A Transformer-based Framework for Neutralizing and Reversing the Political Polarity of News Articles

RUIBO LIU, Department of Computer Science, Dartmouth College, USA
CHENYAN JIA, Moody College of Communication, University of Texas at Austin, USA
SOROUSH VOSOUGHI, Department of Computer Science, Dartmouth College, USA

People often prefer to consume news with similar political predispositions and access like-minded news articles, which aggravates polarized clusters known as “echo chamber”. To mitigate this phenomenon, we propose a computer-aided solution to help combat extreme political polarization. Specifically, we present a framework for reversing or neutralizing the political polarity of news headlines and articles. The framework leverages the attention mechanism of a Transformer-based language model to first identify polar sentences, and then either flip the polarity to the neutral or to the opposite through a GAN network. Tested on the same benchmark dataset, our framework achieves a 3% – 10% improvement on the flipping/neutralizing success rate of headlines compared with the current state-of-the-art model. Adding to prior literature, our framework not only flips the polarity of headlines but also extends the task of polarity flipping to full-length articles. Human evaluation results show that our model successfully neutralizes or reverses the polarity of news without reducing readability. We release a large annotated dataset that includes both news headlines and full-length articles with polarity labels and meta-data to be used for future research. Our framework has a potential to be used by social scientists, content creators and content consumers in the real world.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → *Collaborative and social computing*.

Additional Key Words and Phrases: automatic polarity transfer, echo chamber, journalism, neural networks, political polarization, selective exposure, transformer-based models

ACM Reference Format:
Ruibo Liu, Chenyan Jia, and Soroush Vosoughi. 2021. A Transformer-based Framework for Neutralizing and Reversing the Political Polarity of News Articles. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 65 (April 2021), 26 pages. https://doi.org/10.1145/3449139

1 INTRODUCTION

People often prefer to consume news with similar political predispositions and access like-minded views [32, 64]. In the context of social science research, selective exposure theory suggests that people seek out attitude-consistent information and avoid attitude-challenging news articles [33, 72, 73]. Therefore, even though the advent of social media has brought us an unprecedented variety of content, people generally only select messages, attitudes or decisions congenial with their prior beliefs [93]. Media itself is a perpetrator of such biased consumption: many media outlets selectively

Authors' addresses: Ruibo Liu, Department of Computer Science, Dartmouth College, 9 Maynard St, Hanover, NH, USA, 03755, ruibo.liu.gr@dartmouth.edu; Chenyan Jia, Moody College of Communication, University of Texas at Austin, USA, chenyanjia@utexas.edu; Soroush Vosoughi, Department of Computer Science, Dartmouth College, USA, soroush.vosoughi@dartmouth.edu.

2573-0142/2021/4-ART65 $15.00
https://doi.org/10.1145/3449139

manufacture narrow, ideological viewpoints so as to cater to audience appetite [60, 71, 79]. Such phenomenon aggravates interactive polarized clusters known as “echo chambers", which reinforces political disposition of both the reader and the media [74]. For instance, the news headlines shown in Table 1 reveal different political ideologies[1], even though three media cover the same story. By using the phrase *“illegal aliens”* instead of *“undocumented immigrants”*, conservative media Breitbart News displays an evident negative attitude towards illegal immigrants in United States. On the other hand, liberal media NPR, selects *“people”* as a softened expression of illegal immigrants.

Media (Polarity)	News Title
NPR (L)	*ICE Arrests Nearly 700 People At Agriculture Processing Plants In Mississippi.*
CBS (N)	*ICE rounds up hundreds of undocumented workers in immigration sweeps in Mississippi.*
Breitbart News (C)	*ICE Arrests 680 Illegal Aliens in Largest Single-State Raid in U.S. History.*

Table 1. Three news headlines from NPR, CBS and Breitbart News. L: Liberal; N: Neutral, C: Conservative.

In fact, such polarized cues not only exist in news headlines, but are also ubiquitous in the body texts of news articles [57]. Social science researchers argue that polarized cues have a fundamental influence on social psychology [57] and political decision-making [52], which explains why politicians tend to reinforce their partisan styles in media coverage by using polar expressions for attitude shaping [3, 26, 41] or voting [82]. With the distribution of news increasingly automated [55], news consumers can easily filter out news discordant with their views (e.g. by using news aggregators such as Digg, Slashdot, Reddit). As a result, it's not surprising to witness increasing polarization and political fragmentation around the world in recent years. From a societal level, increased political fragmentation makes it harder for society to find common ground on controversial issues [84, 95]. From an individual perspective, citizens trapped in small, insular information circles loses the opportunity for objective and rational discussions for social good [71, 88], which has been recognized as a crucial ingredient for a healthy democratic society [59, 77]. Thus, it is of importance to expose people to more ideologically diverse news stories [23].

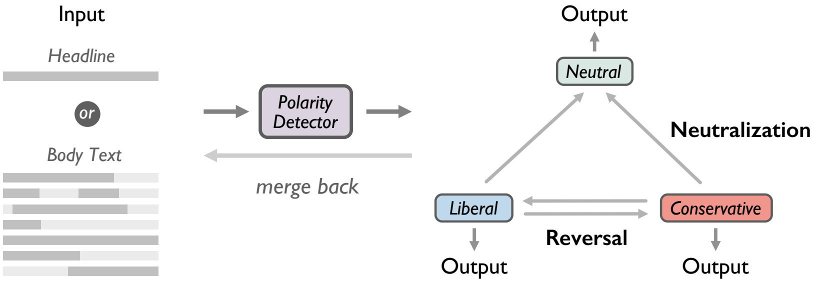


Fig. 1. The overview of our framework. The polar part within headline or body text of the original text will be first detected by the polarity detector, and then reversed or neutralized by the polarity flipper.

One way to tackle this problem is to change the polarity of news articles, without changing the content, so that the content is palatable to audiences with different political ideologies. This helps expose people on different political poles to the same content, which allows for the creation of a

[1] The ideological labels are excerpted from a Pew national survey: *Where News Audiences Fit on the Political Spectrum?* https://www.journalism.org/interactives/media-polarization

common ground between different echo chambers that could potentially lead to depolarization and weakening of the echo chambers. However, doing this manually is time-consuming and can introduce annotator bias. In this paper, we propose an automatic framework that can flip articles into an arbitrary target political polarity. Such a tool can be used by a wide range of users who are interested in reducing (or studying) political polarization.

In the rest of the paper, polarity flipping refers to the general changing of polarity, either by *reversal* (L to C, or C to L) or *neutralization* (L to N and C to N). As shown in Figure 1, the framework can accept either headlines or full-length articles as input. The text is first passed through a polarity detector, which identifies sentences that exhibit political polarity; these sentences are then passed to a polarity flipper. Finally, the output of our framework, that is either reversed polar text or neutralized text, is merged back to the original article. Note that our framework does not change the content of the article. It only changes the way the content is expressed. Though our tool attempts to keep the content of the article unchanged, it cannot guarantee that the underlying message will not be affected as it might depend on carefully chosen expressions by the author.

Developing such a system has three main technical challenges: *First*, political polarity is hard to detect. Unlike lexical features that capture sentiment or emotions (e.g., *"happy"*, *"sad"*, *"angry"*, etc.) that are easily identifiable, political polarity is often implied by the context, or by long spans of text (paragraphs or even the whole article). *Second*, even if after detecting political polarity, it is still a difficult task to separate neutral content and polar expressions in the text. This causes difficulties for ideal polarity flipping that only modifies the polar part of the text without harming the content. *Third*, the current benchmark dataset [20] is limited in size (6,477 articles in total) and diversity (only from five media outlets). This raises concerns about whether the style learned by our model is indeed based on polarity or if the model is just learning the writing style of the media outlets.

We address the above challenges as follows:

- Given the observation that certain phrases or sentences significantly contribute to the polarity of a full-length article, we propose a polarity detector by leveraging the attention weights from a fine-tuned language model (more specifically, XLNet [102]) classifier. It is able to locate polar sentences that need to be transformed within an article, allowing us to avoid unnecessary modification of sentences that do not contribute to the polarity.
- Motivated by the recent success of adversarial networks on text generation [19, 21, 35], we follow the proven Transformer-based encoder-decoder structure [98] and build a discriminator-generator GAN framework [40]. In addition to the adversarial loss (discriminator loss), we use three semantic losses in the framework targeting fluency, content-preservation, and polarity flipping performance of generated sentences. We also overcome the obstacle that certain generation quality metrics (such as fluency) cannot be optimized during training due to them being non-differentiable.
- As we discuss in Section 4.1, the current publicly available benchmark dataset for political polarity flipping has limitations: the size of the dataset is small (6,477 articles) and the data lacks the diversity of source (5 media outlets). As part of this work, we create a much larger (364,986 articles) and more diverse (13 outlets) benchmark dataset for this task. The dataset is annotated with article polarity labels and has additional meta-data, including topics, keywords, paragraphs, and part-of-speech (POS) tags. We make this dataset available to other researchers to enable further development in this area.

2 RELATED WORK

In this section, we examine existing work on political polarity by exploring: political polarization, polarity detection and polarity flipping. We first provide an overview of prior studies on political polarization. Then, we discuss different approaches that have been used in both technical systems and social sciences studies to identify political polarity. Lastly, we explore prior work on polarity manipulation in social sciences, present the polarity flipping task as a special case of text style transfer, and outline previous work on automatic style transfer in the field of natural language processing.

2.1 Political Polarization

Political polarization has been studied for decades [8, 28, 30, 91], going back to at least the 1950s [87]. In the United States, it is well-reported that Americans are deeply divided on controversial issues such as immigration, gun control, and racial inequality; moreover, such divisions have become increasingly aligned with ideological or partisan identities in recent years [1, 12, 31]. Such divisions can be explained by "echo chambers" or selective exposure effects. Partisans often selectively expose themselves to opposite political views that actually reinforces their pre-existing political beliefs [5, 80]. This has been especially true since the advent of the Internet [66].

