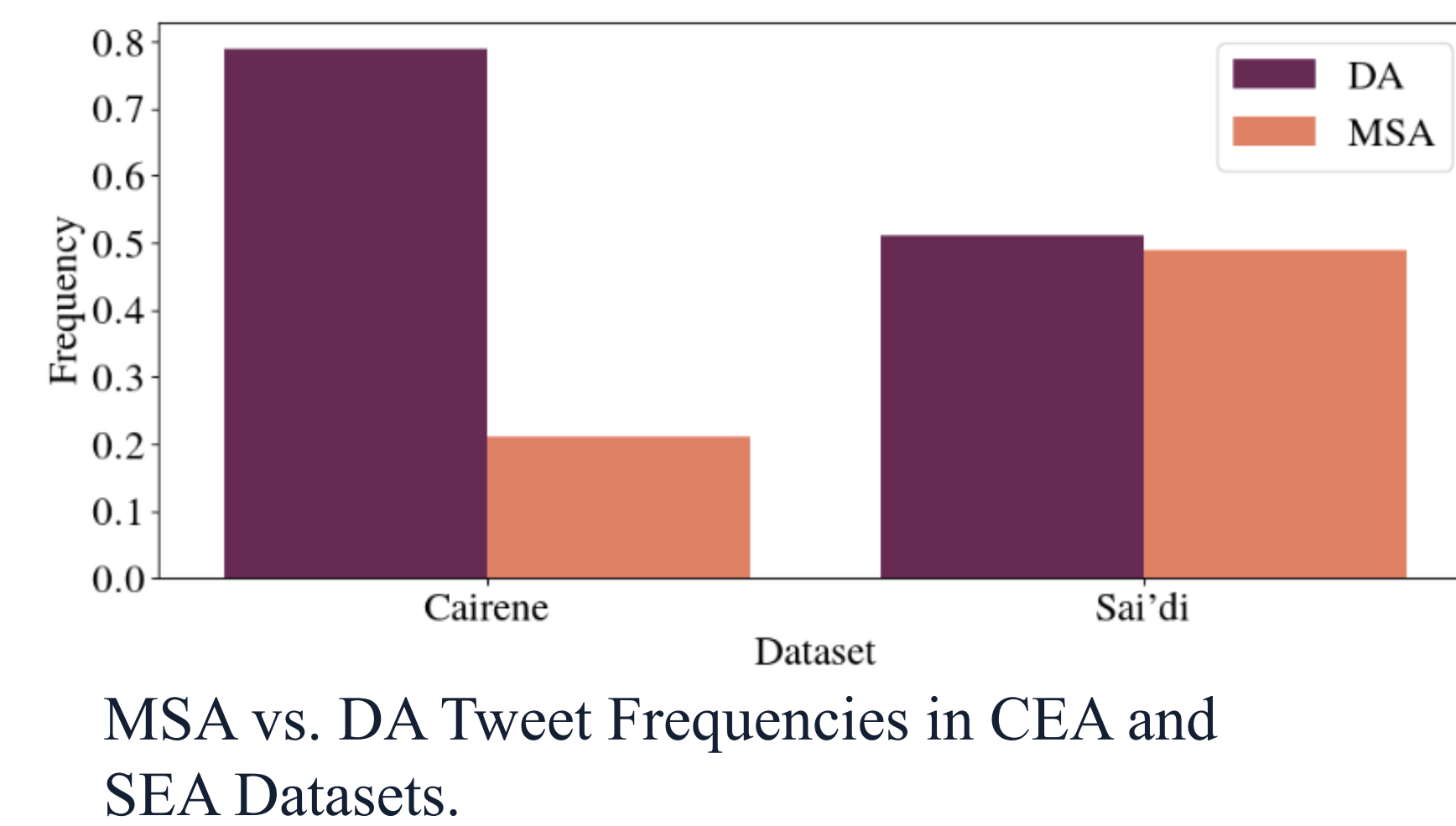
Representation within resources and corpora influence representation across NLP technologies (Dunn, 2020).

Therefore, do geo-tagged EA corpora equally represent both majority varieties (e.g., Cairene Egyptian Arabic: CEA) and minority varieties (e.g., Sa'idi Egyptian Arabic: SEA)?

Spoken vs. Written Register Variation

Modern Standard Arabic is the only standardized orthography in Arabic.

How often do CEA and SEA users prefer MSA to DA?



EA Sub-Dialect Distance Measures

We select morphological and grammatical features of SEA and CEA from ground-truth dialectal surveys (Behnstedt and Woidich, 1985; Khalafallah, 1969; Leddy-Cecere and Schroepfer, 2019).

