Mitigating the Filter Bubble while Maintaining Relevance: Targeted Diversification with VAE-based Recommender Systems

Zhaolin Gao
University of Toronto
Toronto, ON, Canada
zhaolin.gao@mail.utoronto.ca

Tianshu Shen
University of Toronto
Toronto, ON, Canada
tina.shen@mail.utoronto.ca

Zheda Mai
Optimy AI
Toronto, ON, Canada
zheda.mai@mail.utoronto.ca

Mohamed Reda Bouadjenek
Deakin University
Geelong, VIC, Australia
reda.bouadjenek@deakin.edu.au

Isaac Waller
University of Toronto
Toronto, ON, Canada
walleris@cs.toronto.edu

Ashton Anderson
University of Toronto
Toronto, ON, Canada
ashton@cs.toronto.edu

Ron Bodkin
Vector Institute for Artificial Intelligence
Toronto, ON, Canada
rbodkin@gmail.com

Scott Sanner*
University of Toronto
Toronto, ON, Canada
ssanner@mie.utoronto.ca

ABSTRACT

Online recommendation systems are prone to create filter bubbles, whereby users are only recommended content narrowly aligned with their historical interests. In the case of media recommendation, this can reinforce political polarization by recommending topical content (e.g., on the economy) at one extreme end of the political spectrum even though this topic has broad coverage from multiple political viewpoints that would provide a more balanced and informed perspective for the user. Historically, Maximal Marginal Relevance (MMR) has been used to diversify result lists and even mitigate filter bubbles, but suffers from three key drawbacks: (1) MMR directly sacrifices relevance for diversity, (2) MMR typically diversifies across all content and not just targeted dimensions (e.g., political polarization), and (3) MMR is inefficient in practice due to the need to compute pairwise similarities between recommended items. To simultaneously address these limitations, we propose a novel methodology that trains Concept Activation Vectors (CAVs) for targeted topical dimensions (e.g., political polarization). We then modulate the latent embeddings of user preferences in a state-of-the-art VAE-based recommender system to diversify along the targeted dimension while preserving topical relevance across orthogonal dimensions. Our experiments show that our Targeted Diversification VAE-based Collaborative Filtering (TD-VAE-CF) methodology better preserves relevance of content to user preferences across a range of diversification levels in comparison to both untargeted and targeted variations of Maximum Marginal Relevance (MMR); TD-VAE-CF is also much more computationally efficient than the post-hoc re-ranking approach of MMR.

CCS CONCEPTS

• **Information systems** → **Information retrieval diversity**; **Recommender systems**; • **Computing methodologies** → *Neural networks.*

KEYWORDS

Recommendation Systems, Filter Bubble, Diversity

ACM Reference Format:
Zhaolin Gao, Tianshu Shen, Zheda Mai, Mohamed Reda Bouadjenek, Isaac Waller, Ashton Anderson, Ron Bodkin, and Scott Sanner. 2022. Mitigating the Filter Bubble while Maintaining Relevance: Targeted Diversification with VAE-based Recommender Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3477495.3531890

1 INTRODUCTION

Online recommender systems are prone to create filter bubbles [4, 5, 14], where users are increasingly recommended content narrowly aligned with their historical interests due to a feedback loop between data collection and recommendation processes [12]. While the impacts of these recommendation feedback loops are somewhat nuanced (e.g., in some cases they can increase homogeneity due to popularity bias in recommender systems [3, 6]), in the case of media recommendation, it is observed that filter bubble effects arising from feedback loops may restrict user perspectives and viewpoints [1, 13]. As a case in point, we consider the row for VAE-CF (a state-of-the-art recommender system [10]) in Figure 1, which shows a Republican-shifted distribution of politically polarized Reddit community content recommendations (details in Section 3) for a user that has historically consumed Republican-oriented content.

*Affiliate to Vector Institute of Artificial Intelligence, Toronto

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531890

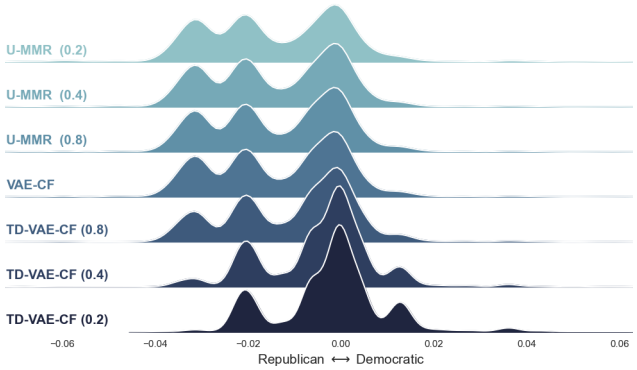

Figure 1: Top 50 recommendations for Republican-oriented users using a Reddit dataset. The x-axis represents the political spectrum of the recommendations; scores are computed by projecting the item embeddings on the Concept Activation Vector (CAV) aligned with the Republican-Democratic political spectrum (Section 3). The recommendations for U-MMR (Section 2) and TD-VAE-CF (Section 3) are computed using 3 different λ diversity levels (0.2, 0.4, 0.8), where diversity increases as λ decreases. Unlike TD-VAE-CF, standard MMR for diversification (U-MMR) is unable to improve diversity (content balance) in the political spectrum.