A large number of studies have demonstrated that political polarization widely exists in today's media. Analysis of newspaper data shows that newsrooms tend to tailor their slants pandering to the potential readers who demand a strong fit between a newspaper's slant and their ideologies [17, 37, 50]. A similar conclusion has been drawn on broadcast and cable TV data [43]. Lab experiments confirm that people consistently prefer congenial information over uncongenial information, especially regarding politics [22]. Behavioral data from a long-term tracking on online consumption of political news also suggests that selective exposure is mostly concentrated among those who regularly consume news from partisan sources [45].

These studies further illustrate the need for a tool that allows the same content to be presented with different polarities so that it is palatable to people with different ideologies.

2.2 Polarity Detection

Groeling [42] categorizes media bias into selective bias and presentation bias, whereas selective bias skews the choices of *what* events to cover, and presentation bias skews the style of *how* to report the news. In this work, we attempt to reduce presentation bias. Early-stage work in this area mainly relies on human annotators [14, 25], where linguistic features are demonstrated to play a crucial role on how humans judge polarity. For example, Yano et al. [103] prepare 1,100 potentially biased sentences and ask annotators to indicate whether a sentence shows bias, and if so, in which political direction and through which word tokens. Each sentence is annotated five times (5,205 judgements total), and nearly half of the sentences are marked as "*not biased*" and only 2.4% of the sentences are marked as "*very biased*". They also present two lists of the strongest weighted bigrams (i.e., phrases with two or less words) for liberal and conservative groups scored by human annotators. Gentzkow and Shapiro [37] record a similar table that contains partisan phrases most often used by congressional Democrats and Republicans based on the 2005 congressional record. Their experiments reveal that these phrases are chosen strategically for their partisan impact. There are several other studies that use human-annotated data to study political polarity [7, 29, 96]. These studies confirm that there are some consistent lexical cues for bias for different political ideologies.

Studies mentioned above all use annotations to manually identify political polarity. This can be costly when dealing with a large quantity of data (where the biased content is shown to be quite sparse [13]). As a result, many researchers opt for automated methods which utilize the power of deep neural networks for text processing. Prior work mostly makes use of RNN [54] or CNN-LSTM [89] to identify the political polarity in term of both syntactic and semantic features. Recent work tries to adopt more advanced network structures, since recurrent neural networks are known for having the limited capability of modelling long text such as full-length news articles [9].

For example, Kulkarni et al. [61] propose an attention based multi-view model to leverage cues from the article's title, content and link structure to identify the ideology of news. Another direction is detecting polarity through topical analysis [38], which is based on the observation that political ideology can be viewed as a distribution of sentiment polarities towards a set of topics [15, 69]. Our polarity detector differs from all above ones since we leverage the power of transfer learning by using an unsupervised language model (LM) trained on a large corpus, which brings qualitative performance advances on feature extraction and classification accuracy.

2.3 Polarity Flipping

Researchers in the field of political science have conducted many experiments on changing the perceived of polarity of news content by manipulating the source attributions. For example, researchers experimentally assigned identical news content to different media outlets with opposite perceived political leanings (such as FOX or CNN) [11, 53]. These studies find that heuristics such as media source can heavily influence perceptions of bias in the content. Though these studies have so far focused on studying the effects of media source on perceived bias, some social scientists suggest further manipulation of the way content is presented in order to find more nuanced signals of perceived bias [11]. Our work provides a supplementary way to manipulate how content is presented via computer-assisted methods.

From the technical perspective, we can find several parallels between our political polarity flipping method and general-purpose style transfer approaches, since in our case, the *style* is the polarity label (*liberal*, *neutral* or *conservative*), and our goal is to transfer the *style* while the *content* is preserved. The mainstream methods in style transfer assume that the style and content information can be perfectly separated in latent space and target style text is generated via conditional generation. Shen et al. [90] incorporate adversarial networks into cross-aligned auto-encoder architecture, encouraging the system to learn the separate style and the content distribution. John et al. [56] artificially divide the latent representation into style and content space, and design auxiliary multi-task loss and adversarial loss, enforcing the separation of style and content latent spaces when training an encoder-decoder network. Based on the style-content separation assumption, many methods have been proposed [34, 51, 63, 105].

There are a few style transfer methods *not* built on the assumption that content and style can be successfully split in latent space. These methods use either step-by-step transfer or reinforcement learning to optimize the generation. Wu et al. [100] propose a hierarchical reinforced sequence operation method to transfer style, which consists of a high-level agent that proposes operation positions and a low-level agent that alters the sentence. Similarly, Wu et al. [101] deploy a simple but effective approach that only replaces the original sentimental tokens in the sentence with target sentimental expressions, instead of building a new sentence from scratch. Gong et al. [39] combine beam search and multinomial sampling in the rewards estimation step. They first use beam search to generate a reference target sentence, and at each step, they draw samples of complete sentences by rolling out the subsequence through multinomial sampling. Liu et al. [68] deploy a two-step procedure to flip political polarity of news articles. They first train a topic and ideology attribute-aware word embedding model, and then use a modified simulated annealing algorithm to pick proper tokens for polarity flipping. This method cannot neutralize polarity and is built upon the assumption that polar text can be detected at the surface level (i.e. token level), which limits its practical application.

Chen et al.'s [20] work, which attempts to flip the bias in the news headlines, bears most similarities to our approach. They use a cross-aligned auto-encoder (similar to Shen et al.'s [90] approach [90]) trained on opposite-ideology articles to generate flipped titles. The main limitation of their work, as stated by the authors, is the loss of content information. The overlap of content between the generated and ground-truth headlines is very low, meaning that the auto-encoder model has the tendency to discard too much non-polar content. Our method, on the other hand, is not built on the assumption that content and style can be successfully split in latent space. Instead, we utilize style (in our case, polarity) embedding and attention mechanism in transformer block [98] to implicitly encode content and style information, so that we avoid content information loss caused by the artificial split of content and style. Also, we extend the task to the depolarization of the entire article, instead of only headlines, which is a more challenging problem since we have to preserve contextual and grammatical correctness across the whole article.

3 METHOD

3.1 Polarity Detector

Given the observation that political polarity does not exist in every sentence, especially when the source text are full-length articles rather than limited-length headlines, a polarity detector is necessary to help us locate the polarity carrying sentences among all sentences in an article. Such a detector should be capable of: (1) detecting polar text in different context (2) accepting input of any length. Existing methods either do not take into account context information by selecting words in fixed contexts [36, 67] or suffer limited modeling power on long sequence input [54] (RNN-based methods). Our approach for automatic detection of polarity in text leverages the power of transfer learning by using an unsupervised language model trained on a large corpus.

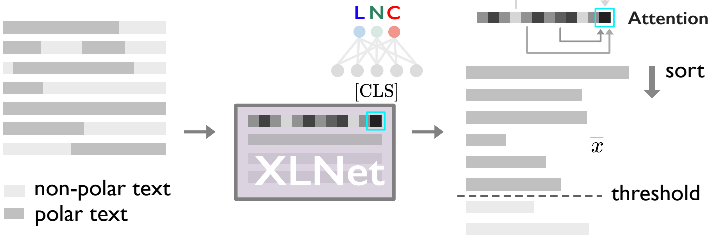


Fig. 2. The structure of our polarity detector, which is built on a XLNet language model. We calculate the polarity score based on the weights from attention layers by Equation 1 and Equation2.

We choose XLNet [102] as our base language model because it can accept long input sequences due to its relative positional encoding scheme and segment-level recurrence mechanism. We stack a linear layer upon the language model so that it is able to do sequence classification on three classes (*liberal*, *neutral* and *conservative*). We set the max sequence length to 1024 tokens, which is long enough to include 99.97% of articles in our dataset.

We first convert the input sequence following the same pattern as mentioned in the original implementation of XLNet [102], whereas two special tokens [SEP] and [CLS] are added to the end of the sequence. We then feed such converted input into the sequence classification model. Note that [CLS] acts as an interface bridging the linear layer and the underlying attention layer. Thus the correlation, or in our case, the attention score between the [CLS] marker and all the other tokens in the sequence could be a reasonable measure of the token's significance. In other words, the contribution of certain token x_i to the classification result ($Y = y$) can be modeled as:

$$\begin{aligned}\mathbb{P}(Y = y|x_i) &= \sum_{h_i \in \text{heads}} \text{Attn}(x_i, [\text{CLS}])_{h_i} \\ &= \sum_{h_i \in \text{heads}} \text{softmax}(\frac{Q_{x_i} K_{x_i}^T}{\sqrt{d_k}}) V_{[\text{CLS}]}\end{aligned} \tag{1}$$

where Q, K refer to the query and key vectors, and d_k is the hidden dimension. *Attn* refers to the multi-head (*heads*) attention mechanism, and here we calculate the summation of the attention scores between [CLS] and other tokens across all the heads.

Besides the attention weights, we also incorporate the classification result (the possibility of the given article belongs to certain class) as confidence (conf) and word frequency (captured using TF-IDF weights) as auxiliary information into the polarity score ($\mathcal{S}$) computation. For a given sentence $\overline{x}$ of article d in the dataset D, its polarity score is proportional to the weighted sum of attention scores of each token x_i:

$$\mathcal{S}_{\overline{x}} = \frac{\text{conf} * \sum_{x_i \in \overline{x}} \mathbb{P}(Y = y|x_i) * \text{TF-IDF}(x_i, d, D)}{|\overline{x}|} \tag{2}$$

We assign a polarity score to each sentence in the articles. Then, we sort all sentences in terms of the polarity score and select those that pass a certain global threshold. This global threshold is a polarity sensitivity hyperparameter set by the user. The sentences whose polarity score is above the threshold are the ones that need to be flipped, the rest of the sentences are left unchanged. Such a method greedily guarantees that the semantic information of those sentences that do not exhibit polarity is well-preserved. The polarity sensitivity hyperparameter, which is picked by the user, is one of the control knobs available to the users to tune the model to their needs, based on empirical observations. Having this control knob is important since different user groups may have different needs, depending on how they are using our model.

