Reducing Cross-Topic Political Homogenization in Content-Based News Recommendation

Karthik Shivaram
Tulane University
New Orleans, USA
kshvaram@tulane.edu

Ping Liu
Illinois Institute of Technology
Chicago, USA
pliu19@hawk.iit.edu

Matthew Shapiro
Illinois Institute of Technology
Chicago, USA
shapiro@iit.edu

Mustafa Bilgic
Illinois Institute of Technology
Chicago, USA
mbilgic@iit.edu

Aron Culotta
Tulane University
New Orleans, USA
aculotta@tulane.edu

ABSTRACT

Content-based news recommenders learn words that correlate with user engagement and recommend articles accordingly. This can be problematic for users with diverse political preferences by topic — e.g., users that prefer conservative articles on one topic but liberal articles on another. In such instances, recommenders can have a homogenizing effect by recommending articles with the same political lean on both topics, particularly if both topics share salient, politically polarized terms like “far right” or “radical left.” In this paper, we propose attention-based neural network models to reduce this homogenization effect by increasing attention on words that are topic specific while decreasing attention on polarized, topic-general terms. We find that the proposed approach results in more accurate recommendations for simulated users with such diverse preferences.

CCS CONCEPTS

• **Information systems** → *Recommender systems*; *Clustering and classification.*

KEYWORDS

News Recommendation Systems, Topic Homogenization, Filter Bubbles

ACM Reference Format:
Karthik Shivaram, Ping Liu, Matthew Shapiro, Mustafa Bilgic, and Aron Culotta. 2022. Reducing Cross-Topic Political Homogenization in Content-Based News Recommendation. In *Sixteenth ACM Conference on Recommender Systems (RecSys '22), September 18–23, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3523227.3546782

1 INTRODUCTION

Personalized news recommenders build models of user preferences based on prior engagement. Recent research suggests that recommendation algorithms can contribute to the formation of “filter bubbles,” in which users are overexposed to homogeneous viewpoints [2, 19, 23, 26, 28, 31, 33]. This form of intellectual isolation can degrade civil discourse, contribute to the spread of misinformation, as well as cause polarization across social media. Given the mix of technological, social, and political factors that contribute to this phenomenon, further research is required to understand the specific mechanisms that cause filter bubbles and design systems that are more resistant to them.

In this paper, we study one specific mechanism of content-based recommendation systems that can lead to filter bubbles and propose methods to mitigate its effects. Specifically, we consider users who have diverse political preferences by topic — e.g., users that prefer to read conservative articles on one topic but liberal articles on another. Based on polling by Pew, such users are a sizable portion of the U.S. population [36]. Because content-based recommenders learn text features that correlate with user engagement, we find that they can have a homogenizing effect by recommending articles with the same political lean on both topics. For example, the phrase “extreme right” may appear in liberal articles a user has liked discussing gun control, as well as in liberal articles on immigration that the user has not yet read. If a user in fact prefers conservative articles on immigration, the recommender may thus incorrectly recommend liberal articles due to the presence of the phrase “extreme right.” This flawed recommendation is further exacerbated when topics are not known *a priori*, which is often the case in political news, where detecting emerging topics is a research challenge of its own [48]. This *cross-topic homogenization* was observed recently in content-based recommenders [23] — our goal in this paper is to study the phenomenon more closely and propose models to reduce its impact.

To do so, we draw upon a collection of 900k news articles annotated with political stance created by Liu et al. [23]. We simulate browsing sessions for users with opposing political preferences for topic pairs, creating a setting in which the system observes more interactions for the first topic than for the second topic. In this way, we are able to measure and focus particularly on the homogenization effect of the first topic on the second. We then propose two attention-based neural network models designed to reduce this homogenization effect. The first model adds a new term to the

RecSys '22, September 18–23, 2022, Seattle, WA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9278-5/22/09…$15.00
https://doi.org/10.1145/3523227.3546782

objective function in order to penalize attention given to topic independent polarized phrases, like "extreme right," that predict stance across many topics. Conversely, the second approach rewards attention placed on topic dependent terms, like "undocumented" versus "illegal" immigrants, resulting in topic-specific models that are less prone to overgeneralize across topics. We also consider a model that combines the two new learning objectives into a single model. In our experiments using 45 topic pairs, we find that the proposed approach improves accuracy by roughly 5% on the second topic, while still maintaining accuracy comparable to the baseline on the first topic. These results provide evidence that recommendation systems can be designed to mitigate cross-topic homogenization.

2 RELATED WORK

Our focus on cross-topic homogenization is motivated in part by sociological theories suggesting that polarization grows when an individual's partisan view on one topic "spreads" to another topic [11]. Furthermore, comprehensive public polling by Pew shows that many Americans do indeed have political stances that vary significantly by topic [36], which is in line with research indicating that the public is less politically monolithic than "elites," and that many citizens do not have fully-formed partisan opinions on many topics [17]. Given these sociological findings, if recommendation systems are systematically biased to show politically homogeneous content across topics, then they may serve as accelerants of partisanship. This is particularly problematic if the user initially does not have fully-formed opinions on a new topic, which is the scenario we aim to simulate in our experiments by having one topic appear less frequently than another. Recent work on content-based recommenders has shown that such cross-topic homogenization can occur in political news recommendation [23]. Our proposed methods are designed to understand how a recommendation system could be trained to reduce the likelihood of this homogenization.

Our work also builds on research studying how partisan bias manifests in news media [5, 30, 37]. For example, Budak et al. [5] find that news sources of different political leanings are distinguished most by "disproportionately criticizing one side." In our data, we observe this in phrases like "far right" and "radical left," topic-independent phrases criticizing the out-group that can lead to cross-topic homogenization. Our proposed methods are specifically designed to reduce the influence of such phrases.

