Neutral Summarization for Framing Biased Articles

Kevin Yao, David Li, Keith Wang, Yuyang Xia

Abstract

Media framing bias, a result of skewed selective writing by journalists, might lead to deepened division between people and undermine social solidarity. Based on previous works, we suggest that automatic neutral summarization could help describe media news in a relatively neutral manner. In this paper, we use the T5 encoder and fine-tune the T5 decoder to perform the Seq2Seq (sequence-to-sequence) generation task. Moreover, we changed the task structure from multi-document summarization to single-document summarization. The input doesn't have to be a triplet of articles from left, neutral, and right standpoints. Instead, the model can take articles from whichever standpoints and output a neutral news summarization, which largely increases neutral summarization generation efficiency and makes it much more convenient for readers to get neutral standpoints from a news report. Our experiment result indicates that our fine-tuned model with a single-document summarization task can achieve better results in framing bias metrics compared with baseline models. We then analyze the advantages of our model over the baseline (Bart-large) model. Furthermore, we propose some possible improvements and future works.

1 Introduction

In media news, the problem of media framing bias often occurs as a result of skewed selective writing by journalists. Different media agencies often write vastly different articles on the same topic or event. Writing articles that better fit the prior belief and ideology of the targeted audience can increase the agencies' profitability and build a reputation for quality (Gentzkow and Shapiro, 2005). However, as media news is crucial in shaping an individual's opinion towards various important issues (de Vreese, 2009), media framing bias might lead to deepened division between people and undermine social solidarity. Worse still, modern media often employ recommender systems, which might further increase the division between people holding different points of view. The problem is widely researched in sociology and journalism.

Neutral summarization, which extracts salient information while remaining neutral politically and sentimentally might be a solution for mitigating the framing bias. From a neutral summarization, viewers can have a grasp of what's the neutral description of the matter, which may help mitigate the division of viewers holding different points of view. However, manually generating a neutral summarization is a laborious task, so the idea may suffer from the difficulty to write a summarization. Auto summarization models could be a possible solution to the task. Yet the ability of summarization models to generate neutral summarization remains largely unexplored.

Lee et al. (2022) presented a new task, a neutral summary generation from multiple news articles of varying political leanings to facilitate balanced and unbiased news reading. They collected a new dataset from Allsides.com (Allsides.com, 2022), in which three articles (one politically left-wing, one right-wing, one centered) are summarized into a neutral summarization written by the journalists. They trained BART (Lewis et al., 2020)'s autoregressive decoder to generate a framing-bias-free summary from news articles with varying degrees and orientations of political bias.

In this work, we fine-tuned a pre-trained transfer-learning model, say T5, to do the task of neutral summarization. We start with processing the dataset by Lee et al. (2022), from Allsides.com (2022). Previously works often viewed this task as a multi-document summarization problem, and applied Multi-document summarization (MDS) models (Lebanoff et al., 2018) to generate neutral sum-

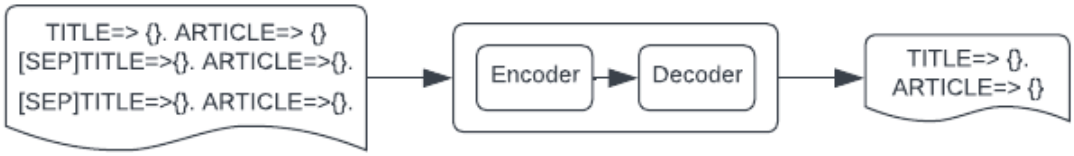

Figure 1: Here is an illustration of a 3-1 task (three input article, left, right, and centered, and one neutral summarization output). The left side is input and the right side is output. The model is a vanilla transformer Bart or T5 (T5-base/T5-large) composed of encoders and decoders.

Figure 2: Here is an illustration of a 1-1 task (one input article, can be left, right, or centered, and one neutral summarization output). The left side is input and the right side is output. The model is a vanilla transformer Bart or T5 (T5-base/T5-large) composed of encoders and decoders.

marization in a larger scale. We modified the design of the task. Instead of summarizing multiple documents at the same time, we designed the task to be a supervised single-document summarization task, or to be more simple, we changed the task structure from a 3-1 (one left-leaning, one right-leaning, one centered) summarization task proposed by Lee et al. (2022) to a 1-1 (one article that can be left-leaning, right-leaning or centered) summarization task. There are two reasons behind the design. Firstly, training a model to provide neutral summarization given one specific article better stimulates the real-world task situation. Secondly, with too much biased information in the input, the model may output text containing more biased information. We suppose the model will generate less biased text with less biased information as the input each time.

