Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-022-01473-1 OPEN

Removing AI's sentiment manipulation of personalized news delivery

Chuhan Wu [1], Fangzhao Wu [2]✉, Tao Qi [1], Wei-Qiang Zhang [1], Xing Xie [2] & Yongfeng Huang [1,3,4]✉

Artificial intelligence (AI) is empowering personalized online news delivery to accommodate people's information needs and combat information overload. However, AI models learned from user data are inheriting and amplifying some underlying human prejudice such as the sentiment bias of news reading, which may lead to potential negative societal effects and ethical concerns. Here, substantial evidence shows that AI is manipulating the sentiment orientation of news displayed to users by promoting the presence chance of negative news, even if there is no human interference. To mitigate this manipulation, a sentiment-debiasing method based on a decomposed adversarial learning framework is proposed, which can reduce 97.3% of sentiment bias with only 2.9% accuracy sacrifice. Our work provides the potential in improving AI's responsibility in many human-centered applications such as online journalism and information spread.

[1] Department of Electronic Engineering, Tsinghua University, 100084 Beijing, China. [2] Microsoft Research Asia, 100080 Beijing, China. [3] Zhongguancun Laboratory, 100094 Beijing, China. [4] Institute for Precision Medicine, Tsinghua University, 100084 Beijing, China. ✉email: fangzwu@microsoft.com; yfhuang@tsinghua.edu.cn

Introduction

News information is essential for people to be informed of the events, characters, and communities in the outside world (Leban et al., 2014; McCombs and Reynolds, 2002). Different from the print and broadcast media, the widespread Web connections have endowed online news information with unprecedented geographic reach and spreading speed (Althaus and Tewksbury, 2000; Wu, 2007). Thus, online platforms such as digital news portals and social media websites have become a primary source for many people to consume news information (Thurman, 2008). To alleviate the information overload brought by the vast amount of news information, only a small set of news picked by online platforms is displayed to their users (Das et al., 2007). Instead of manually choosing news by human editors, many online platforms are employing artificial intelligence (AI) techniques (LeCun et al., 2015) to select news in a personalized way to accommodate individual information needs (Okura et al., 2017), which have achieved notable success in improving the information acquisition efficiency of users (Moller, 2022; Vermeulen, 2022).

Unfortunately, machine-aided news delivery is not as credible as we expect. They can be intentionally intervened by humans to manipulate certain aspects of news delivery, such as sentiment and opinions, as Facebook's "emotional contagion" experiment (Kramer et al., 2014) did. Such a study caused an uproar among the academia and public about the risks of potentially unethical use of AI techniques in human-centered applications (Davies, 2016; Del Vicario et al., 2016; Hallinan et al., 2020; Larson, 2018; Ruxton and Mulder, 2019). More recently, Facebook is accused of using algorithms to amplify hateful or harmful content in the news feed to optimize its profit ("60 Minutes" interview, Facebook whistleblower Frances Haugen; Hemphill and Banerjee, 2021). Beyond financial incentives, intentional manipulation of displayed news sentiment with political motives has shown great power in swaying the outcome of political events like elections (Bovet and Makse, 2019, Gu et al., 2017; Ratkiewicz et al., 2011). Thus, deliberate or malicious manipulation of news sentiment can bring considerable threats to individuals, society, and democracies (Gallotti et al., 2020; Kucharski, 2016; Mihaylov et al., 2018, 2015; Shao et al., 2018).

Although human-involved manipulation of news sentiment has been perceived and can be prohibited by laws in the future (Beridze and Butcher, 2019), personalized news recommender AI itself can manipulate news sentiment without human interference due to the problem of AI's algorithm bias (Gibney, 2020; Zou and Schiebinger, 2018), as shown in Fig. 1. This is mainly because when learning AI models on massive user data, they can inherit and even amplify the biases encoded in human behaviors (Courtland, 2018). As the proverb goes, "for evil news rides fast, while good news baits later" (John Milton), users prefer to interact with negative news articles rather than positive ones (Hornik et al., 2015; Naveed et al., 2011). AI recommender systems may capture this pattern and form their sentiment prejudices in news selection, which leads to the sentiment manipulation of recommended news. As a human-in-the-loop system, the sentiment bias is further magnified during the iterative interactions between users and news feed providers, which may generate unforeseeable negative psychological and societal impacts (Han et al., 2019; Johnston and Davey, 1997).

In fact, researchers are aware of the significant impact of sentiment information on personalized recommender systems. Many methods explore how to incorporate sentiment information from user-generated content, e.g., reviews in Yang et al. (2013) and social media posts (Khattak et al., 2020; Kumar et al., 2020; Sun et al., 2018) into recommendation algorithms, which can bolster the model's ability to model item properties (Huang et al., 2020) and user preferences (Gurini et al., 2013). Some recent studies even successfully encourage the model to enhance recommendation diversity in the sentiment dimension (Wu et al., 2020a). However, the sentiment signal in recommender systems is a mixed blessing, since it may introduce unwanted biases to the recommendation results. Unfortunately, the effects of sentiment bias in recommender systems are rarely studied. Only a few works study the influence of review sentiment on recommendation accuracy (He et al., 2022; Lin et al., 2021), which is the tip of the iceberg of sentiment bias's evil with very limited societal impacts.

In this study, we reveal the sentiment manipulation phenomenon of AI in personalized news delivery. Through extensive experiments on a large-scale real-world news recommendation dataset (Wu et al., 2020b) with one million users, we discover that users' biased preferences for negative news sentiment can be captured by various state-of-the-art AI models when optimizing recommendation accuracy. These models further reinforce the sentiment bias by promoting the presence chance of negative news in the recommendation results, which may pose potential

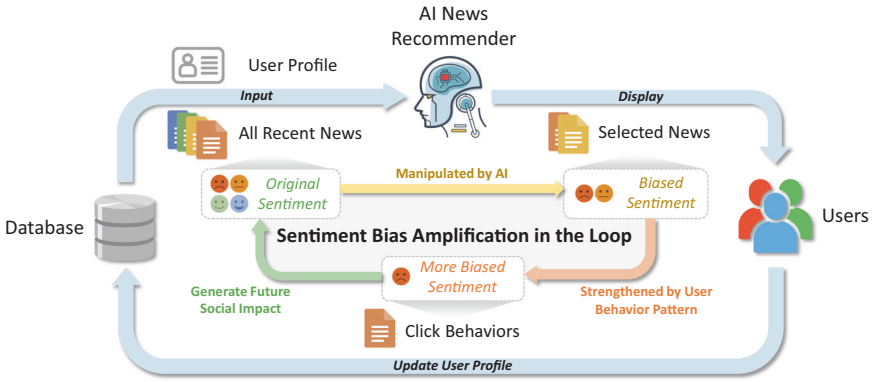

Fig. 1 The amplification of sentiment bias in the loop of human-AI interactions. The AI news recommender selects a few news articles from the full set of recent news according to users' personal interests inferred from the user profile. Users interact with the selected news displayed to them, and their behaviors such as clicks are used to update the user profile in the database. In this loop, since users have biased preferences for negative news sentiment, the recommendation AI learned on user data can inherit and amplify the sentiment bias, which leads to AI's manipulation of the sentiment of selected news. Users' further biased behaviors can strengthen the sentiment bias, and such highly biased sentiment orientation evoked by a large number of users can generate future social impacts and influence the overall sentiment of future news. The dilemma of sentiment bias amplification in the loop can make AI heavily control the sentiment of news displayed to users.

