HERMES

D3.1
Validated Trustworthy RS methods for Multi-objective and multi-stakeholder setting

Project Title	Trustworthy Multi-Objective and Multi-Stakeholder Recommenders
Contract No.	09103-03-V04-00336
Project start date	1.7.2024
Duration	24 months

Funded by the European Union
NextGenerationEU

PLÁN OBNOVY

Grant agreement no.:	09I03-03-V04-00336
Project acronym:	HERMES
Project website:	https://kinit.sk/project/hermes
Project full title:	Trustworthy Multi-Objective and Multi-Stakeholder Recommenders
Project start date:	July 2024 (24 months)
Work Package:	WP3 – Multi-objective and multi-stakeholder RS
Version:	1.0
Authors:	Michal Kompan, Marek Havrila, Martin Olejnik, Matus Tuna

Project funded by VAIA - Research and innovation authority, under the call 09I03-03-V04, Grant agreement no. PPM_09I03-03-V04-00336

Dissemination Level		
PU	Public	x
NP	Non-public, only for members of the consortium (including the Agency Services)	

Table of Contents

1. Introduction..... 4
2. Methods and datasets for enhancing data for fairness and diversity..... 5
3. SMORL4RS..... 10
4. Method for Off-Policy Evaluation..... 12
5. TUNNER: Trustworthy Unbiased Neural Recommender..... 14
6. Conclusions.....17

1. Introduction

This deliverable summarizes the methods and datasets assessed, developed and evaluated with the HERMES project. It leverages the major outcomes and findings, while the detailed results are presented in the scientific publications prepared during the project lifetime.

2. Methods and datasets for enhancing data for fairness and diversity

To demonstrate the methods for fairness and diversity usually the news datasets are used while the task is formulated to deliver balanced news - in terms of the political leaning. One significant part of the project was therefore dedicated to exploring methods for enhancing such datasets with the labels, which are pre-requisite for the evaluation step. We labeled a subset of two datasets (MIND and Reddit) in terms of political leaning of the news / posts. In order for the labels to be trustworthy, we have tried multiple models for labeling the entire datasets, including models available on huggingface which have been trained on different datasets for this task by other researchers. We also explored using an LLM for this task, where we evaluated different prompts for one-shot classification as well as a prompt where we allowed the LLM to generate few sentences to analyze the text and assign the label based on its own analysis - this method is more token and time heavy, but produces better results than the one-shot classification and averaging of one-shot classificators.

For MIND dataset, we considered a one-shot classification using LLM with various prompts, and a prompt that lets the LLM analyze the text in order to come to a conclusion about its label. After these experiments we evaluated the results on their own, but we also looked at averaging multiple one-shot classification results in order to make better predictions, since it is less token heavy than the variant of the LLM analyzing text at first. However we came to the conclusion that for the MIND dataset, the best performing model was the one that used prompt to analyze the dataset first and then draw the conclusion. We also used multiple models for prediction of political leaning of a text from hugging face, as the closed source LLM we used gpt-4.1.

For the case of the Reddit dataset, as this dataset is far bigger than MIND, we had to take into consideration also the financial and time complexity of this task using closed source models. We already ran the labeling for the MIND dataset, so we knew one-shot classification using LLM would most likely not give great results. Due to this, we decided to look more into the models from huggingface, as well as some open source LLMs. Nevertheless we also used gpt-4.1 to evaluate this model on the subset dataset, to see the performance difference if we opted to use this model. For the open source models we tried the Gemma-4-E2B instruct with 4bit quantization provided from unsloth library (gemma-4-E2B-it-unsloth-bnb-4bit), as well as Qwen3 which turned out to be quite poor in generating reasonable texts, so we scrapped experiments with it.

Subset dataset labeling methodology

We considered the political leaning of a text to be a categorical problem with 4 options to choose from - left, right, neutral, nonpolitical. The manual classification process worked in two stages. In the first stage we decide whether the text has anything to do with politics, if not then we label it as “nonpolitical”. If it does have anything to do with politics, then we label it either “left”, “right” or “neutral”. Thus the “neutral” label assigned to text that does have some political element to it, but does not favor either left or right side.

For both of these datasets we employed 3 experts who labeled the sampled texts, independently of each other. Afterwards, we averaged the result such that if a label was

chosen 2 or more times, that is its final label. For cases when no category was chosen twice, we just randomly chose from those categories that were chosen. We also calculated the Cohen's kappa between the labelers, which is a statistic used to measure inter-rater reliability for qualitative or categorical data.

Cohen's kappa between labeler x&y	1&2	1&3	2&3
MIND	0.501	0.451	0.576
Reddit	0.661	0.662	0.687

The higher values for Reddit dataset can be explained by our sampling technique, where we intentionally sampled the texts, for example 100 texts are from "nonpolitical" subreddits, and each expert labeled at least 99 of those as nonpolitical. If we ignored those 100 "nonpolitical" texts, the average Cohen's kappa is 0.451

MIND - This dataset already has category and subcategory features. We used the subcategory feature in order to filter only the political news text for which we only had to consider 3 categories - left, right, neutral - as everything else was from subcategory that was not political so we assume=assign that the label for those is nonpolitical.

Reddit - For this dataset we do not have such subcategories to use for filtering, so we manually chose some of the subreddits that looked like they could have lots of political posts (such as "politics", "The_Donald", "EnoughTrumpSpam", "AskTrumpSupporters", "SandersForPresident", "PoliticalDiscussion", "PoliticalHumor", "Conservative") for which we considered the labels of left,right and neutral. Then we picked other subreddits where we considered left,right,nonpolitical as labels, so that our final dataset would have all 4 labels represented. From the political subreddits we picked 300 texts, sampled based on the corresponding subreddit occurrence of the entire dataset, from the nonpolitical subreddits we picked 100 texts, also sampled based on the occurrence of each subreddit in the entire dataset.