We use regexes to extract features from each corpus, removing false positives, and accounting for orthographic variation.

Feature	MSA	CEA	SEA
Interrogative***	أيضاً	برضه	برضك
Adverb***	‘also’	‘also’	‘also’
Particle*	الآن	دلوقتي	دلوق
Preposition*	‘now’	‘now’	‘now’
Demonstrative***	ليس	مش	مش
	‘not’	‘not’	‘not’
	في	في	ف
	‘in’	‘in’	‘in’
	اولائك	دول	داكها
	‘these’	‘these’	‘these’

Datasets

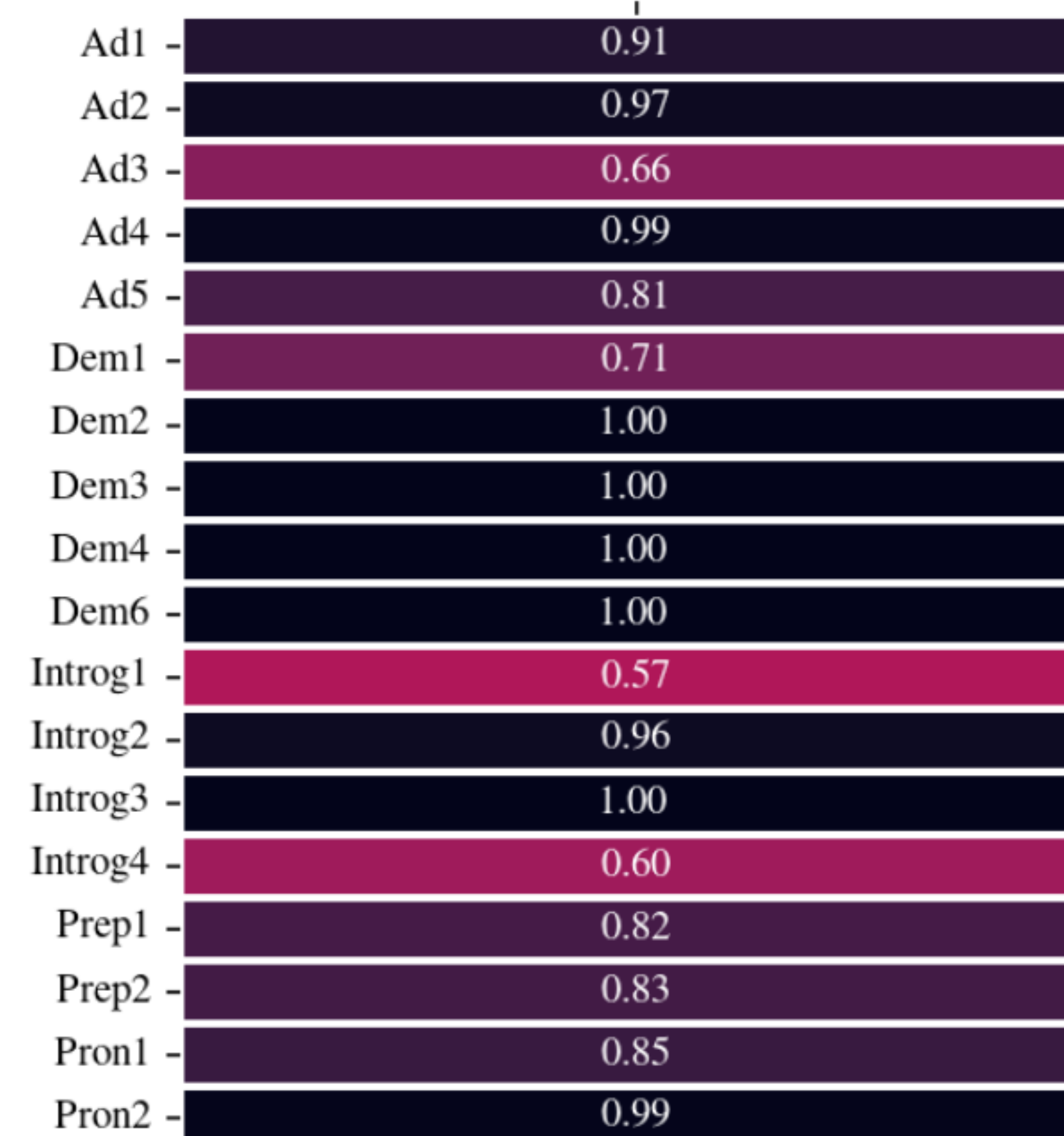
Dataset	Tweets	Tokens	MSA/DA
Baseline	808,312	12,233,632	Both
SEADataset	Tweets	Tokens	MSA/DA
MicroDialect	3,076	39,292	Both
NADI2020	1,862	24,693	Both
NADI2021	1,863	16,507	DA
Total	6,801	80,492	