Our work also adds to the growing study of different types of bias in recommendation systems, such as popularity bias [1] and exposure bias [21]. One major factor that leads to these types of biases is the presence of feedback loops [27], which can contribute to the homogenization of users, causing them to consume similar content while sacrificing utility [7, 23]. Homogenization can also lead to the creation of "filter-bubbles" [33] and "echo-chambers" [15], which may also influence polarization [6, 9, 10]. This is prominent in the case of news recommendation systems [3, 14, 43], where several prototypes have been developed to give users more control of the recommender system to increase diversity [4, 32]. Most prior work focuses on increasing the overall partisan diversity of content exposure, ignoring cross-topic effects; furthermore, most prior work focuses on collaborative filtering recommendation systems [28]. In contrast, we focus here on mitigating cross-topic homogenization in content-based recommenders, filling a key gap in the extant literature.

Recently, a variety of deep learning based approaches have been proposed for news recommendation [16, 20, 38, 39, 45–47, 51, 52]. Most of these methods are content based, using attention-based deep neural networks to learn representations of both the candidate news article and the user's interest based on click logs. The methods predict future click events based on the similarity between these two representations [45, 46, 52]. Some prior work also uses observed topic information to learn user interests in a hierarchical fashion [38] and also to enrich the news article representation learned [16]. Modern pre-trained language models have also recently been used in order to improve news and user representations [47, 51]. To our knowledge, none of these prior approaches directly address the issue of cross-topic homogenization. In our experiments, we compare with a representative baseline from this recent work by Zhang et al. [51], finding that it is also susceptible to this sort of homogenization.

Our technical approach builds on two threads of machine learning for text classification – neural attention mechanisms [8, 22, 50] and multitask learning [25, 41, 49]. We adapt these approaches to the homogenization problem in two ways: (1) by formulating a penalty term to reduce attention given to topic-independent polarized words; (2) by formulating a secondary prediction task to increase attention given to topic-dependent words.

3 PROBLEM SETTING

We assume a user interaction session consists of a sequence of articles $\mathbf{a} = \{a_1 \ldots a_n\}$ and a corresponding sequence of binary feedback labels $\mathbf{y} = \{y_1 \ldots y_n\}$, where $y_i = 1$ means the user liked article a_i, and $y_i = 0$ means they did not. We additionally assume that each article a_i is assigned to exactly one **unobserved** topic $t_i \in \mathcal{T}$. To simulate partisan preferences, we assume that a user's feedback label follows their political preferences for that topic. E.g., if a user prefers conservative articles on topic t_i, then the feedback label will be $y = 1$ for conservative articles shown and $y = 0$ for liberal articles shown.

The phenomenon of interest occurs when a user has opposing political preferences on two topics — e.g., they prefer to read liberal articles on immigration but conservative articles on abortion. This is a challenging case for the recommender — not only are topic assignments unobserved, but topics do not arrive uniformly at random. For example, the system may observe mostly immigration articles and only a few abortion articles. In this setting, the system may incorrectly extrapolate that because the user prefers liberal articles on immigration, they also prefer liberal articles on abortion, leading to poor recommendations. We call this *cross-topic political homogenization*, as the recommender is biased towards showing politically homogeneous articles across the two topics.

To measure system behavior in this setting, we assume we observe a training batch consisting of n_1 article interactions from topic t_1 and n_2 interactions from topic t_2, where $n_2 << n_1$. We assume the user has different political preferences for t_1 and t_2 (e.g., they may prefer liberal articles on t_1 and conservative articles on t_2). Based on these $(n_1 + n_2)$ interactions, the system trains a content-based recommender. We then measure the accuracy of the recommender on a held-out sample of articles from both topics. Accuracy here indicates the fraction of recommended articles that receive positive (simulated) user feedback. We expect overall accuracy to be lower for topic t_2, both because the system observes fewer user interactions for t_2, and also because the user's preferences switch political leanings between topics. This setup can be viewed as a challenging type of cold-start problem; i.e., we have very few training examples from topic t_2, and those examples conflict politically with the training examples from topic t_1.

In our experiments, we consider several binary classifiers that predict user interaction label y given a new article a. We offer models that attempt to reduce cross-topic homogenization both by reducing attention on topic-independent terms and also by increasing attention on topic-dependent terms.

4 METHODS

We propose multiple network architectures trained in both a single task and multitask fashion to mitigate the effect of cross-topic political homogenization. The network architectures are shown in Figure 1. The following subsections discuss these architectures in detail.

4.1 Baseline 1: Single Task Network (STN)

Our first baseline model performs article classification, where each article a_i contains k words $\{w_{i0} \dots w_{ik}\}$. We first pass article a_i through a pre-trained BERT [12] model (uncased, 12-layer, 768-hidden, 12-heads, 110M parameters) to obtain BERT's word level embeddings $\{r_{i0} \dots r_{ik}\}$. We choose r_{i0} (**"CLS"** token's embedding) and pass it through a linear layer $\langle W_q, b_q \rangle$ with a sigmoid activation to compute the corresponding class probability $\hat{y}_i$:

$$\hat{y}_i = \sigma(W_q r_{i0} + b_q) \tag{1}$$

where values of $\hat{y}_i$ close to 1 indicate that the user has high probability of liking article a_i. This network is trained on the $(n_1 + n_2)$ labeled articles from prior user interactions, using binary cross-entropy ($bce(y_i, \hat{y}_i)$) as the loss function.

