VarDial Workshop at EACL 2023

PALI: A Language Identification Benchmark for Perso-Arabic Scripts

Sina Ahmadi, Milind Agarwal, Antonios Anastasopoulos

sahmad46@gmu.edu

https://nlp.cs.gmu.edu/

GEORGE MASON UNIVERSITY

Perso-Arabic Scripts

Used by more than 20 languages/varieties spoken by over 400M speakers in the Middle East and the Subcontinent including Persian, Urdu, Kurdish, Uyghur etc.

NLP GEORGE MASON

Language Identification

Language identification is the task of detecting the language of a text.

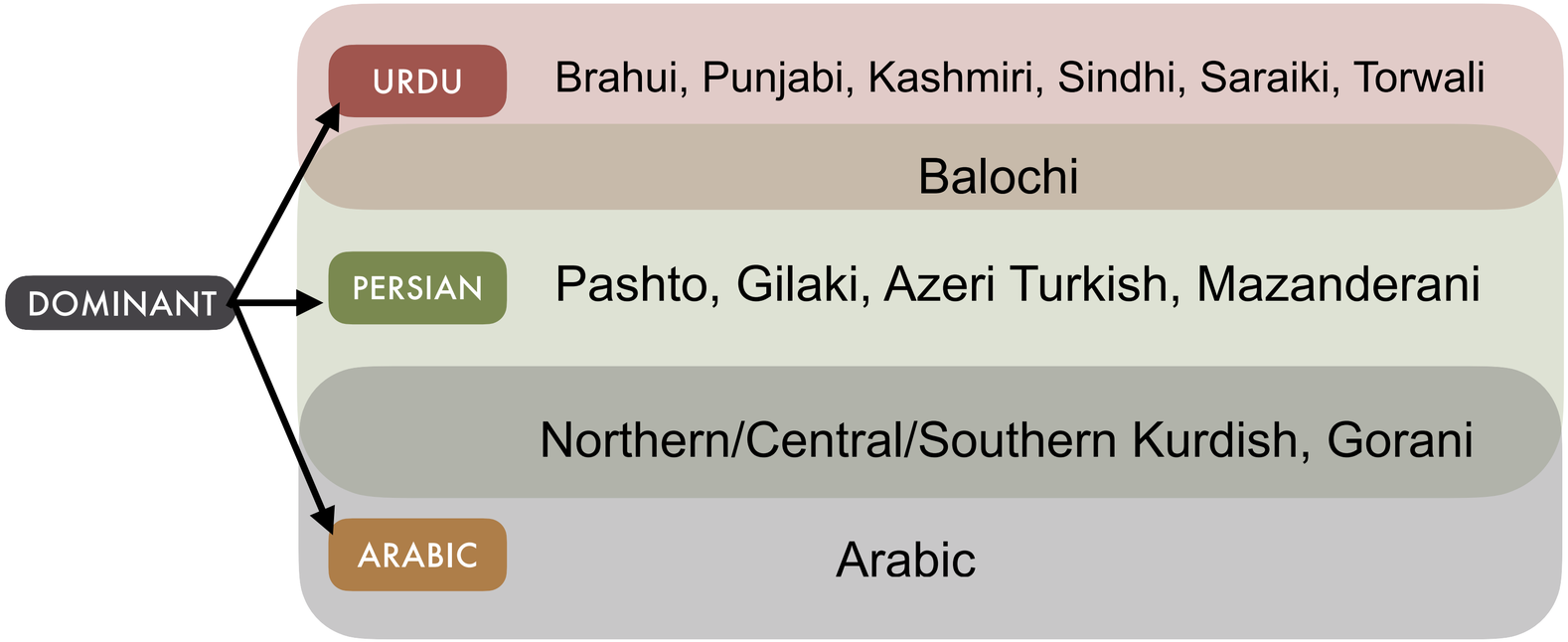
Sentence	Language
اور لادینیت واشتراکیت کو جمہوریت کے حسین لبادہ میں پیش کردیا گیا ۔	Punjabi
کہیں وی زبان وادب تے تحقیق زیادہ تر کیفیتی	Saraiki
گھٹا دفعا هڪ عورت ساڻياڻي جنهن سان ڪوئي افلاطوني	Sindhi
آیانی راہا کہ تئی مہر بوتگ اَنت گنجّ گُوار	Balochi
قوزئی و دوغو سوریه موختار ایداره‌ائتمه‌سی	Azeri
شوراب ایسم ایته روستا ایسه جه راستوپی دهستان	Gilaki
جوانی زمان فرا گرفتن دانایی است. پیری زمان تمرین کردن آن است.	Persian
هەیدەکچیلك تەرتیپنی ئایاغلاشتوُرووُش توغرىسىدا كېسىم چىقىردوُ	Uyghur
ڤایرۆس کۆرۆنا لەڕێ دادوەر و پاریزەرەیل دەوام لە دادگای هەولێر وسان	Southern Kurdish
سودھا رانی چھِ اَکھ ہِندوستاُنے اَداکارہ یوس فِلمَن مَنُز چھِ کاُم کَران.	Kashmiri
سودها رانی چھ اکھ ہندوستانے اداکارہ یوس فلمن منز چھ کام کران.	Kashmiri
ڕێژەی دەرجوانی ئەمسال لە سالی پێشتر زیاترە	Sorani
ريژهي دهرجاني ئهمصال له صالي پيشطر زياطره	Sorani