Huggingface models evaluated:

- https://huggingface.co/mlburnham/Political_DEBATE_large_v1.0
- https://huggingface.co/matous-volf/political-leaning-deberta-large
- https://huggingface.co/mlburnham/deberta-v3-large-polistance-affect-v1.1
- https://huggingface.co/zhezhou1106/political-leaning-classifier
- https://huggingface.co/matous-volf/political-leaning-politics
- https://huggingface.co/Arstacity/political-bias-classifier
- https://huggingface.co/alxdev/echocheck-political-stance
- https://huggingface.co/kartiksrma/roberta-political-ideology-classifier
- https://huggingface.co/bucketresearch/politicalBiasBERT

Some of the results

Below we present the confusion matrices for the models we used for labeling. In the case of the Reddit dataset, only some models are able to predict the "nonpolitical" label, which is what we care about, nevertheless we also present the results if we decided to ignore the

nonpolitical texts. From the confusion matrix and the F1 score printed below the matrix, we can see that the best model for Reddit is “gpt-4.1”, as it has the highest F1 score. Also the accuracy is highest for this model, as it has correctly labeled 241 texts out of 400 (234 for political_debate_large_v1.0, 225 for deberte-v3-large-polistance-affect-v1.1, 223 for gemma-4-E2B-it-unsloth-bnb-4bit). However due to the model being closed source and the inference being quite slow, we opted to use the second best model which is political_debate_large_v1.0.

The way to read the confusion matrix is as follows:
Look at the first row = “left”, the numbers in this row correspond to how the “left” ground truth labels were predicted. So in the first column we have 15, this column corresponds to “left” predicted label, meaning 15 of the ground truth labels that are “left” were also predicted as “left”, next move on to the “neutral” column, where we see the value of 4, meaning that 4 of the actual left ground truth labels were predicted to be neutral and so on.

We can notice that large values are located in the “neutral” row with the “nonpolitical” column, which indicates that the ground truth label was neutral, but the models labeled it as nonpolitical. This is easily explained by the way we labeled our dataset, where we intentionally did not allow the "nonpolitical" label for the texts from the political subreddits, which there were 300 of.

Similar results were obtained for the MIND dataset, where also the best results were from “gpt-4.1” when it was allowed to first analyze the text and then assign a label. In this case, since the MIND “political” subcategory dataset is quite small, we decided to use this model for labeling.

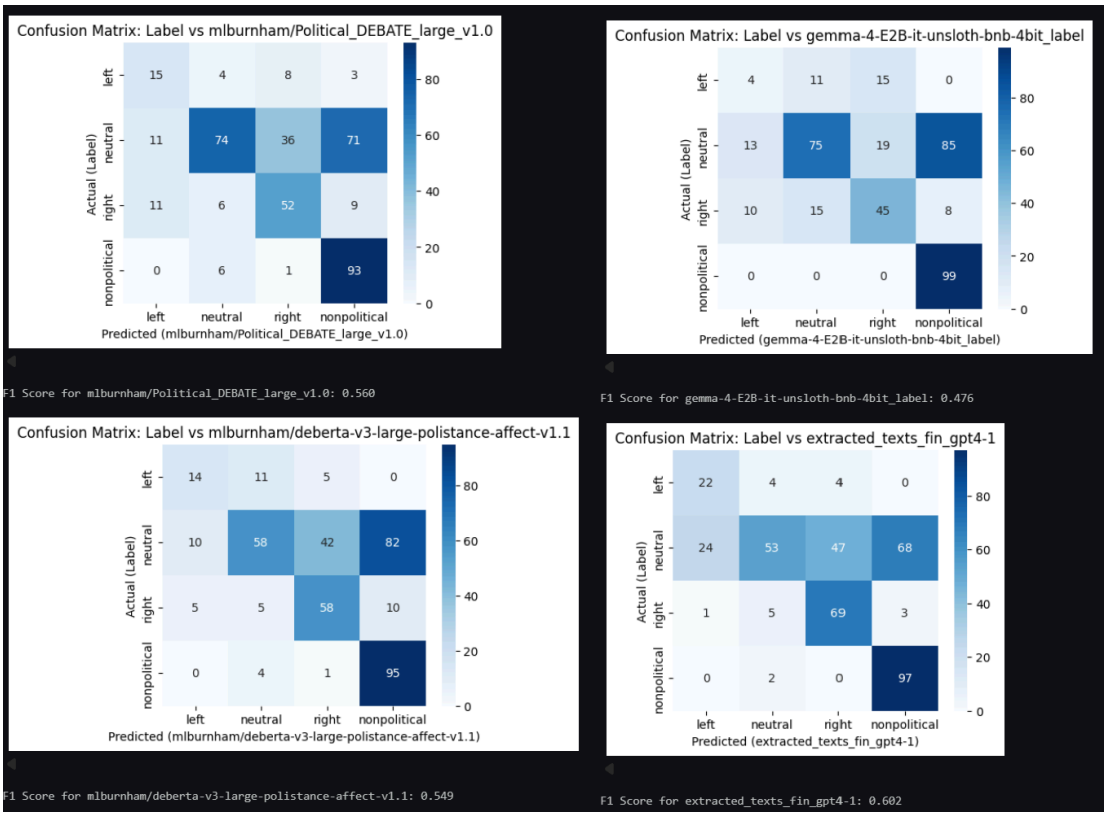


Figure Reddit: Considering all 4 labels - including nonpolitical.