4.2 Baseline 2: Single Task Attention Network (STAN)

Our second baseline augments the prior model with an attention layer. This model is inspired by the approach in [50], but without the hierarchical aspect. In this network an extra linear layer $\langle W_a \rangle$ is used to calculate word attention weights u_{it} given word embedding r_{it} as the input. We next normalize these word attention weights to get $\hat{u}_{it}$ by applying a softmax transformation:

$$u_{it} = W_a r_{it} \tag{2}$$

$$\hat{u}_{it} = \frac{\exp(u_{it})}{\sum_{t=1}^{k} \exp(u_{it})} \tag{3}$$

Next the attention context vector u_i is obtained by taking the weighted average between the word attention weights and the article word embeddings:

$$u_i = \sum_{t=1}^{k} \hat{u}_{it} r_{it} \tag{4}$$

This resulting vector u_i encapsulates all information of the words and their corresponding context in the article. Finally, this vector is passed through an output layer $\langle W_l, b_l \rangle$ with a sigmoid activation to obtain $\hat{y_i} = \sigma(W_l u_i + b_l)$. This network also uses binary cross-entropy loss.

4.3 Proposed Method 1: Single Task Attention Network with Polarization Penalty (STANPP)

Our first proposed approach modifies the STAN model to reduce attention on topic-independent polarized terms. This is accomplished in a two-step process: first, we identify a candidate set of such polarized terms, then we augment the objective function to penalize attention on them and related terms.

In order to identify topic-independent polarized terms, we assume we have access to a large collection of articles labeled by stance but not by topic (e.g., the partisan lean of a news source provides a strong source of such supervision). Terms that predict stance reliably across this collection are likely to be topic-independent. While any number of feature selection approaches could be used here, in the experiments below we simply select the top 200 terms according to a Chi-Squared test, used to measure the dependence between terms and political stance (see Table 1). Terms such as "socialist," "right-wing," and "conservative" exemplify the topic-independent, polarized language we wish to reduce attention towards. Additionally, polarizing figures such as Alexandria Ocasio-Cortez and Rudy Giuliani also appear across many topics while strongly correlating with the political stance of the article. (I.e., conservative articles tend to be critical of Alexandria Ocasio-Cortez, while liberal articles tend to be critical of Rudy Giuliani.)

Given this set of R polarized terms, we next augment the STAN model to reduce the magnitude of attention they and related terms are given. We first embed each of the polarized terms using BERT to obtain word vectors $\{r_1 \dots r_R\}$. Then, for each document a_i, we measure the similarity between the attention context vector u_i from the STAN model with each of the polarized word vectors r_j by taking the sigmoid of their dot product $\sigma(u_i \cdot r_j)$. The loss for a single document is then a linear combination of the bce loss and the average similarity between the attention vector and the polarized words.

$$L_{\text{STANPP}} = (1 - \alpha) bce(y_i, \hat{y}_i) + \alpha \Big(\frac{1}{R} \sum_{j=1}^{R} \sigma(u_i \cdot r_j) \Big) \tag{5}$$

Here α is a hyperparameter tuned on validation data, as described in the experiments below. Thus, the loss function aims to jointly minimize classification error while making the document representation dissimilar to the polarized terms.

4.4 Proposed Method 2: Multitask Attention Network (MTAN)

Rather than penalize topic-independent terms, our second proposed approach instead rewards topic-dependent terms. Since we do not observe topic labels, we cannot use them directly to do so. Instead, we create a multitask model that predicts both the article label as well as a masked word from the article headline. The intuition

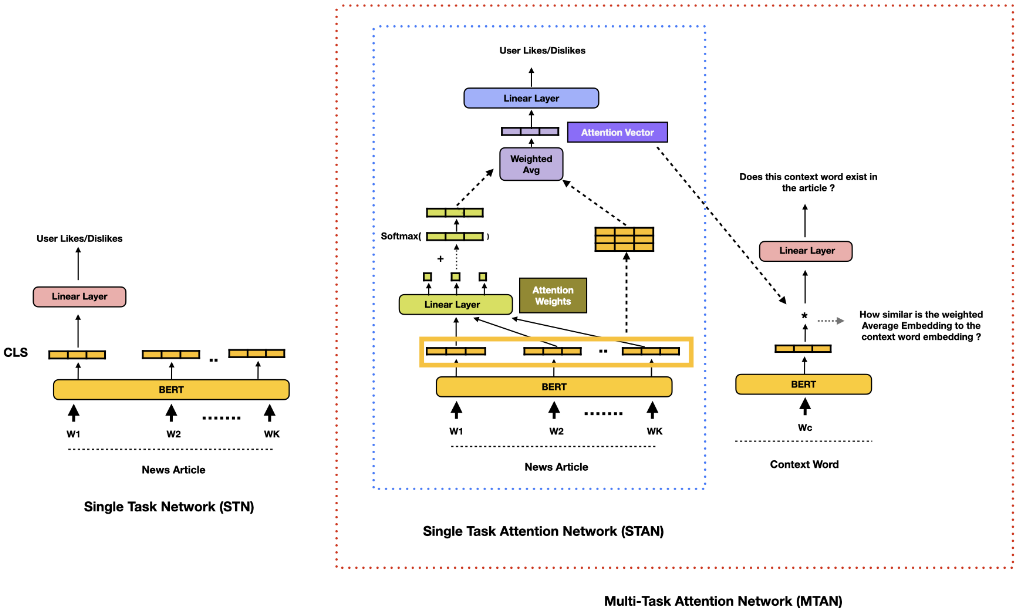
User Likes/Dislikes
Linear Layer
CLS
BERT
W1
W2
WK
News Article
Single Task Network (STN)
User Likes/Dislikes
Linear Layer
Attention Vector
Weighted Avg
Softmax( )
+
Linear Layer
Attention Weights
BERT
W1
W2
WK
News Article
Single Task Attention Network (STAN)
Does this context word exist in the article ?
Linear Layer
*
How similar is the weighted Average Embedding to the context word embedding ?
BERT
Wc
Context Word
Multi-Task Attention Network (MTAN)

Figure 1: Network Architectures for STN, STAN and MTAN

Table 1: Sample of 50 Polarizing Terms used by STANPP

abortion accused adam administration admitted alexandria allegations alleged amy andrew biden bush campaigns chuck conservative conspiracy controversial dca democrat democrats donald emails facts failed fbi foundation fox giuliani gop hunter illegal impeach interference joe kamala liberal nancy ocasiocortez pelosi probe radical republican republicans rightwing rudy scandal schiff socialist terrorist vermont

is that this will encourage the model to pay attention to words that are specific to this article, and that such terms are likely to be topic-dependent.