We also adjusted the evaluation matrix provided by Lee et al. (2022). From our experiment result, our fine-tuned model achieved a better score in the Framing Bias matric.

2 Related Works

2.1 Media Bias

Media bias has always been an important problem in the news area. It shows disparities in news content description from different perspectives, which is harmful to both individuals and society (Hamborg, 2020). Readers are susceptible to media bias in the way of framing (Scheufele, 2000). In general, framing means any factor that has an influence on readers' perception of given reality and information (Goffman, 1974). When it comes to the news area, framing refers to how journalists characterize an issue and how readers interpret the information (Scheufele and Tewksbury, 2007). Media bias is usually caused by choices of facts and nuances in words. For an event, different journalists may select specific facts and ignore some facts to present in the news article. Different words representing the same meaning may vary a lot in terms of readers' perception. In our task, we mainly concentrate on word choices and deliberately added information describing an event that may lead to misperceptions for readers.

2.2 Media Bias Mitigation

While the traditional method to reduce media bias is to combine news from various perspectives on a particular topic (Hamborg et al., 2019), it's time-consuming for readers to read all news. Regardless of readers' reluctance to accept a great amount of information from different views, readers are

likely to be misled by biased news if they cannot distinguish between biased descriptions of an event. Other media bias mitigation models are proposed, mainly focusing on adding extra information to help readers obtain a bird-eye view of specific news. For example, media profiles showing report factuality and hyper-partisanship can be generated and presented together with news stories to readers (Zhang et al., 2019). However, it still requires readers to read additional information and identify biased news by themselves. Recently, the automatic neutralization and summarization model is proposed to remove the burden of mitigating media bias on readers (Lee et al., 2022). Readers do not have to receive additional information to help them reduce media bias. Therefore, we focus on news automatic neutralization and summarization task to further get rid of readers' burden on distinguishing biased news.

2.3 Transfer-Learning Models

Transfer learning language models have revolutionized the field of natural language processing by allowing pre-trained models to be fine-tuned for specific tasks, rather than training from scratch for each new task. This approach has resulted in significant improvements in the performance of natural language processing (NLP) models across a wide range of tasks, including language modeling, text classification, question-answering, summarizations and many others.

One of the most well-known and widely-used transfer-learning language models is BERT (Devlin et al., 2018) (Bidirectional Encoder Representations from Transformers), developed by Google researchers in 2018. BERT is a pre-trained model that uses a transformer architecture to encode contextual relationships between words in a sentence. It has been fine-tuned for a wide range of NLP tasks, including sentiment analysis, named entity recognition, and natural language inference, achieving state-of-the-art performance on many benchmark datasets. It led to the further development of more sophisticated models such as BART (Lewis et al., 2020). BART is a denoising autoencoder

for pretraining sequence-to-sequence models. It is trained by corrupting text with an arbitrary noising function and trying to reconstruct the original text. It uses a standard Transformer-based neural machine translation architecture which, despite its simplicity, can be seen as generalizing BERT. BART is particularly effective when fine-tuned for text generation.

Besides BART, recent advancements in transfer learning language models also include T5 (Raffel et al., 2019) (Text-to-Text Transfer Transformer), developed by Google researchers. T5 is a general-purpose language model that can be fine-tuned for a wide range of tasks, including question-answering, text classification, and summarization. We will further introduce T5 design and fine-tuning in a more detailed way in §4.2.

3 Task Design

3.1 Dataset

Allsides.com (2022) is a unique news platform (mainly focusing on U.S. political topics) that aims to foster balanced and inclusive perspectives on current events by categorizing articles according to their political bias. With a commitment to transparency, Allsides.com (2022) presents news from left, center, and right-leaning sources for each news report, allowing readers to gain a comprehensive understanding of diverse viewpoints and encouraging constructive dialogue across the political spectrum. Moreover, Allsides.com (2022) provides an expert-written neutral title and summary for each reported event on the website.