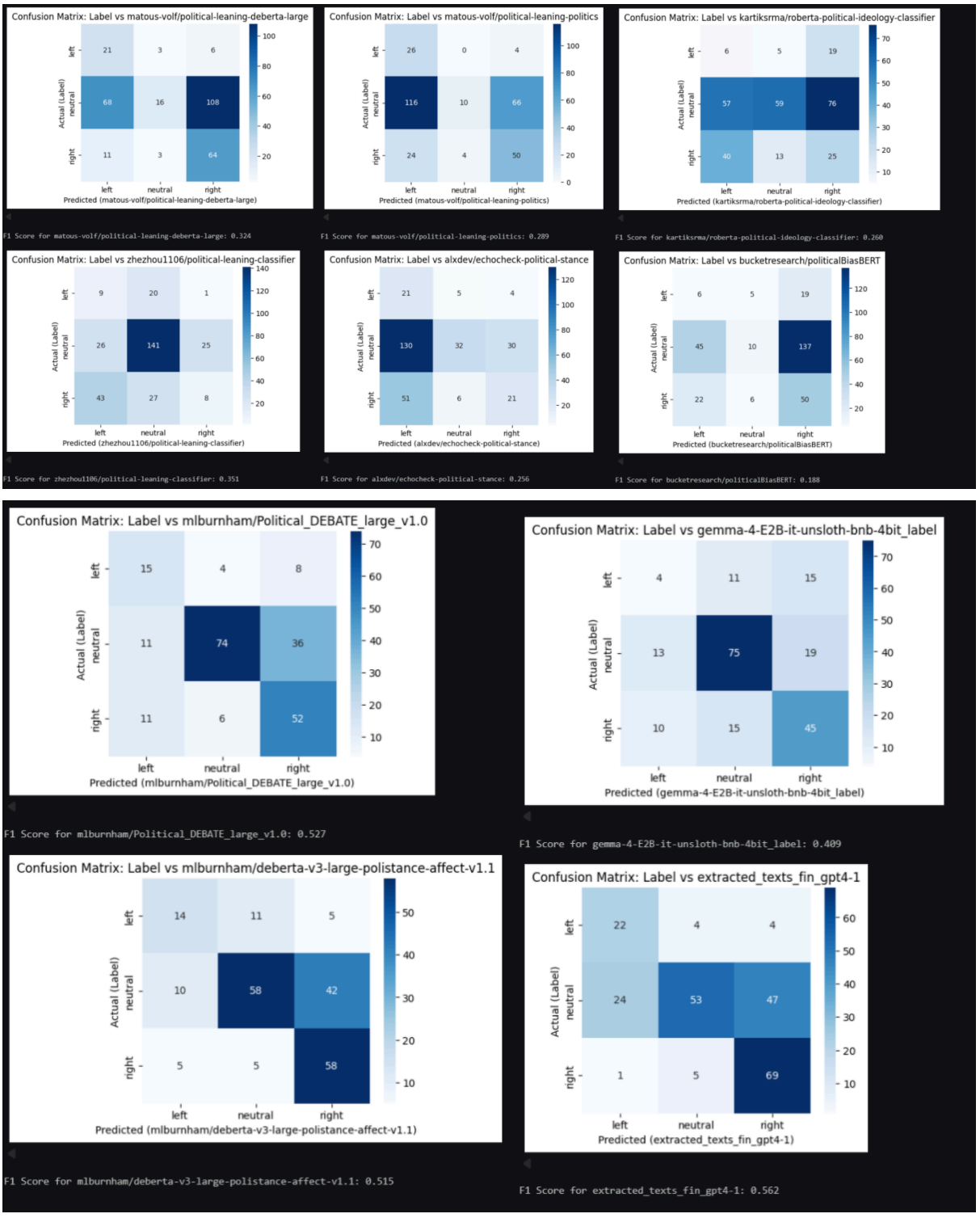
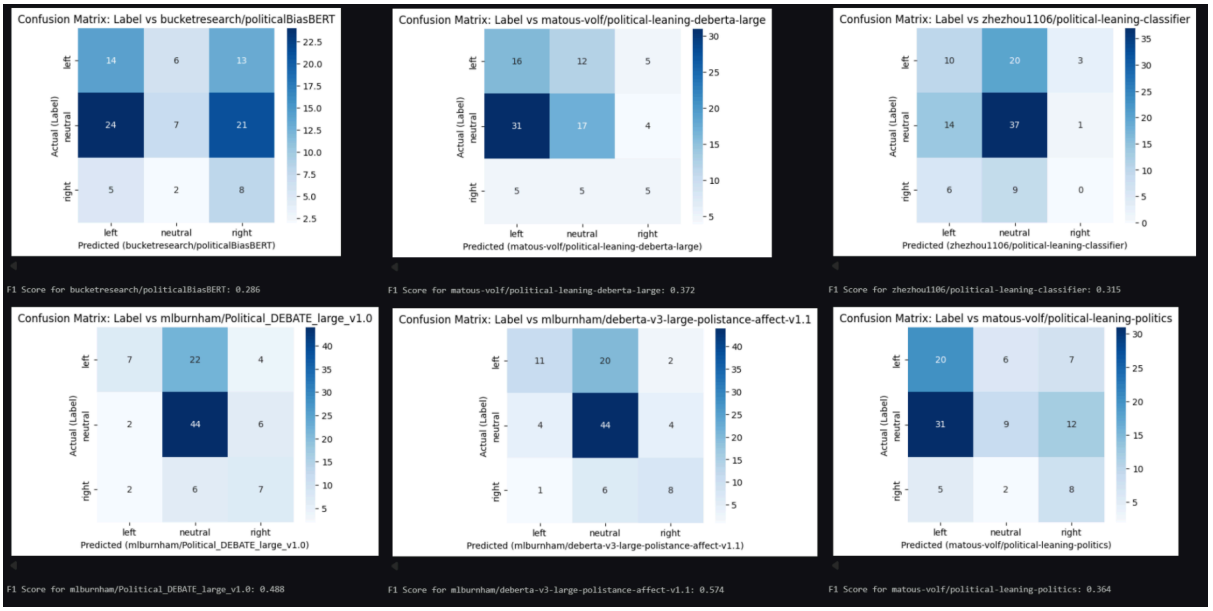


Figure Reddit: Results when ignoring nonpolitical texts.