For the headline word prediction task, we use the "binary negative sampling" approach from word2vec [29]. For each article a_i, we sample a word h_i from the headline of the article and mask it. For a pair (a_i, h_i), we create a binary classification task to determine whether word h_i came from the headline of article a_i.

We then create two samples for each article, one positive (a_i, h_i) and one negative (a_i, h_i'). The negative headline words are sampled from a vocabulary consisting of all headline terms in our dataset, excluding those present in the headline of a_i. The candidate headline words h_i are passed through a pre-trained BERT embedding model to get the corresponding word embedding r_{h_i}. Next we take the dot product g_i between the candidate headline word embedding r_{h_i} and the attention context vector u_i from our STAN subnetwork[1] to measure how similar these two vectors are: $g_i = u_i \cdot r_{h_i}$.

Finally, this dot product g_i is passed through a linear layer $\langle W_c, b_c \rangle$ with sigmoid activation to obtain $\hat{y}_{h_i}$, the predicted probability that the candidate headline word h_i belongs to the headline of the news article a_i: $\hat{y}_{h_i} = \sigma(W_c g_i + b_c)$.

We compute a linear combination of the losses of each of the subnetworks in this architecture as the total loss to optimize:

$$L_{\text{MTAN}} = (1 - \alpha) \cdot bce(y_i, \hat{y}_i) + \alpha \cdot bce(y_{h_i}, \hat{y}_{h_i}) \tag{6}$$

where y_{h_i} is the true binary label for the candidate headline word h_i, and α is a hyperparameter tuned on validation data, as described in 5.

4.5 Proposed Method 3: Multitask Attention Network with Polarization Penalty (MTANPP)

This network combines STANPP and MTAN. It has the same architecture as MTAN but with the extra penalty term from the STANPP network:

$$L_{\text{MTANPP}} = (1 - (\alpha_1 + \alpha_2)) \cdot bce(y_i, \hat{y}_i) + \alpha_1 \cdot bce(y_{h_i}, \hat{y}_{h_i}) + \alpha_2 \left(\frac{1}{R} \sum_{j=1}^{R} \sigma(u_i \cdot r_j) \right) \tag{7}$$

5 EXPERIMENTS

5.1 Data

We use the news article dataset created by Liu et al. [23].[2] This dataset contains 900k news articles collected from 41 different news sources with corresponding political stance scores ranging over a

[1] We remove the masked word prior to embedding.

[2] Data and Code for our experimental results can be found at https://github.com/tapilab/recsys-2022-political

5-point scale (-2,-1,0,1,2) where -2 denotes extremely liberal and +2 denotes extremely conservative. To focus on heterogeneous preferences, we drop neutral articles (0) and collapse +2,+1 articles into a "conservative" class, and -2,-1 articles into the "liberal" class. We uniformly sample 100K of these news articles for this study.

To simulate users with heterogeneous political stances across topics, we first need to assign topics to each document. We adopt a simple, transparent approach by using k-means to cluster the 100k articles into 100 clusters.[3] We first represent each article by concatenating the headline with the first 10 sentences, perform standard tokenization to remove punctuation, then create tf-idf vectors using scikit-learn's [35] tf-idf vectorizer, with *min_df* of 30 and *max_df* of 0.9. We then run k-means clustering with k = 100. To ensure sufficient cluster sizes and sufficient samples from liberal/conservative stances, we filter these clusters to those with at least 400 articles, and sample uniformly so that each cluster has an equal number of liberal and conservative articles. From the clusters that remain, we sample 45 pairs of clusters at random for the basis of our experiments. A manual inspection of these clusters indicates many coherent topics on issues such as immigration, the 2020 election, gun rights, abortion, and healthcare.

5.2 Experimental Setting

We measure the performance of the above networks using a setting where 90% of the articles in the training and validation data are from a randomly sampled topic 1 and 10% are from a randomly sampled topic 2. The small number of training examples from topic 2 makes this a challenging problem, similar to a cold-start setting. The test set is comprised of an equal distribution of articles from topic 1 and topic 2. We simulate user preferences such that their political preferences for topic 1 articles are the opposite of their preferences for topic 2. We repeat experiments for 45 pairs of topics described in the previous section. Thus, each run consists of a different (topic1, topic2) pair, chosen from our list of discovered topics. Throughout, we refer to topic 1 as the majority topic in the training data and topic 2 as the minority topic, though we run experiments for 45 distinct topic pairs.

To tune each network, we hold out 10% of the training data as a validation set. We perform hyperparameter tuning using grid search over each topic pair using values shown in Table 5 and select the best set of parameters based on the accuracy scores on the validation dataset.

After predicting on the test set, we compare the overall accuracy of each approach, as well as investigate how the accuracy varies by topic. Our goal is to improve accuracy on topic 2 without harming accuracy on topic 1. To better assess the ceiling of improvement that is possible, we also fit a model we call the **Topic Oracle**, which, unlike the other methods, is able to observe the topic assignment of each article. To fit this model, we train STAN models separately for topic 1 and topic 2 using the same training data as above. At testing time, we apply the model appropriate for the topic of each test article. The predictions of the Topic Oracle on topic 1 are therefore not influenced by topic 2, and vice versa. This provides a rough upper bound on how well we can expect a model to perform at reducing the impact of cross-topic homogenization. We additionally compare with **UNBERT** [51], a representative example of recent work using BERT for news recommendation. This approach learns a BERT-based representations of a user based on the articles they've liked, then pairs this with an article representation to predict whether they will like a new article. As with the other models, its hyperparameters are tuned on validation data.[4] We use the pytorch [34] and huggingface [44] libraries to implement our networks. All our models are trained using a Nvidia RTX 3090 GPU over a period of 5 days.

