Learning to Flip the Bias of News Headlines

Wei-Fan Chen[†] **Henning Wachsmuth**[‡] **Khalid Al-Khatib**[†] **Benno Stein**[†]

[†] Bauhaus-Universität Weimar
Webis Group, Faculty of Media
<firstname>.<lastname>@uni-weimar.de

[‡] Paderborn University
Computational Social Science Group
henningw@upb.de

Abstract

This paper introduces the task of "flipping" the bias of news articles: Given an article with a political bias (left or right), generate an article with the same topic but opposite bias. To study this task, we create a corpus with bias-labeled articles from *allsides.com*. As a first step, we analyze the corpus and discuss intrinsic characteristics of bias. They point to the main challenges of bias flipping, which in turn lead to a specific setting in the generation process. The paper in hand narrows down the general bias flipping task to focus on bias flipping for news article *headlines*. A manual annotation of headlines from each side reveals that they are self-informative in general and often convey bias. We apply an autoencoder incorporating information from an article's content to learn how to automatically flip the bias. From 200 generated headlines, 73 are classified as understandable by annotators, and 83 maintain the topic while having opposite bias. Insights from our analysis shed light on how to solve the main challenges of bias flipping.

1 Introduction

News portals play a central role in our society in different ways: they keep people informed, bring essential topics into public discussions, and they gradually change the attitudes of communities. Noteworthily in this regard, recent studies have exposed various types of bias in the major media portals in the US (Groseclose and Milyo, 2005). For example, media is able to draw the attention to particular entities or events while ignoring others. Also, the selection of *what* to report about a specific entity (e.g., positive or negative facts) undoubtedly produces bias. And not least, the *way* in which news are phrased can emphasize a positive or a negative impression on certain entities and events.

Among these examples, one can argue that bias becomes more obvious when news articles discriminate against entities — particularly in political news. For illustration, consider the following two headlines on Trump recognizing Jerusalem as the capital of Israel, which have been taken from Fox News and New York Times respectively:

Why Trump is right in recognizing Jerusalem as Israel's capital

Trump is making a huge mistake on Jerusalem

While the two headlines describe the same event, they clearly convey a different stance on it. This difference in stance matches the observation that Fox News is considered to have a right-oriented bias, whereas the New York Times is rather seen as left in general.

To keep a news portal's bias uniform, copy editors possibly rewrite articles after receiving them from journalists or other sources (Einsohn, 2011). As a support of this process, but also as an element of the rhetorical machinery of forthcoming argumentation engines, an automatic "bias flipper" would be a very useful research tool. Moreover, a bias flipper would be helpful in practical application domains such as e-journalism, for instance, to automatically rewrite an article from Fox News and then report it on New York Times.

However, rewriting a text with an opposite bias is a challenging task. It requires to identify and to classify the bias (e.g., as left vs. right), which is anything but trivial. Taking a closer look into the example mentioned above, we also see that, without understanding how the bias is manifested in the texts and what the background of the event is, an automatic bias classifier and flipper will not achieve any reasonable performance.