Thanks to Lee et al. (2022), we use part of the Allsides.com (2022) data provided and processed by them. Lee et al. (2022) stack reports from three perspectives that are categorized as left, center, and right for each news event. They combine titles and articles from the left, center, and right standpoints as the input, and use the expert-written neutral title and summary as the target for their model. They perform the multi-document summarization task. The format of their input and output are as follows:

Input:
$TITLE \Rightarrow T_{x_1}. \ ARTICLE \Rightarrow A_{x_1}.$
$[SEP]TITLE \Rightarrow T_{x_2}. \ ARTICLE \Rightarrow A_{x_2}.$
$[SEP]TITLE \Rightarrow T_{x_3}. \ ARTICLE \Rightarrow A_{x_3}.$
Output Target:
$TITLE \Rightarrow T_C. \ ARTICLE \Rightarrow A_C,$
where $x_i \in \{Left, Center, Right\}$, and $C = Center$. The order of left, right and center will be randomly shuffled and then stacked together.

Most readers, in contrast, rarely have access to news with full perspectives. Normally, they will get exposed to a certain viewpoint of the latest news which is highly possible to be extreme. Publications standing for either side incline to print singu-

SOURCE: <**Left**> **TITLE**=> Supreme Court opens new term at the center of America's bitter political divide. **ARTICLE**=> The Supreme Court opens its new term Monday at the forefront of the national political conversation, but with its future uncertain and the unwelcome prospect of deciding a divisive presidential election on the horizon. With Justice Ruth Bader Ginsburg's seat on the bench still draped in black crepe, the eight remaining justices will gather via teleconference to tackle a docket that, for now, is not nearly as controversial as the last. **SOURCE:** <**Center**> **TITLE**=> U.S. Supreme Court nominee's confirmation hearings on track, McConnell says. **ARTICLE**=> U.S. Senate Majority Leader Mitch McConnell said on Saturday that Supreme Court nominee Amy Coney Barrett's confirmation hearings will proceed as planned even though two Republicans on the Judiciary Committee had contracted the coronavirus. Senators Thom Tillis and Mike Lee, who sit on the panel that will preside over the hearings, tested positive for COVID-19 in the past day. Both were at the White House on Sept. 26 when President Donald Trump introduced Barrett as his nominee to replace Supreme Court Justice Ruth Bader Ginsburg, who died last month. **SOURCE:** <**Right**> **TITLE**=> 'Can move forward': Supreme Court nominee Amy Coney Barrett tests negative for coronavirus. **ARTICLE**=> Amy Coney Barrett, President Trump's nominee for the Supreme Court, has tested negative for COVID-19. White House deputy press secretary Judd Deere told the Washington Examiner on Friday morning that Barrett, who has recently been making the rounds on Capitol Hill and meeting with senators, is coronavirus-free. "Judge Barrett was with the president last on Saturday when she was nominated. Judge Barrett is tested daily for COVID-19 — she has tested negative," Deere, who is also a special assistant to the president, said.
TARGET: TITLE=> Supreme Court to Open New Term While Confirmation Hearings Continue. **ARTICLE**=> The Supreme Court will open a new term via teleconference Monday with one open seat, and the U.S. Senate moves forward on confirmation hearings despite two Republicans on the Judiciary Committee testing positive for coronavirus this week. On Saturday, Senate Majority Leader Mitch McConnell said high court nominee Amy Coney Barrett's confirmation hearings will continue on Oct. 12. Barrett, who was with the president last Saturday during her nomination, tested negative this week, after recovering from the virus earlier this summer. This term, the Supreme Court will likely hear arguments to overturn Obama-era healthcare, address mail-in voting policies and hear arguments over whether Congress can access files from special counsel Robert Mueller's Russia investigation. Coverage from all sides of the political spectrum tended to analyze the issues likely to face the court in the upcoming term. Right-rated outlets tended to highlight Barrett's negative COVID-19 test and her likelihood of being confirmed before the election.
T5-base: TITLE=> U.S. Supreme Court nominee Amy Coney Barrett tests negative for COVID-19. **ARTICLE**=> Senate Majority Leader Mitch McConnell said on Saturday that Barrett's confirmation hearings will proceed as planned even though two Republicans on the Judiciary Committee had contracted the coronavirus.

Table 1: Generation examples for analysis purposes for 3-1 task.

