

Do you MIND? Reflections on the MIND dataset for research on diversity in news recommendations

Sanne Vrijenhoek

Institute for Information Law, University of Amsterdam
s.vrijenhoek@uva.nl

Abstract. The MIND dataset is at the moment of writing the most extensive dataset available for the research and development of news recommender systems. This work analyzes the suitability of the dataset for research on diverse news recommendations. On the one hand we analyze the effect the different steps in the recommendation pipeline have on the distribution of article categories, and on the other hand we check whether the supplied data would be sufficient for more sophisticated diversity analysis. We conclude that while MIND is a great step forward, there is still a lot of room for improvement.

Keywords: news recommendation · dataset analysis · diversity

1 Introduction

Engagement with the news has dropped drastically over the last years, with interest dropping from 63% in 2017 to 51% in 2022, and a significant number of people avoiding news altogether [5]. News recommender systems may have a role to play in alleviating these issues by correctly identifying a reader’s interest and bringing the right content to the right people. An often-heard criticism on news recommender systems is that they increase the risk of locking users in so-called ‘filter bubbles’, where users are consistently presented with items similar to their preferences or items they have interacted with before. The presence of these filter bubbles and their effects has been hard to prove or disprove, exacerbated by the lack of an exact definition [4]. The in 2020 published MIND dataset [12] is at the moment of writing the largest open source dataset for training and evaluating news recommender systems. It also comes with a set of state-of-the-art news recommender systems that can be trained to predict the articles users will click, and can as such be used to investigate the influence news recommender systems have on the distribution of news content and its diversity, often quoted as the antidote for filter bubbles. Recent work has argued for a normative interpretation of diversity that reflects the role news plays in democratic society [9, 10]. The normative diversity metrics proposed here rely on complex metadata that is not readily available without sophisticated analysis of article content. Furthermore,

they are mostly tailored towards so-called ‘hard’ news. In general, “[F]oreign and domestic politics, economy and finance are usually regarded as hard news. News about sports, celebrities, royal families, crime, scandals and service are regarded as soft news.”[6]. In this regard the MIND dataset comes with a number of caveats. MSN News (rebranded Microsoft Start in September 2021) is a news aggregator, and there is very little information available on what news content makes it onto the platform, and how values such as diversity and inclusivity are balanced with financial gains¹. As the MIND dataset is expected to contain a significant amount of soft news, it may not be directly useful for research into normative diversity, and experiments run on it may come back skewed. To investigate this we study the overall content present in MIND, using the article category, which is directly available in the dataset, as the relevant unit of analysis. By comparing the presence of article categories at different stages of the recommendation pipeline we can analyze both the influence the recommender system has on the distribution of content and the datasets’ suitability for more in-depth news diversity research.

2 Method

MIND contains the interactions of 1 million randomly sampled and anonymized users with the news items on MSN News between October 12 and November 22 of 2019. Each datapoint contains an anonymized user id, the user’s reading history at that point in time, a list of which items were presented to the user (which we here refer to as the ‘candidate list’), and which of these items the user ended up clicking. Wu et al. [12] describe the performance of several news recommender algorithms when trained on this dataset, including news-specific recommendation methods NPA, NAML, LSTUR and NRMS. The recommenders rank each candidate based on the likeliness a user will click on it. Unfortunately, how the items for the candidate list are chosen is not discussed in the paper. The data is split among training-, validation- and test sets. We generate the recommendations by running the code in the supplied notebooks² with the large validation set. In a future iteration of this paper, the analysis will be run on the large test set. In total, 376.471 interactions with the system are recorded here, and on average each candidate list consisted of 37 items. 25% of all interactions had 10 items or less in the candidate list. Close to half (179.383) of the anonymized user ids are unique, with roughly 50.000 and 16.000 ids occurring respectively 2 and 3 times, and 10.000 ids returning more frequently. We assume that user ids are static, and that this means that half of the users only access the site once, and roughly 48.000 visits are from recurring users (>4 times). The average time difference between a user id’s first- and last recording in the system is 6 hours and 22 minutes, and the maximum 23 hours and 20 minutes. This correlates with an important caveat of the validation set: it only contains data from November

¹ according to its Support page “[...] the content we show aligns with our values and [...] crucial information features prominently in our experiences”)

² https://github.com/microsoft/recommenders/tree/main/examples/00_quick_start

Table 1. Rank-biased Overlap (RBO) between different neural recommender strategies. The calculation encompasses the complete ranking list.

