MIND Your Language

Andreea Iana[1], Goran Glavaš[2], Heiko Paulheim[1]

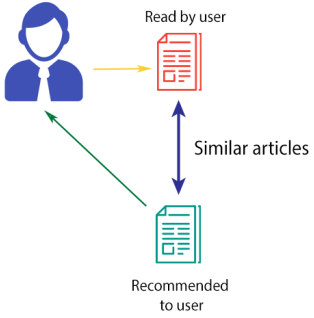
[1]Data and Web Science Group, University of Mannheim
[2]Center for Artificial Intelligence and Data Science, University of Würzburg

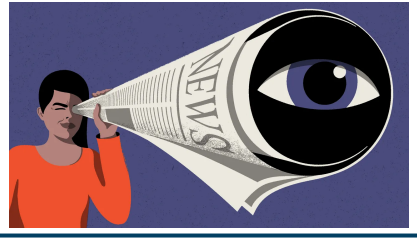
UNIVERSITÄT MANNHEIM

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

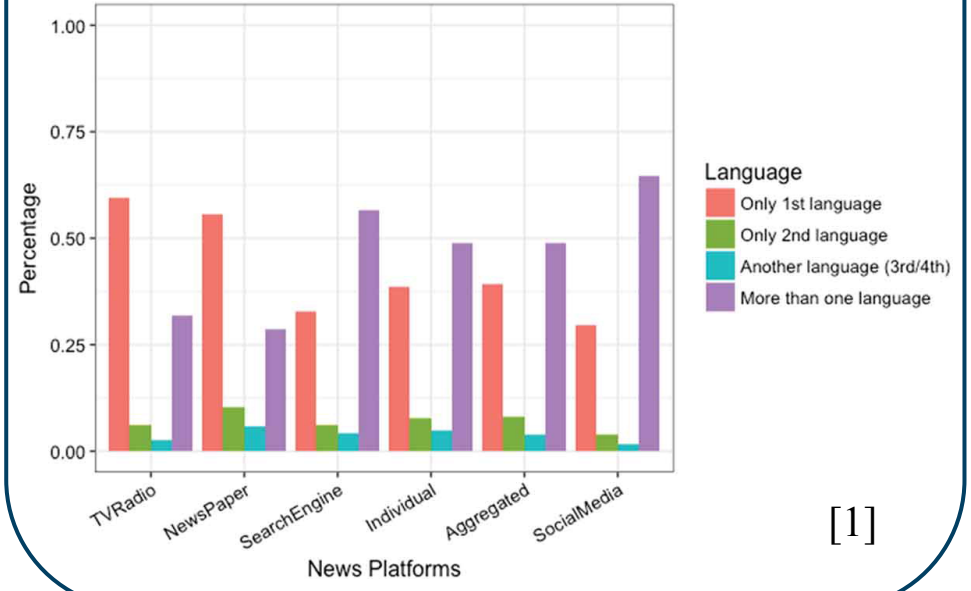
Multilinguality in News & Recommendation

Personalized news recommenders cater the individual needs of readers



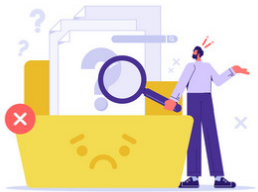


Increasingly **language-diverse** online community & **polyglot** news readers



[1]

Multilinguality in News & Recommendation

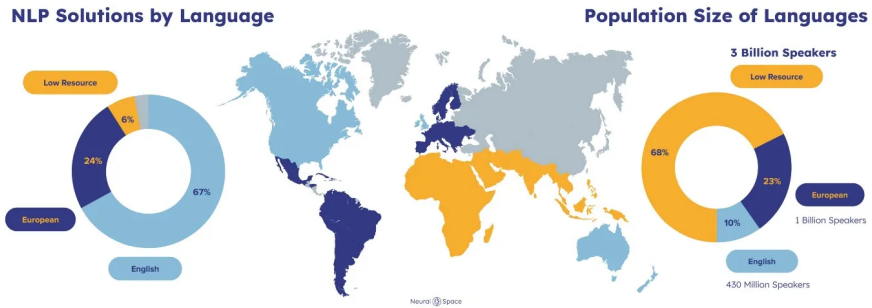


Scarcity of **publicly-available, diverse, multilingual news recommendation datasets**

- Hinders development of efficient multilingual news recommendation and effective cross-lingual transfer to resource-lean languages

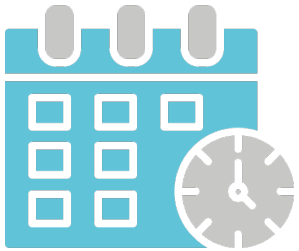
Focus on **monolingual news consumption & high-resource languages**

- Less relevant, less balanced & less diverse recommendations for multilingual news consumers or readers from resource-lean and/or underrepresented languages