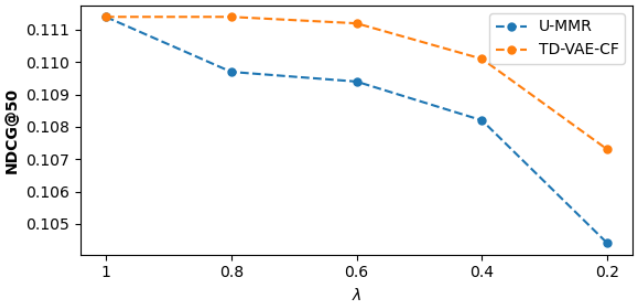

Figure 2: NDCG@50 vs. λ for TD-VAE-CF and U-MMR. λ values are taken from (1, 0.2, 0.4, 0.6, 0.8).

However, considering that topical content of user interest such as "the economy" may be represented in a variety of Reddit communities spanning the political spectrum, one might ask whether it is possible to mitigate this filter bubble effect by shifting the political coverage distribution to a more neutral or balanced position while preserving recommendation relevance for the user?

Historically, Maximal Marginal Relevance (MMR) [2] (cf. Section 2) has been used to diversify content rankings, including recent methods aiming to mitigate filter bubble effects [11]. While such result list diversification may decrease quantitative metrics of recommendation performance, user studies have shown that diversity can also improve overall satisfaction with recommendation lists [20]. However, the drawback of post-hoc reranking methods like MMR is their independent treatment of relevance and diversity [19] that inherently trades off one for the other. Furthermore, MMR typically diversifies across all content and not just targeted dimensions (e.g., political polarization). Finally, MMR's post-hoc reranking approach has quadratic time complexity in terms of the ranked list size.

In this paper, we propose a simple but empirically effective approach to address all of the aforementioned deficiencies of MMR. Our Targeted Diversification VAE-CF (TD-VAE-CF) methodology intrinsically dovetails with the latent user and item representations in state-of-the-art VAE-based collaborative filtering (VAE-CF) [10]. Specifically, we train Concept Activation Vectors (CAVs) [7] for **targeted diversification dimensions** (e.g., political spectrum) and use these to modulate latent embeddings of user preferences in VAE-CF to diversify along that targeted dimension while preserving topical relevance across orthogonal dimensions.

One can observe in Figure 1 that TD-VAE-CF clearly shifts the political polarization distribution to a more neutral range as diversification strength increases (λ decreases) in comparison to standard untargeted MMR (U-MMR), which is unable to shift the political spectrum; later we will see that a targeted version of MMR (T-MMR) performs even worse. Furthermore, as evidenced in Figure 2, the NDCG measure of recommendation relevance does not drop as steeply for TD-VAE-CF as it does for U-MMR when diversification strength is increased. We present more comprehensive experiments in Section 4 that confirm these results in a variety of additional settings, where we additionally show that the latent modulation approach of TD-VAE-CF induces low computation overhead in comparison to the quadratic time complexity of MMR.

In summary, TD-VAE-CF combines CAVs with VAE-based collaborative filtering to enable a novel **targeted** (e.g., political spectrum) diversification approach for recommendation that efficiently and selectively mitigates filter bubble effects while preserving relevance.

2 MAXIMAL MARGINAL RELEVANCE (MMR)

As one of the most popular ranked list diversification methods in the literature and our baseline for comparison, we briefly review Maximal Marginal Relevance (MMR) [2] as it applies to the recommendation setting. Given a set $\mathcal{I}$ of items to select $s_i \in \mathcal{I}$, we aim to build an optimal subset of items $S_k^* \subset \mathcal{I}$ (where $|S_k^*| = k$ and $k < |\mathcal{I}|$) relevant to a given user u. For computational efficiency, we will build S_k^* in a greedy manner by choosing the next optimal selection s_k^* given the previous set of optimal selections $S_{k-1}^* = \{s_1^*, \cdots, s_{k-1}^*\}$ and recursively defining with $S_k^* = S_{k-1}^* \cup \{s_k^*\}$. MMR greedily populates the result set according to the following criteria:

$$s_k^* = \underset{s_k^* \in \mathcal{I} \setminus S_{k-1}^*}{\operatorname{argmax}} \left[\lambda Sim_1(u, s_k) - (1-\lambda) Sim_2(s_i, s_k)\right] \quad (1)$$

Here, similarity metric Sim_1 measures user-item relevance (i.e., recommendation score), metric Sim_2 measures item-item similarity, and the parameter $\lambda \in [0, 1]$ trades off relevance and diversity. In the case of s_1^*, the maximization term is vacuous (=0).