SOURCE: <Left> **TITLE**=> Supreme Court opens new term at the center of America's bitter political divide. **ARTICLE**=> The Supreme Court opens its new term Monday at the forefront of the national political conversation, but with its future uncertain and the unwelcome prospect of deciding a divisive presidential election on the horizon. With Justice Ruth Bader Ginsburg's seat on the bench still draped in black crepe, the eight remaining justices will gather via teleconference to tackle a docket that, for now, is not nearly as controversial as the last.
TARGET: TITLE=> Supreme Court to Open New Term While Confirmation Hearings Continue. **ARTICLE**=> The Supreme Court will open a new term via teleconference Monday with one open seat, and the U.S. Senate moves forward on confirmation hearings despite two Republicans on the Judiciary Committee testing positive for coronavirus this week. On Saturday, Senate Majority Leader Mitch McConnell said high court nominee Amy Coney Barrett's confirmation hearings will continue on Oct. 12. Barrett, who was with the president last Saturday during her nomination, tested negative this week, after recovering from the virus earlier this summer. This term, the Supreme Court will likely hear arguments to overturn Obama-era healthcare, address mail-in voting policies and hear arguments over whether Congress can access files from special counsel Robert Mueller's Russia investigation. Coverage from all sides of the political spectrum tended to analyze the issues likely to face the court in the upcoming term. Right-rated outlets tended to highlight Barrett's negative COVID-19 test and her likelihood of being confirmed before the election.
T5-base: TITLE=> Senate Appoints Amy Coney Barrett to Supreme Court. **ARTICLE**=> Senate Majority Leader Mitch McConnell said on Saturday that Barrett's confirmation hearings will proceed as planned, even though two Republicans on the Judiciary Committee tested positive for COVID-19 in the past day. Both were at the White House on Sept. 26 when President Donald Trump introduced Barrett as his nominee to replace Supreme Court Justice Ruth Bader Ginsburg.

Table 2: Generation examples for analysis purposes for 1-1 task. This example takes the left-positioned title but could be right or center.

lar perspective. More information released tends to be more biased and lacks the neutral and the opposite editions. Hence, we set the input and target output as follows (performing single-document summarization task), to better simulate the actual situation:

Input:
$TITLE \Rightarrow T_\alpha. \ ARTICLE \Rightarrow A_\alpha.$
Output Target:
$TITLE \Rightarrow T_{\hat{\alpha}}. \ ARTICLE \Rightarrow A_{\hat{\alpha}},$
where $\alpha \in \{Left, Center, Right\}$, and $\hat{\alpha} = Center$. For the same topic news, we pair its left, center, and right viewpoint titles and articles to its center. The original training set, therefore, expands three times as the original. The dataset is also more appropriate to our model, which will be discussed more in §3.2.

We perform two tasks in the paper, multi-document summarization task (denoted as 3-1 task) and single-document summarization task (denoted as 1-1 task).

For 3-1 task, the input is composed of article triplets A_L, A_C, A_R and corresponding title triplets T_L, T_C, T_R, where L, C, R means left, center and right standpoints. The target is expert-written neural titles and summaries A_{target}, T_{target}. We have 3562 pieces of data in total. We use 80% of data (2851 pieces) as the training set, 10% data (356 pieces) as the validation set, and 10% data (357 pieces) as the test set.

Differently, for 1-1 task, the input is separated into three single pieces of articles and corresponding titles T_{input}, A_{input} from different standpoints. Unlike 3-1 task, articles and corresponding titles from three different standpoints are put into the model separately (in different rounds), instead of being put into the model as triplets. The target is expert-written neural titles and summaries A_{target}, T_{target}, which is the same as that in 3-1 task. For 1-1 task, we have 10686 pieces of data in total. We use 80% of data (8547 pieces) as the training set, 10% data (1068 pieces) as the validation set, and 10% data (1071 pieces) as the test set.

3.2 Model Design

As mentioned before, our neutral summarization task is a Seq2Seq (sequence-to-sequence) genera-

tion task. We choose the T5 (Raffel et al., 2019) (Text-to-Text Transfer Transformer) base as our base model. It is a powerful natural language processing model developed by Google Research. The T5 model structure is based on the Transformer architecture introduced by Vaswani et al. (2017). It consists of an encoder-decoder framework with multi-headed self-attention mechanisms and position-wise feed-forward networks. Both the encoder and decoder are composed of a stack of identical layers, with each layer containing two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. Additionally, residual connections and layer normalization are employed around each sub-layer to improve training stability and model performance.

