Identifying open challenges in language classification

Task

Enhver har ret til hvile og fritid — dan

Es zayld goah kenn nacht datt sei — pdc

Setup

Textcat (40k)

Naive Bayes (100k)

FastText (4m)

BiLSTM (15m)

GLOT500 (395m)

Scripts/lang. families

#instances

#chars

#langs

wiki

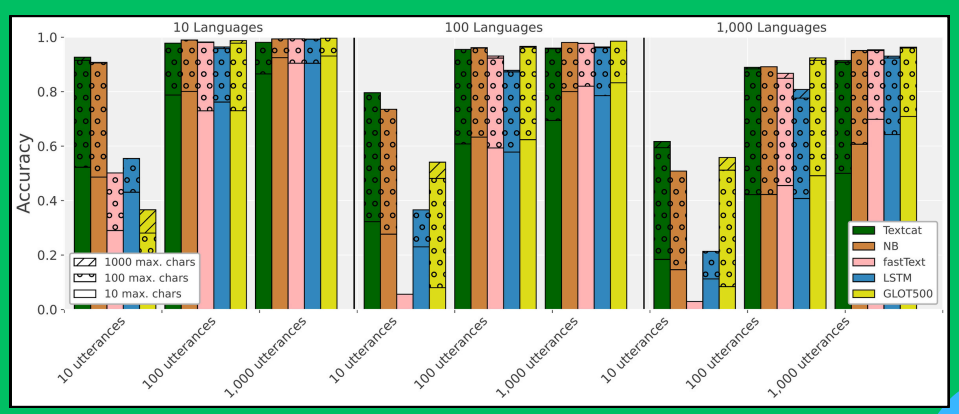
news

religious

rights

social

Size



- 100 characters is enough
- 100 utterances is almost enough
- Improvements with larger models is marginal
- Number of languages is not so important

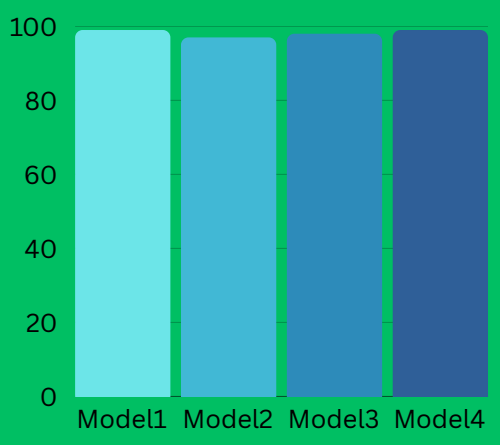
Next: DistaLs

- Language information and distances:

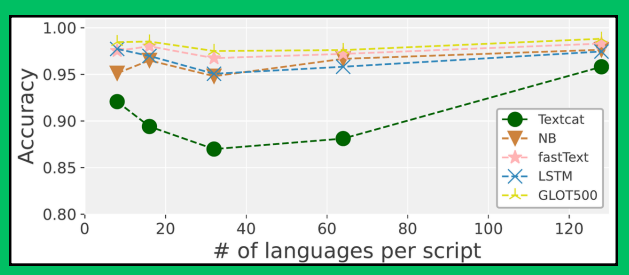
Category	Feature	Source	Coverage
Metadata	wiki_size	Wikipedia	7,856
	nlp_state	state and fate	2,265
	speakers	LinguaMeta	5,539
	AES	Glottolog	7,725
	loc	Glottolog	7,629
Typology	lang2vec	URIEL	3,910
	lang2vec_knn	URIEL	3,910
	PHOIBLE	PHOIBLE	2,024
	grambank_all	Grambank	2,288
	grambank_.*	Grambank	2,288
	glot_fam	Glottolog	7,856
	scripts	LinguaMeta, GlotScript	6,427
Wordlists	**ASJP**	ASJP	5,580
	concepts	Conceptualizer	1,274
Text-driven	whitespace	LTI LangID	2,109
	punctuation	LTI LangID	2,109
	char_distr.	LTI LangID	2,109
	textcat	LTI LangID	2,109

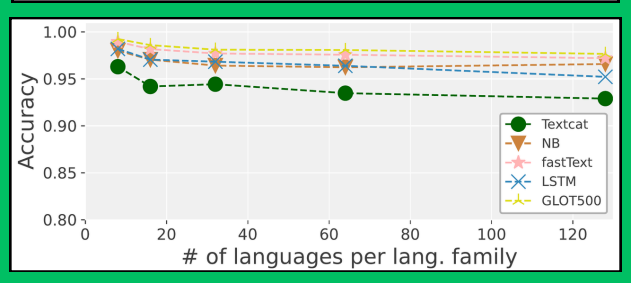
IT UNIVERSITY OF CPH

Previous work



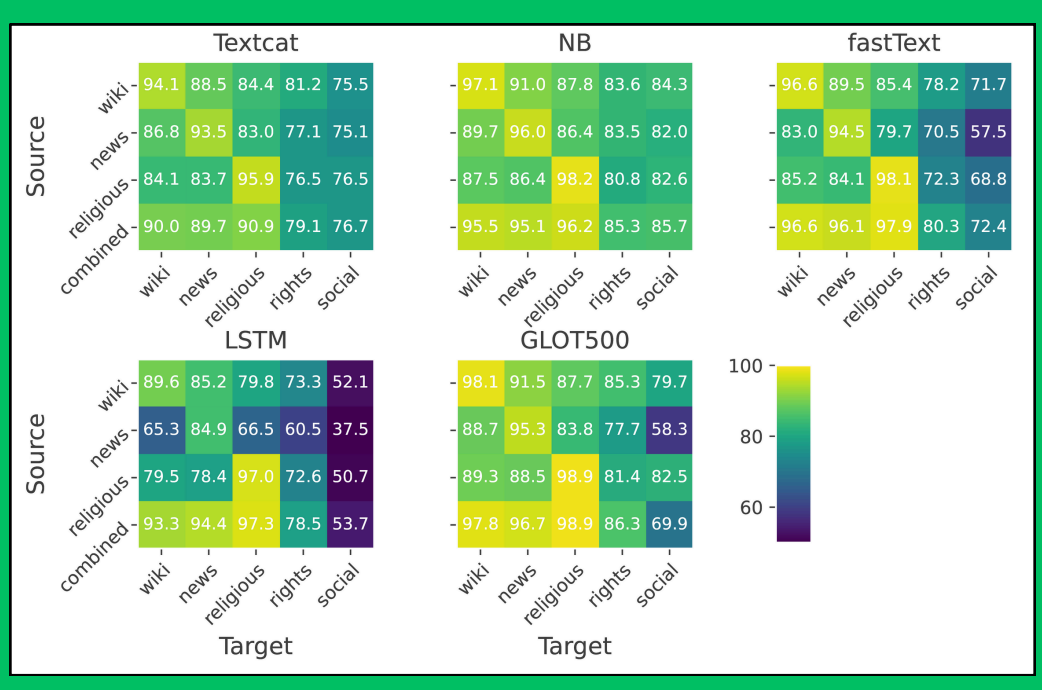
Scripts/families





- Effects are overall small
- Scripts with 16-64 languages are hardest

Domains



- Hardest dimension
- Smaller models are more robust to the test-only domains
- Training on multi-domain data leads to more robust models

Data and models available:
https://bitbucket.org/robvanderg/langid_problems/src