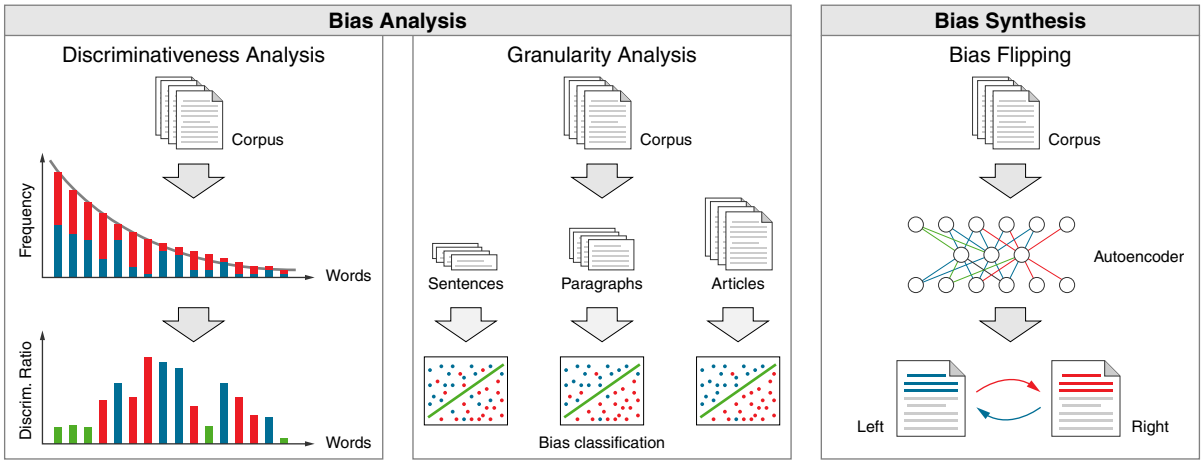

Figure 1: An overview of this paper. Left, we show the discriminativeness analysis of words in the biased text. In the middle, the granularity analysis trains three bias classifiers on different text segments. Right, we use biased articles to train a bias flipper based on autoencoder to flip the bias of headlines.

Accordingly, we approach the bias flipping task with a data-driven approach, addressing the following research questions: (1) How to acquire and sample a reasonable number of biased texts? (2) What kind of bias exists, and how is it manifested in the acquired texts? (3) Given biased texts and a mechanism to understand their bias, how far can we get using the current state-of-art text generation model in trying to flip the bias?

We tackle the first question by exploiting various sources on the web. In particular, we utilize the by-portal article-level bias labels found on *allsides.com*. This platform collects news articles that report on the same event while conveying different bias. Following the distant supervision paradigm, we build a new corpus of 2196 pairs of news article headlines, each of which addresses the same event and opposite bias (i.e., one headline is left-oriented, the other right-oriented).

Using the new corpus, we tackle the second question by analyzing the bias in several experiments (Section 4). Our analysis concentrates on the most discriminative words for identifying the bias, and on how bias is encoded along three granularities of text segments, i.e., in a full article, in a paragraph, and in a single sentence.

Our experiments yield insightful results: While sentimental words play a major role in identifying subjective texts, named entities are shown to be superior for distinguishing left-oriented from right-oriented texts. Moreover, bias often seems to be encoded at article or paragraph level only. In other words, it is hard to capture bias without reading at least a couple of sentences.

Our findings form the ground for tackling the third question, i.e., for developing the first "bias flipper" (Section 5). Considering the difficulty of the task, we focus on flipping news headlines, as a first substantial step in the direction of flipping complete articles. Accounting for recent advances on text generation using deep learning, we study the effectiveness of using autoencoders for flipping. An encoder conditioned on the source bias is used to encode the input text in the semantic representation, while a decoder conditioned on the target bias then decodes the representation into a new text.

We evaluate bias flipping automatically using the Rouge score and manually employing expert annotators. The results of both demonstrate the ability of our model to flip headlines successfully while maintaining the headlines' semantics.

An overview of all experiments carried out in this paper is shown in Figure 1. Our contribution is four-fold: We introduce a new natural language processing task, *bias flipping*; we develop a corpus for investigating this task; we analyze the bias in the developed corpus; and we apply an advanced deep learning model to flip the bias of news headlines. We observe that bias flipping and bias classification are still far from being solved. However, we believe that our bias analysis along with insights from the generation and evaluation experiments will shed light on how to deal with newspaper bias and possibly how to flip the bias of complete articles.

2 Related Work

This section reports on related work regarding the bias datasets, bias analysis, and bias flipping.

Bias Datasets To study the bias in the newspaper domain, several developed corpora include one or more label types related to bias. For example, the news quality corpus created by Arapakis et al. (2016) comprises 561 articles, each of which being labeled with 14 different quality aspects including article's subjectivity. Also, the MPQA corpus contains a label for the subjectivity of its 692 news articles (Wiebe et al., 2005). These two corpora were carefully developed with both article and sentence-level labels. However, they are not large enough to reliably train a supervised learning model.

Recently, a large-scale dataset has been released (Horne et al., 2018). The dataset allows for investigating the news based on various dimensions, including *bias* (the so-called "political impartiality prediction"). Although the dataset is pretty large, it has a major drawback concerning the bias dimension: The articles are not paired according to events, but such a pairing is essential for studying how different news sources report on the same event. To overcome this drawback, we develop a new corpus that aggregates pairs of articles from different news sources. The pairs report on the same event while their sources are said to have an opposite bias. We think that this event-controlled corpus will play a significant role in tackling the tasks of bias analysis and flipping.