The multi-head self-attention mechanism enables the model to process and weigh different features of the input text simultaneously. This is achieved by splitting the input into multiple parts, referred to as "heads", and calculating self-attention weights for each head independently. The final output is generated by concatenating and linearly transforming the attention results from all heads.

The T5 model is pre-trained on a large-scale dataset C4 (Colossal Clean Crawled Corpus) using a text-to-text format and 220 million parameters, wherein both input and output are treated as text sequences. This unified framework simplifies the training process and enables the model to perform tasks such as translation, summarization, and question-answering. To train T5, Raffel et al. (2019) used a process called "unsupervised and supervised learning", which involves retraining the model on unsupervised data, followed by fine-tuning task-specific supervised data. In this work, we use T5-base and T5-large to fine-tune our dataset with specific one-to-one tasks.

In this work, we further fine-tuned the T5 model to do a single-document summarization task. We changed the task structure from a 3-1 summarization (figure 1) task proposed by Lee et al. (2022) to a 1-1 summarization task (figure 2).

4 Evaluation Methods

After we are able to summarize natural summarization by framing biased titles/articles, we need to evaluate our summarization. To assess summaries from various angles, we utilize two metrics (Lee et al., 2022). To address framing bias, we propose a polarity-based metric that takes into account the careful design choices explained in §4.1. To determine if the summaries maintain crucial information, we employ commonly used Information Salient metrics as described in §4.2.

4.1 Framing Bias Metric

To create our metric, we utilize the Valence-Arousal-Dominance (VAD) dataset (Mohammad, 2018), which contains a vast collection of lexicons annotated with valence (v), arousal (a), and dominance (d) scores. Valence, arousal, and dominance respectively indicate the polarity direction (positive or negative), polarity strength (active or passive), and degree of control (powerful or weak).

Given the neutral summary generated from the model $\hat{A}^{neu}$ and the target summary A^{neu}, our metric is calculated using the VAD lexicons by the following steps:

1. To ensure that we are measuring the relative polarity of $\hat{A}^{neu}$ with respect to the neutral target A^{neu} and to produce a calibration effect, we first remove all tokens that appear in the neutral target A^{neu} from $\hat{A}^{neu}$ tokens and thus obtain a set of unique tokens for $\hat{A}^{neu}$.

2. Next, we select tokens with a valence score either greater than 0.65 (they are considered positive valence) or less than 0.35 (they are considered negative valence) to exclude neutral words such as stopwords and non-emotion-provoking words from the metric calculation. This step eliminates tokens that are unlikely to be associated with framing bias.

3. Next, we select tokens with a valence score either greater than 0.65 for positive valence or less than 0.35 for negative valence to exclude neutral words such as stopwords and non-emotion-provoking words from the metric calculation. This step eliminates tokens that are unlikely to be associated with framing bias.

4. We calculate the sum of arousal scores for the identified positive and negative tokens respectfully from Step 2, resulting in a positive arousal score ($Arousal_+$) and a negative arousal score ($Arousal_-$). We separate the positive and negative scores to allow for a more nuanced interpretation.

We also compute a combined arousal score ($Arousal_{sum}$=$Arousal_{+}$+$Arousal_{-}$) for a coarse view.

5. We repeat this process for all A^{neu}, $\hat{A}^{neu}$ pairs in the test set and calculate the average scores to obtain the final metric. We report these scores in the experimental results section (§5).

4.2 Salient Information

To ensure that the generated summary retains important information while minimizing framing bias, we also utilize ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) metrics to compare the generated neutral summary, $\hat{A}^{neu}$, with the human-written summary(target), A^{neu}. ROUGE measures recall by evaluating how frequently the n-grams in the human reference text appear in the machine-generated text, while BLEU measures precision by assessing how often the n-grams in the machine-generated text appear in the human reference text. A higher score for BLEU and ROUGE1-R indicates better convergence of essential information. In our results, we report only Rouge-1 for simplicity.

5 Results

We conducted an evaluation of summarization generated by three models, namely BART-large, T5-base, and T5-large, on 1-1(1 input 1 output) and 3-1 (3 input 1 output) tasks. The results of our evaluation are presented in Table 3, Table 4, and Table 5. However, due to computational limitations, we were unable to complete the computation for the T5-large model in the 1-1 cases. Nevertheless, the results we obtained provide valuable insights.