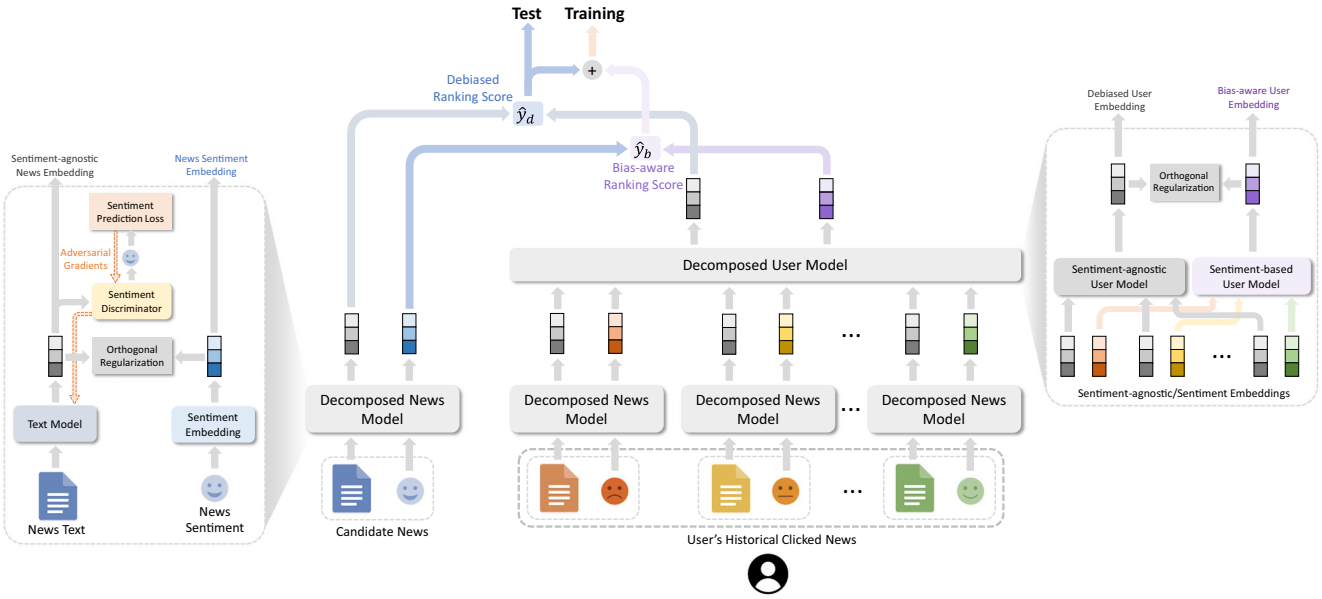

Fig. 2 The framework of our sentiment-debiasing approach. It removes the sentiment manipulation of personalized news recommendations by learning sentiment-agnostic news and user representations via decomposed adversarial learning.

risks to the public. Since such unwanted news sentiment manipulation is mainly brought by the algorithm's sentiment biased learned from user data, we propose a sentiment-debiasing method based on a decomposed adversarial learning framework (Wu et al., 2021) to remove AI's sentiment manipulation. Our approach aims to build up a debiased sentiment-agnostic model from the biased data, to achieve fair news selection concerning different sentiments. Experimental results show that our method can reduce the vast majority of sentiment bias introduced by the AI model to mitigate its sentiment manipulation under minor performance loss. The results also reveal that our approach can further improve the sentiment diversity of news distribution. The insights provided by our study can help the public be aware of the potential risks of AI-empowered news personalization techniques, and inspire researchers to improve the responsibility of AI involved in Internet journalism and other channels of information spread for the well-being of humans.

Methods

Problem formulation. Given a target user u, we denote his/her historical clicked news as $[D_1, D_2, \ldots, D_N]$, where N is the history length. Given a candidate news article D_c, the goal of the recommendation model is to predict a click score $\hat{y}$ that indicates the (non-normalized) probability of the user u clicking D_c. A set of candidate news is ranked according to the corresponding click scores, and the top news with the highest click scores is displayed to the user u. In addition, we denote the sentiment polarity categories of clicked news and candidate news as $[s_1, s_2, \ldots, s_N]$ and s_c, respectively. The goal of our method is to rank clicked candidate news at high positions and meanwhile keep the overall sentiment orientation in top recommendation results to be consistent with the average sentiment of the news corpus.

Framework. Next, we introduce the details of our proposed sentiment-debiasing framework that can remove the model's sentiment manipulation (Fig. 2). The core of this framework is a decomposed news model that aims to learn sentiment-aware and sentiment-independent news information, and a decomposed user model that captures sentiment-related user interests and sentiment-independent user interests. Their details are described as follows.

As shown in the left box in Fig. 2, the decomposed news model takes the news texts and news sentiment as the input. Here the news sentiment is inferred from news texts. We use VADER (Hutto and Gilbert, 2014) to compute a real-valued sentiment score for each news, and then quantize this score and convert it into a discrete sentiment category s as the input. The news texts are processed by a text model that learns a hidden embedding to represent the semantic information of news. Following the text modeling approach in NRMS Wu et al. (2019c), we first convert the word in the news texts into a sequence of word embeddings through a word embedding lookup table, then use a multi-head self-attention (Vaswani et al., 2017) network to learn hidden word representations by capturing the interactions among words, and finally use an attention pooling network to summarize the hidden word representations into a unified news text representation, which is denoted as $\mathbf{h}_t$. The sentiment category is converted into a latent embedding $\mathbf{h}_s$.

Since the text representation $\mathbf{h}_t$ learned from news texts may still contain sentiment information, we apply an additional orthogonal regularization to the text embedding $\mathbf{h}_t$ and the sentiment embedding $\mathbf{h}_s$ to encourage them to be orthogonal. The regularization loss function $\mathscr{L}_R$ is formulated as follows:

$$\mathscr{L}_R = \frac{|\mathbf{h}_t \cdot \mathbf{h}_s|}{||\mathbf{h}_t|| \cdot ||\mathbf{h}_s||}, \tag{1}$$

where $||\cdot||$ means the L_2 norm. By optimizing this regularization loss, the text embedding usually contains less sentiment information. However, this loss usually cannot be perfectly optimized and the sentiment embedding may also have some shifts with the real sentiment space, making the text embedding still encode some sentiment information. To further reduce the sentiment information it contains, we apply adversarial learning to purify it. Specifically, a sentiment discriminator is used to predict the sentiment category s from the text embedding $\mathbf{h}_t$. The soft sentiment category label $\hat{\mathbf{s}}$ is predicted as follows:

$$\mathbf{s} = \text{softmax}(\mathbf{W}\mathbf{h}_t + \mathbf{b}), \tag{2}$$

where $\mathbf{W}$ and $\mathbf{b}$ are linear projection parameters. The loss function $\mathscr{L}_D$ for learning the sentiment discriminator is as