3.2 Polarity Flipper

The framework for our polarity flipper is heavily influenced by the recent success of transformer-based encoder-decoder structure on neural machine translation [98], language modeling [27] and conditional generation [58]. As shown in Figure 3, we first change the input of the encoder to the summation of polarity label and the original input embeddings, so that the decoded output is conditioned on the current polarity label; this turns the decoder into a conditional generator (G). In addition, inspired by the current progress on adversarial networks in different architectures [90, 104, 106], we further modify the encoder-decoder structure to a GAN network [40]: we stack a linear layer with activation on the encoder, converting it to a discriminator (D). We also incorporate one adversarial loss ($\mathcal{L}_D$) in the discriminator and three semantic loss ($\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$), corresponding to the fluency, content-preserving, and polarity flipping success in the conditional generator. The training goal is to minimize a weighted summation of all the loss to obtain a global optimal balance of fluency, content-preserving and polarity flipping.

3.3 Discriminator

In the original structure of the transformer-based encode-decoder, the encoded vectors (k, v) are shared with the decoder, helping it focus on appropriate places in the input text. We add a new mode to the encoder, discriminator, or in other words, classifier, through adding a linear layer to project hidden states to the number-of-class logits. The responsibility of this discriminator is to distinguish true samples of ground truth data from fake (i.e., polarity flipped) samples generated by

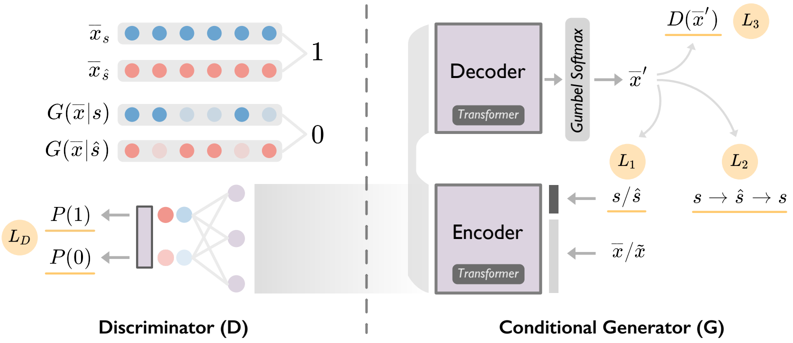


Fig. 3. Political polarity flipper in our framework. Conditional generator (G) generates "fake" samples given specified polarity labels (L, N or C), and Discriminator (D) distinguishes original data from the generated ones by G. We incorporate three types of semantic loss (fluency loss, content preserving loss, and polarity controlling loss) to guarantee the generation quality during training.

the G. We use adversarial loss $\mathcal{L}_1$ to express the object (assuming the input text is $\overline{x}$ and its original polarity label is s):

$$\mathcal{L}_D^{\theta_D}(\overline{x}) = \mathbb{E}_{\overline{x}}(\log D(\overline{x})) + \mathbb{E}_{\overline{x}' \in G(\overline{x},s)}(1 - \log D(\overline{x}')) \tag{3}$$

where $\overline{x}'$ denotes the fake sentences with polarity label s generated by conditional generator G. Following such form, the discriminator D is encouraged to minimize the classification loss for true samples while maximizing the loss for generated samples (i.e., polarity flipped samples). This trains the discriminator to distinguish between fake and true samples. (The $\mathbb{E}$ in the equations corresponds to the expectation of the random variable) .

3.4 Conditional Generator

Following the standard GAN framework [40], we need a conditional generator G to generate sentences given different polarity labels. We do this by switching the discriminator back to its original primary role, an encoder, and again incorporating three semantic loss into the generator training.

3.4.1 Fluency Loss. For any sentence $\overline{x} = \{x_1x_2x_3...\}$, we randomly reorder tokens in the sentence as noisy input ($\widetilde{x} = \{x_3x_1x_2...\}$), and the generator G is forced to reorder the noisy input to recover the original input sentence. We use negative log-likelihood to measure the difference between the ground-truth input $\overline{x}$ and the reconstructed output tokens $\overline{x}'$, and use it as the fluency loss (assuming the original polarity label of $\overline{x}$ is s):

$$\mathcal{L}_1^{\theta_G}(\overline{x}) = -\overline{x}\ \log \overline{x}', \text{ where } \overline{x}' = G(\widetilde{x}, s) \tag{4}$$

Some prior work choose to randomly drop or replace tokens [24] in sentences as noisy input. We decide not to follow these methods because: (1) It is possible that content-related tokens are dropped, so it is problematic for the generator to precisely recover those tokens (2) Replacing tokens with random ones from the vocabulary set is an uncontrolled procedure which may introduce unnecessary polarity style shift into the input.

3.4.2 Content Preserving Loss. To implicitly indicate the content preserving loss during training stage, we borrow an idea from CycleGAN [106] and Back-Translation [78], that if one sentence (or picture) is transformed from style s to style $\hat{s}$ and then back transformed to style s, the input and output tokens should be very similar. In our case, for both polarity flipping and neutralization modes, we perform such back-transformation and calculate the content preserving loss. We train our network by minimizing the negative log-likelihood of input $\overline{x}$ and back-transformed sentence $\overline{x}'$ (assuming the original polarity label of $\overline{x}$ is s, and the transferred style is $\hat{s}$):

$$\mathcal{L}_2^{\theta_G}(\overline{x}) = -\overline{x}\ \log \overline{x}',\ \text{where } \overline{x}' = G(G(\overline{x}, \hat{s}), s) \tag{5}$$

Due to the discrete and non-differentiable nature of the tokens, one problem we face during training is that gradient propagation will be broken down when we feed the output $G(\overline{x}, \hat{s})$ into the generator again with opposite style s. One of the mainstream solutions to this problem is to use policy gradient from reinforcement learning, thus avoiding the gradient-based optimization method [39, 104]. The main problem of such a method is that it relies on estimated rewards rather than true rewards because the true rewards can only be precisely computed when the generation of current sentence terminates. Even though it can lead to a good estimate of the complete sentence rewards, it comes at a significant computational cost.

Instead, another mainstream method is to use continuous approximation based on softmax so that the objective function becomes differentiable and tokens selection is converted to drawing samples from a continuous distribution. Prabhumoye et al. [78] use temperature parameter controlled Softmax to approximate the real Softmax.

We decide to use Gumbel Softmax [62] as a continuous alternative of the softmax function:

$$\text{softmax}\Big(\frac{\log(\mathbb{P}(x_i)) - \log(-\log \epsilon_i)}{\tau}\Big)_{\epsilon_i \sim \mathcal{U}[0,1]}^{x_i \in \widetilde{x}} \tag{6}$$

where $\mathbb{P}(x_i)$ is the probability of certain token being selected in each step decoding, and τ is the annealing parameter used to control the approximation strength ($\tau \longrightarrow 0$ means Gumbel Softmax degenerates to the non-differentiable normal softmax, while larger τ means more randomness. We slowly decrease τ from 1 to 0.01 during our training). Note that now the randomness is generated only from a known uniform distribution $\mathcal{U}[0, 1]$ and we are able to do gradient-based optimization on such continuous estimation.

3.4.3 Polarity Controlling Loss. If we solely rely on the previous back-transformation loss, the generator will tend to copy and paste tokens from the input sequence. Therefore, we track the polarization-reversal/depolarization loss in each step, and leverage the fixed polarity discriminator (D) to judge the quality of newly generated sentences with reversed/neutralized polarity (depending on whether we are flipping the polarity or neutralizing). We compute the classification loss through comparing target polarity label and output of the discriminator (assuming the input text is $\overline{x}$ and $\hat{s}$ is the target polarity label):

$$\mathcal{L}_3^{\theta_G}(\overline{x}) = -\hat{s}\ \log(\mathbb{P}(D(\overline{x}') = \hat{s}),\ \text{where } \overline{x}' = G(\overline{x}, \hat{s}) \tag{7}$$

Basically, this loss measures how well the generator is able to convince the discriminator with regards to the target polarity of the generated text.

3.4.4 Combining the Loss. Combining all above loss, we propose the following global loss function for optimizing the conditional generator (**G**):

$$\min_{\theta_G} \mathcal{L}_G = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 + \lambda_3 \mathcal{L}_3 \tag{8}$$

where λ_1, λ_2 and λ_3 are balancing hyperparameters to control the weights of each kind of loss. We choose $\lambda_1 = 0.1, \lambda_2 = 0.5, \lambda_3 = 0.4$ as our setting through experimentation. The experimentation was done by evaluating the performance of our model for different combinations of these hyperparameters. In general, we found that the fluency loss should be weighted lower than the two other losses, since it encourages the generation to repeat the original text and even the order of tokens. Meanwhile, content preserving loss should be weighted slightly higher than the polarity controlling loss to guarantee the generation quality.

3.5 Training Procedure

The training procedure is summarized by Algorithm 1 (The algorithm shows the procedure for training the polarity neutralizer, similar procedure is used for the polarity reversal model). In general, we follow the classical GAN [40] training process: for each iteration, we first train the discriminator until we reach the max D training steps in order to obtain a qualified discriminator, and then we train the conditional generator until reaching max G training steps focusing on three main goals of our task: the generated sentence should be fluent ($\mathcal{L}_1$), content-preserved ($\mathcal{L}_2$) and have either lower or reversed polarity ($\mathcal{L}_3$). After several alternate training on D and G, we expect the generator to produce high-quality "fake" target polarity sentences and correspondingly the discriminator should be able to distinguish fake samples from true samples.