From an inspection of Equation 1, one can easily identify the key weaknesses of MMR that we seek to address in this work: (1) it directly sacrifices relevance to achieve diversity [19]; (2) in its standard form, the item diversification Sim_2 is generic and untargeted (U-MMR) [11] though we later define a targeted variant (T-MMR) for additional comparison; (3) it requires a quadratic complexity computation of pairwise similarity measures between items.

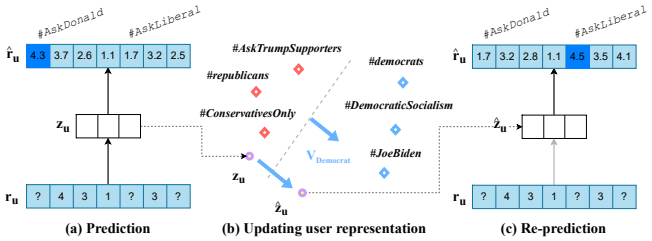

Figure 3: Step-by-step flow of TD-VAE-CF architecture. (a) We first obtain the user latent preference representation from the off-the-shelf VAE-CF model. (b) Next we find the CAV of two subtopics and update the user latent representation. (c) Re-predict to obtain CAV-shifted user preference ratings.

3 METHODOLOGY

VAEs for Collaborative Filtering: The impressive generalization and reconstruction ability of the VAE model is particularly attractive to the recommendation community and has inspired many recent deep learning-based recommendation models [9, 10, 16]. Figure 3(a) shows the basic VAE-CF model for recommendation, where a (sparse) vector of user preferences $\vec{r}_u$ over n items are encoded by the VAE [8] into a Gaussian-distributed latent preference embedding $\vec{z}_u$ of width d. $\vec{z}_u$ is then stochastically decoded to a (dense) reconstruction $\hat{\vec{r}}_u$ that generalizes user preferences to unobserved items. Formally, VAE-CF optimizes the following objective over the respective parameters ϕ and θ of the encoder and decoder:

$$\sum_u \log p(\vec{r}_u) \geq \sum_u \left[E_{q_\phi(\vec{z}_u|\vec{r}_u)} [\log p_\theta(\vec{r}_u|\vec{z}_u)] - KL[q_\phi(\vec{z}_u|\vec{r}_u)||p(\vec{z}_u)] \right], \tag{2}$$

In practice, the approximation of user distribution $q_\phi(\vec{z}_u|\vec{r}_u)$ is usually a Normal distribution with learned parameters μ_u and Σ_u.

In our implementation, we use a one-layer decoder such that the weights of the decoder can be directly used as item embeddings, $\mathbb{X}^{n\times d}$, where the i^{th} vector $\vec{x}_i$ is the i^{th} weight of the decoder that corresponds to the i^{th} item and n is the number of items. From the latent embedding space, we can obtain the user embeddings as $\mathbb{Z}^{m\times d}$ where the u^{th} vector $\vec{z}_u$ is the latent preference embedding $\vec{z}_u$ of user u and m is the number of users.

Concept Activation Vectors: Existing work for conversational recommender systems [17] leveraged the CAV [7] methodology with the VAE-CF framework to determine the alignment of keyphrase embeddings with user embeddings and applied a Bayesian update to user beliefs after each critique [17].

Here, we propose two methods for generating CAVs: I-CAVs and U-CAVs. We define a CAV as the normal to a hyperplane that separates two opposing subtopics (e.g. Republican vs. Democratic) in the embedding space as shown in Figure 3(b). For I-CAVs, we sample k items from each subtopic to form two subsets of items, $\mathcal{I}_k{}^1$ and $\mathcal{I}_k{}^2$, from decoder matrix $\mathbb{X}^{n\times d}$. To obtain the activation vector $\vec{v}_i \in \mathbb{R}^d$, multiple linear classifiers are trained to distinguish $\mathcal{I}_k{}^1$ and $\mathcal{I}_k{}^2$ and the averaged classifier is used as $\vec{v}_i$. U-CAVs sample from the user embeddings, $\mathbb{Z}^{m\times d}$, to form two subsets of users, $\mathcal{U}_k{}^1$ and $\mathcal{U}_k{}^2$. Similarly, the activation vector, $\vec{v}_u$, is generated using the averaged linear classifier between $\mathcal{U}_k{}^1$ and $\mathcal{U}_k{}^2$.

Table 1: Dataset statistics.