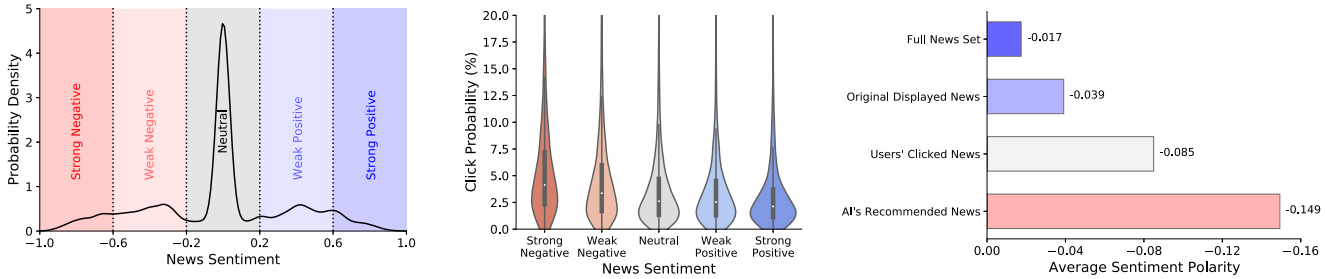

Fig. 3 Sentiment bias and AI's sentiment manipulation. Left: the sentiment distribution of news in the dataset. We categorize the sentiment polarities according to the real-valued sentiment scores. Most news has very weak or neutral sentiment while other has positive or negative sentiment orientation with stronger intensities. The overall sentiment of the news corpus is nearly neutral (the average sentiment score is −0.0174). Middle: the click probability of news with different sentiment polarities. News articles with more negative sentiments are more likely to be clicked by users, which is the core source of sentiment bias. The differences in click probabilities among different sentiment polarity categories are significant ($p < 0.001$ according to two-sided t-tests). Right: average sentiment scores of the full news set, the news displayed to users in the training data, users' clicked news in the training data, and top 50 recommendation results given by a state-of-the-art news recommendation model NRMS (Wu et al., 2019c). The negative sentiment is amplified in a cascaded way due to users' biased news reading choices and AI's algorithm biases learned from user data. This provides evidence of AI's news sentiment manipulation in the loop of human-machine interactions on news delivery platforms.

follows:

$$\mathscr{L}_{\mathrm{D}} = -\sum_{i=1}^{C} \mathbf{s}_i \log(\hat{\mathbf{s}}_i), \tag{3}$$

where C is the number of sentiment categories, $\mathbf{s}_i$ and $\hat{\mathbf{s}}_i$ are the real and predicted labels for the ith class. The negative gradients inferred by the sentiment discriminator are used to learn the text model in an adversarial way to encourage it to remove sentiment information. When the discriminator and the text model achieve a Nash equilibrium, most sentiment information encoded in the text embedding $\mathbf{h}_{\mathrm{t}}$ can be effectively removed. Thus, $\mathbf{h}_{\mathrm{t}}$ can be regarded as a sentiment-agnostic news embedding. We apply the decomposed news model to the user's clicked news and candidate news to learn their sentiment-agnostic embeddings and sentiment embeddings. We denote the sentiment-agnostic embeddings of clicked news and candidate news as $[\mathbf{h}_{\mathrm{t},1}, \mathbf{h}_{\mathrm{t},2}, \mathbf{h}_{\mathrm{t},N}]$ and $\mathbf{h}_{\mathrm{t,c}}$, respectively. The sentiment embeddings of them are denoted as $[\mathbf{h}_{\mathrm{s},1}, \mathbf{h}_{\mathrm{s},2}, \mathbf{h}_{\mathrm{s},N}]$ and $\mathbf{h}_{\mathrm{s,c}}$, respectively.

The decomposed user model takes the sentiment-agnostic and sentiment embeddings of clicked news as the input. It contains a sentiment-agnostic user model to learn a debiased user embedding $\mathbf{u}_{\mathrm{d}}$ from sentiment-agnostic news embeddings and a sentiment-based user model to learn a bias-aware user embedding $\mathbf{u}_{\mathrm{b}}$ (right box in Fig. 2). The debiased user embedding is mainly used to capture sentiment-independent user interest, and the bias-aware user embedding aims to encode sentiment biases. Following NRMS (Wu et al., 2019c), we use two independent multi-head self-attention networks with attention pooling modules to capture the relatedness between different news and learn unified user embeddings. Although the sentiment-aware and sentiment-independent information is nearly decomposed in the news model, the user model may further encode sentiment information into the user embedding. Thus, we apply an additional orthogonal regularization loss $\mathscr{L}'_{\mathrm{R}}$ to the user embeddings learned by the two user models, which is formulated as follows:

$$\mathscr{L}'_{\mathrm{R}} = \frac{|\mathbf{u}_{\mathrm{d}} \cdot \mathbf{u}_{\mathrm{b}}|}{||\mathbf{u}_{\mathrm{d}}|| \cdot ||\mathbf{u}_{\mathrm{b}}||}. \tag{4}$$

By optimizing this loss, the user interest information can also be effectively decomposed into sentiment-aware and sentiment-independent components.

After learning the decomposed news and userembeddings, we compute two ranking scores based on them. One score is a debiased ranking score (denoted as $\hat{y}_{\mathrm{d}}$), which measures the relevance between debiased user embedding and the sentiment-agnostic candidate news embedding via the inner product (i.e., $\hat{y}_{\mathrm{d}} = \mathbf{u}_{\mathrm{d}} \cdot \mathbf{h}_{\mathrm{t,c}}$). This score reflects the matching degree of candidate news content and debiased user interest. Another score is a bias-aware ranking score (denoted as $\hat{y}_{\mathrm{b}}$), which is computed by the relevance between bias-aware user embedding and the sentiment embedding of candidate news using their inner product (i.e., $\hat{y}_{\mathrm{b}} = \mathbf{u}_{\mathrm{b}} \cdot \mathbf{h}_{\mathrm{s,c}}$). This score reflects the impact of sentiment bias on users' click behaviors. To capture the sentiment bias patterns in the training data, both scores are added into a unified score $\hat{y}$ for model training. Following many prior studies (Wu et al., 2019b, c), we use the negative sampling method to construct representative training samples. More specifically, for each clicked news D_{c}^{+} (regarded as a positive sample), we sample T non-clicked news $[D_{\mathrm{c},1}^{-}, D_{\mathrm{c},2}^{-}, ..., D_{\mathrm{c},T}^{-}]$ (regarded as negative samples) and jointly predict their click scores (the choice of T is discussed in Supplementary Fig. 6). The loss function $\mathscr{L}_{\mathrm{p}}$ for learning the recommendation model is formulated as follows:

$$\mathscr{L}_{\mathrm{p}} = -\log\left(\frac{\exp(\hat{y}^{+})}{\exp(\hat{y}^{+}) + \sum_{i=1}^{T} \exp(\hat{y}_i^{-})}\right), \tag{5}$$

