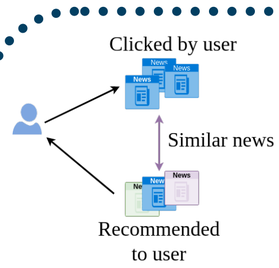
MIND Your Language: A Multilingual Dataset for Cross-lingual News Recommendation

Andreea Iana[1], Goran Glavaš[2], Heiko Paulheim[1]
[1]Data and Web Science Group, University of Mannheim, Germany
[2]Center for Artificial Intelligence and Data Science, University of Würzburg, Germany

xMIND

News Recommendation Needs (More) Diverse Multilingual Datasets

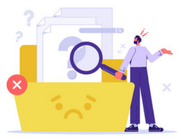


Algorithmic news curation shapes readers' views

IN

An increasingly language-diverse online user community

BUT



Scarcity of **publicly available, diverse, multilingual** news recommendation datasets

AND

Focus on **monolingual news consumption & high-resource languages**

RESULT IN

Less relevant & less balanced & less diverse recommendations for polyglots and readers of low-resource languages

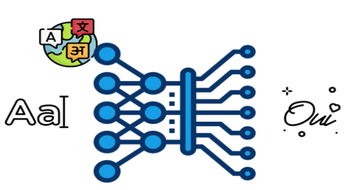
xMIND: Diverse, Multi-parallel, and Open-source

MIND: most used news recommendation benchmark

1 million users

130,379 unique articles

> 24 million clicks



Neural Machine Translation (NLLB 3.3B)

xMIND

Code	Language	Script	Macro-area	Family
SWH	Swahili	Latin	Africa	Niger-Congo
SOM	Somali	Latin	Africa	Afro-Asiatic
CMN	Mandarin Chinese	Han	Eurasia	Sino-Tibetan
JPN	Japanese	Japanese	Eurasia	Japonic
TUR	Turkish	Latin	Eurasia	Altaic
TAM	Tamil	Tamil	Eurasia	Dravidian
VIE	Vietnamese	Latin	Eurasia	Austro-Asiatic
THA	Thai	Thai	Eurasia	Tai-Kadai
RON	Romanian	Latin	Eurasia	Indo-European
FIN	Finnish	Latin	Eurasia	Uralic
KAT	Georgian	Georgian	Eurasia	Kartvelian
HAT	Haitian Creole	Latin	North-America	Indo-European
IND	Indonesian	Latin	Papunesia	Austronesian
GRN	Guarani	Latin	South America	Tupian

Diverse

✓ linguistically

✓ geographically

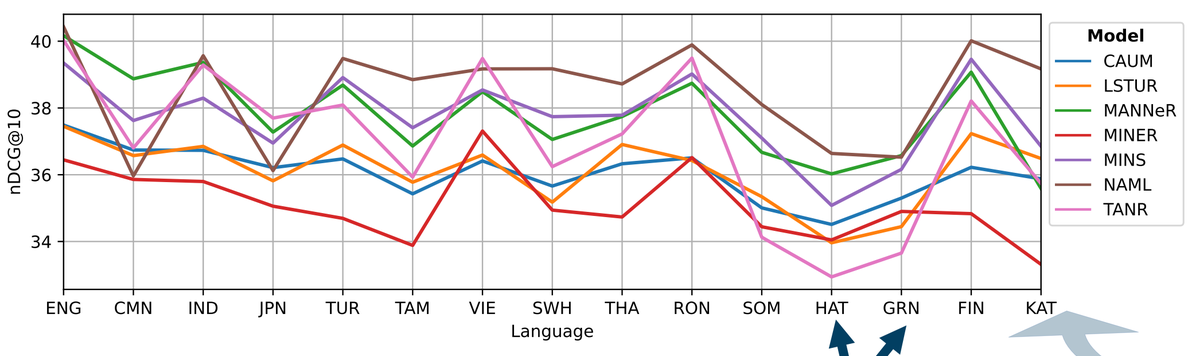
✓ w.r.t. digital footprint size

But

✗ US-centric content

Performance Deteriorates Substantially in Zero-shot Cross-lingual Transfer Recommendation

- Training: **monolingually** on **ENG** as **source**
- Testing: **monolingually** on **target** from xMIND