MOSTLY LESS-RESOURCED LANGUAGES SPOKEN IN BILINGUAL COMMUNITIES

NLP GEORGE MASON

Unconventional Writing

Unconventional writing refers to the usage of the script of another language, presumably that of a dominant language.



Kashmiri	سودھا رانی چھِ اَکھ ہِندوستانٕے اَداکارہ یوس فِلمَن مَنٛز چھِ کأم کَران.
Kashmiri	سودها رانی چھ اکھ ہندوستانے اداکارہ یوس فلمن منز چھ کام کران.
Sorani	ڕێژەی دەرجوانی ئەمسال لە سالی پێشتر زیاترە
Sorani	ريژەي دەرجاني ئەمصال له صالي پيشطر زياطرە

Methodology

1. Data collection
2. Script mapping
3. Synthetic data generation
4. Benchmarking
5. Hierarchical modelling

Data Collection

1. **Collection:** Not an easy task for low-resourced languages!
   - Various sources of data were explored:
     - Wikipedia (in a Perso-Arabic script):
       - Central Kurdish, Kashmiri, Pashto, Mazanderani, Gilaki, Azeri Turkish, Sindhi, Saraiki and Uyghur
     - Crawling local news websites:
       - Northern Kurdish, Southern Kurdish, Balochi and Brahui
     - Existing datasets and corpora for Central Kurdish, Gorani, Punjabi and Torwali
2. **Preprocessing:**
   - Normalization of Unicode encoding
   - Removing script-switched text
   - Unifying numerals

Script Mapping

Map the Perso-Arabic script used by a language to the script used by the dominant language

- Common characters
- Visual resemblance of graphemes (<چ> <ڇ> <ج> <ڃ>)
- Orthographic rules
- Uyghur is not mapped to any script!