Algorithm 1: Our training procedure for the polarity neutralizer

```
Data: Batches of polar ($x_s$) or neutral ($x_{\hat{s}}$) sentences $\overline{x}$ in the dataset.
while not reaching the end of dataset do
    repeat
        generate fake polar $x'_s = G(\overline{x}, s)$ ;
        generate fake neutral $x'_{\hat{s}} = G(\overline{x}, \hat{s})$ ;
        Fix $G$, use $D$ to discriminate:
          fake polar $x'_s$ and true polar $x_s$ as 1;
          fake neutral $x'_{\hat{s}}$ and true neutral $x_{\hat{s}}$ as 0;
          Compute $\mathcal{L}_D$ given true and fake pairs;
          Optimize $D$ with $\mathcal{L}_D$;
    until reaching max D training steps;
    repeat
        randomly reorder tokens in $\overline{x} \rightarrow \widetilde{x}$;
        Fix $D$, use $G$ to generate:
          $\overline{x}' = G(\widetilde{x}, s)$ or $G(\widetilde{x}, \hat{s})$ given noisy input $\widetilde{x}$
          $\overline{x}'' = G(\overline{x}, s \rightarrow \hat{s} \rightarrow s)$ or $G(\overline{x}, \hat{s} \rightarrow s \rightarrow \hat{s})$;
          use $D$ to classify $G(\overline{x}, s)$ and $G(\overline{x}, \hat{s})$;
        Compute $\mathcal{L}_1$ given $\overline{x}'$ and $\overline{x}$ by Eq. (2);
        Compute $\mathcal{L}_2$ given $\overline{x}''$ and $\overline{x}$ by Eq. (3);
        Compute $\mathcal{L}_3$ given $D(G(\overline{x}, s))$ and $D(G(\overline{x}, \hat{s}))$ by Eq. (4);
        Optimize $G$ with $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$;
    until reaching max G training steps;
```

Polarity Reversal & Polarity Neutralization Generation Samples

Type	Label	Sample
L → C	L	Trump Accuses FBI, Justice Department of *Favoring Democrats*.
	C	Trump Accuses FBI, Justice Department of *bias against Republicans*.
C → L	C	Trump Administration policy that could *split up families who enter US illegally*.
	L	Trump Administration Considers *Tearing Families Apart In who enter US*.

Type	Label	Sample
L → N	L	Trump *Dreamers* Plan *path to citizenship*, Concessions From Democrats
	N	Trump *Immigration* Plan Demands *Tough* Concessions From Democrats
C → N	C	Senate approves budget in *crucial step for Trump's tax overhaul*
	N	Senate *narrowly* Approves budget, *paving way for tax reform*

Table 2. Sample generation of our framework. We *underline* the main differences between the sentence pairs.

4 IMPLEMENTATION

4.1 Datasets

4.1.1 Allsides. The benchmark dataset we use for the headline polarity flipping task is from the current state-of-the-art work [20]. They collected 6,447 news articles including metadata (title, author, publication date) based on 2,781 events from June, 2012 to Feb, 2018 from allsides.com[2]. Each news article in allsides.com is labeled with a political polarity label by an editing expert. We treat *left* and *lean left* as *liberal*; *right*, and *lean right* as *conservative*. We use this dataset for flipping headlines only to fairly compare our approach with Chen et al. [20].

4.1.2 Media Cloud. Media Cloud[3] is an academic research project led by MIT and Harvard University. The Media Cloud collects articles from a large number of media outlets. Using Media Cloud, we collected and parsed around 360k full-length articles from May 1st, 2018 to May 1st, 2019 (around 6G plain text). The articles are from 22 media outlets. We assign an ideological polarity label to each outlet using data from the Pew Research Center. The data from Pew is based on a survey of news consumption by people with different political affiliation[4] [18]. The Pew survey has five ideological groups: consistent liberals, mostly liberals, mixed, mostly conservatives, and consistent conservatives. These labels are assigned by Pew based on responses to questions about a range of political values. Pew's methodology is explained in detail in [83] [5]. The labels we assigned for polarity are *liberal* (corresponding to Pew's *consistent liberal* and *mostly liberal* labels), *neutral* (corresponding to Pew's *mixed* label), and *conservative* (corresponding to Pew's *consistent conservative* and *mostly conservative* labels). This dataset is used for full article flipping.

The articles are further annotated with 11 topics and corresponding keywords. We use the survey-based website[6] to choose 11 topics. For each topic, we use the words under the corresponding title on isidewith.com as query keywords (e.g. topic: Social Issue → query terms: *abortion*, *gay*

[2] https://www.allsides.com/unbiased-balanced-news
[3] https://mediacloud.org
[4] https://www.journalism.org/interactives/media-polarization/table/overall/
[5] https://www.pewresearch.org/politics/2014/06/12/appendix-a-the-ideological-consistency-scale/
[6] https://www.isidewith.com/polls

marriage, *death penalty*, etc). Though the topic information is not used in this paper, it could be a valuable signal in extensions of this work as polarity can be topic dependent. This dataset and its corresponding annotations will be made publicly available to encourage replication and extensions of this work.

4.2 Data Preprocessing

The data goes through preprocessing before it is fed into our model. Since our data is parsed directly from published articles, there is a sizeable amount of media-related content that exist in the data (such as html tags, image links, ads, etc). We remove all these content and keep only the text of the article. All punctuation except for , .?! are also removed. Punctuations are also separated from the adjacent words (e.g., "is it?" becomes "is it ?"). Digits lower than ten are converted to letter representation (e.g. 1 → *one*), and the rest stay unchanged. All paragraph information (like newline markers) is kept since it could potentially be useful in our task .

4.3 Training

We split the data from Allsides and Media Cloud into training, validation and test sets (70%, 10%, and 20% correspondingly). The XLNet-based polarity detector was trained on the Media Cloud dataset with parameter fine-tuning to obtain the best performance ($F1 = 0.934$). During training, we used early stop when we observe no improvements after 2000 iterations. We trained our flipping model using two RTX 2080 GPUs for one hour on the Allsides dataset and ten hours on the Media Cloud dataset.

5 AUTOMATED EVALUATION

Based on prior work, we use several automatic metrics to evaluate our model's performance on polarity flipping, content preservation, and fluency. Each evaluation is explained below. We show the full result in Table 3. We show example generations in Table 2.

5.1 Polarity Flipping

We train a FastText [70] polarity classifier on our dataset as an external independent polarity flipping judgement. The classifier has 3 classes (conservative, neutral, and libera), has 300 hidden dimensions, uses n-grams of up to 3, and was trained for 200 epochs. The classifier achieves an F1 score of 0.91 on the validation dataset. We use this classifier to predict the polarity label of the generated sentences. The rate at which the generated sentences "fool" this classifier (i.e., have their polarity be classified as the target polarity label) corresponds to the success of our polarity flipping. We use this classifier instead of the discriminator D in our GAN model in order to have independent evaluation of the polarity flipping. We define the flipping accuracy s_p as the percentage of the flipped/neutralized sentences that are classified as having the target polarity labels. From the results, shown in Table 3, it can seen that for headlines, our method outperforms the prior work for the headlines [20, 81] by at least 10% for all flipping directions. (Chen et al.'s [20] model only works on polarity reversal so we use Pryzant et al.'s [81] model as a baseline for neutralization). There are no prior studies for fully body flipping to compare against. However, we can see that our model's performance drops when working on fully body vs the headlines. One possible explanation is that there is weak structural parallelism between headlines is missing in the body text. This makes it harder for our model to learn polarity invariant information in the full text.

We also do an ablation study to figure out whether our polarity detector (PD) benefits our model. As can be seen in Table 3, the detector can greatly improve the accuracy of the flipping (reversal or neutralization). This makes sense since with the PD, the model focuses the sentences that manifestly carry polarity.

Automatic Evaluation
N = 36,636

		L → C				C → L			
	Methods	Accuracy	Content	Overall	PPL	Accuracy	Content	Overall	PPL
headlines	Chen et al.	0.539	**0.832**	0.654	**89.14**	0.302	**0.931**	0.456	**93.58**
	Ours	**0.693**	0.841	**0.756**	89.14	**0.528**	0.895	**0.664**	109.76
body text	Ours	0.615	0.546	0.289	129.36	0.577	0.756	0.327	134.63
	Ours (no PD)	0.537	0.733	0.309	157.81	0.355	0.903	0.254	148.32

		L → N				R → N			
	Methods	Accuracy	Content	Overall	PPL	Accuracy	Content	Overall	PPL
headlines	Pryzant et al.	0.478	**0.753**	0.585	**101.22**	0.446	0.652	0.530	**120.33**
	Ours	**0.588**	0.648	**0.617**	100.04	**0.612**	**0.774**	**0.683**	91.78
body text	Ours	0.531	0.522	0.526	110.61	0.425	0.542	0.476	103.61
	Ours (no PD)	0.374	0.837	0.517	149.76	0.353	0.478	0.406	142.36

Table 3. Automatic evaluation of our framework, compared with two prior state-of-the-art models (Chen et al. [20] on polarity reversal and Pryzant et al. [81] on polarity neutralization). We report polarity flipping accuracy (§5.1), content preservation (§5.1), overall score (§5.3) and PPL for fluency (§5.4). PD refers to the Polarity Detector.

5.2 Content Preservation

We use the measurement proposed by Fu et al., [34] that uses embedding-based sentence similarity to compare the original and generated sentences. We first use a pre-trained FastText model (wiki-news-300d-1M from FastText [70]) to generate embedding vectors for each token in the original and generated sentences. We then compute the sentence embeddings by using max, min, and mean pooling of the embeddings of the words in a sentence. Finally the semantic similarity between the original and generated sentences are computed through the cosine similarity between of their embeddings: $s_c = \sum_{i=1}^{D_{val}} cos(v^-, v^+)_i$. We compute the semantic similarity between the original text and the flipped text in the test set.

The results are shown in Table 3 as the “Content” score. For the headline polarity flipping task, our model has comparable performance with the state-of-the-art model in content preservation. For the body text flipping task, we find that the content preserving score of our model without the PD is generally higher than the one with the PD. This is because without the PD, our model is working on the whole text. The majority of the text of an article does not carry polarity and thus is minimally changed. Since the part of the text that is not changed has perfect content preservation, the content preserving score of our model without PD is trivially high. In general, the polarity flipping and content preserving scores are at odds with each other. Better flipping score generally leads to lower content preserving score and vice versa. This is why we combine these two scores into one score called the “Overall” score.

5.3 Overall

It is easy for the polarity flipper to achieve high accuracy if it does not have to preserve the content. Similarly, it can achieve perfect content preservation by simply copying the original input into the flipped generation. To capture this trade-off relationship in the evaluations, we use the harmonic mean of the polarity flipping and the content preservation scores as the overall score for evaluating the performance of our framework: $s_o = \frac{2s_c s_p}{s_c + s_p}$. As shown in Table 3, for the headline flipping task our model outperforms the prior work in the overall score for both tasks. The ablation study on the body text shows that the presence of the PD does indeed improve the overall score.