Dataset	#User	#Item	#Interactions	Density
Yelp	7,000	4,997	151,456	0.433%
Reddit Politics	9849	9,892	449,660	0.462%
Reddit Gender	9779	9,892	365,307	0.377%

Targeted Diversification VAE-CF (TD-VAE-CF): TD-VAE-CF applies a targeted CAV direction to update the user-embeddings from VAE-CF in the latent space. After we generate CAV $\vec{v}$ using either user or item embeddings, we update a user-embedding by subtracting its projection on the CAV with a parameter λ:

$$\hat{\vec{z}}_u = \vec{z}_u - (1-\lambda)\frac{\vec{z}_u \cdot \vec{v}}{||\vec{v}||^2}\vec{v} \tag{3}$$

where $\hat{\vec{z}}_u$ is the updated user embedding for user u. As shown in Figure 3(b), the updated user embedding moves closer to the hyperplane that separates two subtopics. The parameter λ is taken from $[0, 1]$ which controls the degree of update. Finally, the updated user-embedding is decoded using the one-layer decoder in VAE-CF to the (dense) reconstruction, $\hat{\vec{r}}_u$, where the user preference rating for the opposing subtopic would increase as shown in Figure 3(c).

4 EXPERIMENTS

We now experimentally compare the recommendation and diversification quality of our proposed TD-VAE-CF with a baseline VAE-CF model (i.e., TD-VAE-CF with $\lambda = 1$) and MMR-diversified variants of VAE-CF to address the following research questions:

- **RQ1 – Relevance vs. Diversity:** Is TD-VAE-CF able to maintain relevance and achieve diversity better than MMR?
- **RQ2 – Targeted Diversification:** Does TD-VAE-CF effectively distribute items in the targeted latent direction?
- **RQ3 – Running Time:** Is TD-VAE-CF efficient vs. MMR?

Appendix A provides a methodology and experimental results for a variation of the TD-VAE-CF methodology to "flatten" the filter bubble rather than "neutralize" it as we currently do in Section 3. All code to reproduce the experimental results is available on github.[1]

4.1 Datasets

We conduct experiments on two datasets: **Reddit** for recommendation of communities, and **Yelp** for recommendation of restaurants. We follow the same preprocessing steps as in previous work [15, 18] and randomly select 80% for training, 10% for validation, and 10% for testing. For Reddit, we select two spectra to diversify along with two subtopics for each spectrum: politics (*Republican* vs. *Democratic*) and gender (*Men* vs. *Women*). Since the Reddit communities are sparse and there are, on average, less than ten communities corresponding to one subtopic, U-CAVs are used for Reddit by selecting the users that have interacted exclusively with the communities within one subtopic in the spectrum (e.g., *DemocraticSocialism* for

[1] https://github.com/ZhaolinGao/TD-VAE-CF

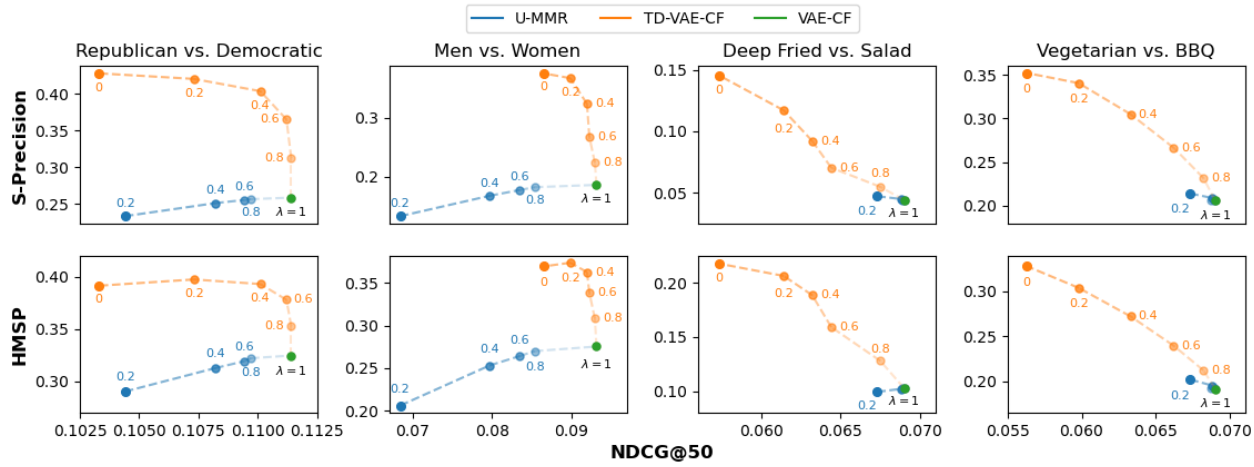

Figure 4: NDCG@50 vs. S-Precision and HMSP for four spectra of two datasets. λ values for TD-VAE-CF and U-MMR are (0, 0.2, 0.4, 0.6, 0.8, 1) and (0.2, 0.4, 0.6, 0.8, 1) respectively. Smaller λ values (higher diversity) are marked with higher opacity.