where $\hat{y}^{+}$ and $\hat{y}_i^{-}$ stand for the click scores of the positive sample and its associated ith negative sample. In the test stage, only the debiased ranking score $\hat{y}_{\mathrm{d}}$ is used for ranking. In this way, the influence of sentiment bias is removed from the recommendation results. To learn the entire model, the unified loss function $\mathscr{L}$ on each training sample $(D_{\mathrm{c}}^{+}, D_{\mathrm{c},1}^{-}, D_{\mathrm{c},2}^{-}, ..., D_{\mathrm{c},T}^{-})$ is formulated as follows:

$$\mathscr{L} = \mathscr{L}_{\mathrm{p}} - \frac{\alpha}{N+T+1} \sum_{d \in \mathscr{D}} \mathscr{L}_{\mathrm{D}}^{d} + \beta(\mathscr{L}'_{\mathrm{R}} + \frac{1}{N+T+1} \sum_{d \in \mathscr{D}} \mathscr{L}_{\mathrm{R}}^{d}), \tag{6}$$

where $\mathscr{D}$ means the union of historical clicked news, positive sample and negative samples, $\mathscr{L}_{\mathrm{D}}^{d}$ and $\mathscr{L}_{\mathrm{R}}^{d}$ represent the adversarial loss and regularization loss on the news d, and α and β are two coefficients that control the intensity of the adversarial loss and the orthogonal regularization loss, respectively (the selection of these coefficients is shown in Supplementary Fig. 5). The loss function for training the discriminator is $\frac{1}{N+T+1}\sum_{d \in \mathscr{D}} \mathscr{L}_{D}^{d}$. By training the discriminator and the recommendation model towards convergence, our model can be effectively debiased to get rid of the sentiment manipulation issue. Since the recommendation model and the sentiment discriminator are two adversaries, they cannot be optimized

simultaneously. Thus, we adopt a batch-wise training method to learn them in turn on each batch of training samples, as shown in Algorithm 1. In this way, the two adversaries can be jointly trained on the same data.

Algorithm 1. Training algorithm of our approach
1: Initialize the recommendation model parameter set Θ_m and the sentiment discriminator parameter set Θ_d
2: **repeat**
3: Randomly select a batch of samples s from the entire training set $\mathcal{S}$
4: Freeze the recommendation model parameter set Θ_m
5: Compute $\mathcal{L}_D$ on s
6: Optimize Θ_d based on $\mathcal{L}_D$
7: Freeze the sentiment discriminator parameter set Θ_d
8: Compute $\mathcal{L}$ on s
9: Optimize Θ_m based on $\mathcal{L}$
10: **until** model convergence

Result

AI's manipulation of news delivery sentiment. We perform analysis and experiments on a public large-scale news recommendation dataset named MIND (Wu et al., 2020b), which is constructed by real interaction logs of 1 million users collected on the Microsoft News platform during 6 weeks from October 12 to November 22, 2019. The sentiment of each news article is indicated by a real value from −1 to 1 (see the "Methods" section). We classify news sentiment into five categories according to polarity and intensity. From the sentiment distribution of news in the corpus (Fig. 3 left), we observe that most news has neutral sentiment, and the overall sentiment orientation of the full news set is nearly neutral (the average sentiment score is −0.0174). However, the click probabilities of news with different sentiments have significant differences (Fig. 3 middle), where $p < 0.001$ among different sentiment categories. It verifies users' biased behavior patterns of news reading, i.e., more negative news is more likely to attract clicks. In fact, many news categories with strong negative sentiment (Supplementary Table 2) involve common topics, such as health, crime, and disaster, which can be consumed by a broader audience than topics with specific interests (e.g., soccer and basketball).

To investigate AI's sentiment manipulation phenomenon, we compare the average sentiment of the full news set, the news displayed to users in this dataset, users' clicked news, and the top news recommended by a state-of-the-art (SOTA) AI-based news recommendation approach (Wu et al., 2019c) (Fig. 3 right). The results indicate that the displayed news articles amplify the negative sentiment by 124% compared with the full news set, which is mainly due to the sentiment bias of the original recommender system for generating this dataset. The negative sentiment orientation is strengthened by users' click behaviors (+117%) because of the biased user preferences for negative news sentiment. The SOTA news recommendation AI learned on such click data further magnifies the negative sentiment 1.76 times in its top recommendation results. The cascaded amplification of negative sentiment reveals the worrying increase of sentiment bias in the loop of human–machine interactions, where news sentiment may be heavily manipulated by AI after multiple rounds of biased data accumulation and biased AI model learning.

Results of sentiment-debiasing. To verify the effectiveness of our proposed sentiment-debiasing method in removing AI's sentiment manipulation, we compare it with several SOTA AI-empowered news recommendation methods (An et al., 2019; Liu et al., 2020; Okura et al., 2017; Wang et al., 2018, Wu et al., 2019a, c) in terms of sentiment bias and recommendation accuracy. The recommendation accuracy is indicated by Area under the ROC Curve (AUC) score and the normalized Discounted Cumulative Gain (nDCG) score of the top 10 recommended news (Wu et al., 2020b), which are formulated as follows:

$$\mathrm{AUC} = \frac{\sum_{p\in\mathscr{P}}\sum_{n\in\mathscr{N}} I[P(p) > P(n)]}{|\mathscr{P}||\mathscr{N}|}, \tag{7}$$

$$\mathrm{nDCG@K} = \frac{\sum_{i=1}^{K}(2^{r_i}-1)/\log_2(1+i)}{\sum_{i=1}^{N_p} 1/\log_2(1+i)}, \tag{8}$$

where $P(\cdot)$ is the predicted click score of a sample, $\mathscr{P}$ and $\mathscr{N}$, respectively denote the positive and negative sample sets, and $I[\cdot]$ is an event indicator function. The symbol N_p represents the number of positive samples, and r_i is a relevance score of news with the ith rank, which is 1 for clicked news and 0 for non-clicked news. Note that nDCG@10 is an instance of nDCG@K that computes the metric based on the top 10 recommendation results. Since the MIND dataset provides the real impression logs, we use the candidate news in each impression to compute the metrics of recommendation accuracy. The sentiment bias can be reflected by the average sentiment of top K-recommended news. Since the original impression data in the dataset already contained some sentiment bias, it cannot be used to evaluate the removing degree of sentiment bias. Instead, we use the entire news set as the candidate news set to be ranked and use the average sentiment of top K-ranked news as the sentiment bias measurement. In our experiments, we repeat each experiment 5 times, and the average performance with 0.95 confidence intervals (if applicable) is illustrated. The ideally minimal bias is benchmarked by the average sentiment of randomly ranked news (i.e., the average sentiment of a full news set), and the absolute difference between this benchmark and the average sentiment of top recommendation results generated by AI algorithms is used as the metric for quantitatively evaluating AI's sentiment bias, where smaller sentiment biases indicate lighter sentiment manipulations.