Table 2: Average model accuracy over 45 topic pairs

Network	Topic 1 Accuracy	Topic 2 Accuracy	Total Accuracy
UNBERT	0.647	0.447	0.547
STN	0.613	0.489	0.551
STAN	0.682	0.498	0.590
STANPP (ours)	0.664	0.525	0.594
MTAN (ours)	**0.693**	0.531	0.612
MTANPP (ours)	0.687	**0.552**	**0.619**
Topic Oracle	0.701	0.596	

Table 3: Average Network Performance across 45 Topic Pairs with Additional Metrics

Network	Score Type	F1	Precision	Recall	AUC
UNBERT	**Topic 1**	0.629	0.635	0.634	0.646
	Topic 2	0.410	0.429	0.404	0.447
	Total	0.522	0.535	0.519	0.547
STN	**Topic 1**	0.602	0.620	0.610	0.613
	Topic 2	0.461	0.482	0.483	0.489
	Total	0.535	0.555	0.546	0.551
STAN	**Topic 1**	0.670	0.693	0.660	0.682
	Topic 2	0.426	0.476	0.433	0.498
	Total	0.563	0.608	0.547	0.590
MTAN (ours)	**Topic 1**	0.676	**0.716**	0.653	**0.693**
	Topic 2	0.473	0.522	0.466	0.531
	Total	0.583	**0.637**	0.559	0.612
STANPP (ours)	**Topic 1**	0.661	0.673	0.663	0.664
	Topic 2	0.466	0.521	0.466	0.525
	Total	0.573	0.612	0.564	0.594
MTANPP (ours)	**Topic 1**	**0.681**	0.696	**0.676**	0.687
	Topic 2	**0.522**	**0.563**	**0.509**	**0.552**
	Total	**0.605**	0.630	**0.592**	**0.619**

[3] More complicated clustering methods could be used, but our approach is independent of how the topics are determined.

[4] Following the implementation [51], this model is trained using only article headlines, due to its high computational complexity by sequence length.

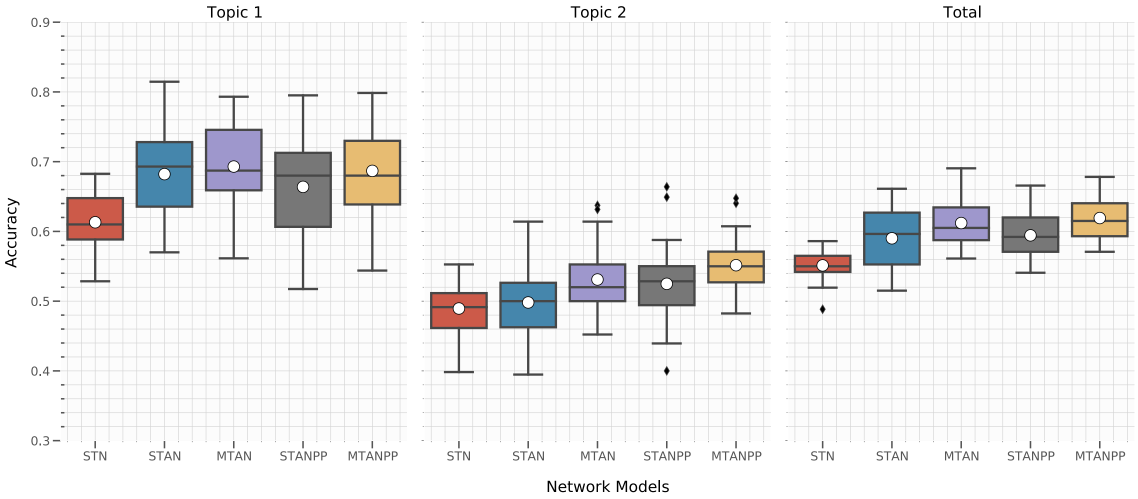

Figure 2: Distribution of test accuracies across 45 topic pairs for STN, STAN, MTAN and MTANPP

5.3 Experimental Results

Table 2 reports the accuracy of each approach averaged across 45 different topic pairs. Figure 2 shows boxplots of the same results to visualize the variance across topic pairs and Table 3 shows additional measures including precision, recall, and F1.

By comparing STAN and Topic Oracle, we can see the considerable impact cross-topic homogenization can have. For Topic 2, which has fewer training samples, accuracy drops from .596 to .498 when training data from topic 1 is included, indicating that the content from topic 1 is prohibiting an accurate model for topic 2 articles. This shows how the recommendations for an emerging topic can be quite poor, as the system defaults to recommendations in line with preferences on prior topics.

We observe that on average the proposed STANPP, MTAN, and MTANPP networks tend to have higher accuracy (3%-6%) for recommending topic 2 articles compared to the baseline STN and STAN networks. We also observe an increase in accuracy across topic 1 recommendations for the MTAN, STAN and STANPP networks compared to the STN network (1%-8%). Furthermore, combining STANPP and MTAN into MTANPP appears to do as well or better than each in isolation.

We computed pairwise t-tests for each pair of models. For topic 2 accuracy, all results are significant at the 5% level except for the differences between STN and STAN and between MTAN and STANPP. For total accuracy, all results are significant except for STAN and STANPP. For topic 1 accuracy, three differences are insignificant: STAN vs MTAN, STAN vs MTANPP, MTAN vs MTANPP. We also see that compared to UNBERT, our proposed approaches perform better across all metrics of comparison.

By comparing with the Topic Oracle, we see that the best of the proposed models approaches the accuracy of the topic aware oracle (topic 1: .693 vs .701; topic 2: .552 vs .596). These results also highlight the difficulty of this problem setting, which we attribute to two factors: First, the training data have few examples from topic 2 (often less than 100). Second, the article collection contains a wide variety documents, most of which are not opinion pieces. Thus, the difference between -1 and +1 articles can be difficult to discern based on linguistic evidence, requiring instead a nuanced understanding of the political and policy landscape.