Out-of-sample for the language model

Language & content-agnostic

Average over all xMIND languages

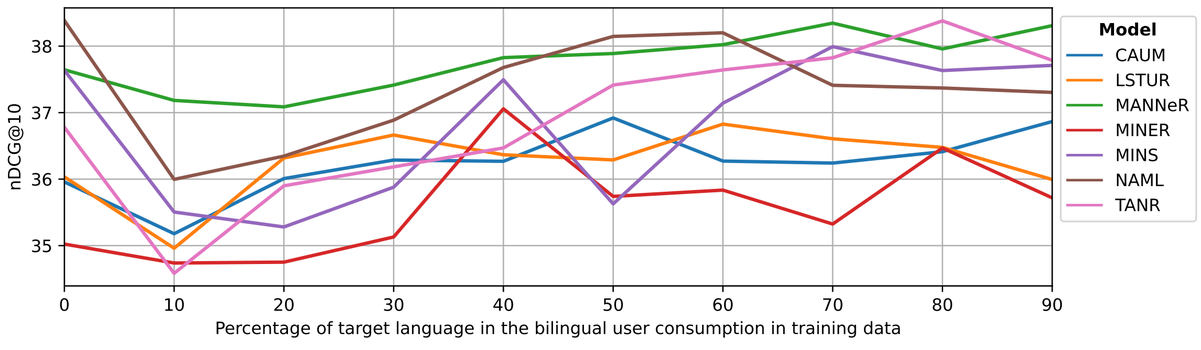
Model	AUC			nDCG@10		
	ENG	AVG	%Δ	ENG	AVG	%Δ
$NAML_{CAT}$	55.46±0.18		0.0	35.81±0.59		0.0
CAUM-PLM	57.82±3.01	55.90±1.75	-3.32	37.49±1.71	35.96±1.58	-4.08
LSTUR-PLM	56.80±1.36	56.28±1.68	-0.92	37.45±0.54	36.03±0.85	**-3.78**
MANNeR	50.00±0.00	50.00±0.00	**0.00**	40.17±0.21	37.64±0.44	-6.28
MINER	57.73±7.33	55.81±4.33	-3.32	36.45±4.84	35.02±3.51	-3.90
MINS-PLM	**59.89±0.29**	**56.94±1.40**	-4.93	39.35±0.20	37.64±0.50	-4.35
NAML-PLM	52.85±2.27	52.49±2.60	-0.68	**40.43±0.39**	**38.38±1.02**	-5.06
TANR-PLM	54.18±5.91	53.27±1.91	-1.68	40.03±0.86	36.78±0.88	-8.11

Relative performance under zero-shot XLT w.r.t. ENG

Few-shot Target-language Injection During Training Shows Limited Benefits

(A) Monolingual News Consumption

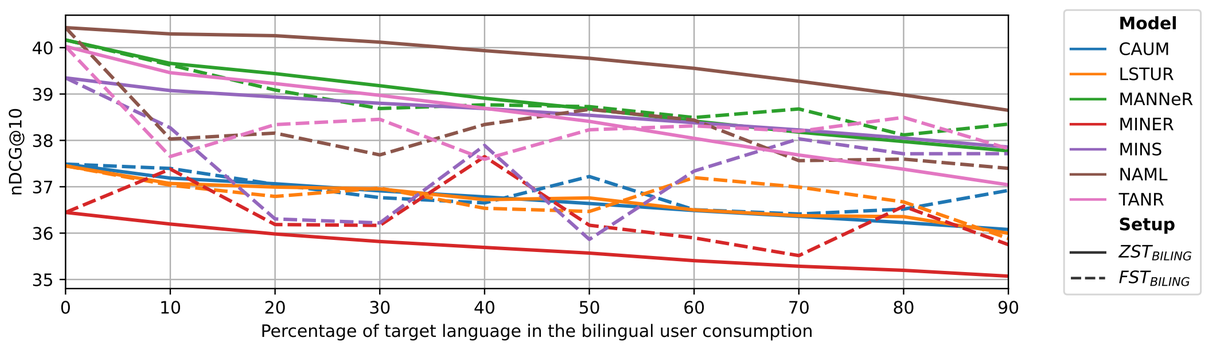
- Training: **injection of x% target language examples**
- Testing: **monolingually** on **target** from xMIND



(B) Bilingual News Consumption

***Simulated* bilingual consumption**: replace x% of source-language news in user history with target-language news

- Testing: **bilingually** on **ENG as source + target** from xMIND



Julius-Maximilians-
UNIVERSITÄT WÜRZBURG

xMIND @GitHub

xMIND @HuggingFace

Contact

UNIVERSITÄT MANNHEIM