Bias Analysis The analysis of media bias has been a subject of investigation for decades (Groseclose and Milyo, 2005; Fang et al., 2012; Arapakis et al., 2016). Various aspects of bias have been studied from different perspectives.

In particular, Groseclose and Milyo (2005) explored the bias on a sample of 20 news sources in the US. The bias was quantified based on the number of citations that were used by the think tanks and policy groups. Their work is one of the first that provided clear evidence of the presence of bias in media. Furthermore, Lin et al. (2011) proposed a scheme for bias categorization. The scheme includes the political party, frequently mentioned legislators, region, ideology, and gender. In a comparison study between the bias in news and blogs, the authors found blogs to be more sensitive to bursting events. In another related work, Yano et al. (2010) focused on liberal and conservative bias. Most notably, they conducted a manual annotation of the bias at the sentence-level. Their study showed that bias indicators usually include named entities of opposing bias. As for our work, we deal with right and left bias, e.g., the democrats' and republicans' bias, or conservative and liberal bias. Also, we conduct an analysis to find the terms that frequently indicate left or right bias.

Bias Flipping Over the few last years, several deep neural networks models have been proposed for text generation. In these models, a variational autoencoder (VAE) has often been used to impose a prior distribution on the hidden vector (Kingma and Welling, 2013; Rezende et al., 2014; Bowman et al., 2016; Yang et al., 2017).

A related research line that addresses rewriting texts is *controlled generation* (Guu et al., 2017; Mueller et al., 2017; Zhou and Neubig, 2017). Controlled generation studies how to rewrite a text with a given attribute. Examples of controlled models include the multi-space VAE of Zhou and Neubig (2017), which modifies a word for a given tense and a part-of-speech tag, and the model of Guu et al. (2017), which generates a sentence given a template vector and an edit vector. This model is shown to be able to paraphrase a given template instead of re-generating a sentence entirely.

Among the collection of VAE models, our work is most closely related to text style transfer (Shen et al., 2017; Hu et al., 2017; Li et al., 2018; Fu et al., 2018); The VAE of Hu et al. (2017) generates sentences with a given style aspect, such as a sentiment or tense. Moreover, the model of Shen et al. (2017) modifies the sentiment of restaurant reviews while aiming to preserve their meaning. However, none of these models has considered bias.

In contrast, this paper employs the cross-aligned autoencoder from Shen et al. (2017). The choice of this model was made based on the results we obtained in our analysis experiments. In particular, we "transfer" the bias of news article headlines using the content of the articles, i.e., we rewrite the headline while flipping the embedded bias from left to right or the other way round.

3 A Corpus of Biased News Articles

This section introduces our new corpus of news articles with different political bias, based on existing bias labels from a news aggregator. The corpus is freely available at https://webis.de/data/corpus-webis-bias-flipper-18.html.

3.1 The News Aggregator allsides.com

The news aggregation platform allsides.com lists news events as of June 1st, 2012; about two to three events per day, focusing on American politics. Each event comes with a title and a short summary, providing information to readers that is said to be free of bias. In addition, one selected news article is given for each of three biases: *left*, *center*, *right* (sometimes, only two articles are available).

The provided bias labels are not article-specific but portal-specific.[1] At the time we collected the data, 247 news portals were assigned one out of six labels each: *left*, *lean left*, *center*, *lean right*, *right*, and *mixed*. We see both the left and the lean left portals as left candidate news sources, and both the right and lean right portals as right candidate news sources. The center and mixed portals are preserved for future applications.

Since the labels are portal-specific, news articles with a particular bias are selected from all portals that have the respective label. Conversely, no portal contains articles with different biases.

3.2 Corpus Construction

We first collected all 2781 events available on the aggregator on February 10th, 2018 (spanning a period of about five and a half years).[2] For each event, the title, the summary, all news portals belonging to the event, and the links to the news portals with respective bias were recorded. After that, we crawled the news portals with the given links to retrieve their headlines and the content of all articles, because the content is not provided on allsides.com. Metadata such as an article's author and its publication time were also collected for future applications. Since some news articles were not available anymore, we retrieved 6447 news articles in the end.