The sentiment bias comparison (left Fig. 4) shows that all compared SOTA baseline methods introduce heavy sentiment bias, which provides consistent evidence of AI's sentiment manipulation by amplifying the ratio of negative content in news delivery. The average sentiment of our approach is very close to random ranking, which represents that most sentiment bias is eliminated. Specifically, the sentiment bias in the top 50 recommended news is reduced by 97.3% (compared with its basic model NRMS; Wu et al., 2019c) and is reduced by 96.7% compared with the least biased method DKN (Wang et al., 2018). From the recommendation accuracy results (right Fig. 4), our approach can achieve comparable performance with other SOTA methods. It has only 2.9% absolute AUC and 2.5% nDCG@10 sacrifice compared with the best-performed NRMS model. These results verify the effectiveness of our methodology in reducing sentiment bias without heavy performance loss.

To further understand the impact of sentiment debiasing on the recommendation results, we compare our approach with its basic model NRMS (Wu et al., 2019c) in terms of the sentiment distributions of their recommended news as well as the sentiment correlations between recommended news and users' historical clicked news (Fig. 5). We find in debiased recommendation results, the ratio of negative news is reduced while positive news is promoted (upper left Fig. 5). In addition, the overall sentiment intensity is slightly decreased (from 0.3311 to 0.3286, t-test $p < 0.01$), which means that our debiased model tends to recommend less emotional content (upper middle Fig. 5). In addition, we observe a huge sentiment standard deviation

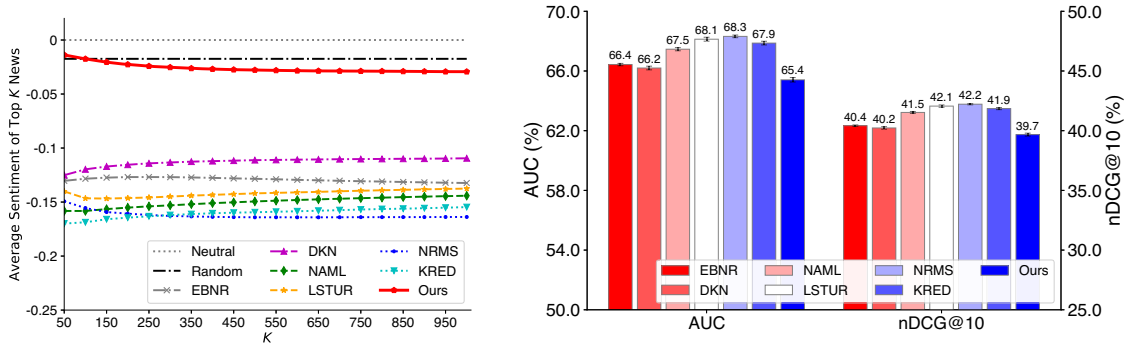

Fig. 4 The sentiment bias and recommendation performance of different methods. Left: the average sentiment scores of top K news recommended by different methods. The "Random" dashed line (black) represents recommending news randomly, and the expectation of its average sentiment is the average sentiment of the full news set. We use this score as an unbiased benchmark, and the distance to it is regarded as sentiment bias. The average sentiments of all compared SOTA deep learning-based news recommendation methods: EBNR (Okura et al., 2017), DKN (Wang et al. (2018), NAML (Wu et al., 2019a), LSTUR (An et al., 2019), NRMS (Wu et al., 2019c), and KRED (Liu et al., 2020) are much more negative than the unbiased benchmark, which indicates their sentiment manipulation phenomenon. The average sentiment of news recommended by our approach (red line) is very close to the benchmark, especially for the top 50 news articles that are preferentially displayed to users. It shows that our approach effectively mitigates AI models' sentiment manipulation. Right: the recommendation accuracy is evaluated by the AUC and nDCG@10 scores of news ranking. The results show that our approach achieves comparable results with other SOTA methods (the maximal performance drop is 2.9% AUC and 2.5% nDCG@10). The error bars represent mean scores with 0.95 confidence intervals ($n = 5$ independent experiments).

difference (t-test $p < 0.001$) between the original and debiased models (upper right Fig. 5). This shows that our debiased approach tends to recommend news with various sentiments, which can promote the sentiment diversity (Wu et al., 2020a) of news distribution to individuals. From lower Fig. 5, we find that the sentiment of recommended news given by the original biased model is correlated to the average sentiment of users' clicked news significantly (Pearson $r = 0.5109$, $p < 0.001$), while there is no such significant correlation in debiased recommendation results (Pearson $r = -0.0030$, $p = 0.7569$). These results reveal that biased AI models may tend to provide users with content with homogeneous sentiment, which may strengthen the polarization of social opinions. Our approach has a greater ability in recommending news with diverse sentiments, which can help mitigate the filter-bubble problem (Bergstrom and Bak-Coleman, 2019) to better satisfy users' diverse needs on news information (see Supplementary Fig. 4 for an example).

Recommendation topic analysis. We then analyze the high-frequency topic categories in the original news set and the recommendation results (Fig. 6, the topic categories are sorted in descending order by their frequencies). The "newscrime" category has a strong negative sentiment orientation, but its rank is promoted in the recommendation results without debiasing, which is an indication of the amplification of negative sentiment. Although crime news can effectively attract users' attention, it may be inappropriate to display crime news excessively because of its potential societal impacts (Mastro et al., 2009). By contrast, in the debiased recommendation results generated by our approach, the position of the "newscrime" category is degraded. In addition, topics with relatively strong positive sentiment such as "recipes" and "lifestyleroyals" gain more display chances. These results further support the effectiveness of our sentiment-debiasing approach in reducing the sentiment bias related to the amplification of negative sentiment.

Model component analysis. Next, we verify the effectiveness of the decomposed adversarial learning framework in our approach (see the "Methods" section for more details). We use the leave-one-out scheme to evaluate the contributions of the core techniques in our approach, including the adversarial learning mechanism, orthogonal regularization, and the decomposition framework. From the results of recommendation accuracy and sentiment bias (Fig. 7), we observe that the adversarial learning mechanism plays the most important role in reducing sentiment bias, though it has some sacrifice on recommendation accuracy. The orthogonal regularization can improve accuracy and meanwhile eliminate sentiment bias. This is because it encourages the model to disentangle sentiment-aware and sentiment-independent information, which can aid the elimination of sentiment bias. The decomposition framework shows great importance, especially in keeping recommendation accuracy. Since removing sentiment bias and optimizing user clicks can be contradictory, it can be difficult for the canonical adversarial training method (Zhang et al., 2018) without information decomposition to balance debiasing and performance. These experimental results corroborate the effectiveness of our methodology in alleviating AI's sentiment manipulation without heavy performance decreases.