5.3.1 *Shift in Attention.*

To further understand model behavior, we examine how attention varies by model to confirm whether the loss functions are having the intended effects. To do so, we analyze the change in the attention rank of terms, where the ranking is done based on cumulative attention scores. For a term w_t, let $\hat{u}_{it}$ be the corresponding normalized word attention weight for the term w_t contained in the news article d_i. Assume there are V unique terms in the vocabulary. Then, the cumulative attention C_t for term w_t across n documents is calculated as:

$$C_t = \frac{\sum_{i=1}^{n} \hat{u}_{it}}{\sum_{j=1}^{V} \sum_{i=1}^{n} \hat{u}_{ij}} \tag{8}$$

For illustrative purposes, we analyze the shift in ranks based on these cumulative attention scores for a topic pair where topic 1 discusses **gun control** and topic 2 discusses **climate change**. Table 4 shows the top 30 terms with the highest cumulative attention scores using our attention network models. For the STAN network, most of the top terms are either very specific to topic 1 (e.g., gun, shooting, firearm) or are terms that are polarized and occur across documents (e.g., trump, democrats, left). For the STANPP network we see that terms that are ranked highly are more topic specific (e.g., gun, violence, rifle) and have more focus on topic 2 (e.g., fossil, protection, environmental, climate, fuel, energy, emissions). We see similar trends for the MTAN and MTANPP networks. This indicates that both the single task attention network with the updated loss and the multitask attention network seem to shift attention away from more polarized terms that occur across topics and towards terms that are more topic specific.

5.3.2 *Effect of topic similarity.*

We next investigated how the models perform based on the similarity between topic 1 and topic 2.

Table 4: Top 30 terms with highest attention scores for a topic pair discussing climate change and gun control.

STAN				STANPP				MTAN			
Terms	Avg Attention	Terms	Avg Attention	Terms	Avg Attention	Terms	Avg Attention	Terms	Avg Attention	Terms	Avg Attention
gun	0.144	democrat	0.010	fossil	0.256	second	0.012	gun	0.564	rifle	0.006
rourke	0.044	people	0.010	protection	0.137	fuel	0.011	rourke	0.059	joe	0.005
percent	0.024	firearm	0.010	environmental	0.087	white	0.011	guns	0.030	work	0.005
trump	0.018	united	0.010	gun	0.084	rifle	0.010	firearm	0.026	energy	0.005
background	0.018	fossil	0.009	rights	0.052	trump	0.009	sanders	0.025	bernie	0.005
guns	0.017	carbon	0.009	control	0.048	energy	0.008	said	0.020	left	0.005
said	0.015	mass	0.008	violence	0.046	donald	0.006	rep	0.013	wednesday	0.005
sanders	0.014	rights	0.008	political	0.030	emissions	0.005	firearms	0.011	senator	0.004
private	0.013	left	0.008	democratic	0.026	national	0.005	activists	0.009	thursday	0.004
industry	0.012	government	0.008	amendment	0.022	carbon	0.004	trump	0.009	deal	0.004
shooting	0.012	democratic	0.008	world	0.019	right	0.004	weapons	0.008	activist	0.004
firearms	0.011	joe	0.008	change	0.019	fuels	0.004	just	0.008	doesn	0.004
democrats	0.011	thursday	0.008	climate	0.019	assault	0.004	background	0.007	rights	0.003
rep	0.010	change	0.007	public	0.016	elizabeth	0.003	released	0.006	republican	0.003
gov	0.010	dead	0.007	nearly	0.015	years	0.002	don	0.006	rifles	0.003

Table 5: Network hyperparameters considered.

Hyperparameter	Values
learning rate	0.01, 0.001, 0.0001,0.00001
epochs	3, 5, 10, 20, 30, 50
batch size	8,16, 32
dropout	0.0, 0.1, 0.3, 0.5
l2 penalty	0.01, 0.05, 0.1
loss weight (α)	0.01, 0.03, 0.05, 0.07, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9

Intuitively, we expect that if the topics are very different, and share few terms, then there is little opportunity for homogenization, and thus we do not expect our models to provide much improvement. On the other hand, if the topics are too similar, then disentangling them will prove challenging. To measure this, we use a simple method to quantify the overlap in predictive terms across two topics. We fit two logistic regression classifiers, one per topic, to predict the political stance of each article. We then select the top terms from each classifier by picking those whose coefficient has magnitude greater than 0.01.[5] Given these two sets of terms, we compute their Jaccard similarity to measure the overlap of each cluster pair. Thus, topics are similar if they share terms predictive of political stance. Figure 3 shows the results for 20 cluster pairs, fit with lowess regression to visualize trends. While there is noticeable variance across cluster pairs, the trends generally match our expectations. The biggest gains occur in the middle of the x-axis, where the topics are neither too similar nor too dissimilar. In future work, it may be helpful to develop diagnostics to determine the divergence between the training and testing set to guide model tuning.

5.3.3 Effect of loss weights. Both the STANPP and MTAN models use a linear interpolation of loss terms (Equations 5 and 6). While these α weights are tuned on the validation set, in order to understand their impact on accuracy, we additionally plot results as we vary the α terms in each equation. We first fix all other hyperparameters in Table 5 found by optimizing on the validation set. Then, we enumerate α values and plot accuracy on the test set in Figure 4. We observe that STANPP performs best with small values of α. When α is too large, the accuracy on topic 1 begins to drop. While topic 2 accuracy continues to increase, the cost to topic 1 accuracy begins to overwhelm the tradeoff. In contrast, MTAN appears relatively stable over a range of α values, until a drop-off once α is greater than 0.7. This suggests that MTAN may be more suitable in settings where it is difficult to carefully tune α.