3.3 Corpus

The distribution of news portals and articles in our corpus is shown in Table 1. To validate the accuracy of the by-portal bias, we asked one editing expert to label the bias of all headlines from major left-oriented (New York Times and Huffington Post) and right-oriented portals (Fox News and Townhall). The expert is familiar with American politics and he works as a news editor in the US. His labels are based on the headline only, and the judgments follow the notion of political bias from an American's point of view.

	News Portals		News Articles	
Bias	**Most Common**	**Total**	**Most Common**	**Total**
Left	Huffington Post	21	479	641
Lean left	NY Times	18	688	1747
Center	CNN (web)	24	776	1517
Lean right	Fox News	6	1061	1616
Right	Townhall	28	279	926

Table 1: News portals and articles in our corpus for each bias in total and in the most common portal.

The expert assigned *left* to the headlines of left-oriented portals 3.4 times more than *right*, while the headlines from right-oriented portals have 1.9 times *right* more than *left*. Given that we only looked at the headlines, we conclude that the by-portal labels from the aggregator seem trustable.

The portal labels on allsides.com are created based on different methods including blind surveys, academic research, feedback from the community, and in-depth editorial reviews from allsides.com editors[3]. The final portal labels consider the strength and the consistency of the labels from the different methods. The most common portal contributes at least 30 percent of articles of each bias. The total number of right-oriented news slightly exceeds the number of left-oriented (2542 vs. 2388).

According to the community feedback on the website, the provided labels are agreed by the website's users in general. Thus, we argue that the labeling can be seen as being of high quality.

4 Bias Analysis

In this section, we describe experiments for analyzing biased text, whose results will later be discussed in Section 6. As in the example in Section 1, we observe that bias can be found if we can identify sentiment towards a given entity. Hence, it is worth studying whether the application of sentiment analysis techniques helps on biased text. We seek to identify words which discriminate either sentimental or biased text, and to classify the type of bias using standard features from sentiment analysis.

4.1 Discriminativeness Analysis

We capture the fundamental difference between biased and sentimental text based on the words that

[1] https://www.allsides.com/media-bias/media-bias-ratings

[2] https://www.allsides.com/story-list

[3] https://www.allsides.com/media-bias/media-bias-rating-methods

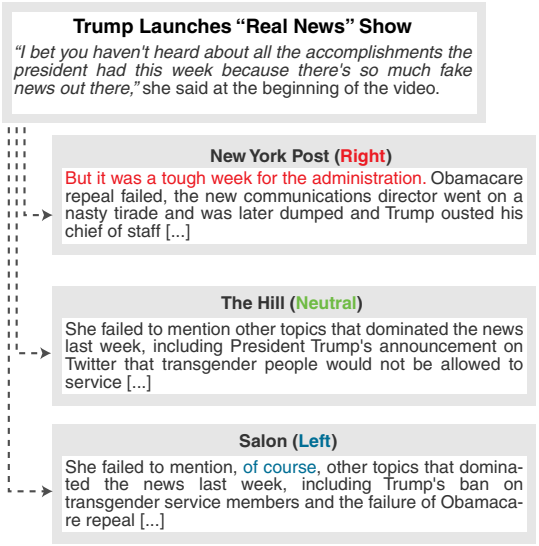

Figure 2: Three news articles on the event *Trump launches "real news" show*. Some bias indicators in the articles are highlighted. Representing three different points of view, the articles provide completely different interpretations of the event.

discriminate the two respective types best. Specifically, the discriminativeness of a word w can be measured in terms of the *discriminativeness ratio*

$$\frac{occ(w, D_t)}{occ(w, D_{\bar{t}})}, \quad (1)$$

where $occ(w, D)$ is the frequency of w in text D and t and $\bar{t}$ are the types of text. In biased text, t and $\bar{t}$ correspond to *right* and *left*. In sentimental text, t and $\bar{t}$ are *positive* and *negative* respectively. We normalize the occurrence by the total numbers of words of the respective type of texts.