5.4 Fluency

We use perplexity as a measure of fluency. Given a word sequence of M words $w_1, ..., w_M$ and the sequence probability $p(w_1, ..., w_M)$ estimated by a language model (LM), the perplexity is defined as:

$$PPL = p(w_1, ..., w_M)^{-\frac{1}{M}}$$

Perplexity is a widely used measure that captures how well a LM predicts a given sentence. In other words, if a LM is trained on a large English corpus, it can capture how well a given sentence fits the English language (i.e., the English fluency of that sentence). Using kenLM [49] (chosen for its speed and efficiency), we train a LM on our dataset. The LM is trained on n-grams of up to 5. We use this LM to estimate the perplexity of generated sentences. Lower perplexity scores mean better fluency of generated sentences. The results are shown in Table 3. In general, the perplexity of generated body texts is higher than that of headlines. This implies the model has more difficulties in handling long text. In the headline polarity flipping task, our method does not show an obvious advantage over the prior methods in terms of the perplexity. A possible reason could be that the fluency loss (one of the three semantic loss) is a weak training objective for the perplexity of long text, since we only permute 30% of the total tokens in the input text.

6 HUMAN EVALUATION

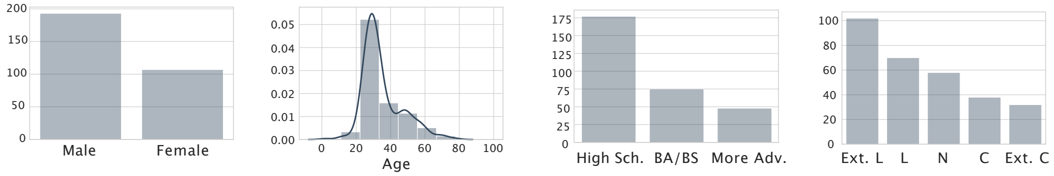
To further evaluate whether our method can successfully alter the polarity of news articles as perceived by human readers, we conducted an online experiment on Amazon Mechanical Turk (MTurk). MTurk is an online crowd-sourcing platform for online data collection with a diverse population of workers [16, 75].

6.1 Participants

We recruited 300 participants (*N*=300) from Amazon Mechanical Turk (MTurk). Participants were randomly assigned to one of the two tasks (neutralization task: *n*=152; reversal task: *n*=148). Participants were all from the United States and above 18 years old. Participants were required to have a HIT approval rate greater than 95%. Before the task, participants were asked to answer a demographic questionnaire about their gender, age, level of education, media consumption preference, party affiliation, and where they fell on a political ideological spectrum (1-Extremely liberal; 4-Moderate; 7-Extremely conservative). The mean age of the participants was about 34.8 years-old (*SD*=11.6, *Median*=31). Among 300 participants, 193 (64.3%) were male, and 107 (35.7%) were female. Participants on average received about 14.9 years of education. When asked to self-report their party affiliation, 156 of the participants self-reported as Democrats (52%), 110 (37%) as self-reported Republican, and 34 (11%) as independent. Participants spent an average of 12.5 minutes to finish the task and were compensated one US dollar for completing the experiment. The participants' demographic distribution was shown in Figure 4.

6.2 Experiment Design

6.2.1 Stimuli. We selected ten sets (five originally liberal and five originally conservative) of full-length articles (body text) and ten sets of headlines from the test set as our evaluation samples. Each task was a 2 (version: original vs. neutral or reversed) × 2 (headline vs. body text) within-subjects



(a) Gender. (b) Age. (c) Education. (d) Party Affiliation.

Fig. 4. The gender, age, education and party tendency information of the 300 participants (shown in Figure 4(a), Figure 4(b), Figure 4(c) and Figure 4(d) respectively) based on their answer to the demographic questions at the beginning of the survey. Education is reported by the highest academic level. High Sch.: High School; BA/BS: Bachelor of Art/Science; More Adv.: More Advanced. For party affiliation, Ext. = Extreme.

design. In the neutralization task, each set has one originally biased article or headline (*Liberal* or *Conservative*) and one article or headline neutralized by our framework. In the reversal task, each set has one originally biased article or headline as well as one article or headline with reversed ideology. Each participant was randomly assigned to read two sets of body text and two sets of headlines, without being informed which ones are the original and which ones are polarity reversed or neutralized. The order of stories and headlines were randomized.

6.2.2 Procedure. Within each set, first, participants were asked to read and rate each version separately. They were asked to rate perceived bias (or political polarity), political ideology, and readability of each stimulus. Then, two versions (original and flipped) appeared again together without disclosing the type. Participants needed to judge which version is more unbiased, whether the two versions have similar content and hold similar political views.

6.2.3 Measurements. Every measurement used 7-point Likert scales [2]. Perceived bias was measured by asking "How biased do you think this headline (or story) is? 1-Extremely unbiased; 7-Extremely biased." The ideology of the headline or story was measured by asking "What ideology do you think this headline or story has? 1-Extremely liberal; 4-Moderate; 7-Extremely conservative." Content preservation was measured by three perspectives of agreement on the preservation of topic, political views, and semantic meaning (note that we want to preserve the political views as the point of our tool is not to change the views but to change the way of expression). The readability scale included five items adapted from prior work [48]: well-written, concise, comprehensive, coherent, and clear. Participants were asked to rate each item from 1- Very low to 7- Very high. We also evaluated the performance of the polarity detector by asking whether the sentences highlighted by the polarity detector carried more polarity than other sentences. We included fake/randomly highlighted sentences to reduce priming effect and make sure the answers were not biased due to the highlights.

6.3 Results

6.3.1 Polarity Detector. We asked the participants the following question about the polarity of the two sets of highlighted sentences (true highlights: highlighted by the polarity detector; fake highlights: randomly highlighted by researchers):

- *Do the highlighted sentences carry more political polarity than the unhighlighted sentences?*

83% of the participants agreed that the "true" highlighted sentences carried more political polarity, while only 31% of participants agreed to that for the "fake" highlighted sentences.

	Neutralization Task			**Reversal Task**		
	L → N			L → R		
	Mean (SD)	*t*	Sig.	Mean (SD)	*t*	Sig.
headline	3.55(1.90) → 4.24(1.59)	-2.02	0.054†	3.60(1.76) → 4.84(1.75)	2.92	0.008*
body text	3.69(1.67) → 4.10(1.86) -	-0.20	0.053†	3.48(1.48) → 4.27(1.68)	-2.74	0.010**
	R → N			R → L		
	Mean (SD)	*t*	Sig.	Mean (SD)	*t*	Sig.
headline	4.72(1.91) → 4.16(1.78)	1.74	0.098†	4.18(1.68) → 3.21(1.71)	-0.20	0.078†
body text	4.53(1.80) → 4.38(1.81)	0.42	0.681	4.41(2.01) → 3.96(1.65)	1.25	0.222

Table 4. Paired samples t-tests of polarity flipping (neutralization task and reversal task). († corresponds to $p < 0.10$, * to $p < 0.05$, ** to $p < 0.01$, and *** to $p < 0.001$)

6.3.2 Polarity Neutralization. We used paired sample t-tests to examine whether there existed significant differences in the extent of polarity between the original and neutralized headlines and body texts. Results were shown in Tables 4: Neutralization Task. It is worth noting that for originally liberal text, the expected political polarity score should be lower than 4, and for originally conservative text, it should be higher than 4. An expected polarity score of neutralized text is close to 4. Results of the neutralization task in the two tables showed that while participants agreed that our framework succeeded in neutralizing the polarity for all types of of input, the difference before and after neutralized was marginal significant. Results showed that our framework was more successful in neutralizing originally liberal headline and text. When seeing both versions, participants rated the neutralized versions as more neutral than the original versions. On average, 86% of participants agreed the neutralized version are more unbiased than the original version for both headlines and body text.

6.3.3 Polarity Reversal. Same as the neutralization analysis, we used paired sample t-tests to examine whether there existed significant differences between the polarity of the original and reversed headlines and body text. The results were shown in Tables 4: Reversal Task. Results of the polarity reversal task showed that participants agreed that our framework succeeded in reversing the polarity for all types of of input, though the effects were only statistically significant for originally liberal input. For originally conservative headline, the difference before and after reversing was marginal significant. As was the case with the neutralization, our tool was more successful in reversing the polarity of originally liberal text. After seeing both versions, on average 76% of the participants agreed that our framework successfully reversed the political polarity of the headlines, while 81% of the participants agreed with that for body.

6.3.4 Party Affiliation Effect. To further examine whether party affiliation had a main effect on perceived bias scores, we also conducted a multivariate analysis of variance (MANOVA) to measure the effect of the participants' political party affiliation on the bias scores before and after polarity flipping. Analysis from the MANOVA showed that for the liberal headline neutralization task, party affiliation had a significant main effect on the bias score before neutralization (F=5.52, p= .01, *Wilk's* Λ= .04, *partial* η= .21) whereas party affiliation did not have a significant main effect on the bias score after neutralization. Though these results were not significant in other setups, they still indicate that our neutralization tool has the potential to reduce the bias gap brought by party difference. The overall shift in perceived bias for all tasks is shown in Figure 5.