Our findings indicate that the 1-1 task outperform the 3-1 task in terms of media bias mitigation. Specifically, the $Arousal_{sum}$ metric decreases from 3.12 in the 3-1 BART-large model to 2.62 in the 1-1 BART-large model, and from 2.62 in the 3-1 T5-base model to 2.39 in the 1-1 T5-base model. This outcome is consistent with our expectations, as the 3-1 task receive more information as input, which may include non-neutral words that could be present in the output. In contrast, the 1-1 task receive less information as input, resulting in fewer non-neutral words in the output.

Moreover, we observed that the T5-large and T5-base model outperforms the BART-large model, as evidenced by a decrease in the $Arousal_{sum}$ metric from 3.12 to around 2.6 in the 3-1 task and from 2.62 to 2.39 in the 1-1 task. Our discussion section will provide further insights into why the T5-base/T5-large model outperforms the BART-large model.

Our model scored lower in Salient Information Metrics, as expected. We figured the main reason was the dataset. As the dataset is a neutral summarization of three articles written by journalists, it intentionally omitted some sentimentally strong content in pursuit of neutrality. Moreover, as we changed the 3-1 task to 1-1 task, the 1-1 task received less information, and some of the information in referenced summary will not appear in the auto-summarization generated by the 1-1 task. For further improvement, if we could find a dataset that pairs a neutral summarization with exactly one article, say a website that asks its journalists to write one neutral summarization for each news article, we expect the model's performance in Salient Information Metrics to further improve.

6 Discussion

The conspicuous success is that T5-base/T5-large model outstrips the BART-large model on the framing bias metric which has 220 million/770 million and 400 million parameters respectively. We attempt to analyze the potential reasons that could explain why T5-base model's superior performance.

1. Model Architecture:
   T5 is based on the Transformer architecture, similar to BART. However, it adopts a unified text-to-text approach, meaning that both input and output are processed as sequences of tokens. This allows T5 to generalize across various tasks more effectively, potentially leading to better bias-neutral summarization performance.

2. Pretraining Objective:
   T5 is pretrained using a denoising autoencoder objective, which involves reconstructing the original text from a corrupted version. By learning to fill in masked or reordered tokens, T5 gains a more thorough understanding of the context, which might contribute to its better performance in generating bias-neutral summarizations.
   BART is also pre-trained using a denoising autoencoder objective but differs in the specific method of corruption. It masks tokens by replacing them with a special token and

Models	Avg. Framing Bias Metric			Salient Info	
	Arousal$_+$ ↓	Arousal$_-$ ↓	Arousal$_{sum}$ ↓	BLEU↑	ROUGE1-R↑
All Source input	3.14	1.64	5.78	9.32	**51.08%**
BART-large	2.01	1.11	3.12	12.8	35.40%
T5-base	1.77	**0.85**	2.62	9.71	32.79%
T5-large	**1.66**	0.94	**2.6**	**13.7**	37.38%

Table 3: 3-1 task: Experimental results for ALLSIDES test set. We provide the level of framing bias inherent in "source input" from the ALLSIDES test set to serve as a reference point for the framing bias metric. For framing bias metric, the lower number is the better (↓). For salient info score, the higher number is the better (↑). Bold numbers are the highest score in terms of its section.

Models	Avg. Framing Bias Metric			Salient Info	
	Arousal$_+$ ↓	Arousal$_-$ ↓	Arousal$_{sum}$ ↓	BLEU↑	ROUGE1-R↑
All Source input	2.64	1.34	3.98	**10.88**	**32.32%**
BART-large	1.71	0.91	2.62	9.55	30.40%
T5-base	**1.56**	**0.83**	**2.39**	7.66	28.64%
T5-large	-	-	-	-	-

Table 4: 1-1 task: Experimental results for ALLSIDES test set. We provide the level of framing bias inherent in "source input" from the ALLSIDES test set to serve as a reference point for the framing bias metric. For framing bias metric, the lower number is the better (↓). For salient info score, the higher number is the better (↑). Bold numbers are the highest score in terms of their section. Note: due to computational limitations, we were unable to complete the computation for the T5-large model in the 1-1 task.