The discriminativeness ratio will make function words and type-unrelated words have values close to one, because these words are expected to occur similarly often in both types. On the other hand, words that often appear in one type but rarely in the other will have a high value (in case of type t) or a low value (type $\bar{t}$). To demonstrate the differences in discriminativeness ratios, we analyze biased texts from the corpus introduced in the previous section and compare them to sentimental texts from the public yelp review corpus.[4]

4.2 Granularity Analysis

As in the example shown in Figure 2, we are also aware that some biased text segments can be identified just by looking at its preceding and/or following segments. In this figure, all three sources quote the same utterance, and later give three different interpretations in order to comment on why the woman referred to failed to mention some weakness points of the president during the show. The sentences by The Hill and by Salon are almost the same, but the phrase *of course* in the Salon article is an obvious clue of political bias in it. In contrast, the New York Post gives a reason to explain why the woman failed.

To account for such observations, we train bias models for classifying left and right, based on different lengths of text segments. For each model, we use a support vector machine with word trigram features—a standard yet powerful baseline in sentiment analysis (Liu and Zhang, 2012).

We use the left-right article pairs along with their label from the aggregator as the gold standard. To know whether bias is already recognizable in short text segments, we train and test the model on the article, the paragraph, and the sentence level (for uniform handling, a paragraph is approximated as a continuous sequence of 10 sentences). In case bias is less clear in smaller text segments, we should see a lower classification performance in the paragraph and sentence level results.

We point out, though, that other factors besides this *cross-segment* bias, can influence the performance as well. For example, the different writing style of portals may play an important role, because our dataset is dominated by certain portals (see Table 1). To account for this factor, we decided to upsample our data to balance sources. Since some portals appear only a few times in our dataset, we upsampled only the top-10 most frequent sources in both left and right text.

We expect the performance of classification after the conducted upsampling to be lower than before. However, we should be able to figure out that the performance of smaller text is lower.

5 Bias Flipping

In this section, we introduce a model from related work to generate right-biased headlines given left-biased headlines and vice versa. However, we observed that not all headlines in our corpus show bias. To enrich bias information in the training set, we added the content of each article, split into sentences. We use these sentences as supplemental information during learning. Since we do not have a "flipped" version of each sentence in the content, we do not use the content for the validation and test

[4] https://www.yelp.com/dataset

set, and we evaluate the results only based on the headlines. Knowing that two sentences in a training pair may have different semantics, we need a model that learns to flip bias, but at the same time infers the semantics of a sentence.

Formally, given a source sentence s_o along with its bias label b_o and its content z_o, during training, our goal is to generate the target sentence s_t with label b_t and content z_t, while z_o and z_t could be different. We are interested in flipping the bias from b_o to b_t and from b_t to b_o, so we train two encoders $E(s_k, b_k)$, $k \in \{o, t\}$, that learn to infer z_k:

$$z_k \sim E(s_k, b_k) \tag{2}$$

Analogously, we train two generators G to generate s_k given b_k and z_k:

$$\hat{s_k} \sim G(z_k, b_k) = p(s_k|b_k, z_k) \tag{3}$$

Given the parameters in E and G, θ_E and θ_G, the two autoencoders (one flips from source to target, the other from target to source) are then optimized to minimize the reconstruction error from s_k to $\hat{s_k}$:

$$\mathcal{L}_{rec}(\theta_E, \theta_G) = \mathbb{E}_{s_k \sim S_k}[-\log p(s_k|s_k, E(s_{\bar{k}}, b_{\bar{k}}))],$$

where $\bar{k}$ is o when k is s, and $\bar{k}$ is s when k is o.

As in other generative approaches, we also learn to maximize the loss of the adversarial discriminator as follows:

$$\mathcal{L}_{adv} = -\log D_k(s_k)] - \mathbb{E}[\log -D_k(\hat{s_{\bar{k}}}))], \tag{4}$$

where D_k is the discriminator used to distinguish s_k from the flipped version $s_{\bar{k}}$.

Finally, the loss function aims to minimize the loss from reconstruction and the adversarial discriminators from two directions:

$$\mathcal{L}_{rec_{o \to t}} + \mathcal{L}_{rec_{t \to o}} - (\mathcal{L}_{adv_{o \to t}} + \mathcal{L}_{adv_{t \to o}}),$$

where $o \to t$ means flipping from source to target and $t \to o$ from target to source. To train the model, the architecture of Shen et al. (2017) fits our needs (see Section 2). We thus replicate their cross-alignment setting: During training, we choose the same number of left and right sentences randomly and then train the autoencoder from two directions in one batch. Even though the pairing information is saved by this architecture, the results are promising: Modifying the sentiment while maintaining semantics worked correctly in 41.5% of all cases.