	LSTUR	NPA	NAML
NRMS	0,614	0,626	0,746
NAML	0,616	0,639	-
NPA	0,635	-	-

15, 2019. We calculate the overlap between the generated recommendations using Rank-Biased Overlap (RBO) [11], reported in Table 1. This shows a strong overlap in the results of the neural recommenders of $0.61 - 0.63$ between most recommenders, with a much higher score of 0.746 between NRMS and NAML. As Rank Biased Overlap weighs matches at the beginning of the lists more heavily than those at the end, the recommendations should be divergent enough to observe differences in their outcomes. Interestingly, LSTUR and NRMS are reported by [12] to perform best in terms of accuracy, and show comparatively little overlap in Table 1. To avoid redundancy we will only include LSTUR and NRMS in further analysis and comparison with the content in the dataset.

3 Results

The different article categories present at different stages of the recommendation pipeline are counted and averaged, the results of which are displayed in Table 2. For the recommended items, only the top 8 are considered.³ As the goal of the neural recommenders is to predict which items have been clicked, the category distributions for the neural recommenders often resembles the distribution in the ‘clicked’ column.

Table 2 shows a general overview of the distribution of article categories among all the articles that were in the dataset, the result after the first candidate selection, what was in users’ history, and in the set of articles recommended by LSTUR and NRMS. Furthermore, we aggregate the categories present into *hard* and *soft* news following the distinction described in the Introduction. In this dataset, this means that the categories ‘news’ and ‘finance’ are considered hard news, whereas the rest is soft. One major discrepancy can already be observed after candidate selection: the ‘lifestyle’ category, which in the complete dataset only accounts for 4.4% of the articles, has a comparatively big representation (17%) in the set of candidates. The news and sport categories are the most inversely affected, with a 30% and 31% representation in the overall dataset and 23% and 16% after candidate selection. Because the recommender strategies evaluated here have no influence over the candidate selection, this is an important observation to take into account.

³ The dataset also contains a few items with categories ‘kids’, ‘middleeast’ and ‘games’, but as these appear less than 0.1% in the full dataset and never in the recommendations they are left out of the analysis.

Table 2. Distribution of the different article categories (the whole dataset, what was in the users’ reading history, the dataset after candidate selection, and what the user clicked), and the recommender approaches. For the recommendations the top 8 items are selected. The distribution shown does not account for ranking.

	MIND				Recommendations	
	all	candidate	history	clicked	LSTUR	NRMS
hard	0,363	0,302	0,348	0,269	0,261	0,253
soft	0,636	0,698	0,622	0,730	0,739	0,747
news	0,305	0,233	0,279	0,235	0,215	0,224
sports	0,314	0,163	0,142	0,245	0,209	0,207
finance	0,058	0,069	0,069	0,034	0,046	0,029
travel	0,049	0,027	0,030	0,020	0,024	0,021
video	0,045	0,020	0,019	0,021	0,021	0,016
lifestyle	0,044	0,171	0,105	0,178	0,174	0,185
foodanddrink	0,043	0,068	0,050	0,057	0,078	0,077
weather	0,040	0,028	0,012	0,015	0,028	0,021
autos	0,030	0,030	0,036	0,024	0,019	0,019
health	0,028	0,034	0,047	0,034	0,047	0,041
music	0,013	0,044	0,027	0,035	0,042	0,057
tv	0,013	0,047	0,082	0,048	0,046	0,041
entertainment	0,008	0,029	0,034	0,024	0,029	0,034
movies	0,008	0,038	0,038	0,030	0,024	0,028

In general, the two news-specific recommender strategies seem to behave largely similar. Given that the neural recommenders take the items in users’ reading history into account, we would expect the recommenders to reflect similar patterns as the history; however, this does not seem to be the case. On the contrary, while the list of candidate items consisted of 23% news items, and the reading history almost 28%, the neural recommenders are further downplaying the share of news items in the recommendations, containing only about 22% news. The opposite happens for the sport and lifestyle category: where the candidate selection contains 16% sport and 17% lifestyle, and the reading history respectively 14% and 10%, the LSTUR recommender is increasing the presence of these categories to 21% and 17%. It does, however, very closely resemble the distribution of items that users have clicked, which is also not surprising given that this is what the recommender is optimized on.