Kurdish	Arabic
ئێ	ا
ئا	ا
ئی	ای
ئۆ	ا
چ	ج
گ	ك
ژ	ز
ئ	ئ
ا	ا
ب	ب

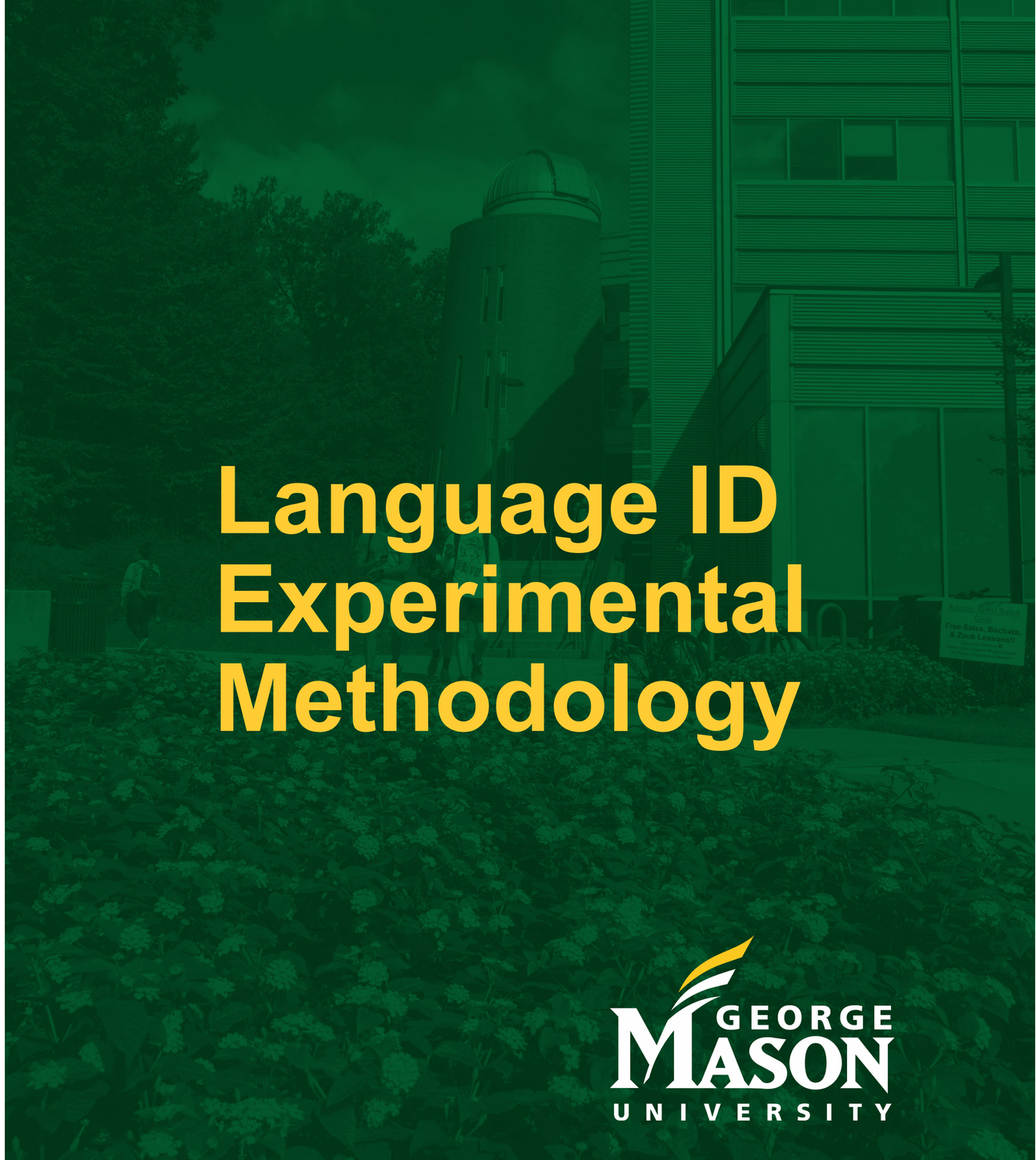
Synthetic Data Generation

Mimic unconventional writing by generating synthetic sentences based on the ‘clean’ ones

- Replace characters based on the script mapping
- Synthesize data at various levels starting from 20% noise up to 100%
- 10,000 sentences for each language
- Three datasets: Clean, Noisy & Merged

Noise %	Sentence
Clean	دووەمین پێشانگەها فۆتۆگرافەرێن کورد ل بەلجیکا Second Kurdish photographers’ exhibition in Belgium
20	دووهمین پێشانکهها فۆتۆکرافهرێن کورد ل بهلجیکا
40	دووه مین بشانکه ها فطکرافه رن کورد ل به لجیکا
60	دووة مین بشانکة ها فوتوکرافة رن کورد ل بة لجیکا
80	دووةمین بیشانکةها فؤتؤکرافةرین کورد ل بةلجیکا
100	دووهمین بیشانکهها فوتوکرافهرین کورد ل بهلجیکا