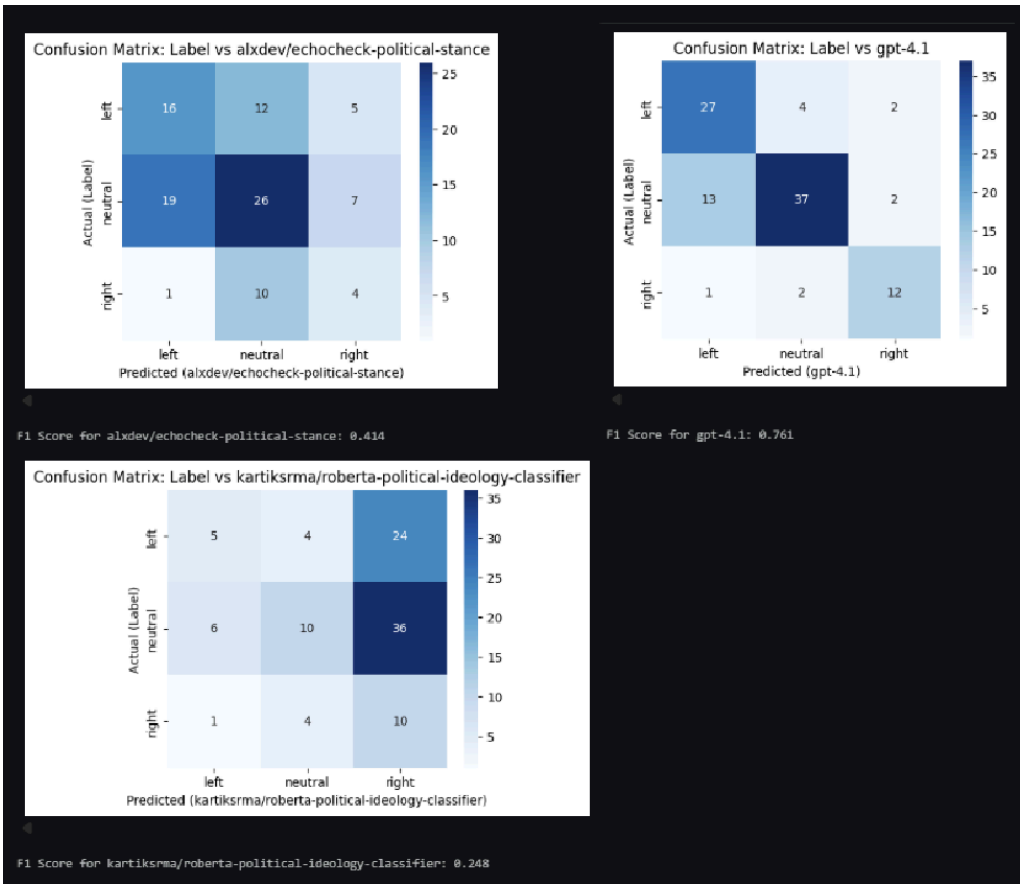


Figure MIND: Results when ignoring nonpolitical texts.

3. SMORL4RS

Scalarized multi-objective reinforcement learning (SMORL[1]) is an approach for training reinforcement learning models in settings where several objectives should be optimized simultaneously. Each of the objectives, such as recommendation accuracy, diversity, novelty or fairness, is represented by an objective-specific reward function and Q-value function. The individual rewards are combined by a scalarization function into one aggregated training signal, which allows optimization using standard reinforcement learning algorithms.

SMORL4RS[2], proposed by Stamenković et al., applies this principle to sequential recommendation. The method builds on an existing neural next-item prediction model (base model) and augments it with a SMORL component. The base model processes a user interaction sequence and produces a hidden representation of the current user state. This representation is used both by the standard self-supervised prediction head, trained with cross-entropy loss, and by the SMORL component, which provides an auxiliary regularization signal for the base recommender.

The SMORL component is thus used only during training. Recommendations are generated by the trained base model through its ordinary self-supervised prediction head. In this way, SMORL4RS aims to transfer the multi-objective training signal into the parameters of the underlying recommender model without increasing inference complexity.

This combination aims to address a practical limitation of standalone reinforcement learning in implicit-feedback recommendation, where observed interactions provide positive signals but reliable negative feedback is usually missing. The self-supervised next-item prediction objective stabilizes learning with a strong accuracy-oriented signal, while the reinforcement learning component allows additional objectives to be expressed in flexible mathematical form. Moreover, changing weights of individual objectives during scalarization should in theory allow control over the extent to which each objective is emphasized.

Within the project we have focused on exploring the method and correcting issues and methodological flaws present in the method. The corrected implementation[3] improves the original SMORL4RS codebase by addressing issues in both the core learning mechanism and the evaluation pipeline. First, we fixed a major implementation error in the accuracy reward function, where all actions were rewarded instead of only the correct action. Second, we improved data preparation and evaluation by removing padding-only input sequences introduced by the original sliding-window procedure, as these introduce substantial noise into the evaluation. Together, these corrections provide a more reliable basis for assessing the original method, which reinforced accuracy, diversity, and novelty. The implementation

[1] Hossam Mossalam, Yannis M. Assael, Diederik M. Roijers, and Shimon Whiteson. 2016-10-09. Multi-Objective Deep Reinforcement Learning. arXiv:1610.02707 doi:10.48550/arXiv.1610.02707

[2] Dusan Stamenkovic, Alexandros Karatzoglou, Ioannis Arapakis, Xin Xin, and Kleomenis Katevas. 2022-02-11. Choosing the Best of Both Worlds: Diverse and Novel Recommendations through Multi-Objective Reinforcement Learning. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (Virtual Event AZ USA). ACM, 957–965. doi:10.1145/3488560.3498471

[3] https://github.com/kinit-sk/sqn-smorl-reproducibility

was further complemented by a Pareto-aware evaluation perspective that considers the whole training process rather than relying only on a single selected evaluation point.