6 DISCUSSION AND CONCLUSIONS

In this paper, we have identified a specific mechanism that can lead to political homogenization in news recommendation systems, and we have proposed attention-based neural networks to reduce this behavior. The proposed approach exhibits reduction in the impact of political homogenization for simulated users with opposing political leanings across topics. While promising, a considerable amount of work is needed to better understand this phenomenon. First of all, user studies are required to both confirm the propensity of such homogenization as well as to better measure the impact of the proposed approaches, the user study can resemble a randomized control trial where we would have a control and treatment group of users and the treatment effect would be to expose these users to our adapted models (STANPP,MTAN and MTANPP). Second, there is a need to focus on the existing debate about the role of attention in explaining model decisions [18, 42], although these issues appear to be more important in tasks of greater complexity than text classification. Finally, news sources are not monolithic in the viewpoints they publish, which can introduce some bias in the article labels [13], although in aggregate we expect this to have a limited effect.

In future work, it would be valuable to perform algorithmic auditing to quantify the extent to which cross-topic homogenization

[5] This is a somewhat arbitrary threshold; similar trends were found with different thresholds.

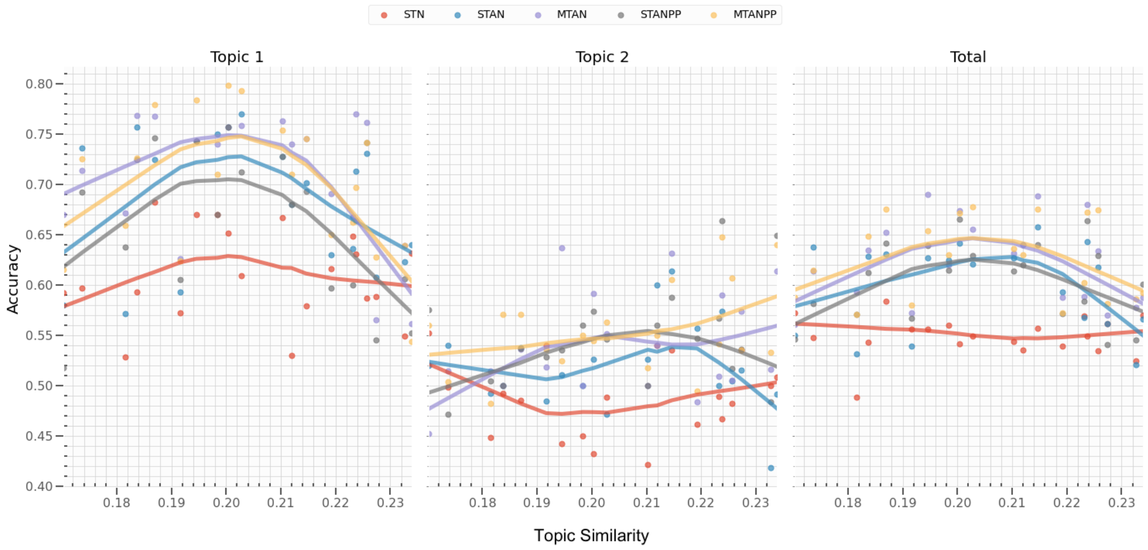

Figure 3: Topic similarity vs test accuracy for 20 topic pairs. The topic similarity is measured using Jaccard similarity between sets of overlapping terms for a given topic pair. The trend lines are generated using a lowess regression model.

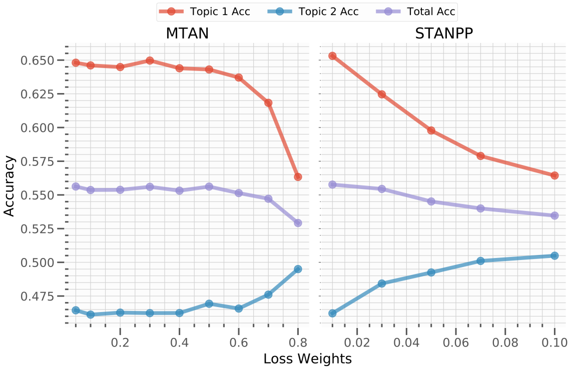

Figure 4: Average test accuracy of 45 topic pairs vs loss weights (α) used in STANPP and MTAN

occurs in deployed recommender systems [40]. We will also experiment with a wider set of recommendation approaches (e.g., collaborative filtering, reinforcement learning) to better understand the variation in algorithmic effects. Recent work to rewrite articles with depolarized terms may also be helpful here [24]. In tandem, these will help us better understand the mechanisms underlying unique forms of user engagement with the news.

7 ETHICAL CONSIDERATIONS

As described in the motivation, news recommendation systems have the potential to exacerbate hyper-partisanship, promote misinformation, and thus degrade civic discourse. While the intent of this paper is to reduce this risk, the risk still remains. Any attempts to personalize news recommendation must weigh these risks with the potential utility of increased user satisfaction achieved by showing articles more relevant to their interests.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation Award #1927407

REFERENCES

[1] Himan Abdollahpouri. 2019. Popularity Bias in Ranking and Recommendation. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 529–530. https://doi.org/10.1145/3306618.3314309

[2] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[3] Jack Bandy and Nicholas Diakopoulos. 2020. Auditing news curation systems: A case study examining algorithmic and editorial logic in apple news. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 36–47.

[4] Rahul Bhargava, Anna Chung, Neil S Gaikwad, Alexis Hope, Dennis Jen, Jasmin Rubinovitz, Belén Saldías-Fuentes, and Ethan Zuckerman. 2019. Gobo: A system for exploring user control of invisible algorithms in social media. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 151–155.

[5] Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* 80, S1 (2016), 250–271.

[6] L. Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. 2019. Controlling Polarization in Personalization: An Algorithmic Framework. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 160–169. https://doi.org/10.1145/3287560.3287601

[7] Allison JB 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*. 224–232.