Discussion

With the explosion of online information, people's daily lives depend heavily on personalized services to alleviate information overload (Littman, 2015). Among them, personalized news delivery is a special one that can generate huge impacts on users' emotions, decisions, and views on the world outside (Fischer et al., 2020). Although AI techniques have been successfully incorporated into many news recommender systems to improve user experiences, their potential ethical risks and intrinsic causes are not fully identified nor addressed. Our work provides quantitative empirical evidence that news recommendation AI can be manipulating the sentiment orientation of news for display by increasing the recommendation chances of news with stronger negative sentiments. Since users have biased behaviors towards news with different sentiments, AI models learned on big user data will encode these sentiment biases and generate more biased recommendation results. The sentiment bias can be amplified in the loops of human–AI interactions, which leads to heavier sentiment manipulation by news recommender models. Since users are vulnerable to the sentiment manipulation of news feeds

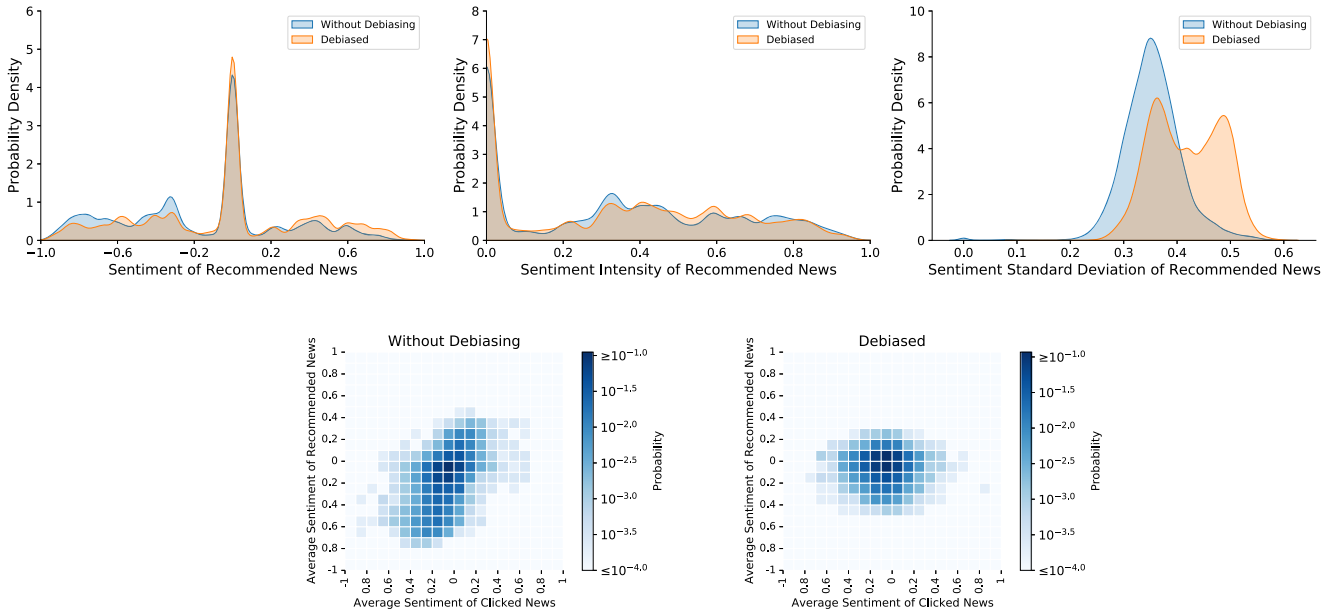

Fig. 5 Impact of sentiment debiasing on the sentiment of recommended news. Upper: the distributions of sentiment orientation, sentiment intensity, and sentiment variance of the biased or debiased recommendation results. The left plot shows that negative news is demoted in debiased recommendation results while positive and neutral news articles are promoted. The middle plot shows the sentiment intensity of debiased recommendations is slightly weaker than biased ones ($p < 0.01$). The right plot shows that the sentiment standard deviation of debiased recommendations is much larger than biased ones, indicating that our sentiment-debiasing method improves sentiment diversity. Lower: the correlations between the average sentiment of clicked news and recommended news given by biased or debiased models. Darker colors indicate higher probability densities. The left plot shows the sentiments of news recommended by biased models have significant correlations with historically clicked news ($r = 0.5109$, $p < 0.001$). The right plot indicates that in debiased recommendation results such correlation is not significant ($r = -0.0030$, $p = 0.7569$).

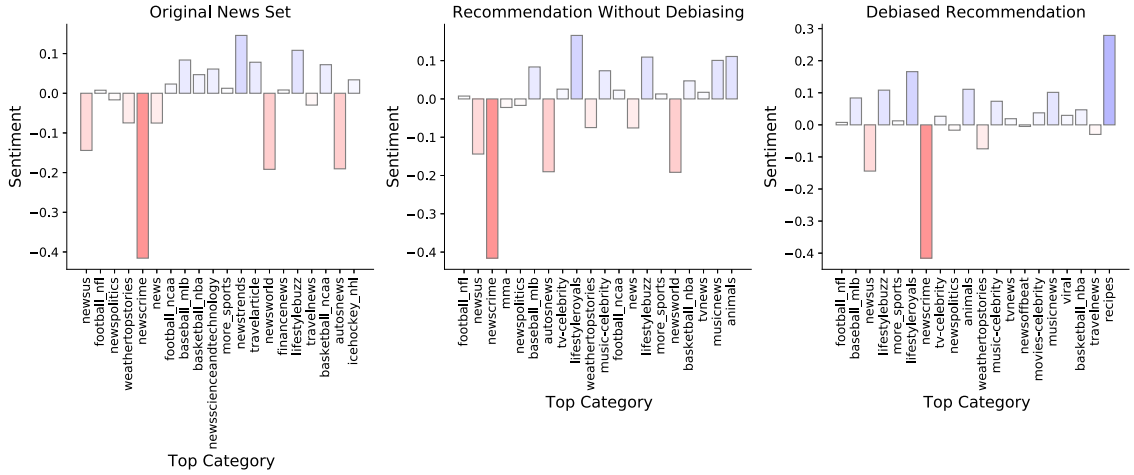

Fig. 6 Sentiment analysis of news topics. The top-frequency fine-grained news topic categories with their average sentiment orientations in the original news set, recommendations without debiasing, and debiased recommendations. The topic categories are sorted by their frequencies in descending order (from left to right). The results show that some topics with strong negative sentiment orientation are promoted by the biased recommender, while our debiased model demotes some negative news topics such as "newscrime" and promotes news topics with positive sentiment such as "recipes".

(Chen et al., 2021), using biased AI for news selection has great risks of generating unforeseeable negative societal impacts. We should be vigilant about AI's sentiment manipulation brought by unwanted algorithm biases when developing and using personalized news feed services.