Language ID Experimental Methodology



Comparing Language Identification Systems

OFF-THE-SHELF

- fastText model-lid.176
- Google's CLD3
- langid.py
- Franc

CUSTOM-TRAINED

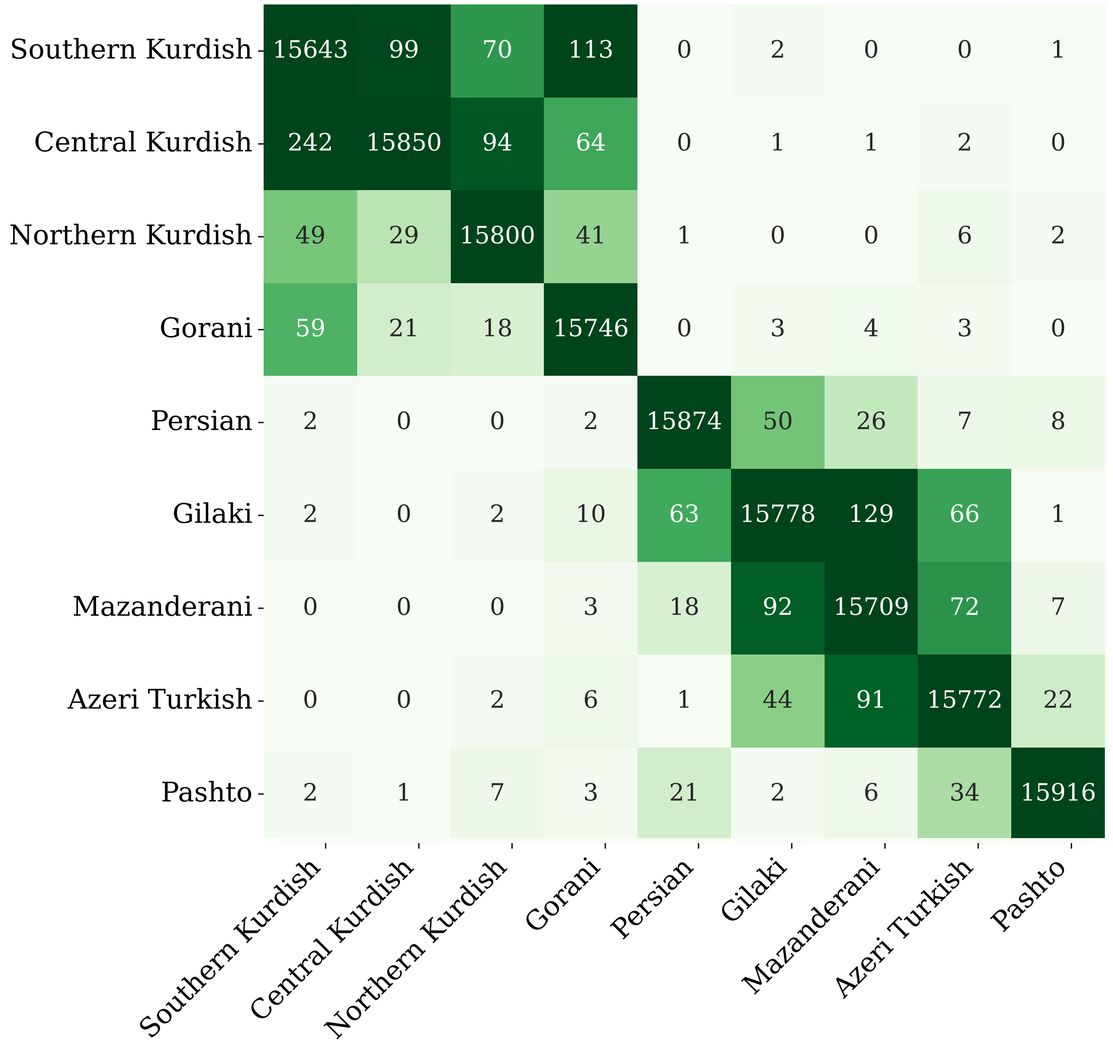
- Custom fastText model
- Multinomial Naive Bayes
- Multilayer Perceptron