Cairo Baseline Corpus (Dunn, 2020)

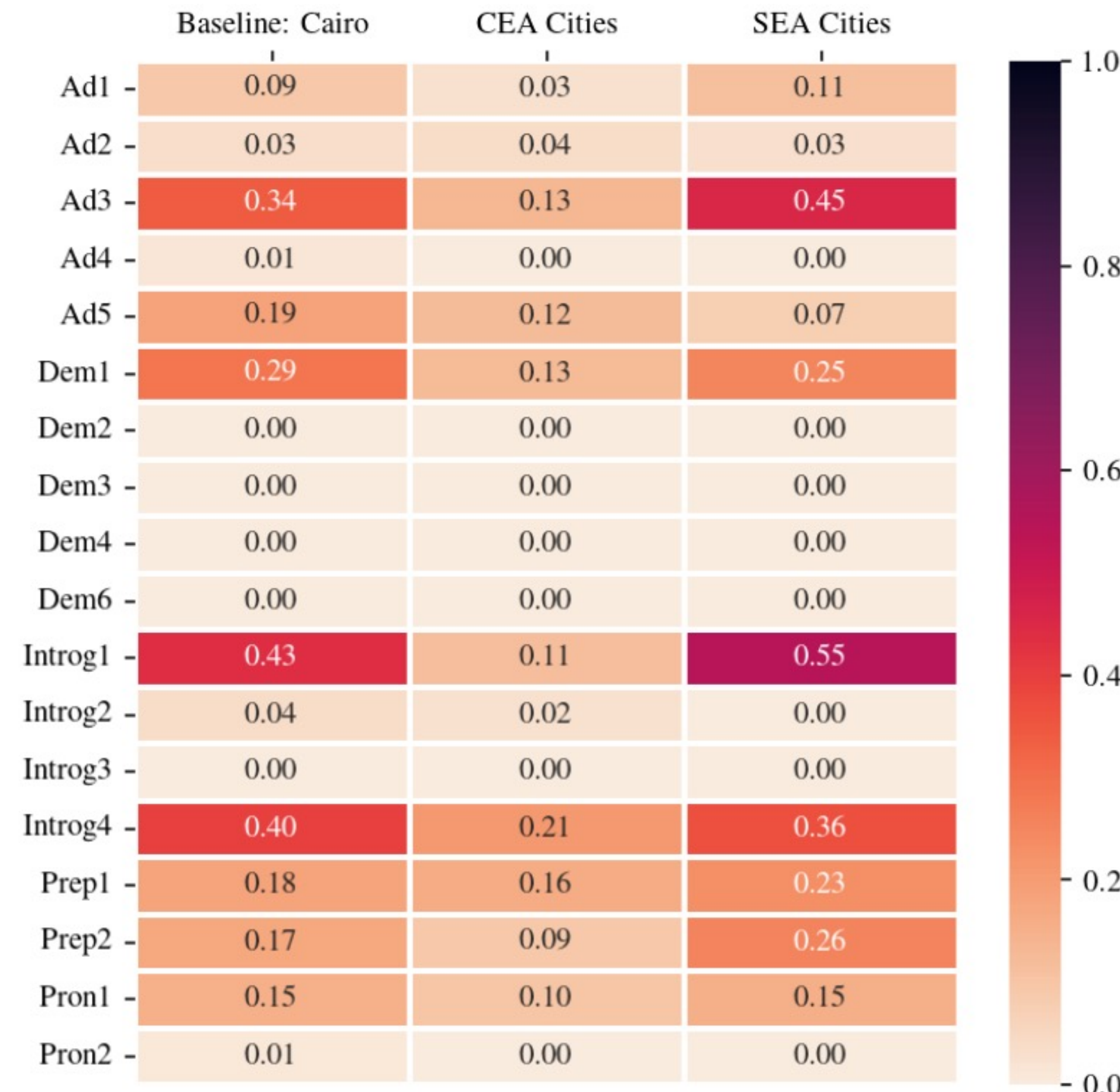
(Abdul-Mageed et al., 2020a, 2020b, 2021)