To get rid of AI's sentiment manipulation of personalized news delivery, in this work we propose a sentiment-debiasing method to eliminate the model's sentiment bias inherited from user data. We decompose news information into a sentiment-aware component and a sentiment-independent component and regularize them to be orthogonal. By applying adversarial learning to the sentiment-independent part, its encoded sentiment bias can be effectively removed, and thereby the recommendation results are sentiment-agnostic. Our approach can reduce most of AI's sentiment bias with minor accuracy loss, which indicates that the sentiment manipulation problem is effectively mitigated without severely harming user experiences. Our work can promote the responsibility of AI-empowered news delivery to provide users with both effective and trustworthy information acquisition resources. In addition, our

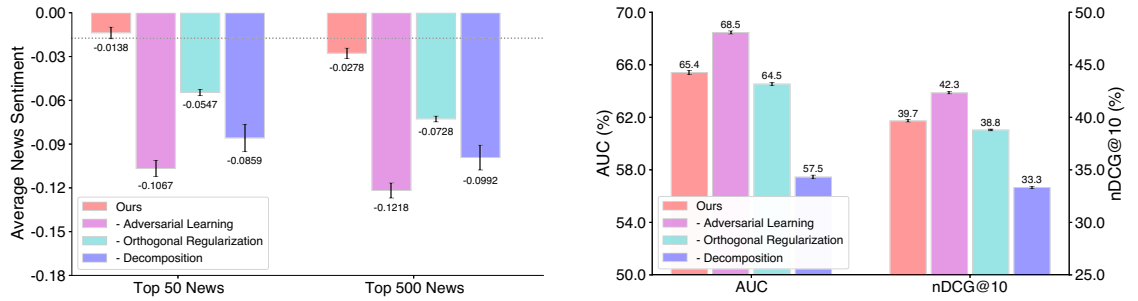

Fig. 7 Effectiveness of the core techniques used in our approach. The contribution of each module is evaluated by the changes in sentiment bias and recommendation performance when removing it from our approach. Left: the sentiment bias is indicated by the average sentiment of the top 50 and top 500 news. The dashed line represents the unbiased benchmark. Right: the recommendation performance is indicated by the AUC and nDCG@10 scores. The adversarial learning mechanism contributes most to the removal of sentiment bias. The orthogonal regularization technique improves model performance and decreases sentiment bias. The decomposition framework has a major contribution to the model performance. Differences between different bars are significant ($p < 0.01$ in the left figure and $p < 0.001$ in the right figure according to the two-sided t-test). Error bars stand for mean scores with 0.95 confidence intervals ($n = 5$ independent experiments).

proposed methodology can be generalized to reduce other types of biases in AI systems, such as gender (Park et al., 2018) and racial (Obermeyer et al., 2019) biases, to build more controllable, inclusive, and fair machine intelligence for the good of humanity.

However, we still need to be cautious when handling sentiment biases in news recommendations, since removing them can change the impacts of other types of biases (e.g., gender bias, see Supplementary Fig. 5) on the recommendation results. This chain reaction may amplify (or fortunately alleviate) the bias effects on the news information delivered to users. In our future work, we would like to study how to jointly mitigate the effects of multiple types of biases on the personalized recommendations.

Data availability

The MIND dataset used by this study is publicly available at https://msnews.github.io/. The use of this dataset adheres to the Microsoft Research License Terms (on the same webpage).

Code availability

Code used for this study has been publicly available at https://github.com/wuch15/Sentiment-debiasing. Experiments and implementation details are described in sufficient detail in the Methods section or in the Supplementary Information.

Received: 30 June 2022; Accepted: 30 November 2022;
Published online: 20 December 2022

References

Althaus SL, Tewksbury D (2000) Patterns of internet and traditional news media use in a networked community. Political Commun 17(1):21–45
An M, Wu F, Wu C, Zhang K, Liu Z, Xie X (2019) Neural news recommendation with long-and short-term user representations. In: Korhonen An, Traum DR, Màrquez L (Eds.) Proceedings of the ACL, ACL, pp. 336–345.
Bergstrom CT, Bak-Coleman JB (2019) Information gerrymandering in social networks skews collective decision-making. Nature 573(7772):40–41
Beridze I, Butcher J (2019) When seeing is no longer believing. Nat Mach Intell 1(8):332–334
Bovet A, Makse HA (2019) Influence of fake news in Twitter during the 2016 us presidential election. Nat Commun 10(1):1–14
Chen W, Pacheco D, Yang K-C, Menczer F (2021) Neutral bots probe political bias on social media. Nat Commun 12(1):1–10
Courtland R (2018) The bias detectives. Nat. 558(7710):357–360
Das AS et al (2007) Google news personalization: scalable online collaborative filtering. In: Williamson CL, Zurko ME, Patel-Schneider PF, Shenoy PJ (Eds.) Proceedings of the WWW, ACM, pp. 271–280.
Davies J (2016) Program good ethics into artificial intelligence. Nature 538(7625):311–313
Del Vicario M, Vivaldo G, Bessi A, Zollo F, Scala A, Caldarelli G, Quattrociocchi W (2016) Echo chambers: emotional contagion and group polarization on Facebook. Sci Rep 6(1):1–12
Fischer S, Jaidka K, Lelkes Y (2020) Auditing local news presence on google news. Nat Hum Behav 4(12):1236–1244
Gallotti R, Valle F, Castaldo N, Sacco P, De Domenico M (2020) Assessing the risks of 'infodemics' in response to covid-19 epidemics. Nat Hum Behav 4(12):1285–1293
Gibney E (2020) The battle for ethical ai at the world's biggest machine-learning conference. Nature 577(7791):609–610
Gu L, Kropotov V, Yarochkin F(2017) The fake news machine: how propagandists abuse the internet and manipulate the public, vol 5. Trend Micro, pp. 1–85.
Gurini, DF, Gasparetti, F, Micarelli A, Sansonetti G (2013) A sentiment-based approach to Twitter user recommendation, vol 1066. RSWeb@RecSys, CEUR-WS.org.
Hallinan B, Brubaker JR, Fiesler C (2020) Unexpected expectations: public reaction to the Facebook emotional contagion study. New Media Soc 22(6):1076–1094
Han L, Sun R, Gao F, Zhou Y, Jou M (2019) The effect of negative energy news on social trust and helping behavior. Comput Hum Behav 92:128–138
He M, Chen X, Hu X, Li C (2022) Causal intervention for sentiment de-biasing in recommendation. In: Al Hasan M, Xiong L (Eds.) Proceedings of the CIKM. ACM, pp. 4014–4018.
Hemphill TA, Banerjee S (2021) Facebook and self-regulation: efficacious proposals–or 'smoke-and-mirrors'? Technol Soc 67:101797
Hornik J, Satchi RS, Cesareo L, Pastore A (2015) Information dissemination via electronic word-of-mouth: good news travels fast, bad news travels faster! Comput Hum Behav 45:273–280
Huang C, Jiang W, Wu J, Wang G (2020) Personalized review recommendation based on users' aspect sentiment. ACM TOIT 20(4):42:1–42:26
Hutto C, Gilbert E (2014) Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Adar E, Resnick P, Choudhury MD, Hogan B, Oh A (Eds.) Proceedings of the ICWSM, AAAI Press, vol 8.
Johnston WM, Davey GCL (1997) The psychological impact of negative TV news bulletins: the catastrophizing of personal worries. Br J Psychol 88(1):85–91
Khattak AM, Batool R, Satti FA, Hussain J, Khan WA, Khan AM, Hayat B (2020) Tweets classification and sentiment analysis for personalized tweets recommendation. Complexity 2020:8892552:1–8892552:11
Kramer ADI, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. PNAS 111(24):8788–8790
Kucharski A (2016) Study epidemiology of fake news. Nature 540(7634):525–525
Kumar S, De K, Roy PP (2020) Movie recommendation system using sentiment analysis from microblogging data. IEEE Trans Comput Soc Syst 7(4):915–923
Larson HJ (2018) The biggest pandemic risk? Viral misinformation. Nature 562(7726):309–310
Leban G, Fortuna B, Brank J, Grobelnik M (2014) Event registry: learning about world events from news. In: Chung C-W, Broder AZ, Shim K, Suel T (Eds.) Proceedings of the WWW, ACM, pp. 107–110.
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