Models	Avg. Framing Bias Metric			Salient Info	
	Arousal$_+$ ↓	Arousal$_-$ ↓	Arousal$_{sum}$ ↓	BLEU↑	ROUGE1-R↑
T5-base(3-1)	1.77	0.85	2.62	**9.71**	**32.79%**
T5-base(1-1)	**1.56**	**0.83**	**2.39**	7.66	28.64%

Table 5: 3-1 & 1-1 task: Experimental results for ALLSIDES test set. For framing bias metric, the lower number is the better (↓). For salient info score, the higher number is the better (↑). Bold numbers are the highest score in terms of its section.

then learns to reconstruct the original text. This different corruption strategy might lead to variations in performance when it comes to bias-neutral summarization.

3. Fine-tuning Procedure:
   T5 follows a multi-task learning framework during fine-tuning, which allows it to adapt to various tasks simultaneously. This approach enables the model to leverage shared knowledge across tasks, which can improve its ability to generate bias-neutral summarizations.
   BART, in contrast, is typically fine-tuned on a single task. While it can still achieve strong performance, this difference in the fine-tuning procedure might make T5 more effective in generating neutral summaries.

4. Dataset and Training Style:
   The choice of dataset and training style can significantly impact a model's performance. T5 is pre-trained on a large-scale dataset called C4, which is specifically designed to minimize potential biases. This focus on creating a clean dataset might contribute to T5's better performance in generating bias-neutral summaries.
   BART is pre-trained on a large-scale dataset as well, but the specific dataset used could introduce differences in performance. The quality and diversity of the data used during pretraining and fine-tuning are crucial factors that influence a model's ability to generate neutral summarizations.

T5's ability to generate neutral summarizations from either biased titles and articles or one single biased title and article can be attributed to its unified text-to-text approach, pretraining objective, fine-tuning procedure, and dataset choices. By leveraging these factors, T5 can learn to focus on the most objective aspects of the input text and

generate summaries that are less influenced by the biases present in the original article. While BART-large's generation outgoes T5-base on BLEU and ROUGE1-R wherewith its almost double parameters of T5-base model, T5-large model exhibits a higher and more promising figure with its 770 million parameters. Therefore, it is reasonable that the result of T5 model will be much better once we increase model parameters.

7 Conclusion

We improved the task of Neutral Multi-News Summarization (NEUS), proposed by Lee et al. (2022), to mitigate media framing bias by providing a neutral summary of articles. We proposed a 1-1 T5 model, instead of the original BART-large model, to generate neutral summarization and mitigate media bias. The model takes one input (which could be positioned left, right, or center) and output a natural generalization of the article.

We use the dataset ALLSIDES to train and a set of metrics to evaluate our summarization performance. Throughout the work, we share insights to understand the challenges and future directions in the task. We studied the related work on media bias and come across the idea of transfer-learning models. We change the model from BART-large to T5-base and T5-large, and proposed a one input one output model structure. We examine our models using metrics and get satisfying results. We hope that our work can encourage researchers to proactively address political framing bias in texts that are generated by both humans and machines.

The code of our project is open source and can be found here.

References

Allsides.com. 2022. Balanced news via media bias ratings for an unbiased news perspective.

Claes de Vreese. 2009. The effects of strategic news on political cynicism, issue evaluations, and policy support: A two-wave experiment. *Mass Communication and Society*, 7.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Matthew Aaron Gentzkow and Jesse M. Shapiro. 2005. Media bias and reputation. *SSRN Electronic Journal*.

Erving Goffman. 1974. *Frame analysis: An essay on the organization of experience*. Harvard University Press.

Felix Hamborg. 2020. Media bias, the social sciences, and NLP: Automating frame analyses to identify bias by word choice and labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 79–87, Online. Association for Computational Linguistics.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium. Association for Computational Linguistics.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. NeuS: Neutral multi-news summarization for mitigating framing bias. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3131–3148, Seattle, United States. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Dietram A. Scheufele. 2000. Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. *Mass Communication and Society*, 3(2-3):297–316.

Dietram A. Scheufele and David Tewksbury. 2007. Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of Communication*, 57(1):9–20.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Yifan Zhang, Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, Kareem Darwish, James Glass, and Preslav Nakov. 2019. Tanbih: Get to know what you are reading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 223–228, Hong Kong, China. Association for Computational Linguistics.