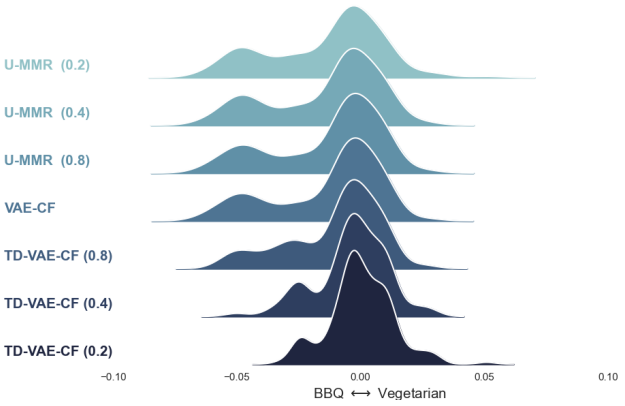

Figure 5: Top 50 recommendations for BBQ lovers using Yelp dataset. X-axis represents the meat consumption spectrum of the recommendations. The scores are computed by projecting the item embeddings on the Concept Activation Vector (CAV). The recommendations for TD-VAE-CF and U-MMR are computed using three different λ values (0.2, 0.4, 0.8). U-MMR fails to diversify along the Vegetarian vs. BBQ dimension while TD-VAE-CF can equally balance such content.

Democratic and *Republicans* for Republican). Similarly, the two spectra we choose for Yelp are health (*Deep Fried* vs. *Salad*) and meat consumption (*Vegetarian* vs. *BBQ*). The labels for each restaurant are generated using the ten most common key phrases in the user reviews (e.g. *DeepFried*, *BBQ*, etc.). We use I-CAVs for Yelp since there is a sufficient number of items for each subtopic to generate a representable CAV. Since there is no need to perform targeted diversification on users who haven't interacted with the targeted direction, we used two subsets of Reddit dataset on politics and gender. The statistics for the datasets are summarized in Table 1.

4.2 Metrics

We evaluate the relevance of top-k ranking performance using Normalized Discounted Cumulative Gain (NDCG) as done in previous work [10, 16]. The diversity of top-k items along the targeted spectrum is evaluated using S-Precision [19] and Harmonic Mean of Subtopic Probabilities (HMSP) since these two metrics can capture the distribution difference of the two ends of the targeted spectrum.

NDCG: NDCG is a measure of ranking quality using Discounted Cumulative Gain (DCG). Formally, it is defined as:

$$DCG@K = \sum_{i=1}^{K} \frac{rel_i}{log_2(i+1)} \qquad NDCG@K = \frac{DCG@K}{IDCG@K} \tag{4}$$

where rel_i is the graded relevance score of the i^{th} item and $IDCG@K$ is ideal discounted cumulative gain.

S-Precision: S-Precision reflects the number of subtopics covered in the recommended items and is derived from S-Recall that measures the number of subtopics covered among the top-k items:

$$S\text{-}Recall@K \equiv \frac{|\cup_{i=1}^{K} subtopics(d_i)|}{n_A}, \tag{5}$$

where n_A is the number of subtopics. If S is a recommendation system, we define $minRank(S, r)$ as the minimal rank K at which the ranking produced by S has S-Recall r. Then, we can define:

$$S\text{-}Precision@r \equiv \frac{minRank(S_{opt}, r)}{minRank(S, r)}, \tag{6}$$

where S_{opt} is a system that produce the optimal ranking (i.e., $minRank(S_{opt}, r)$ is the smallest K for a S-Recall r). Since we focus

on two subtopics, we use *S-Precision*@1 such that the $minRank(S, r)$ would be the minimum rank that covers both subtopics.

Harmonic Mean of Subtopic Probabilities (HMSP): Harmonic Mean of Subtopic Probabilities (HMSP) measures the empirical prevalence of each subtopic in the recommendation list; we choose the harmonic mean since it is strictly greater than or equal to the minimum subtopic probability and thus guarantees at least that much coverage for *each* subtopic. In contrast, the arithmetic mean could be relatively high even if one subtopic probability is 0.

Given empirical probabilities p_A and p_B of subtopics A and B in the recommendation list, HMSP = $\frac{2p_Ap_B}{p_A+p_B}$.

4.3 Methods Compared

We compare the following (diversified) recommendation methods:

- **TD-VAE-CF**: Our Targeted Diversification VAE-CF as defined in Section 3 with U-CAVs (Reddit) and I-CAVs (Yelp). Gaussian negative log-likelihood loss is used for VAE-CF. 100 CAVs are generated using 10 embeddings for each of the two subtopics which are randomly sampled and the final CAV is the mean of the 100 CAVs.
- **VAE-CF**: The undiversified baseline collaborative filtering system defined in Section 3 (VAE-CF = TD-VAE-CF @ $\lambda = 1$). Gaussian negative log-likelihood loss is used.
- **U-MMR**: Untargeted MMR defines Sim_1 as the VAE-CF user-item embedding dot product and Sim_2 as VAE-CF item-item embedding dot product (VAE-CF = U-MMR @ $\lambda = 1$).
- **T-MMR**: To see if we can achieve a Targeted MMR, we first project the user and item embeddings on the CAV to obtain their preference scores in the targeted spectrum. Then, the similarity metrics Sim_1 and Sim_2 are computed by taking the negative absolute value of the difference between the user-item and item-item scores, respectively.