Lin C, Liu X, Xv G, Li H (2021) Mitigating sentiment bias for recommender systems. In: Diaz F, Shah C, Suel T, Castells P, Jones R, Sakai T (Eds.) Proceedings of the SIGIR. ACM, pp. 31–40.
Littman ML (2015) Reinforcement learning improves behaviour from evaluative feedback. Nature 521(7553):445–451
Liu D, Lian J, Wang S, Qiao Y, Chen J-H, Sun G, ie X (2020) Kred: knowledge-aware document representation for news recommendations. In: Santos RLT, Marinho LB, Daly EM, Chen L, Falk K, Koenigstein N, de Moura ES (Eds.) Proceedings of the Recsys, ACM, pp. 200–209.
Mastro D, Lapinski MK, Kopacz MA, Behm-Morawitz E (2009) The influence of exposure to depictions of race and crime in tv news on viewer's social judgments. J Broadcast Electron Media 53(4):615–635
McCombs M, Reynolds A (2002) News influence on our pictures of the world. In: Proceedings of the media effects. Routledge, pp. 11–28.
Mihaylov T, Mihaylova T, Nakov P, Màrquez L, Georgiev GD, Koychev IK (2018) The dark side of news community forums: opinion manipulation trolls. Internet Res 28(5):1292–1312
Mihaylov T et al (2015) Finding opinion manipulation trolls in news community forums. In: Alishahi A & Moschitti A (Eds.) Proceedings of the CoNLL, ACL, pp. 310–314.
Møller LA (2022) Between personal and public interest: how algorithmic news recommendation reconciles with journalism as an ideology. Digit Journalism 1–19.
Naveed N, Gottron T, Kunegis J, Alhadi AC (2011) Bad news travel fast: a content-based analysis of interestingness on Twitter. In: Roure DD and Poole MS (Eds.) Proceedings of the 3rd international web science conference. ACM, pp. 1–7.
Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453
Okura S, Tagami Y, Ono S, Tajima A (2017) Embedding-based news recommendation for millions of users. In: Proceedings of the KDD. ACM, pp. 1933–1942.
Park JH, Shin J, Fung P (2018) Reducing gender bias in abusive language detection. In: Riloff E, Chiang D, Hockenmaier J & Tsujii J (Eds.) Proceedings of the EMNLP. ACL, pp. 2799-2804.
Ratkiewicz J, Conover M, Meiss M, Gonçalves B, Flammini A, Menczer F (2011) Detecting and tracking political abuse in social media. In: Adamic LA, Baeza-Yates R & Counts S (Eds.) Proceedings of the ICWSM, vol. 5, AAAI Press.
Ruxton GD, Mulder T (2019) Unethical work must be filtered out or flagged. Nature 572(7768):171–172
Shao C, Ciampaglia GL, Varol O, Yang K-C, Flammini A, Menczer F (2018) The spread of low-credibility content by social bots. Nat Commun 9(1):1–9
Sun Y, Fang M, Wang X (2018) A novel stock recommendation system using Guba sentiment analysis. Pers Ubiquitous Comput 22(3):575–587
Thurman N (2008) Forums for citizen journalists? adoption of user generated content initiatives by online news media. New Media Soc 10(1):139–157
Vaswani A, Shazeer N, Parmar P, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg Uv, Bengio S, Wallach HM, Fergus R, Vishwanathan S. V. N & Garnett R (Eds.) Proceedings of the NIPS. pp. 5998–6008.
Vermeulen J (2022) To nudge or not to nudge: news recommendation as a tool to achieve online media pluralism. Digit Journalism 1–20.
Wang H, Zhang F, Xie X, Guo M (2018) Dkn: deep knowledge-aware network for news recommendation. In: Champin P-A, Gandon F, Lalmas M & Ipeirotis PG (Eds.) Proceedings of the WWW. ACM, pp. 1835–1844.
Wu C, Wu F, An M, Huang J, Huang Y, Xie X (2019a) Neural news recommendation with attentive multi-view learning. In: Korhonen A, Traum DR & Màrquez L (Eds.) Proceedings of the IJCAI. AAAI Press, pp. 3863–3869.
Wu C, Wu F, An M, Huang J, Huang Y, Xie X (2019b) Npa: neural news recommendation with personalized attention. In: Teredesai A, Kumar V, Li Y, Rosales R, Terzi E & Karypis G (Eds.) Proceedings of the KDD. ACM, pp. 2576–2584.
Wu C, Wu F, Ge S, Qi T, Huang Y, Xie X (2019c) Neural news recommendation with multi-head self-attention. In: Inui K, Jiang J, Ng V & Wan X (Eds.) Proceedings of the EMNLP-IJCNLP. ACL, pp. 6390–6395.
Wu C, Wu F, Qi T, Huang Y (2020a) Sentirec: sentiment diversity-aware neural news recommendation. In: Wong K-F, Knight K & Wu H (Eds.) Proceedings of the AACL. ACL, pp. 44–53.
Wu C et al (2021) Fairness-aware news recommendation with decomposed adversarial learning. In: Proceedings of the AAAI. AAAI, vol. 35, pp. 4462–4469.
Wu F, Qiao Y, Chen J-H, Wu C, Qi T, Lian J, Liu D, Xie X, Gao J, Wu W et al (2020b) Mind: a large-scale dataset for news recommendation. In: Jurafsky D, Chai J, Schluter N & Tetreault JR (Eds.) Proceedings of the ACL. ACL, pp. 3597–3606.
Wu HD (2007) A brave new world for international news? Exploring the determinants of the coverage of foreign news on us websites. Int Commun Gaz 69(6):539–551
Yang D, Zhang D, Yu Z, Wang Z (2013) A sentiment-enhanced personalized location recommendation system. In: Stumme G, Hotho A (Eds.) Proceedings of the ACM HT. ACM, pp. 119–128.
Zhang BH, Lemoine B, Mitchell M (2018) Mitigating unwanted biases with adversarial learning. In: Furman J, Marchant GE, Price H & Rossi F (Eds.) Proceedings of the AIES. ACM, pp. 335–340.
Zou J, Schiebinger L (2018) Ai can be sexist and racist–it's time to make it fair. Nature 559(7714):324–327

Acknowledgements

This work was supported by the National Key Research and Development Project of China under Grant number 2022YFC3302100 (YH), Tsinghua University Initiative Scientific Research Program of Precision Medicine under Grant number 2022ZLA007 (YH), and the National Natural Science Foundation of China under Grant numbers U1936208 (YH), U1836204 (YH), U1936216 (YH), 61862002 (YH), and 6200197 (YH).

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-022-01473-1.

Correspondence and requests for materials should be addressed to Fangzhao Wu or Yongfeng Huang.