Evaluation of corrected SMORL4RS method

Evaluation of the corrected SMORL4RS implementation revealed two main findings: the method does not reliably control individual recommendation objectives, and its previously reported positive results are strongly affected by the choice of evaluation procedure. In particular, a more appropriate Pareto-aware evaluation does not indicate consistent improvement in beyond-accuracy objectives across the evaluated base methods.

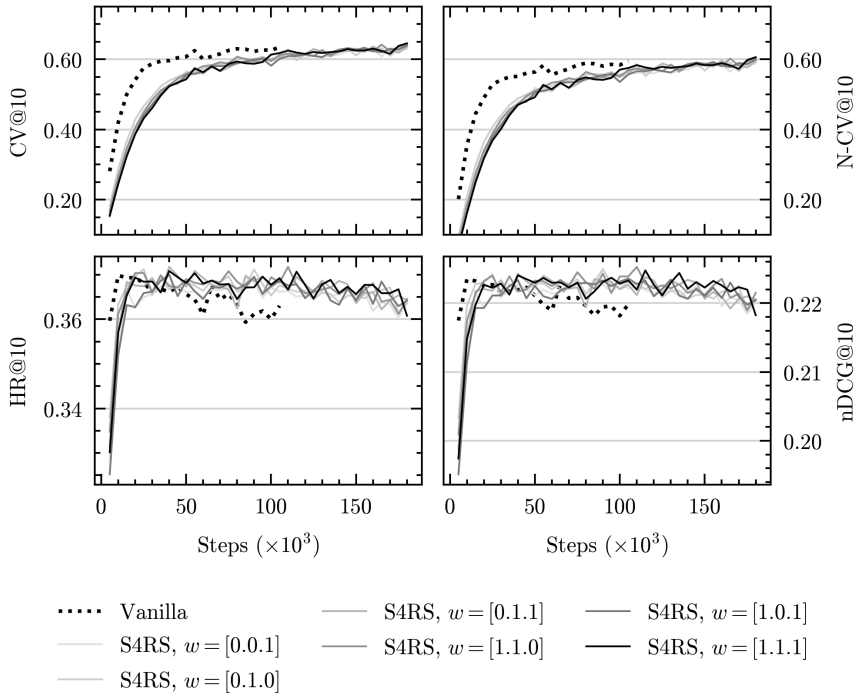


Figure Weights: Inability of SMORL4RS method to control individual objectives for Caser-SMORL4RS models on RC15 dataset. Plot shows development of utilized metrics for six different combinations of reinforced objectives defined by weight vector w = [w_{acc} , w_{div} , w_{nov}], where w_{acc} , w_{div} , w_{nov} are scalarization weights of accuracy, diversity, and novelty. Note, that manipulating with w does not lead to significant systematic difference in measured metrics, and that diversity and novelty values of S4RS models are not better than vanilla.CV@10 stands for diversity measured as catalog coverage at 10; N-CV@10 stands for novelty measured as coverage of less-popular catalog items at 10; S4RS stands for SMORL4RS.

The first major finding concerns the limited ability of SMORL4RS to control individual recommendation objectives. Changing the scalarization weights of accuracy, diversity, and novelty should influence the extent to which the corresponding objectives are optimized. However, our experiments did not confirm this behavior. Even after correcting the major implementation error in the accuracy reward function, changing the weights of individual objectives did not lead to systematic and interpretable changes in the corresponding evaluation metrics (***Figure Weights***). The method was therefore not able to reliably distinguish between the identity of reinforced objectives. Instead, the observed behavior was explained mainly by the overall strength of the reinforcement learning loss relative to the standard self-supervised next-item prediction loss.

The second major finding concerns the evaluation procedure. The positive conclusions in the original work appear to be strongly influenced by an overly simplistic evaluation procedure based on selecting a single evaluation point. In a multi-objective setting, and especially in multi-stakeholder recommender systems, such evaluation is insufficient because different objectives may improve at different stages of training and may involve trade-offs rather than simultaneous improvement. A model selected only according to one criterion, such as accuracy, can therefore give a misleading impression of improvement in other objectives.

We therefore complemented the corrected implementation with a Pareto-aware evaluation. Instead of evaluating the method only at a single selected checkpoint, this perspective considers the development of multiple metrics across the whole training process and examines whether the method provides a better trade-off between objectives. Under this evaluation, we found no evidence that SMORL4RS reliably improves or controls diversity and novelty as intended. We further show that a small trade-off in accuracy may lead to considerable improvements in beyond-accuracy metrics even for the base methods, making their performance comparable to the more complex SMORL4RS approach (***Table Trade-off***). These results suggest that multi-objective recommender systems should be evaluated through objective trade-offs rather than through isolated metric values at a single selected point. It also shows a prominent example of phantom-progress in the field of recommender systems.