		Same Event (Q3)			
		Same	**Changed**	**Not Sure**	**All**
Bias (Q4)	**Flipped**	57	1	0	58
	Same	28	1	0	29
	Not Sure	10	1	2	13
	All	95	3	2	100

Table 2: Counts of all possible combinations in the manual evaluation of whether the ground-truth headlines capture the same event with flipped bias.

Besides, generative models are known to often produce *UNK* (the out-of-vocabulary word), which is especially harmful in understanding the meaning of short sentences, as given in our task. In order to reduce the frequency of *UNK* in the generated outputs, we set the size of beam search to 10, and keep the candidates with the fewest *UNK*.

6 Results and Discussion

In this section, we try to answer our three research questions from Section 1 by analyzing the results of our experiments. Firstly, to study the appropriateness of our corpus for the given task, we verify that the corpus headlines are informative and have the expected bias. Then, we discuss the result of bias analysis. Later, we evaluate headlines generated by the approach against this ground-truth, both automatically and manually. Finally, a general discussion of the bias flipping task is given.

6.1 Ground-truth Headlines

From our corpus, we took all 2196 opposite headline pairs *(left-oriented, right-oriented)*. Both headlines of a pair are about the same event. We randomly selected 100 pairs as the validation set, another 100 pairs as the test set, and the remaining as the training set. To verify the test set, we hired three experts in journalism editing to annotate all 100 test pairs. For each pair, the annotators had to answer four questions:

Q1. Do you understand headline 1?
{*yes* | *partially yes* | *no* | *not sure*}

Q2. Do you understand headline 2?
{*yes* | *partially yes* | *no* | *not sure*}

Q3. Do both headlines report on the same event?
{*same* | *mostly same* | *changed* | *not sure*}

Q4. Do the headlines have opposite bias?
{*flipped* | *partially flipped* | *same* | *not sure*}

Sentimental Text		Biased Text	
Word	**Ratio**	**Word**	**Ratio**
excellent	220.22	Chad	9.52
gem	183.99	Maduro	5.56
wonderful	183.66	purportedly	7.81
delicious	156.72	Chechnya	6.80
fantastic	142.52	Bethlehem	6.04
...	...	...	...
mushrooms	1.01	victorious	1.01
breadsticks	1.01	oppressive	1.01
dresser	0.99	tragedy	0.99
...	...	...	...
unfortunately	< 0.01	Shawn	0.04
terrible	< 0.01	incarceration	0.04
rude	< 0.01	album	0.03
horrible	< 0.01	valuable	0.03
worst	< 0.01	N.S.A	0.02

Table 3: The five words each with the highest and lowest discriminativeness ratio, and words with a ratio close to one, in sentimental and in biased text.

Text segment	**Original**	**Source-normalized**
Article	0.94	0.89
Paragraph	0.82	0.73
Sentence	0.76	0.59

Table 4: Bias classification accuracies on different size of text segments, once on the original data and once for normalized (upsampled) sources.

		Same Event (Q3)			
		Same	**Changed**	**Not Sure**	**All**
Bias (Q4)	**Flipped**	83	17	4	104
	Same	21	10	0	31
	Not Sure	23	33	9	65
	All	127	60	13	200

Table 5: Counts of all combinations in the manual evaluation of the generated compared to the ground-truth headlines in terms of event and bias.

The resulting Fleiss'κ values were 0.97 (Q1), 0.97 (Q2), 0.62 (Q3), and 0.30 (Q4). All annotators understood almost all headlines, except for one with only two words: "Lerner speaks". The agreement for Q3 was substantial and fair for Q4. Majority voting was used for the final decision.

Table 2 shows the annotations of Q3 and Q4, combining *same* and *mostly same* for Q3, and *flipped* and *partially flipped* for Q4. From the 100 pairs, 95 were labeled as being on the same event, while only five pairs confused the annotators. For the bias label, 58 headline pairs have opposite bias, while the rest did not show any clear difference.

