Annotation-Efficient Language Identification Using Weak Labels

Shriphani Palakodety*
Onai
(Speaker)

Ashiqur KhudaBukhsh*
Carnegie Mellon University

* Equal contribution

Indian Social Media Text

- Noisy, Romanization, Multilingual
- ***Language Identification***
  - ***Noisy*** Text
  - ***Romanization***

Language Identification

- Support
  - Noisy, Romanized text
- **Low annotation budget**
  - ***Weak Labels***

Weak Labels

- Polyglot language model:
  - Monolingual clusters [Palakodety*, KhudaBukhsh*, Carbonell, 2020]
- Each cluster:
  - Label small set
  - Expand to other cluster members
- ***260 Labels*** -> ***2.4 million Weak Labels***

Indic Language Identification

Github:

https://github.com/onai/indic-language-identification

Paper:

Language use by geography

Bi/Multi-linguality statistics