Model	Dataset											
	RC15						RetailRocket					
	Accuracy (nDCG)			Diversity (CV)			Accuracy (nDCG)			Diversity (CV)		
	@10	$@10_{95}$	Δ%	@10	$@10_{95}$	Δ%	@10	$@10_{95}$	Δ%	@10	$@10_{95}$	Δ%
GRU4Rec	0.245	0.235	−4.1	0.523	0.624	19.1	0.220	0.220	0.0	0.573	0.573	0.0
GRU4Rec-S4RS	0.251	0.245	−2.5	0.487	0.552	13.5	0.237	0.230	−2.7	0.560	0.625	11.6
Caser	0.223	0.220	−1.6	0.420	0.635	51.4	0.188	0.181	−3.7	0.508	0.582	14.7
Caser-S4RS	0.225	0.218	−2.9	0.619	0.645	4.3	0.207	0.200	−3.2	0.589	0.642	8.9
SASRec	0.263	0.260	−1.3	0.654	0.683	4.4	0.231	0.223	−3.6	0.597	0.653	9.4
SASRec-S4RS	0.270	0.268	−0.8	0.622	0.652	4.9	0.246	0.238	−3.3	0.553	0.634	14.6
NextItNet	0.248	0.243	−2.3	0.567	0.644	13.5	0.228	0.228	0.0	0.644	0.644	0.0
NextItNet-S4RS	0.256	0.252	−1.6	0.524	0.667	27.4	0.236	0.226	−4.3	0.689	0.713	3.5

Table Trade-off: Comparison of the highest-accuracy models (*@10*) with models selected for higher diversity ($@10_{95}$) on the RC15 and RetailRocket datasets. Shifting the best-model selection criterion from a purely accuracy-oriented strategy to a Pareto-aware strategy (by allowing the selection of models that retain at least 95% of the maximum achieved accuracy while maximizing diversity) leads to considerably higher diversity with only a small loss in accuracy. The table is based on data published by Paparella et al. NG and CV stand for nDCG and catalog coverage, respectively. S4RS stands for SMORL4RS.

4. Method for Off-Policy Evaluation

Off-policy evaluation (OPE) estimates how a policy would perform without deploying it, which is crucial in high-risk domains like recommender systems, advertising, or healthcare where

online testing can be costly or harmful. While inverse propensity scoring (IPS) provides unbiased estimates, it suffers from high variance when the logging policy rarely chooses certain actions, motivating lower-variance alternatives that rely on hyper-parameters but lack principled tuning methods. Unlike supervised learning where cross-validation thrives thanks to unbiased samples, OPE has long been assumed incompatible with such data-driven model selection because rewards are biased by the logging policy. Our work challenges that belief by showing that cross-validation is feasible in OPE: using IPS on held-out data yields unbiased estimates of a target policy's value, enabling reliable comparison of estimators. The authors introduce a simple cross-validated estimator-selection procedure requiring only one logged dataset, analyze its variance, propose improvements, and demonstrate across real-world datasets that it outperforms prior methods while remaining computationally efficient.

We present our method in Algorithm 1. It works as follows. First, we estimate the variance of the validation estimator V~ (line 2). Second, we estimate the variance of each evaluated estimator class $\mathscr{V} \in \mathbf{V}$ (line 4). Third, we repeatedly split $\mathscr{D}$ into the training and validation sets (line 6), estimate the policy value with V^ on the training set (line 7), and calculate the loss against the validation estimator V~ on the validation set (line 8). Finally, we select the estimator V^∗ with the lowest one-standard-error upper bound on the estimated loss (line 12). We call our algorithm Off-policy Cross-Validation and abbreviate it as OCV.

Algorithm 1: Off-policy evaluation with cross-validated estimator selection.

1: **Input:** Evaluated policy π, logged dataset $\mathcal{D}$, set of estimators $\mathbf{V}$, number of random splits K
2: $\tilde{\sigma}^2 \leftarrow$ Empirical estimate of var $\left[\tilde{V}(\pi; \mathcal{D})\right]$
3: **for** $\hat{V} \in \mathbf{V}$ **do**
4: $\quad \hat{\sigma}^2 \leftarrow$ Empirical estimate of var $\left[\hat{V}(\pi; \mathcal{D})\right]$
5: $\quad$ **for** $k = 1, \ldots, K$ **do**
6: $\quad\quad \hat{\mathcal{D}}_k, \tilde{\mathcal{D}}_k \leftarrow$ Split $\mathcal{D}$ such that $|\hat{\mathcal{D}}_k|/|\tilde{\mathcal{D}}_k| = \hat{\sigma}^2/\tilde{\sigma}^2$
7: $\quad\quad L_{\hat{V},k} \leftarrow (\tilde{V}(\pi; \tilde{\mathcal{D}}_k) - \hat{V}(\pi; \hat{\mathcal{D}}_k))^2$
8: $\quad$ **end for**
9: $\quad \bar{L}_{\hat{V}} \leftarrow \frac{1}{K} \sum_{k=1}^{K} L_{\hat{V},k}$
10: **end for**
11: $\hat{V}_* \leftarrow \arg\min_{\hat{V} \in \mathbf{V}} \bar{L}_{\hat{V}} + \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} (L_{\hat{V},k} - \bar{L}_{\hat{V}})^2}$
12: **Output:** $\hat{V}_*(\pi; \mathcal{D})$

Results

Cross-validation consistently chooses a good estimator

Figure Selector demonstrates that our methods consistently avoid choosing the worst-performing estimator and, on average, outperform both Slope and PAS-IF. OCVDR stands out by achieving significantly better results than all alternatives on two datasets, while never performing significantly worse. Slope behaves reasonably well here because the bias–variance assumptions it relies on are satisfied. In contrast, PAS-IF shows a systematic bias toward selecting DM even when DM performs poorly. We hypothesize that PAS-IF's data-splitting strategy (driven by a learned neural network) introduces bias into its validation

estimates. A biased validation estimator tends to favor similarly biased estimators, making it unreliable for robust estimator selection.