Feature by Feature Representation

Baseline Cairo - CEA Features



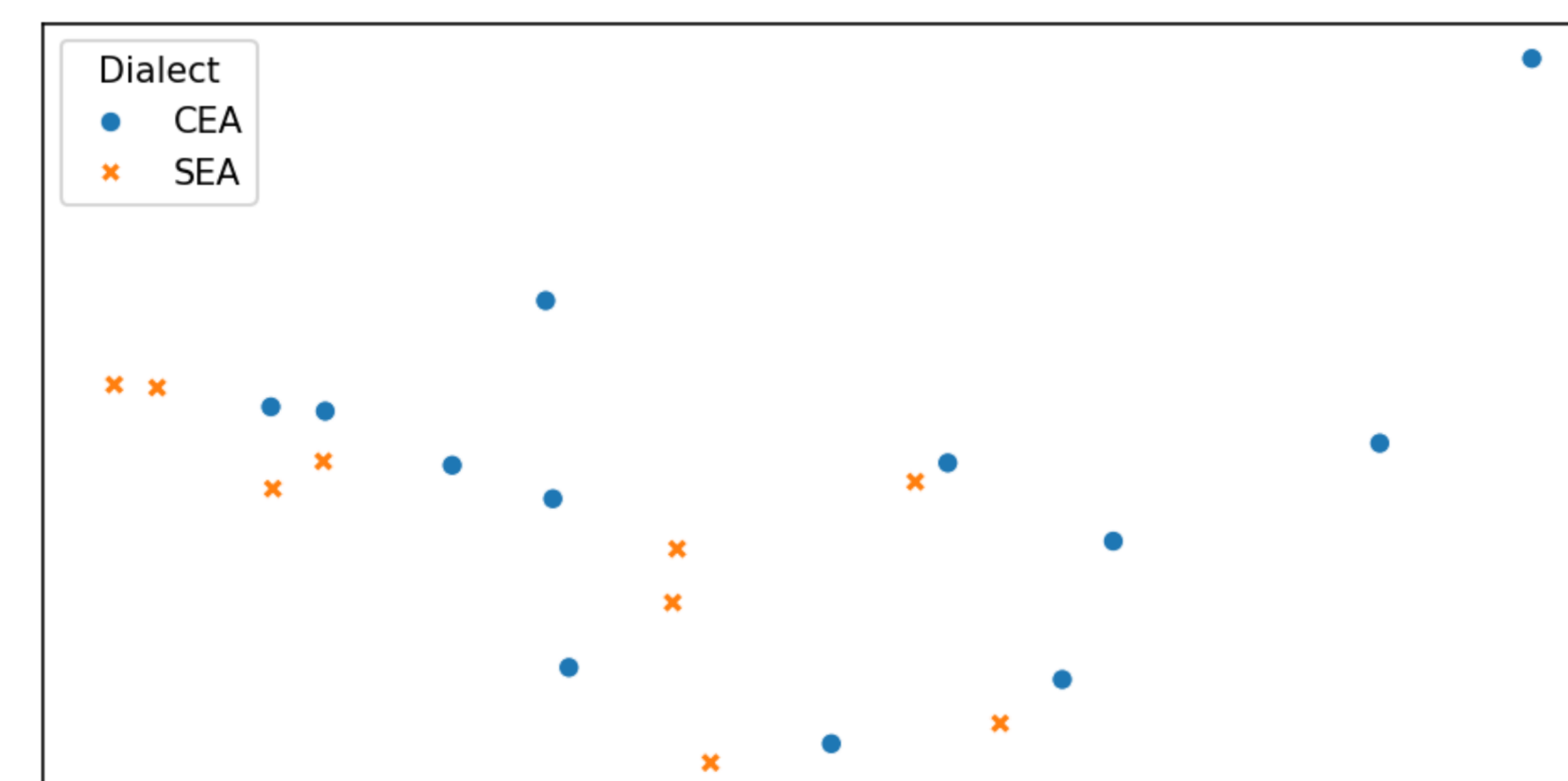
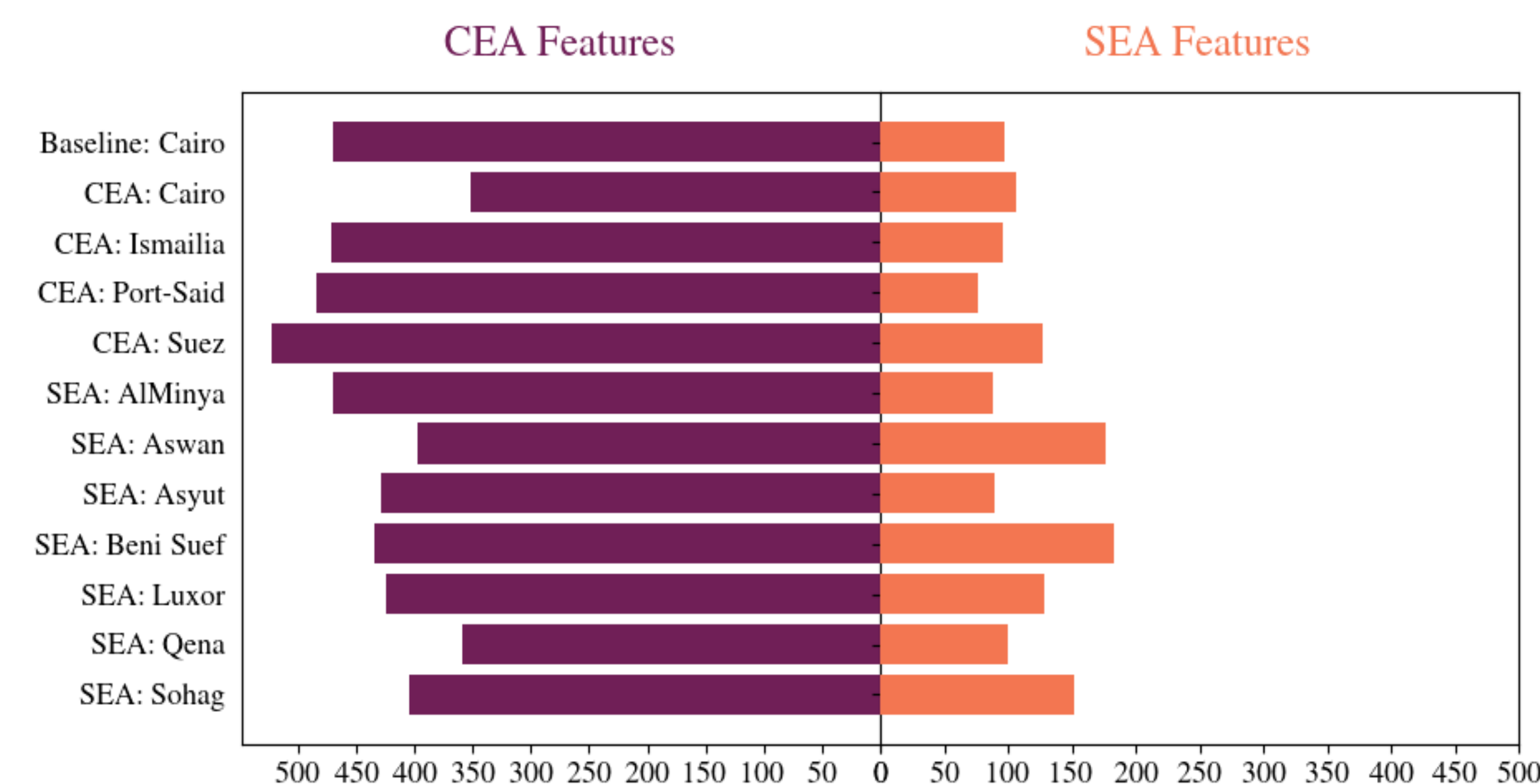
Share of Expected CEA Variants for each Alternation for Cairo Baseline Corpus. MSA and DA Corpus. Features are complementary.