4.4 Performance Evaluation

RQ1 – Relevance vs. Diversity: Overall, the targeted variation of MMR (T-MMR) demonstrates relevance scores that are extremely low (0.0043 for Yelp, 0.0036 for Reddit Politics, and 0.0038 for Reddit Gender on NDCG@50). *Therefore, in the remaining results, we omit T-MMR and only report results of the three other methods that have acceptable relevance scores.* The results of relevance and diversity on TD-VAE-CF, VAE-CF, and U-MMR are shown in Figure 4. For any fixed level of NDCG recommendation performance (x-axis), the TD-VAE-CF method *strictly dominates* U-MMR's diversity metrics. Since U-MMR performs untargeted diversification while the metric for diversity is measured along the targeted direction, U-MMR's diversity and relevance may *both* decrease for some datasets.

RQ2 – Targeted Diversification: The distribution of recommended items on the target CAV are shown in Figure 1 and 5. We see a clear shift of the distribution from one side to the middle, showing the diversified recommendations from TD-VAE-CF are well-balanced between the two targeted subtopics of the spectrum, unlike U-MMR.

RQ3 – Running Time: Empirical running times for TD-VAE-CF and U-MMR are shown in Figure 6 (T-MMR would be identical to U-MMR). U-MMR reranks items according to Equation 1. At the Kth rank selection step U-MMR compares all non-selected items with

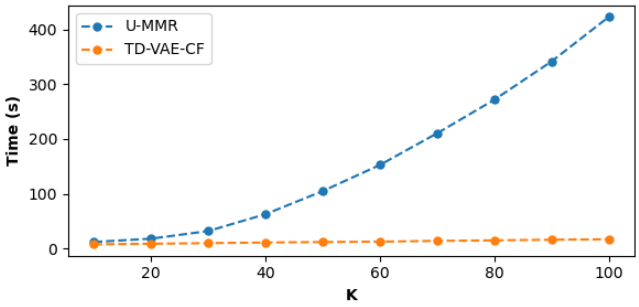

Figure 6: Time for generating top-K recommendations.

$K-1$ selected items. Therefore, the time complexity for U-MMR is $O(mnK^2)$ where m is the number of users, n is the number of items, and K is the number of recommendations to generate. For TD-VAE-CF, the time complexity can be divided into two parts: the user embedding update requires $O(m)$, and user-item rating generation and sorting need $O(mn)$ to generate the ratings and $O(m(K + Klog(K)))$ to select and sort the top-K ratings. Then, the total time complexity for TD-VAE-CF is:

$$O(m + mn + m(K + Klog(K))) = O(m(K + K\log(K) + n)), \quad (7)$$

Since the complexity of U-MMR has a quadratic growth with respect to K while TD-VAE-CF has a dominant term of $K\log(K)$ growth, the time difference of each increases drastically with higher K.

Conclusion: Overall, in comparison to popular MMR-based diversification methods, these results collectively confirm that our novel TD-VAE-CF can mitigate filter bubble effects via targeted modulation of a user's latent preference embeddings, while maintaining relevance and having lower computational complexity than MMR.

REFERENCES

[1] Guy Aridor, Duarte Goncalves, and Shan Sikdar. 2020. *Deconstructing the Filter Bubble: User Decision-Making And Recommender Systems.* Association for Computing Machinery, New York, NY, USA, 82–91.

[2] Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98).* Association for Computing Machinery, New York, NY, USA, 335–336. https://doi.org/10.1145/290941.291025

[3] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18).* Association for Computing Machinery, New York, NY, USA, 224–232.

[4] Seth Flaxman, Sharad Goel, and Justin M. Rao. 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly* 80 (2016), 298–320.

[5] Yingqiang Ge, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou, and Yongfeng Zhang. 2020. Understanding Echo Chambers in E-Commerce Recommender Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2261–2270.

[6] Kartik Hosanagar, Daniel Fleder, Dokyun Lee, and Andreas Buja. 2014. Will the Global Village Fracture Into Tribes? Recommender Systems and Their Effects on Consumer Fragmentation. *Management Science* 60, 4 (2014), 805–823.

[7] Been Kim, M. Wattenberg, J. Gilmer, C. J. Cai, James Wexler, F. Viégas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *ICML.* Stockholm, Sweden.

[8] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1312.6114

[9] Xiaopeng Li and James She. 2017. Collaborative variational autoencoder for recommender systems. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 305–314.

[10] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. *arXiv preprint arXiv:1802.05814* (2018).

[11] Gabriel Machado Lunardi, Guilherme Medeiros Machado, Vinicius Maran, and Jose Palazzo M. de Oliveira. 2020. A metric for Filter Bubble measurement in recommender algorithms considering the news domain. *Applied Soft Computing* 97 (2020), 106771.