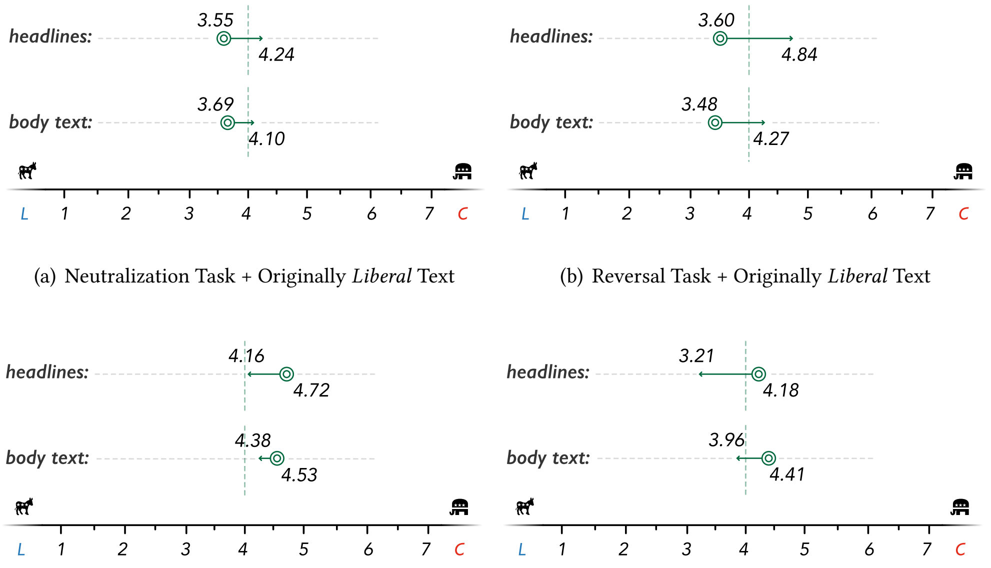
	Model L → C			**Model** C → L			**Model** L → N			**Model** C → N		
	B	SE	Sig.	B	SE	Sig.	B	SE	Sig.	B	SE	Sig.
Constant	2.20	1.25	0.09†	4.39	2.02	0.03*	2.02	1.24	0.11	2.22	1.29	0.09†
Original Bias	0.35	0.12	0.006**	0.37	0.16	0.02*	0.58	0.12	0.00***	0.52	0.12	0.00***
Original Ideology	0.15	0.12	0.23	0.01	0.16	0.93	-0.08	0.13	0.54	-0.02	0.12	0.88
Gender (*Male*)	-0.65	0.42	0.13	-0.45	0.52	0.39	0.29	0.47	0.54	0.47	0.41	0.26
Race (*White*)	0.21	0.43	0.63	0.10	0.68	0.88	-0.003	0.41	0.99	-0.14	0.52	0.80
Education	-0.01	0.05	0.90	-0.11	0.12	0.34	-0.03	0.06	0.62	-0.03	0.04	0.49
Age	0.04	0.02	0.03*	0.003	0.02	0.89	-0.01	0.02	0.39	0.00	0.02	0.86
Participant's Ideology	-0.66	0.23	0.006**	0.003	0.26	0.99	0.08	0.19	0.69	-0.07	0.16	0.66
Party Affiliation	0.21	0.11	0.06†	0.07	0.16	0.65	0.08	0.15	0.63	0.09	0.10	0.37
Media Consumption	0.36	0.24	0.15	-0.05	0.30	0.88	0.10	0.17	0.58	0.05	0.17	0.76
R^2		0.34			0.17			0.53			0.33	
Adjusted R^2		0.21			0.003			0.44			0.22	
Sig. F Change		0.02*			0.44			0.00***			0.006**	

Table 5. Multiple Linear Regression Results. The outcome variable in each model is the perceived bias score after flipped or neutralized. Gender (*Male* = 1, *Female* = 0) and Race (*White* = 1, *Non-white* = 0) are categorical variables. Other predictors in the model are all continuous variables († corresponds to $p < 0.10$, * to $p < 0.05$, ** to $p < 0.01$, and *** to $p < 0.001$).

6.3.5 Multiple Linear Regression. We further investigated the effects of certain demographic variables on the performance evaluation. Multiple linear regression was conducted to predict the perceived bias after being neutralized or flipped (on the same 7-point Likert scale as before) by predictors including the original perceived bias, the original article ideology, and participants' demographic variables. Except for Model C → L, significant regression equations were found in other three models (L → C, L → N, C → N). Table 5 showed that perceived bias of the original version was a significant predictor in each model. Party ideology was a significant predictor in Model L → C (p= .006). Bias and ideology were all measured as continuous variables with the lower score corresponding to more liberal and the higher score corresponding to more conservative. The result showed that in Model L → C, the increase of participants' ideology score can lead to the decrease of the flipped perceived bias score when controlling for other predictors. This result indicates that the more liberal the participants, the more likely they are to think that our model successfully flipped the polarity from liberal to conservative.

6.3.6 Readability. The readability scale was adapted from Haim et al. [48] by asking "how readable the headlines/full-length is?" (1-Very low to 7-Very high). Five items were used to measure the readability of each stimulus:

- Q1: *Is the headlines/full-length article well-written?*
- Q2: *Is the headlines/full-length article concise?*
- Q3: *Is the headlines/full-length article comprehensive?*
- Q4: *Is the headlines/full-length article coherent?*
- Q5: *Is the headlines/full-length article clear?*



(a) Neutralization Task + Originally *Liberal* Text

(b) Reversal Task + Originally *Liberal* Text

(c) Neutralization Task + Originally *Conservative* Text

(d) Reversal Task + Originally *Conservative* Text

Fig. 5. The averaged polarity shifts after we perform polarity neutralization and reversal on originally *liberal* or *conservative* text.

Five items were highly correlated (*Cronbach's* $\alpha = .92$, $M = 23.69$, $SD = 7.06$). Thus, we averaged five items into one reliable readability index. For the polarity neutralization task, the results showed that the neutralized version did not lose readability. As shown in Table 6: Readability, paired sample t-tests showed that no statistically significant differences exist between the readability of the original versions and neutralized versions of the texts for both headlines and full-length articles. In terms of the polarity-reversed versions, the readability was slightly lower than the original versions but this difference was not statistically significant. Results showed that articles and headlines flipped by our framework were as readable as the original ones.

6.3.7 Content Preservation. Participants were asked the following three questions about content preserving performance of our tool:

- Q1: *Do you agree that the three texts above talk about the same topic?*
- Q2: *Do you agree that the three texts above hold the same political views?*
- Q3: *Do you agree that the three texts above are semantically similar?*

The first question was designed as a soft content preserving check, since it is not hard for the tool to maintain the general topic of an article. The second questions was a harder check as it asked about the underlying political views contained in the text. Note that we do not want the tool to change the political views (i.e., turn an anti-immigration article to a pro-immigration article). The third question was designed to measure the preservation at the semantic level. Results were shown in Table 6: Content Preservation. For all tasks and inputs, the vast majority of the participants agreed that the content of the texts were similar for all three questions. In general, more participants thought that the framework preserved the content better for the body than the headline. This could be explained by the fact that full-length articles are longer and have more context than the headlines, making them more robust to changes.

	Neutralization Task			**Reversal Task**		
	Readability					
	M(SD)-*before*	M(SD)-*after*	Sig.	M(SD)-*before*	M(SD)-*after*	Sig.
headline	5.17(1.33)	5.28(1.42)	0.62	5.44(1.25)	5.24(1.31)	0.64
body text	5.33(1.17)	5.38(1.29)	0.63	5.28(1.05)	5.12(1.06)	0.37
	Content Preservation					
	Topic	Political	Semantic	Topic	Political	Semantic
headline	91.46%	81.27%	87.60%	92.22%	79.17%	87.78%
body text	97.52%	87.13%	92.57%	94.53%	74.63%	93.03%

Table 6. **Readability**: Paired samples t-tests of readability were used to determine whether there is statistically significant differences in readability between original and flipped text. **Content Preservation**: The percentage of participants who agree that our flipped text successfully preserves the content in terms of topic, political views and semantic similarity. († corresponds to $p < 0.10$, * to $p < 0.05$, ** to $p < 0.01$, and *** to $p < 0.001$)

7 DISCUSSION

7.1 Theoretical Contribution

Our work adds to the previous literature in both social science and computer science fields. Many studies on selective exposure have found that people are subconsciously inclined to read the news that is congruent with their prior opinions due to confirmation bias [4]. According to our human evaluation, the original perceived bias is a strong predictor of the flipped polarity in the regression models, which further confirms the important role of people's preexisting beliefs.

Social scientists have demonstrated that news content has significant effects on shaping readers' political attitudes [10, 37]. Previous work constructs an index of media slant to automatically identify most partisan phrases used by either Democrats or Republicans. For example, one study is based on counts of all phrases used by Congressional Record and newspapers [37]. The frequency counting method is meaningful yet not capable of reflecting different contexts. Our work, however, addresses this issue by using a context-aware polarity detector.

Apart from detecting bias, our work also provides an effective way to reduce readers confirmation bias. Prior research has explored how to expose people to diverse political news by developing a news service which provides readers with multiple viewpoints [76] or manipulating simple presentation techniques such as highlighting agreeable items via browser widgets [72]. These highlighting presentation techniques, however, appear to be not effective for challenge-averse individuals who dislike challenging messages [72]. Adding to previous literature, our work further examines whether presenting news either neutralized or flipped by our transformer-based framework can reduce people's perceived bias and is shown through evaluations to be effective in mitigating bias.

One finding of our human evaluation is that the more liberal people's orientation are, the more they believe that our model successfully flips the polarity from liberal to conservative. This result is interesting in that it shows the effect of people's political ideology on their perception of news bias. Previous studies on hostile media effect have found that partisans tend to perceive the bias of slanted news coverage differently depending on their own political ideology [46]. Our results are consistent with previous findings. Furthermore, our work provides a new approach to examine hostile media effect by examining how partisans react differently when seeing the polarity of news flips. Another interesting finding of our study is that our neutralization tool can reduce the perceived bias differences between two parties. Future work can further explore whether the partisan difference can be reduced by exposing neutral news content to people with different party affiliations.

7.2 Practical Implications

Our framework has great potential to be used by both scholars and practitioners in the real world. Many scholars have pointed out that without intervention, media will inevitably keep facilitating the proliferation of online echo chambers due to cognitive bias and other social reasons [4, 85]. In order to maximize profits, certain media organizations may even cater to a partisan audience by suppressing information that partisan readers do not like hearing [14]. By proposing our novel framework to detect polarity, and then reverse or neutralize the polarized text, we offer an alternative way to mitigate political polarization and echo chambers.