6.2 Bias Analysis

In Table 3, we list the words having the highest and the lowest discriminativeness ratio in sentimental and in biased text respectively. We see that, the top-5 words in sentimental text are positive words and the bottom-5 words are negative words. Entities such as *mushrooms* or *dresser* have values close to one. The results fit the intuition that people usually use positive words in a positive review, such as "great breakfast place", and negative words in a negative review. While sometimes negative expressions use positive words by negating ("my experience here was not great at all"), the ratio of words clearly shows this tendency.

In contrast, we observe that this is not the case in biased text. There, both positive and negative sentiment words have a frequency ratio close to one. This is expected, because we observe that both sides use positive (negative) words to support (oppose) some entities. Moreover, many of the top-5 and the bottom-5 words are named entities, such as *Maduro* and *N.S.A*. This indicates that articles with either bias tend to criticize or approve different entities, but that they do not use different sentiment words to do so. In line with this, a previous analysis on bias language showed that many bias indicators include named entities (Yano et al., 2010).

The results of bias classification is shown in Table 4, and the distribution of bias is balanced. In general, we observe that bias classification on the article level appears not to be very difficult. Even though we only employ rather simple models and features, we achieve a very high accuracy of 0.94. Also, the shorter the segments that we use for training and testing, the lower the classification performance we get (although it always remains higher than chance). As expected, when we upsample the sources, performance is reduced. However, our hypothesis is still supported: a part of bias is conveyed by longer text segments only.

6.3 Generated Headlines

Besides the model we propose in the paper, we also experimented with other approaches that generate a text given another text. Specifically, we tried (1) training our model only with headline pairs, (2) the pointer generator (See et al., 2017) trained only with headline pairs, and (3) the sentiment and style transfer from Li et al. (2018). The pointer generator originally focuses on abstractive summarization where it achieved high Rouge scores. It learns to copy words from the source to han-

Ground-truth headline pair		Generated versions of the headlines		Evaluation	
Headline	Bias	Headline	Bias	Event	Bias
John McCain urges republicans not to filibuster gun control.	left	*John McCain has elected to avoid gun control.*	right	same	flipped
White House looks to salvage gun-control legislation.	right	*White House got to get bipartisan change.*	neutral	mostly same	partially flipped
Obama accepts nomination, says his plan leads to a "better place".	left	*Obama blasted re-election, saying it a "very difficult" to go down.*	right	mostly same	flipped
Lackluster Obama: change is hard, give me more time.	right	*Real GOP: debate is right, and more Trump.*	left	changed	flipped

Table 6: Two left-right headline pairs, along with the rewritten versions generated by our approach. The bias of the ground-truth headlines is given in our corpus. The bias of the generated headlines is from the human annotators.

dle out-of-vocabulary issues. The sentiment and style transfer focuses on detecting the attribute (the sentiment words for instance), trying to alter it by looking for the best candidates in a corpus.

However, even when fine-tuning their parameters, neither of these approaches generated readable outputs. Mostly, they just repeated words or phrases, such as in "the the the" or "trump he same he for trump". So, without sufficient content in the training data, it seems hard to obtain a language model that generates meaningful sentences.

In particular, the pointer generator requires paired training samples, hence training with sentences from the content is not possible. The sentiment and style transfer does not require paired training samples, but its attribute detection mechanism requires an unequal distribution of sentiment words. From the experiment in bias analysis, we know that this assumption does not hold in our corpus. The model described in the approach section is an end-to-end model without any strong assumption. Although it has higher amount of parameters, it can produce more readable sentences.

For automatic evaluation, we measured the similarity between the generated and the ground-truth headlines via Rouge-1, Rouge-2, and Rouge-L, resulting in F-scores of 15, 3, and 12. In an additional manual evaluation, another three editing experts answered Q2 to Q4 by comparing the original and generated headlines, with a Fleiss' κ of 0.61 (Q2), 0.51 (Q3), and 0.29 (Q4). Out of 200 generated headlines (100 left-to-right, 100 right-to-left), 73 were seen as understandable (Q2), which we see as a good result for a generative model. For Q3 and Q4, Table 5 details the results. For those headlines, where the content was kept (127), the bias was flipped in 83 cases (65%). Even for those with changed meaning, 28% got the opposite bias.