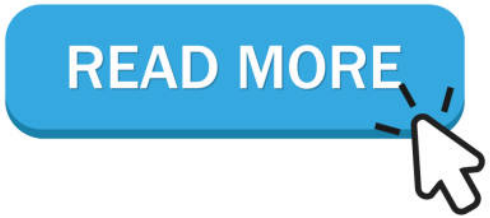
MIND: MIcrosoft News Dataset

Most-used standard benchmark in the news recommendation community [2]

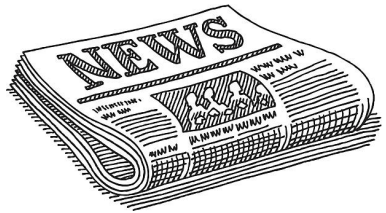


Oct 12th – Nov 22nd 2019

1 million users



> 24 million clicks



130,379 unique articles

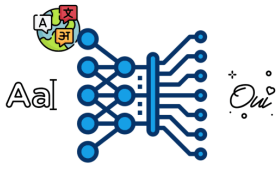
xMIND: A Multilingual News Dataset

Considerations

- Diversity
  - Linguistic
  - Geographic
  - Digital footprint size
- Multi-parallel data
- Open source

Methodology

MIND articles
(titles & headers)



Neural Machine
Translation (NMT)

xMIND

✓ Open-source NMT [3]

✓ Extendable with more languages

xMIND: A Multilingual News Dataset

Code	Language	Script	Macro-area	Family	Genus	Total Speakers (M)	Res.
SWH	Swahili	Latin	Africa	Niger-Congo	Bantu	71.6	high
SOM	Somali	Latin	Africa	Afro-Asiatic	Lowland East Cushitic	22.0	low
CMN	Mandarin Chinese	Han	Eurasia	Sino-Tibetan	Sinitic	1,138.2	high
JPN	Japanese	Japanese	Eurasia	Japonic	Japanesic	1,234.5	high
TUR	Turkish	Latin	Eurasia	Altaic	Turkic	90.0	high
TAM	Tamil	Tamil	Eurasia	Dravidian	Dravidian	86.6	low
VIE	Vietnamese	Latin	Eurasia	Austro-Asiatic	Vietic	85.8	high
THA	Thai	Thai	Eurasia	Tai-Kadai	Kam-Tai	60.8	high
RON	Romanian	Latin	Eurasia	Indo-European	Romance	24.5	high
FIN	Finnish	Latin	Eurasia	Uralic	Finnic	5.6	high
KAT	Georgian	Georgian	Eurasia	Kartvelian	Georgian-Zan	3.9	low
HAT	Haitian Creole	Latin	North-America	Indo-European	Creoles and Pidgins	13.0	low
IND	Indonesian	Latin	Papunesia	Austronesian	Malayo-Sumbawan	199.1	high
GRN	Guarani	Latin	South America	Tupian	Maweti-Guarani	(L1 only) 6.7	low

✓ 14 languages from 13 families

✓ 5 / 6 macro-areas

✓ 5 low-resource languages

✓ 6 scripts from 3 families

✗ US-centric content

✗ Cultural nuances ignored

Applications of xMIND

News Recommendation

- ➔ Efficient multilingual systems
- ➔ Robust cross-lingual transfer methods
- ➔ More language diversity & inclusion of low-resource languages
- ➔ Algorithmic bias analysis across languages
- ➔ Recommenders for polyglots

Beyond Recommendation

- ➔ Multilingual topic classification
- ➔ Democratize access to information for underrepresented communities

Thank you!

Dataset

Email

https://github.com/andreeaiana/xMIND

andreea.iana@uni-mannheim.de

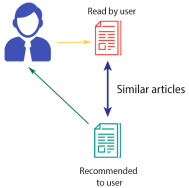
References

[1] Chenjun Ling, Ben Steichen, and Silvia Figueira. 2020. Multilingual news–an investigation of consumption, querying, and search result selection behaviors. *International Journal of Human–Computer Interaction* 36, 6 (2020), 516–535.

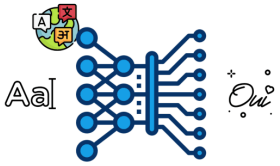
[2] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting ofthe Association for Computational Linguistics*. 3597–3606.

[3] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672* (2022).

Image Sources



https://towardsdatascience.com/brief-on-recommender-systems-b86a1068a4dd



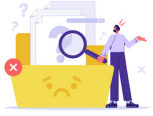
https://pyimagesearch.com/2022/08/15/neural-machine-translation/

https://avenuemail.in/role-of-press-in-democracy/

https://www.pinterest.com/pin/116249234102961966/

https://www.newyorker.com/news/the-future-of-democracy/how-can-the-press-best-serve-democracy

https://stock.adobe.com/de/search?k=multilingual&asset_id=512613046



https://www.vectorstock.com/royalty-free-vectors/no-data-found-vectors

https://medium.com/neuralspace/low-resource-language-what-does-it-mean-d067ec85dea5