7.2.1 Social Science Research. In recent decades, political polarization has become a central focus of social scientists [5]. Social scientists can use our framework to test different theories. For instance, selective exposure scholars suggest that individuals often prefer attitude-consistent information due to habitual news use and cognitive dissonance [47, 59]. Other research show that people also seek counter-attitudinal information when a) they have great interests in politics and strong party preference; b) when they perceive the issue as highly important; c) when the articles have accessible attitudes [59]. Previous studies often use news content from real media outlets as experimental stimuli (e.g. [99]). For instance, many researchers use national outlets the *New York Times* and the *Wall Street Journal* to respectively represent liberal- and conservative-leaning newspapers [44]. Even though many social scientists try to pre-test the stimuli in order to select news articles or headlines with different political ideology (e.g. [97]), it's still hard to avoid potential bias caused by different writing quality or source credibility [65, 94]. By using our framework to manipulate the ideology of stimuli, such potential bias can be reduced since our framework is capable of flipping the ideology without shifting semantic meaning and readability of news articles. Social scientists can further explore whether selective exposure effects can be reduced when using phrases and expressions that are either neutral or leaning towards participants' preexisting political beliefs, while controlling for other factors. Other social science theories such as confirmation bias and hostile media effect can also be tested by using our framework to generate comparable stimuli.

7.2.2 Practical Application. Besides the usage in academic research, our framework can become a promising tool for a variety of practical applications. The potential user groups could include:

- *Content creators* such as journalists or editors who have the intention to appeal to larger audiences so that audience of different ideological persuasion would be more receptive of their messages. Studies have demonstrated that news producers or media watchdog organizations (e.g., *Ad Fontes Media*[7]) have made considerable efforts, such as establishing journalism ethics and standards to mitigate media bias [6]. Our tool can be a valuable addition to the arsenal of such content creators.
- *Content consumers* who seek out a broad range of political opinions and information, or so-called "diversity-seeking" individuals [92]. Existing tools on the market either simply score the political bias in terms of media source (such as *NoBias*[8] and *NewsGuard*[9]), or act as news aggregators that expose diverse news to readers (such as *AllSides*[10]). Laboratory tools [76, 92] mostly focus on finding the optimal ratio of congenial and uncongenial information to satisfy

[7] https://www.adfontesmedia.com

[8] https://www.nobias.com

[9] https://www.newsguardtech.com

[10] https://www.allsides.com

readers to the utmost extent. Our tool does not simply label articles with bias scores or prepare a set of unbiased news for the consumer. It empowers news consumers with the ability to flip or neutralize the polarity of any article they choose by themselves so as to make the content more palatable to their tastes.

- *Social media platforms* whose recommendation algorithms try to amplify the information that users are likely to engage with in the future [86]. By changing the polarity of information to better match the users' tastes, social media platforms can serve the same content, albeit with different polarity, to users on different sides of the political spectrum. This will not only help social media platforms by increasing engagement, but will also help reduce filter bubbles.

7.3 Limitations and Ethical Concerns

Admittedly, our framework can raise ethical concerns if used by people with ulterior motives. For instance, computational propagandist may use the framework to campaign and attract supporters. Another concern might be the misperception of the original author's intentions by transferring the style. Note that our framework aims to provide an alternative way to engage people with different pre-existing opinions and beliefs. Our framework can potentially reduce confirmation bias and intrigue people with different beliefs to expose themselves to topics and content they would selectively avoid before, which is the main goal of the framework.

A main limitation of our framework is that while it can maintain the content of a piece of text, it cannot guarantee that the underlying message is unchanged when the polarity is modified. That is because in some cases the style used by the author of a text is deliberately chosen to convey a particular message. Since our tool is essentially modifying the polarity by changing the writing style, it could inadvertently affect the underlying message.

8 CONCLUSION AND FUTURE WORK

This work extends the current political polarity headline flipping task to full-length articles and proposes a novel framework aimed at solving two key problems in this task: (1) Detecting and locating politically polar text in a large article; (2) Reversing or neutralizing the polar text without harming content and fluency. We proposed a polarity detector and a polarity flipper, all based on the Transformer architecture. We used automatic and human evaluations to measure the performance of our framework. We outperform the current state-of-the-art model in several evaluation tasks where direct comparison can be made. Human evaluations confirm that from the view of independent observers, our tool is capable of reversing/neutralizing both headline and full-length articles while maintaining overall fluency and preserving the content. While evaluations show that our model is performing well, there is still much room for improvement. This task, flipping the political polarity of articles, suffers from a lack of clear definition, and more importantly, a lack of standard benchmark datasets. We have made an attempt in this paper to rectify this problem by clearly defining the task and its sub-tasks and collecting and releasing a large annotated dataset for this task.

We hope our framework can be used by social scientists and practitioners to break the echo chamber by reducing political polarization. Though outside the scope of this paper, an immediate extension of this work is to recruit communicators to use this tool and use randomized controlled trials to evaluate the effectiveness of the tool in reaching filter-bubbles. Future work can expand our dataset by including news articles from more diverse outlets (such as independent, non-profit or startup media) to cover a wider spectrum of political views. Considering partisans may have diverse attitudes towards different issues, future work can also include a more balanced distribution of topics so that the effects of different topics (such as gun control and abortion) on partisans' perceptions of flipped news stories can be better understood. Finally, another direct extension to our work could be an investigation on what is the proper ratio of flipped and original sentences that can produce the most satisfying reading experience.

ACKNOWLEDGEMENT

We sincerely thank the reviewers for their insightful comments and suggestions that helped improve the paper. This research was supported in part by the Dartmouth Burke Research Initiation Award and the Amazon Research Award.

REFERENCES

[1] Christopher H Achen and Larry M Bartels. 2017. *Democracy for realists: Why elections do not produce responsive government.* Vol. 4. Princeton University Press.

[2] Gerald Albaum. 1997. The Likert scale revisited. *Market Research Society. Journal.* 39, 2 (1997), 1–21.

[3] Kevin Arceneaux. 2008. Can partisan cues diminish democratic accountability? *Political Behavior* 30, 2 (2008), 139–160.

[4] Karl Ask and Pär Anders Granhag. 2005. Motivational sources of confirmation bias in criminal investigations: The need for cognitive closure. *Journal of investigative psychology and offender profiling* 2, 1 (2005), 43–63.

[5] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[6] Brent H Baker, Tim Graham, and Steve Kaminsky. 1994. *How to identify, expose & correct liberal media bias.* Media Research Center Alexandria, VA.

[7] R. Bakker, C. D. de Vries, Erica R. Edwards, L. Hooghe, S. K. Jolly, Gary Marks, J. Polk, Jan Rovný, M. Steenbergen, and M. Vachudova. 2015. Measuring party positions in Europe. *Party Politics* 21 (2015), 143 – 152.

[8] Delia Baldassarri and Andrew Gelman. 2008. Partisans without constraint: Political polarization and trends in American public opinion. *Amer. J. Sociology* 114, 2 (2008), 408–446.

[9] R. Baly, Georgi Karadzhov, Dimitar Alexandrov, James R. Glass, and Preslav Nakov. 2018. Predicting Factuality of Reporting and Bias of News Media Sources. *ArXiv* abs/1810.01765 (2018).

[10] David P Baron. 2006. Persistent media bias. *Journal of Public Economics* 90, 1-2 (2006), 1–36.

[11] Matthew A Baum, Phil Gussin, et al. 2008. In the eye of the beholder: How information shortcuts shape individual perceptions of bias in the media. *Quarterly Journal of political science* 3, 1 (2008), 1–31.

[12] W Lance Bennett and Shanto Iyengar. 2008. A new era of minimal effects? The changing foundations of political communication. *Journal of communication* 58, 4 (2008), 707–731.

[13] K. Benoit, D. Conway, B. Lauderdale, M. Laver, and Slava Mikhaylov Jankin. 2016. Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data. *American Political Science Review* 110 (2016), 278–295.

[14] Dan Bernhardt, Stefan Krasa, and Mattias Polborn. 2008. Political polarization and the electoral effects of media bias. *Journal of Public Economics* 92, 5-6 (2008), 1092–1104.

[15] Sumit Bhatia and P Deepak. 2018. Topic-Specific Sentiment Analysis Can Help Identify Political Ideology. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.* 79–84.

[16] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science* 6, 1 (2011), 3–5.

[17] Daniel M Butler and Emily Schofield. 2010. Were newspapers more interested in pro-Obama letters to the editor in 2008? Evidence from a field experiment. *American Politics Research* 38, 2 (2010), 356–371.

[18] Pew Research Center. 2014. Where News Audiences Fit on the Political Spectrum. (October 2014). https://www.journalism.org/interactives/media-polarization/table/overall/

[19] Mark S Chen, Xinyi Lin, Chen Wei, and Rui Yan. 2019. BoFGAN: Towards A New Structure of Backward-or-Forward Generative Adversarial Nets. In *International World Wide Web Conference '19.*

[20] Wei-Fan Chen, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2018. Learning to Flip the Bias of News Headlines. In *Proceedings of the 11th International Conference on Natural Language Generation.* 79–88.

[21] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems.* 2172–2180.

[22] Harris Cooper. 2003. *Psychological bulletin.* American Psychological Association, 555.

[23] Lincoln Dahlberg. 2001. The Internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. *Information, communication & society* 4, 4 (2001), 615–633.

[24] Ning Dai, Jianze Liang, Xipeng Qiu, and Xuan-Jing Huang. 2019. Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation. In *2019 Annual Conference of the Association for Computational Linguistics*.

[25] Dave D'Alessio and Mike Allen. 2000. Media bias in presidential elections: A meta-analysis. *Journal of communication* 50, 4 (2000), 133–156.

[26] Russell J Dalton, Paul A Beck, and Robert Huckfeldt. 1998. Partisan cues and the media: Information flows in the 1992 presidential election. *American Political Science Review* (1998), 111–126.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[28] Paul DiMaggio, John Evans, and Bethany Bryson. 1996. Have American's social attitudes become more polarized? *American journal of Sociology* 102, 3 (1996), 690–755.

[29] Holger Döring and Sven Regel. 2019. Party Facts: A database of political parties worldwide. *Party Politics* 25, 2 (2019), 97–109.