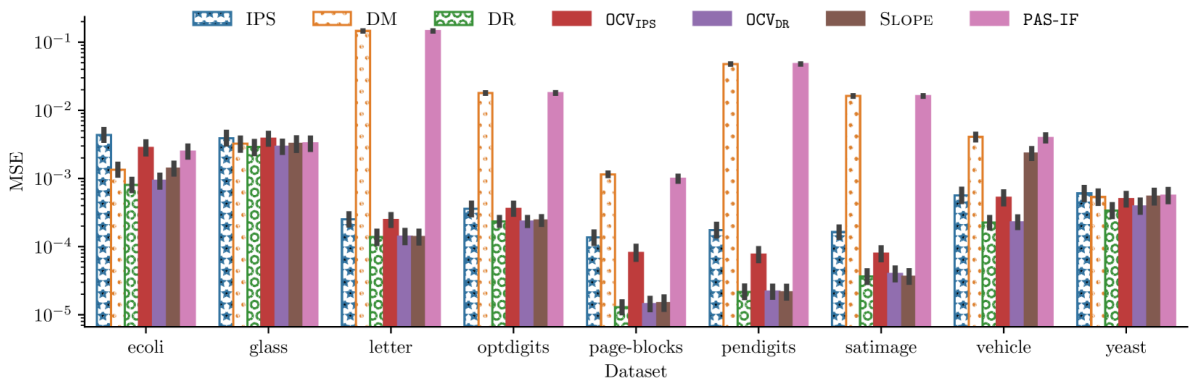


Figure Selector: MSE of our estimator selection methods, OCVIPS and OCVDR, compared against two other estimator selection baselines, Slope and PAS-IF. The methods select the best estimator out of IPS, DM, and DR. We report 95% confidence intervals for all results.

Our work introduces a simple, data-driven procedure for estimator selection and hyper-parameter tuning in off-policy evaluation by adapting cross-validation, effectively narrowing the long-standing gap between OPE and supervised learning. Whereas OPE has traditionally relied on theory-based methods due to the unknown value of the target policy, we show that this obstacle can be overcome by applying an unbiased estimator to held-out data - mirroring how supervised learning uses samples from an unknown distribution. Across nine real-world datasets and both estimator-selection and tuning tasks, their approach proves broadly applicable, easy to implement, familiar to practitioners, and consistently superior to existing techniques. They highlight promising future directions, including extending the method to off-policy learning (where hyper-parameters must generalize across all policies) and to reinforcement learning, though naïve worst-case tuning may lead to overly conservative choices.

5. TUNNER: Trustworthy Unbiased Neural Recommender

TUNNeR is a multiobjective trustworthy recommender system for news domain, based on MANNeR[4] recommender system. MANNeR implements a multiobjective recommender system as two separately trained modules called CR-Module and A-Module (***Figure MANNER***). CR-Module (Content Relevance) predicts the most relevant item from a candidate pool of news items based on the click history of a given user. A-Module (Aspect) is trained separately to predict the category of given candidate news. Both CR-Module and A-Module are trained using Supervised Contrastive Loss (SCR). Final classification is implemented as a nonparametric layer that weights the content relevance logits and individual aspects logits and sums them into final prediction that considers individual users' relevance, as well as individual aspects of news.

[4] Iana, A., Glavaš, G., & Paulheim, H. (2024, November). Train once, use flexibly: A modular framework for multi-aspect neural news recommendation. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 9555-9571).

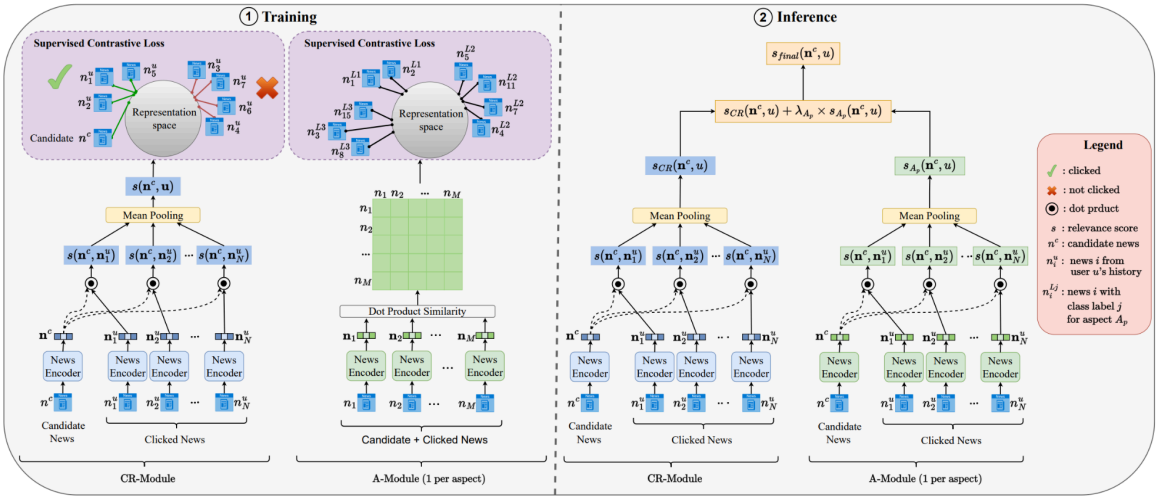


Figure MANNER: Illustration of the MANNeR framework.

These aspects might include categorical diversity, sentiment diversity, political leaning diversity and various other aspects that are important to serve diverse and balanced news. Our TUNNeR model improves upon existing model by:

- precomputing news embeddings by fixed LLM model which saves computational resources,
- because we use fixed LLM model for computing news embeddings we can use much smaller transformer-based CR module instead of finetuning large general-purpose LLM model which further reduces computational requirement and might improve diversity of recommendations,
- multilayer perceptron-based A-Module instead of finetuned LLM-based module,
- introducing batch class balancing in A-module training,
- We added a political leaning objective to reduce political polarization in recommendations.