More interesting patterns can be observed when considering the length of the recommendation, as shown in Table 3. At 1, only the item with the highest predicted relevance is included, continuing on until all items in the recommendations are. At this point, the recommendation is equal to the full candidate list, as ordering is not taken into account in this analysis. The table is ordered on the category’s share in position 1, which is of extra importance given an average user’s tendency towards position bias [2]. Both LSTUR and NRMS are very likely to recommend sports and news at the beginning of a recommendation. Finance only appears much later: despite it’s relatively large presence in both the overall dataset and the candidate list (7% and 6%), finance does not even appear

among NRMS’ top 10 categories. They also both prominently feature category ‘foodanddrink’ in first position (10% and 15%, versus only 7% in the candidate list). But we see also more distinct differences: NRMS comparatively often recommends items from it’s top categories in first place, whereas for LSTUR this is more spread out. At the first position, NRMS recommends more than 83% of the content out of 4 most frequently occurring categories, whereas for LSTUR this is around 74%. NRMS is also much more likely than LSTUR to recommend news in the first position (28% vs. 22%). At position 10, both recommenders actually list *less* news than in the overall dataset. It seems here that news either gets recommended in the top positions, or not at all.

Table 3. Distribution of the top 8 article categories at different recommendation lengths, ordered by frequency at recommendation length 1. Category ‘food’ is short for category ‘foodanddrink’.

		1	2	5	10	20	∞
LSTUR	sports	0,24	0,25	0,23	0,20	0,18	0,16
	news	0,22	0,21	0,21	0,22	0,22	0,23
	lifestyle	0,17	0,17	0,17	0,17	0,17	0,17
	food	0,10	0,09	0,08	0,08	0,07	0,07
	health	0,04	0,05	0,05	0,05	0,04	0,03
	travel	0,04	0,03	0,03	0,02	0,02	0,03
	enter	0,03	0,03	0,03	0,03	0,03	0,03
	finance	0,03	0,04	0,04	0,05	0,06	0,07
NRMS	news	0,28	0,26	0,23	0,22	0,22	0,23
	sports	0,22	0,22	0,22	0,20	0,18	0,16
	lifestyle	0,17	0,18	0,18	0,18	0,18	0,17
	food	0,15	0,13	0,09	0,07	0,07	0,07
	music	0,04	0,06	0,06	0,06	0,05	0,04
	enter	0,03	0,03	0,03	0,03	0,03	0,03
	health	0,02	0,03	0,04	0,04	0,04	0,03
	tv	0,02	0,02	0,03	0,04	0,05	0,05

4 Conclusion

Analyzing the results of the different recommendation strategies reveals characteristics of the recommendations that are not visible when purely reporting on performance statistics such as NDCG or AUC. The neural recommenders have a distinct impact on the dissemination of content, especially considering what content is present in the overall dataset and the type of content users have clicked in the past. As expected, the neural recommenders largely reflect the types of clicks that have been recorded. The candidate list reduces the presence of frequently occurring categories while inflating that of less frequent ones. However, this behavior is to be expected; the candidate selection ought to contain a wide range

of content, so that the recommender system can correctly identify content that is specifically relevant to that particular user. It does however raise questions about the granularity of the categories chosen. In the design of MIND the choice was made to distinguish between ‘movies’, ‘music’, ‘tv’ and ‘entertainment’, even though these account for less than 10% of all items in the dataset. A dataset that is more focused on news content could instead split this top-level category into subcategories such as ‘local news’, ‘global news’ or ‘politics’. MIND does contain subcategories, such as ‘newsus’, ‘newspolitics’, and ‘newsworld’ (respectively 47%, 17% and 8% of all news items), which could be more relevant for future research.

The neural recommenders also behave differently when compared to each other, with NRMS prominently recommending news and food in top positions, and LSTUR favoring sports and other, less common categories. With the neural recommenders largely focusing on lifestyle and entertainment, and downplaying news and finance, one could argue they mostly promote soft news. This is not to say these personalized recommendations are bad; there can be value in bringing the right soft news to the right people, as Andersen [1] notes that consuming soft news may serve as a stepping stone to more active participation and engagement. This does warrant a more in-depth discussion about the purpose of the recommender system, a thorough investigation into the mismatch between produced content, user reading history and user clicking behavior, and an editorial decision on the balance between ‘quality’ and ‘fun’ [8].

In terms of research into normative diversity, MIND leaves a few things to be desired. With only 20% of articles in the recommendations being news articles, there is only little information to determine whether users receive a balanced overview of the news. This is strengthened by the lack of metadata that is present in the dataset: only the article title, (sub)category and url are directly supplied. Automatic stance- or viewpoint detection based on article fulltext, which could be retrieved by following the url to the MSN News website, may be a direction for future research [7]. For example, Mascarell et al. [3] published a detailed annotation of different stances and emotions present in German news articles. They do, however, lack the scale and user interactions that MIND has.