[30] Morris P Fiorina and Samuel J Abrams. 2008. Political polarization in the American public. *Annu. Rev. Polit. Sci.* 11 (2008), 563–588.

[31] James S Fishkin. 2011. *When the people speak: Deliberative democracy and public consultation.* Oxford University Press.

[32] Andrew J Flanagin and Miriam J Metzger. 2000. Perceptions of Internet information credibility. *Journalism & Mass Communication Quarterly* 77, 3 (2000), 515–540.

[33] Dieter Frey. 1986. Recent research on selective exposure to information. In *Advances in experimental social psychology*. Vol. 19. Elsevier, 41–80.

[34] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *AAAI Conference on Artificial Intelligence 2018*.

[35] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016), 2096–2030.

[36] Matthew Gentzkow and Jesse M Shapiro. 2006. What drives media slant. *Evidence from US daily newspapers. University of Chicago Technical report* (2006).

[37] Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? Evidence from US daily newspapers. *Econometrica* 78, 1 (2010), 35–71.

[38] Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Cross-Lingual Classification of Topics in Political Texts. In *Proceedings of the Second Workshop on NLP and Computational Social Science*. 42–46.

[39] Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement Learning Based Text Style Transfer without Parallel Training Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 3168–3180.

[40] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[41] Paul Goren, Christopher M Federico, and Miki Caul Kittilson. 2009. Source cues, partisan identities, and political value expression. *American Journal of Political Science* 53, 4 (2009), 805–820.

[42] Tim Groeling. 2013. Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science* 16 (2013).

[43] Tim Groeling and Matthew A Baum. 2009. Journalists' incentives and media coverage of elite foreign policy evaluations. *Conflict Management and Peace Science* 26, 5 (2009), 437–470.

[44] Tim Groseclose and Jeffrey Milyo. 2005. A measure of media bias. *The Quarterly Journal of Economics* 120, 4 (2005), 1191–1237.

[45] Andrew M Guess. 2018. everything in moderation: New evidence on Americans' online media diets. *Unpublished manuscript). Retrieved from https://webspace. princeton. edu/users/aguess/Guess_OnlineMediaDiets. pdf* (2018).

[46] Albert C Gunther, Cindy T Christen, Janice L Liebhart, and Stella Chih-Yun Chia. 2001. Congenial public, contrary press, and biased estimates of the climate of opinion. *Public Opinion Quarterly* 65, 3 (2001), 295–320.

[47] Joseph Graf Christian Sandvig Kyu Sup Hahn Steven H. Chaffee, Melissa Nichols Saphir. 2001. Attention to counter-attitudinal messages in a state election campaign. *Political Communication* 18, 3 (2001), 247–272.

[48] Mario Haim and Andreas Graefe. 2017. Automated news: Better than expected? *Digital journalism* 5, 8 (2017), 1044–1059.

[49] Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation.*

[50] Daniel E Ho, Kevin M Quinn, et al. 2008. Measuring explicit political positions of media. *Quarterly Journal of Political Science* 3, 4 (2008), 353–377.

[51] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1587–1596.

[52] Robert Huckfeldt, Jeffrey Levine, William Morgan, and John Sprague. 1999. Accessibility and the political utility of partisan and ideological orientations. *American Journal of Political Science* (1999), 888–911.

[53] Shanto Iyengar and Kyu S Hahn. 2009. Red media, blue media: Evidence of ideological selectivity in media use. *Journal of communication* 59, 1 (2009), 19–39.

[54] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1113–1122.

[55] Chenyan Jia and Jacek Gwizdka. 2020. An Eye-Tracking Study of Differences in Reading Between Automated and Human-Written News. In *NeuroIS Retreat*. Springer, 100–110.

[56] Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2018. Disentangled representation learning for text style transfer. *arXiv preprint arXiv:1808.04339* (2018).

[57] Kate Kenski and Kathleen Hall Jamieson. 2017. Political Messages and Partisanship. In *The Oxford handbook of political communication*. Oxford University Press.

[58] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* (2019).

[59] Silvia Knobloch-Westerwick and Jingbo Meng. 2009. Looking the other way: Selective exposure to attitude-consistent and counterattitudinal political information. *Communication Research* 36, 3 (2009), 426–448.

[60] Silvia Knobloch-Westerwick and Jingbo Meng. 2011. Reinforcement of the political self through selective exposure to political messages. *Journal of Communication* 61, 2 (2011), 349–368.

[61] Vivek Kulkarni, Junting Ye, Steven Skiena, and William Yang Wang. 2018. Multi-view Models for Political Ideology Detection of News Articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3518–3527.

[62] Matt J Kusner and José Miguel Hernández-Lobato. 2016. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051* (2016).

[63] Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-Attribute Text Rewriting. In *International Conference on Learning Representations*.

[64] Paul Felix Lazarsfeld, Bernard Berelson, and Hazel Gaudet. 1944. The people's choice. (1944).

[65] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[66] Yphtach Lelkes, Gaurav Sood, and Shanto Iyengar. 2017. The hostile audience: The effect of access to broadband internet on partisan affect. *American Journal of Political Science* 61, 1 (2017), 5–20.

[67] Yu-Ru Lin, James P Bagrow, and David Lazer. 2011. More voices than ever? quantifying media bias in networks. *Fifth International AAAI Conference on Weblogs and Social Media.*

[68] Ruibo Liu, Lili Wang, Chenyan Jia, and Soroush Vosoughi. 2021. Political Depolarization of News Articles Using Attribute-aware Word Embeddings, In Proceedings of the 15th International AAAI Conference on Web and Social Media (ICWSM 2021). *arXiv preprint arXiv:2101.01391.*

[69] Stefano Menini, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. 2017. Topic-Based Agreement and Disagreement in US Electoral Manifestos. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2938–2944.

[70] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *The International Conference on Language Resources and Evaluation 2018.*

[71] Sean A Munson, Stephanie Y Lee, and Paul Resnick. 2013. Encouraging reading of diverse political viewpoints with a browser widget. In *Seventh international aaai conference on weblogs and social media.*

[72] Sean A Munson and Paul Resnick. 2010. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1457–1466.

[73] Diana C Mutz and Paul S Martin. 2001. Facilitating communication across lines of political difference: The role of mass media. *American political science review* (2001), 97–114.

[74] Dimitar Nikolov, Mounia Lalmas, Alessandro Flammini, and Filippo Menczer. 2019. Quantifying biases in online information exposure. *Journal of the Association for Information Science and Technology* 70, 3 (2019), 218–229.

[75] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making* 5, 5 (2010), 411–419.

[76] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2009. NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 443–452.

[77] Nathaniel Persily. 2017. The 2016 US Election: Can democracy survive the internet? *Journal of democracy* 28, 2 (2017), 63–76.

[78] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000* (2018).

[79] Markus Prior. 2007. *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press.

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

[81] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2019. Automatically Neutralizing Subjective Bias in Text. *arXiv preprint arXiv:1911.09709* (2019).

[82] Wendy M Rahn, John H Aldrich, and Eugene Borgida. 1994. Individual and contextual variations in political candidate appraisal. *American Political Science Review* 88, 1 (1994), 193–199.

[83] Pew Research. 2014. *Political Polarization in the American Public*.

[84] Fred Rowland. 2011. The Filter Bubble: What the Internet is Hiding from You. *portal: Libraries and the Academy* 11, 4 (2011), 1009–1011.

[85] Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2019. On the inevitability of online echo chambers. *arXiv preprint arXiv:1905.03919* (2019).

[86] Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2020. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science* (2020), 1–22.

[87] Margaret Scammell, Holli Semetko, and David Manning White. 1950. The "Gate Keeper": A Case Study In the Selection of News. In *The Media, Journalism and Democracy*. 119–126.

[88] Philip Seargeant and Caroline Tagg. 2019. Social media and the future of open debate: A user-oriented approach to Facebook's filter bubble conundrum. *Discourse, Context & Media* 27 (2019), 41–48.

[89] Hyoung Ju Seo, Hyeon Seok Tom Yu, and GSB Stanford. 2019. Analyzing the US Federal Legislation Texts: A Deep Learning Approach. *Standford University Technical report* (2019).

[90] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*. 6830–6841.

[91] John Sides and Daniel J Hopkins. 2015. *Political polarization in American politics*. Bloomsbury Publishing USA.

[92] Jennifer Stromer-Galley. 2003. Diversity of political conversation on the Internet: Users' perspectives. *Journal of Computer-Mediated Communication* 8, 3 (2003), JCMC836.

[93] Natalie Jomini Stroud. 2011. *Niche news: The politics of news choice*. Oxford University Press on Demand.

[94] S Shyam Sundar and Clifford Nass. 2001. Conceptualizing sources in online news. *Journal of communication* 51, 1 (2001), 52–72.

[95] Cass R Sunstein. 1999. The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper* 91 (1999).

[96] Astrid Van Aggelen, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders. 2017. The debates of the european parliament as linked open data. *Semantic Web* 8, 2 (2017), 271–281.

[97] Emily Van Duyn and Jessica Collier. 2019. Priming and fake news: The effects of elite discourse on evaluations of news media. *Mass Communication and Society* 22, 1 (2019), 29–48.

[98] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[99] Magdalena Wojcieszak. 2019. What predicts selective exposure online: Testing political attitudes, credibility, and social identity. *Communication Research* (2019), 0093650219844868.

[100] Chen Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. 2019. A Hierarchical Reinforced Sequence Operation Method for Unsupervised Text Style Transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4873–4883.

[101] Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. " Mask and Infill": Applying Masked Language Model to Sentiment Transfer. *arXiv preprint arXiv:1908.08039* (2019).

[102] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*. 5754–5764.

[103] Tae Yano, Philip Resnik, and Noah A Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Association for Computational Linguistics, 152–158.

[104] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI Conference on Artificial Intelligence 2017*.

[105] Ye Zhang, Nan Ding, and Radu Soricut. 2018. SHAPED: Shared-Private Encoder-Decoder for Text Style Adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1528–1538.

[106] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

Received June 2020; revised October 2020; accepted December 2020