Share of Expected SEA Features for each Alternation across SEA cities and CEA cities in NADI2020, NADI2021, and MicroDialect Corpus compared with baseline Cairo corpus.

City by City vs. User by User Representation

Prevalence of CEA and SEA features by city, using frequency per 10k words. MicroDialect Corpus.



User-by-user plots of feature usage, visualized using PCA for dimension reduction. The original vectors undergoing PCA are the relative frequency of each dialectal feature.

Discussion

As highlighted by the results, geo-referenced SEA datasets are not representative of SEA sub-dialect.

- One possibility could be that language change has taken place since the dialect survey was undertaken.
- Another possibility is that older or less connected speakers retain the SEA features but are not represented on social media.
- A third possibility is that SEA speakers do not produce SEA variants in this digital written setting. While there is a lack of recent SEA dialectal surveys, there is evidence on low attitudes and stigmatized perceptions of the SEA dialect (Bassiouny, 2014; Bassiouny, 2017).
- SEA users could be avoiding SEA markers to position themselves differently across digital platforms.

A larger geo-referenced written digital corpus is needed to explore these possibilities further. Regardless, the examined SEA datasets are not representative of the SEA sub-dialect, and register variation is significant across SEA spoken and written registers.

Conclusion

This paper finds that EA sub-dialects (except CEA) are low-resourced, and existing Tweet datasets are not representative of EA sub-dialects. Further, register variation between SEA speech and naturally-occurring digital written tweets is significant, therefore, these results call into question the validity of relying on geo-referenced tweets alone to represent dialectal differences.

Correspondence

mائم2@illinois.edu