[12] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback Loop and Bias Amplification in Recommender Systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

[13] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity. In *Proceedings of the 23rd International Conference on World Wide Web*. Association for Computing Machinery, New York, NY, USA, 677–686.

[14] Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin Group , The.

[15] Isaac Waller and Ashton Anderson. 2021. Quantifying social organization and political polarization in online platforms. *Nature* 600, 7888 (2021), 264–268. https://doi.org/10.1038/s41586-021-04167-x

[16] Ga Wu, Mohamed Reda Bouadjenek, and Scott Sanner. 2019. One-Class Collaborative Filtering with the Queryable Variational Autoencoder. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-19)*. Paris, France.

[17] Hojin Yang, Scott Sanner, Ga Wu, and Jin Peng Zhou. 2021. Bayesian Preference Elicitation with Keyphrase-Item Coembeddings for Interactive Recommendation. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 55–64.

[18] Hojin Yang, Tianshu Shen, and Scott Sanner. 2021. Bayesian Critiquing with Keyphrase Activation Vectors for VAE-based Recommender Systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-21)*. Online.

[19] ChengXiang Zhai, William W. Cohen, and John Lafferty. 2015. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. *SIGIR Forum* 49, 1 (jun 2015), 2–9. https://doi.org/10.1145/2795403.2795405

[20] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving Recommendation Lists through Topic Diversification. In *Proceedings of the 14th International Conference on World Wide Web* (Chiba, Japan) *(WWW '05)*. Association for Computing Machinery, New York, NY, USA, 22–32.

A FLATTENING THE FILTER BUBBLE BY INJECTING NOISE IN THE TARGETED CAV DIMENSION

One side effect of diversifying with the TD-VAE-CF methodology of Section 3 is that it tends to neutralize the targeted spectrum by shifting the user's latent preference towards the center of the spectrum. One caveat is that this method tends to recommend fewer items that are at the extremes of the spectrum, which can alternately be seen as a benefit or drawback according to the intent of the diversification. We can observe this effect in Figure 1 and 5: while the distributions of the recommended items shifted to the middle, the range of the distributions also decrease. In this section, we present a method that can diversify across both sides of the targeted spectrum by effectively "flattening" the distribution via the injection of Gaussian noise.

To achieve this, first we update the user-embedding, $\vec{z}_u$, following the same method in Section 3 with $\lambda = 0$ which removes preference information of the user in the targeted CAV direction $\vec{v}$:

$$\hat{\vec{z}}_u = \vec{z}_u - \frac{\vec{z}_u \cdot \vec{v}}{||\vec{v}||^2}\vec{v} \tag{8}$$

Then, we compute perturbations of the user's latent embedding $\hat{\vec{z}}_u^j$ n times along the targeted CAV direction $\vec{v}$ by computing

$$\hat{\vec{z}}_u^j = \hat{\vec{z}}_u - \alpha^j \vec{v} \tag{9}$$

where $j \in \{1, \ldots, n\}$ and $\alpha^j \sim \mathcal{N}(0, \sigma^2)$ randomly determines the magnitude of noise injection in the CAV dimension (sometimes sampling more extreme ends of the CAV spectrum). Clearly, as σ increases, the amount of extreme content from both ends of the spectrum that is sampled also increases.

Finally, the n perturbed user-embeddings $\hat{\vec{z}}_u^j$ are decoded to produce n VAE-CF recommendation predictions r_{ij} for each item i and perturbation j. The final rating score r_i for item i is computed by averaging the n scores produced by each perturbed user embedding:

$$r_i = \frac{1}{n}\sum_{j=1}^{n} r_{ij}\,. \tag{10}$$

The resulting targeted spectrum distributions are shown in Figure 7 and the results of relevance and diversity analysis are shown in Figure 8. We fixed n at 10 and varied σ within $(0, 1, 2, 5, 10, 20)$. When $\sigma = 0$, the perturbation process does not change the user embeddings and it is the same method as TD-VAE-CF in Section 3 with $\lambda = 0$.

We can clearly see that the distribution is flattened with higher σ and a wider range of items on the targeted spectrum, although still balanced and thus yielding relatively high diversity measures. The diversity continues to increase with higher σ while the *relevance decreases more rapidly than TD-VAE-CF since extreme recommendations are less likely to be relevant for most users.*

Whether one should "neutralize" the curve (and remove extreme content for potentially sensitive users) using TD-VAE-CF as proposed in Section 3 or "flatten" the curve as shown here (and recommend content at all points on the spectrum – even extreme points of view from both sides) is a decision that ultimately rests with the recommendation system designer and depends on the desiderata that motivate the need for diversification w.r.t. the target audience.