PROPOSED

Confusion-resolution Hierarchical model

Identifying Confusion Between Languages

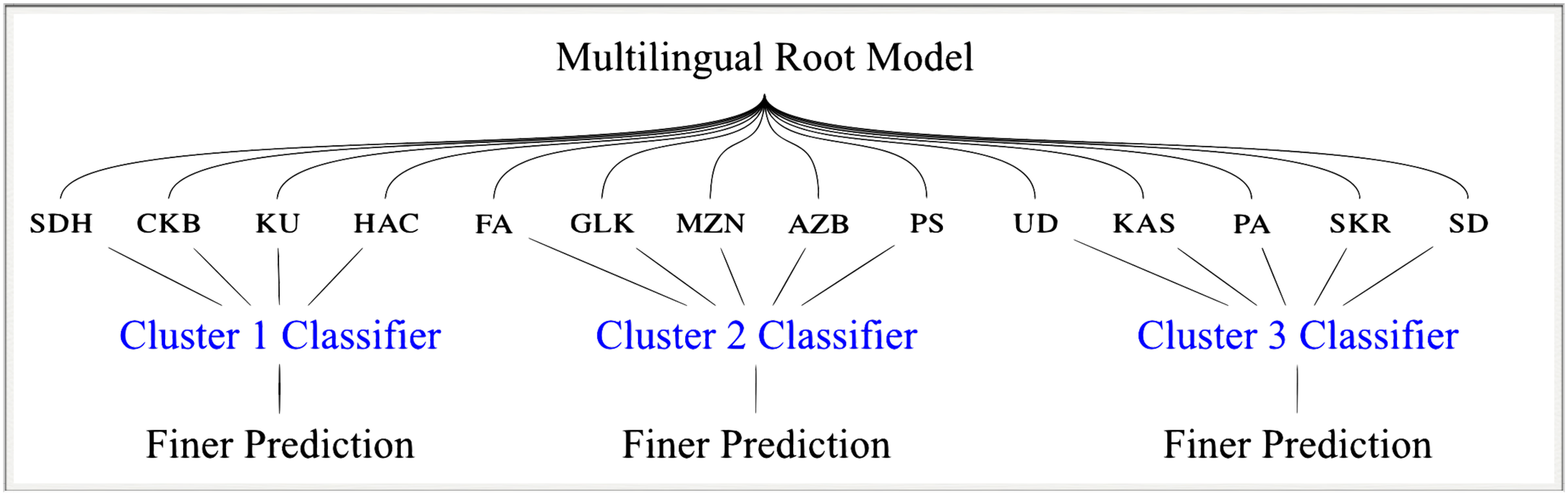
- Confusion matrix can be analyzed to identify clusters of closely related languages, often confused by the model
- We identify 3 clusters among our languages:
  - Southern/Central/Northern Kurdish, Gorani
  - Persian, Gilaki, Mazanderani, Azeri Turkish, Pashto
  - Urdu, Kashmiri, Punjabi, Saraiki, Sindhi
- Small classifiers are trained to distinguish between each cluster