[8] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. 2019. An attentive survey of attention models. *arXiv preprint arXiv:1904.02874* (2019).

[9] Uthsav Chitra and Christopher Musco. 2020. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 115–123.

[10] P. Dandekar, A. Goel, and D. T. Lee. 2013. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences* 110, 15 (Mar 2013), 5791–5796. https://doi.org/10.1073/pnas.1217220110

[11] Daniel DellaPosta. 2020. Pluralistic collapse: The "oil spill" model of mass opinion polarization. *American Sociological Review* 85, 3 (2020), 507–536.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*

preprint arXiv:1810.04805 (2018).

[13] Soumen Ganguly, Juhi Kulshrestha, Jisun An, and Haewoon Kwak. 2020. Empirical Evaluation of Three Common Assumptions in Building Political Media Bias Datasets. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 939–943.

[14] Kiran Garimella, Tim Smith, Rebecca Weiss, and Robert West. 2021. Political Polarization in Online News Consumption. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 152–162.

[15] R Kelly Garrett. 2009. Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of computer-mediated communication* 14, 2 (2009), 265–285.

[16] Linmei Hu, Chen Li, Chuan Shi, Cheng Yang, and Chao Shao. 2020. Graph neural news recommendation with long-term and short-term interest modeling. *Information Processing & Management* 57, 2 (2020), 102142.

[17] William G Jacoby. 2018. Neither Liberal Nor Conservative: Ideological Innocence in the American Public. *Political Science Quarterly* 133, 4 (2018), 758–761.

[18] Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186* (2019).

[19] M Jurkowitz and A Mitchell. 2020. About one-fifth of Democrats and Republicans get political news in a kind of media buble. *Pew Research Center* (2020).

[20] Dhruv Khattar, Vaibhav Kumar, Vasudeva Varma, and Manish Gupta. 2018. Weave&rec: A word embedding based 3-d convolutional network for news recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1855–1858.

[21] Sami Khenissi and Olfa Nasraoui. 2020. Modeling and Counteracting Exposure Bias in Recommender Systems. arXiv:2001.04832 [cs.IR]

[22] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).

[23] Ping Liu, Karthik Shivaram, Aron Culotta, Matthew A Shapiro, and Mustafa Bilgic. 2021. The Interaction between Political Typology and Filter Bubbles in News Recommendation Algorithms. In *Proceedings of the Web Conference 2021*. 3791–3801.

[24] Ruibo Liu, Lili Wang, Chenyan Jia, and Soroush Vosoughi. 2021. Political Depolarization of News Articles using Attribute-Aware Word Embeddings. In *ICWSM*.

[25] Shikun Liu, Edward Johns, and Andrew J. Davison. 2019. End-to-End Multi-Task Learning with Attention. arXiv:1803.10704 [cs.CV]

[26] Emma Lurie and Eni Mustafaraj. 2019. Opening Up the Black Box: Auditing Google's Top Stories Algorithm. In *The Thirty-Second International Flairs Conference*.

[27] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. *Feedback Loop and Bias Amplification in Recommender Systems*. Association for Computing Machinery, New York, NY, USA, 2145–2148. https://doi.org/10.1145/3340531.3412152

[28] Farzan Masrour, Tyler Wilson, Heng Yan, Pang-Ning Tan, and Abdol Esfahanian. 2020. Bursting the Filter Bubble: Fairness-Aware Network Link Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 841–848.

[29] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[30] Jonathan Mummolo. 2016. News from the other side: How topic relevance limits the prevalence of partisan selective exposure. *The Journal of Politics* 78, 3 (2016), 763–773.

[31] 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*. 677–686.

[32] Ruchi Ookalkar, Kolli Vishal Reddy, and Eric Gilbert. 2019. Pop: Bursting News Filter Bubbles on Twitter Through Diverse Exposure. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 18–22.

[33] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[35] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[36] Pew Research Center. 2021. Beyond red vs. blue: The political typology. https://www.pewresearch.org/politics/2021/11/09 /beyond-red-vs-blue-the-political-typology-2/. Accessed: 2021-12-01.

[37] Markus Prior. 2013. Media and political polarization. *Annual Review of Political Science* 16 (2013), 101–127.

[38] Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. 2021. HieRec: Hierarchical user interest modeling for personalized news recommendation. *arXiv preprint arXiv:2106.04408* (2021).

[39] Zhaopeng Qiu, Yunfan Hu, and Xian Wu. 2022. Graph Neural News Recommendation with User Existing and Potential Interest Modeling. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16, 5 (2022), 1–17.

[40] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 131–141.

[41] Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* (2017).

[42] Sofia Serrano and Noah A Smith. 2019. Is Attention Interpretable?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2931–2951.

[43] Yotam Shmargad and Samara Klar. 2020. Sorting the news: How ranking by popularity polarizes our politics. *Political Communication* 37, 3 (2020), 423–446.

[44] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs.CL]

[45] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2576–2584.

[46] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 6389–6394.

[47] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Empowering news recommendation with pre-trained language models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1652–1656.

[48] Guixian Xu, Yueting Meng, Zhan Chen, Xiaoyu Qiu, Changzhi Wang, and Haishen Yao. 2019. Research on topic detection and tracking for online news texts. *IEEE access* 7 (2019), 58407–58418.

[49] Yongxin Yang and Timothy M Hospedales. 2016. Trace norm regularised deep multi-task learning. *arXiv preprint arXiv:1606.04038* (2016).

[50] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 1480–1489.

[51] Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang, and Xiuqiang He. 2021. UNBERT: User-News Matching BERT for News Recommendation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*. 3356–3362.

[52] Qiannan Zhu, Xiaofei Zhou, Zeliang Song, Jianlong Tan, and Li Guo. 2019. Dan: Deep attention neural network for news recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 5973–5980.