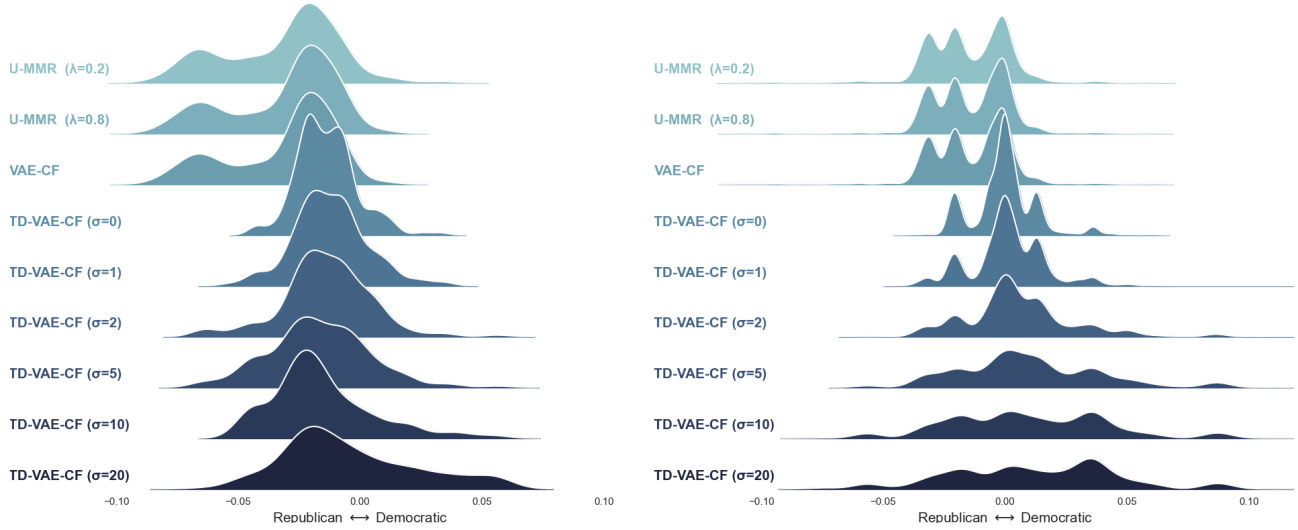

Figure 7: Top 50 recommendations for Republicans and BBQ lovers using Reddit and Yelp dataset respectively. X-axis represents the political spectrum or the meat consumption spectrum of the recommendations. The scores are computed by projecting the item embeddings on the Concept Activation Vector (CAV). The recommendations for TD-CAE-CF and U-MMR are computed using three different σ values (0, 1, 2, 5, 10, 20).

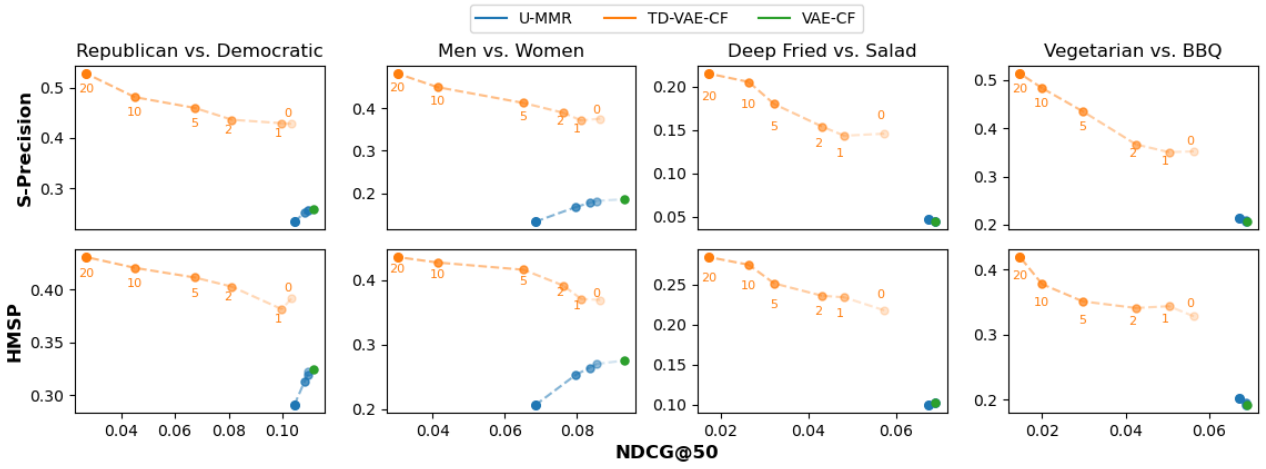

Figure 8: NDCG@50 vs. S-Precision and HMSP for four spectra of two datasets with different σ values (0, 1, 2, 5, 10, 20). Higher σ values (higher standard deviation) are marked with higher opacity. $\sigma = 0$ for TD-VAE-CF is the same as $\lambda = 0$ without any remapping. U-MMR results are the same as Section 4 since its not possible to perform the same remapping process for U-MMR.