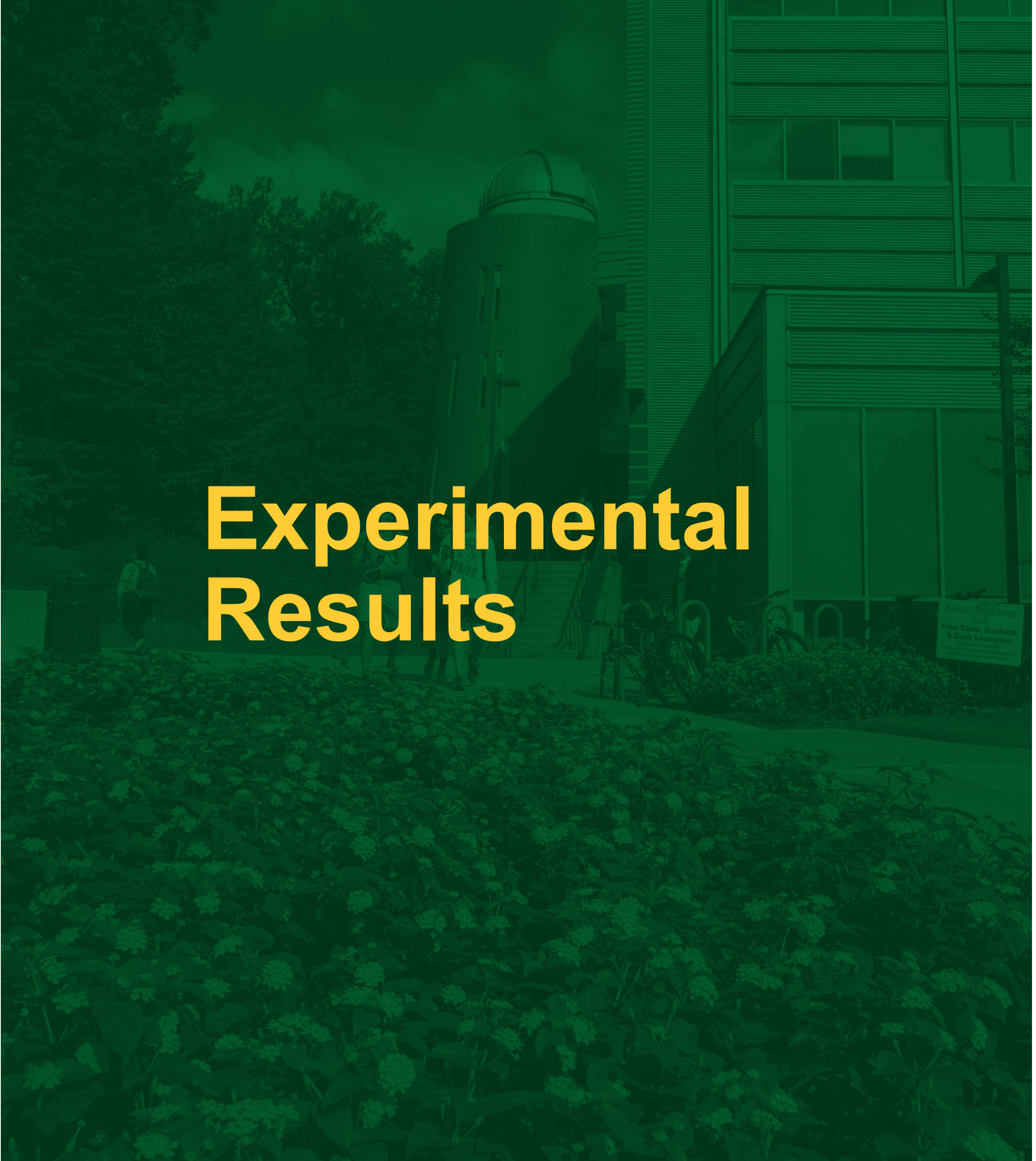


Confusion matrix of predictions (rows) and ground truth (columns)

Hierarchical Modelling

Resolve a model's confusion between highly-related languages by training expert classifiers that specialize in distinguishing between a small set of languages





Experimental Results

NLP GEORGE MASON

Evaluating Language Identification Systems

- Despite coverage of high-resource languages like Urdu, Persian and Arabic, off-the-shelf models' performance remains low overall
- Custom-trained models perform better overall than any off-the-shelf system like Google CLD3, Franc, Langid.py
- A confusion-resolution approach provides further insight into training data and model's shortcomings
- Hierarchical models are easy to train and provide statistically significant improvements

	Precision	Recall	F1 Score
Hier	**0.95**	**0.94**	**0.95**
Root	0.95	0.94	0.94
fastText	0.28	0.27	0.27
CLD3	0.06	0.16	0.09
langid.py	0.11	0.16	0.13
Franc	0.11	0.16	0.13
MNB	0.15	0.08	0.10
MLP	0.15	0.07	0.10

Macro-results for all languages o the Merged (noisy + clean) data

Language-Specific Performance Insights

- On the merged dataset (clean + noisy), the confusion-resolution model brings improvements across clusters
- The proposed approach, with the exception of Saraiki, doesn't reduce the F1 score of the root model on any language
- Complete results across all noise settings are available in Table 5 in the paper

	Root	Hier
Southern Kurdish	0.95	**0.96**
Central Kurdish	0.95	0.95
Northern Kurdish	0.95	0.95
Gorani	0.94	0.94
Farsi	0.97	**0.98**
Gilaki	0.92	**0.94**
Mazanderani	0.92	0.92
Azeri Turkish	0.91	0.91
Pashto	0.96	0.96
Urdu	0.96	**0.97**
Kashmiri	0.94	**0.95**
Punjabi	0.91	0.91
Sindhi	0.93	**0.94**
Saraiki	**0.92**	0.91

PALI: A Language Identification Benchmark for Perso-Arabic Scripts

Sina Ahmadi, Milind Agarwal, Antonios Anastasopoulos

Contact Us

sahmad46@gmu.edu

GitHub Repository

https://github.com/sinaahmadi/PersoArabicLID

Thank you!