The majority of interactions recorded in MIND are (assumed to be) unique visits, though it does contain a considerable amount of returning users: almost 10.000 access the system 4 times or more, resulting in a total of 48.000 visits from recurring users. If we combine this with the average length of the candidate list of 37 items, and the fact that 22% of recommended items is news, this yields us about 400.000 news items shown. However, even when the users return to the system more frequently, the validation set only contains information on the interactions users had with the system on one specific day, making it impossible to see how the users and the recommender’s behavior towards those users change over time [4]. While the large test set does contain data over six days rather than just one, this would still not be enough to actually see differences in users’ behavior, even if they do use the system intensively. Ideally, if one were to research the effect of a recommender system on the diversity of consumed

news, they would want to do this based on a system with 1) a large number of frequently returning users (though a smaller number of unique users compared to MIND would be acceptable), 2) a focus on hard news, and 3) over a longer period of time, allowing for both the users and the recommender system to evolve over time. In conclusion: the MIND dataset is, especially given the fact that it is open source, a great step forward in the research on news recommender systems and their effects. However, when the goal is to move the discussion beyond recommender accuracy and towards news recommender diversity, there are still several points of improvement necessary.

Acknowledgements

I thank Mateo Gutierrez Granada for his help in generating the recommendations used in this analysis. I also thank Savvina Daniil, Lien Michiels and an anonymous reviewer for their critical comments on earlier versions of this work, and in doing so their contributions to improving the end product.

Bibliography

- [1] Andersen, K.: An entrance for the uninterested: Who watches soft news and how does it affect their political participation? *Mass Communication and Society* **22**(4), 487–507 (2019)
- [2] Craswell, N., Zoeter, O., Taylor, M., Ramsey, B.: An experimental comparison of click position-bias models. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*, p. 87–94, WSDM '08, Association for Computing Machinery, New York, NY, USA (2008), ISBN 9781595939272, <https://doi.org/10.1145/1341531.1341545>, URL <https://doi.org/10.1145/1341531.1341545>
- [3] Mascarell, L., Ruzsics, T., Schneebeil, C., Schlattner, P., Campanella, L., Klingler, S., Kadar, C.: Stance detection in german news articles. In: *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pp. 66–77, Association for Computational Linguistics (2021)
- [4] Michiels, L., Leysen, J., Smets, A., Goethals, B.: What are filter bubbles really? a review of the conceptual and empirical work. In: *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 274–279 (2022)
- [5] Newman, N., Fletcher, R., Robertson, C.T., Eddy, K., Nielsen, R.K.: Reuters institute digital news report 2022. Reuters Institute for the study of Journalism (2022)
- [6] Reinemann, C., Stanyer, J., Scherr, S., Legnante, G.: Hard and soft news: A review of concepts, operationalizations and key findings. *Journalism* **13**(2), 221–239 (2012)
- [7] Reuver, M., Mattis, N., Sax, M., Verberne, S., Tintarev, N., Helberger, N., Moeller, J., Vrijenhoek, S., Fokkens, A., van Atteveldt, W.: Are we human, or are we users? the role of natural language processing in human-centric news recommenders that nudge users to diverse content. In: *Proceedings of the 1st Workshop on NLP for Positive Impact*, pp. 47–59, Association for Computational Linguistics, Online (Aug 2021), <https://doi.org/10.18653/v1/2021.nlp4posimpact-1.6>, URL <https://aclanthology.org/2021.nlp4posimpact-1.6>
- [8] Smets, A., Hendrickx, J., Ballon, P.: We're in this together: A multi-stakeholder approach for news recommenders. *Digital Journalism* pp. 1–19 (2022)
- [9] Vrijenhoek, S., Bénédict, G., Gutierrez Granada, M., Odijk, D., De Rijke, M.: Radio-rank-aware divergence metrics to measure normative diversity in news recommendations. In: *Proceedings of the 16th ACM Conference on Recommender Systems*, pp. 208–219 (2022)
- [10] Vrijenhoek, S., Kaya, M., Metoui, N., Möller, J., Odijk, D., Helberger, N.: Recommenders with a mission: assessing diversity in news recommendations. In: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pp. 173–183 (2021)

- [11] Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)* **28**(4), 1–38 (2010)
- [12] Wu, F., Qiao, Y., Chen, J.H., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., Gao, J., Wu, W., Zhou, M.: Mind: A large-scale dataset for news recommendation. *ACL* (2020)