Results

We compared the TUNNeR model to the original MANNeR model and several other multiobjective trustworthy news recommenders including SentideBias[5] and SentiRec[6]. We used MIND[7] news recommendation dataset to test and compare TUNNeR model to baselines. Considering MIND dataset does not contain political leaning labels, we used LLM-model (GPT-4.1) to add political leaning labels to the dataset. We used standard left, neutral, right and nonpolitical labels to label the dataset. Main findings are shown in the table below.

[5] Wu, C., Wu, F., Qi, T., Zhang, W. Q., Xie, X., & Huang, Y. (2022). Removing AI's sentiment manipulation of personalized news delivery. Humanities and Social Sciences Communications, 9(1), 459.

[6] Wu, C., Wu, F., Qi, T., & Huang, Y. (2020, December). Sentirec: Sentiment diversity-aware neural news recommendation. Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing (pp. 44-53).

[7] Wu, F., Qiao, Y., Chen, J. H., Wu, C., Qi, T., Lian, J., ... & Zhou, M. (2020, July). Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3597-3606).

	model	Dataset	Aspect	Lambda	auc	mrr	ndcg@10	pll_div@10	pll_pers@10	categ_div@10	categ_pers@10
SentideBias	SentideBias-pll-modif	MIND-small	PLL		0.5404	0.2887	0.3368	0.0561	0.4032	0.4599	0.2104
SentiRec	SentiRec-pll-modif	MIND-small	PLL		0.4866	0.2463	0.2860	0.0842	0.3993	0.5465	0.2113
MANNeR	CR - 0.2 Amodule-pll-modif	MIND-small	PLL	-0.2	**0.6722**	**0.3654**	**0.4104**	0.0694	0.4054	0.4900	0.2492
	CR - 0.2 A module (cat)	MIND-small	CTG	-0.2	0.6613	0.3552	0.3996	0.0451	**0.4075**	0.5144	0.2305
	CR - 0.2 A module (cat)	MIND-large	CTG	-0.2	0.6751	0.3760	0.4211			0.4961	0.2342
	TUNNeR CR module	MIND-small	none		0.6313	0.2956	0.3854	0.0694	0.4048		
	TUNNeR CR module	MIND-small	none		0.6309	0.2984	0.3871			0.5111	**0.2370**
	TUNNeR CR + A module	MIND-small	PLL	-0.2	0.6335	0.2997	0.3890	**0.0834**	0.4033		
TUNNeR	TUNNeR CR + A module (diversity optimized weights)	MIND-small	PLL	-0.2	0.6077	0.2778	0.3650	**0.0904**	0.4007		
	TUNNeR CR + A module	MIND-small	CTG	-0.2	0.6194	0.2932	0.3785			**0.5253**	0.2234
	TUNNeR CR + A module (diversity optimized weights)	MIND-small	CTG	-0.2	0.5692	0.2429	0.3265			**0.5463**	0.1980
	TUNNeR CR + A module	MIND-large	CTG	-0.2	0.6321	0.2964	0.3860			**0.5185**	0.2329

Table TUNNeR: The proposed TUNNeR method improves the diversity of generated recommendations.

We did more extensive testing on MIND-small dataset due to computational constraints. For both MIND-small and MIND-large datasets, the original MANNeR model exhibits better recommendation accuracy as measured by AUC, MRR and NDGC metrics but TUNNeR outperforms both SentiRec and SentiDebias models. The main advantage of TUNNeR model is in diversity metrics, which we measured the same way as the authors of MANNeR paper using normalized entropy of the aspect distribution. Category diversity is labeled as categ_div@10 and political diversity as pll_div@10. Both political leaning and categorical diversity are better for TUNNeR compared to MANNeR model when tested on MIND-small dataset (0.0834 vs 0.0694 for political leaning diversity and 0.5253 vs 0.5144 for categorical diversity). SentiDebias model is worse than both TUNNeR and MANNeR. SentiRec model is slightly better than TUNNeR model when it comes to both political leaning diversity (0.0842 vs 0.0834) and categorical diversity (0.5465 vs 0.5253) but at the expense of accuracy metrics which are worse across the board. When weights that optimize diversity instead of recommendation accuracy are used in TUNNeR model, TUNNeR beats SentiRec by a very small margin in diversity metrics while still exhibiting better recommendation accuracy.

6. Conclusions

During the project lifetime we have addressed various aspects of the Trustworthy RS. Comparison and methods for labeling datasets used for the evaluation have been conducted, while we have evaluated the state-of-the-art methods. Then we have explored the RL methods designed for multi-objective recommendations, showing that RL may not be the optimal solution for selected problems, specifically proposed SOTA methods showed significant shortcomings. We have also explored off-policy evaluation for the hyperparameter search and off-line evaluation of the RS methods. Last but not least, the proposed TUNNeR method improved the diversity of generated recommendations while reducing the computation cost and offering a reasonable trade-off in precision.

Detailed results of the are presented in respective scientific publications:
Cief, Matej, Branislav Kveton, and Michal Kompan. 2025. Cross-Validated Off-Policy Evaluation. Proceedings of the AAAI Conference on Artificial Intelligence 39 (15):16073-81. https://doi.org/10.1609/aaai.v39i15.33765.

Marek Havrila, Martin Olejník, Matúš Tuna, Michal Kompan. 2026. The Illusion of Recommender Systems Reproducibility: Replication Studies Copy Bugs and Flaws from Original Works Leading to Confirmation of Incorrect Findings. ACM Transactions on Recommenders Systems. (submitted)

Adrian Gavornik, Katarína Marcinčinová, Marek Havrila. Open artifacts, closed research: how shared code can undermine replicability. AISB 2026 Symposium.
https://aisb.org.uk/wp-content/uploads/2025/12/Gavornik_AIBC_abstract.pdf