6.4 Analysis

Table 6 shows selected pairs of ground-truth and generated headline. They demonstrate that our model keeps the event similar by using the same words, and flips bias by replacing or adding bias words. The generated headlines contain some grammar errors, but we see these as tolerable in machine-generated text on limited data.

In the first pair, the original headline states that McCain was pro gun control, while the rewritten one implies he was against — a successful flip. The ground-truth bias-flipped headline in the second row mostly uses other words while being pro gun control. The generated headline also keeps most words, but turns out rather neutral. In the second pair, the original headline shows a positive opinion on Obama, the generated headline a negative opinion on him. When rewriting the ground-truth bias-flipped headline (last row), the meaning is not kept. However, it is visible that the generated headlines is pro Trump.

We point out that there is a difference between bias flipping and fact changing. For example in the first pair, without knowing what John McCain stood for, we could neither guess his real opinion on gun control nor could we conclude what he supported or not. In fact, bias can be conveyed by emphasizing facts supporting a claim, as well as by hiding facts attacking a claim. In other words, we might see different facts about the same event with different types of bias. A news headline may be a conclusion, while the news content shows the facts supporting this conclusion. In such cases, no computational model will be able to flip the content only using the text itself, as it is hardly possible to simply generate new facts. Including more articles reporting on the same event will be useful to help

the model learn the unseen information. We see this as future work on article-level bias flipping.

Finally, we found that an automatic evaluation of bias flipping is limited. In the discussed examples, we see that even for a successful flipping, the overlapping of generated and ground-truth headlines are very low. In fact, the successful cases have a mean Rouge-1 score of 17, unsuccessful ones of 15. Furthermore, if we divide the test pairs into those labeled as *same event* and *flipped bias* (57 pairs) and the rest (43), we find that the former are more often rewritten successfully (43% vs. 20%). This suggests that filtering out noisy cases with the help of experts will help improve the performance.

7 Conclusion

This paper has introduced the challenging task of rewriting news articles with flipped political bias as well as a bias-labeled corpus to study the task. As a first step, we have tackled the analysis of biased text and compared biased with sentimental text. We have found that (1) the types of discriminative words for biased and sentimental text are entirely different, and (2) some bias is visible on paragraph level only or even article level only. We have then applied a cross-aligned autoencoder to rewrite article headlines with flipped bias, incorporating content information from the article. Our experiments suggest that current state-of-the-art approaches struggle with this task. While our best tested model performed considerably well, there is still much room for improvement. Regarding the evaluation of the model, the Rouge score turned out insufficient to assess bias flipping quality.

In the future, we aim to employ the knowledge from bias analysis in the generation process, to rethink existing automatic evaluation metrics, and to study how to flip the bias of complete articles.

References

Ioannis Arapakis, Filipa Peleja, Barla Berkant, and Joao Magalhaes. 2016. Linguistic benchmarks of online news article quality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1893–1902.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *CoNLL 2016*, pages 10–21.

Amy Einsohn. 2011. *The copyeditor's handbook: A guide for book publishing and corporate communications*. University of California.

Yi Fang, Luo Si, Naveen Somasundaram, and Zhengtao Yu. 2012. Mining contrastive opinions on political texts using cross-perspective topic model. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 63–72. ACM.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the Thirty-Second Conference on Artificial Intelligence*.

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

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2017. Generating sentences by editing prototypes. *arXiv preprint arXiv:1709.08878*.

Benjamin D Horne, William Dron, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. *arXiv preprint arXiv:1803.10124*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. In *The 2nd International Conference on Learning Representations*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *North American Association for Computational Linguistics*.

Yu-Ru Lin, James P. Bagrow, and David Lazer. 2011. More voices than ever? quantifying media bias in networks. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Jonas Mueller, David Gifford, and Tommi Jaakkola. 2017. Sequence to better sequence: continuous revision of combinatorial structures. In *International Conference on Machine Learning*, pages 2536–2544.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning*, pages 1278–1286.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083.

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*, pages 6833–6844.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3881–3890.

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*, pages 152–158. Association for Computational Linguistics.

Chunting Zhou and Graham Neubig. 2017. Multi-space variational encoder-decoders for semi-supervised labeled sequence transduction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 310–320.