NeuS: Neutral Multi-News Summarization for Mitigating Framing Bias

Nayeon Lee **Yejin Bang** **Tiezheng Yu** **Andrea Madotto*** **Pascale Fung**
Hong Kong University of Science and Technology
nyleeaa@connect.ust.hk, pascale@ece.ust.hk

Abstract

Media framing bias can lead to increased political polarization, and thus, the need for automatic mitigation methods is growing. We propose a new task, a *neutral* summary generation from multiple news headlines of the varying political leanings to facilitate balanced and unbiased news reading. In this paper, we first collect a new dataset, obtain insights about framing bias through a case study, and propose a new effective metric and model (NEUS-TITLE) for the task. Then, we conduct experimental analyses to provide insights about remaining challenges and future directions. One of the most interesting observations is that generation models can hallucinate not only factually inaccurate or unverifiable content, but also politically biased content.

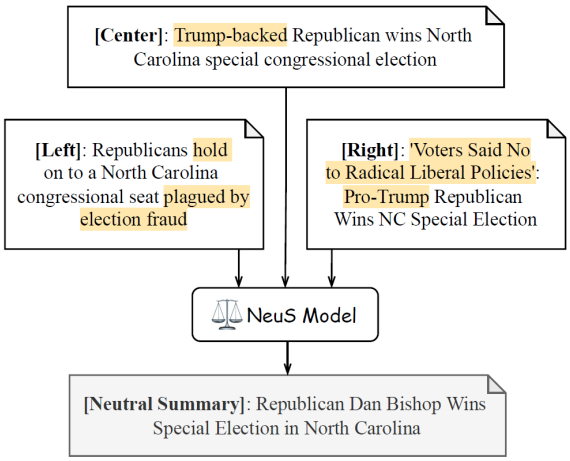

Figure 1: Illustration of the proposed task. We want to generate neutral summarization of news headlines from varying of political orientations. Orange highlights indicate phrases that can be considered framing bias.

1 Introduction

Media framing bias occurs when journalists make skewed decisions regarding which events or information to cover (informational bias) and how to cover them (lexical bias) (Entman, 2002; Groeling, 2013). Even if the reporting of the news is based on the same set of underlying issues or facts, the framing of that issue can convey a radically different impression of what actually happened (Gentzkow and Shapiro, 2006). Since the news media plays a crucial role in shaping public opinion toward various important issues (De Vreese, 2004; McCombs and Reynolds, 2009; Perse and Lambe, 2016), bias in media can reinforce the problem of political polarization.

Allsides.com (Sides, 2018) mitigates this problem by displaying articles from various media in a single interface along with an expert-written roundup of news headlines. This roundup is a neutral summary for readers to grasp a bias-free understanding of an issue before reading individual articles. Although Allsides fights framing bias, scalability still remains a bottleneck due to the time-consuming human labor needed for composing the roundup. Multi-document summarization (MDS) models (Lebanoff et al., 2018; Liu and Lapata, 2019) could be one possible choice for automating the roundup generation as both multi-document summaries and roundups share a similar nature in extracting salient information out of multiple input articles. Yet the ability of MDS models to provide *neutral* understanding of an issue – a crucial aspect of the roundup – remains unexplored.

In this work, we propose to fill in this research gap by proposing a task of Neutral multi-news Summarization (NEUS), which aims to generate a framing-bias-free summary from news headlines with varying degrees and orientation of political bias (Fig. 1). To begin with, we construct a new dataset by crawling Allsides.com, and investigate how framing bias manifests in the news so as to provide a more profound and more comprehensive understanding of the problem. The first important insight is a close association between framing bias

* This work was done when the author was studying at The Hong Kong University of Science and Technology.

and the polarity of the text. Grounded on this basis, we propose a polarity-based framing-bias metric that is simple yet effective in terms of alignment with human perceptions. The second insight is that titles serve as a good indicator of framing bias. Thus, we propose NEUS models that leverage news titles as an additional signal to increase awareness of framing bias.

Our experimental results provide rich ideas for understanding the problem of mitigating framing bias. Primarily, we explore whether existing summarization models can already solve the problem and empirically demonstrate their shortcomings in addressing the stylistic aspect of framing bias. After that, we investigate and discover an interesting relationship between framing bias and hallucination, an important safety-related problem in generation tasks. We empirically show that the hallucinatory generation has the risk of being not only factually inaccurate and/or unverifiable but also politically biased and controversial. To the best of our knowledge, this aspect of hallucination has not been previously discussed. We thus hope to encourage more attention toward hallucinatory framing bias to prevent generations from fueling political bias and polarization.

We conclude by discussing the remaining challenges to provide insights for future work. We hope our work with the proposed NEUS task serves as a good starting point to promote the automatic mitigation of media framing bias.

2 Related Works

Media Bias Media bias has been studied extensively in various fields such as social science, economics, and political science. Media bias is known to affect readers' perceptions of news in three main ways: priming, agenda-setting, and framing[1](Scheufele, 2000). Framing is a broad term that refers to any factors or techniques that affect how individuals perceive certain reality or information (Goffman, 1974; Entman, 1993, 2007; Gentzkow and Shapiro, 2006). In the context of news reports, framing is about how an issue is characterized by journalists and how readers take the information to form their impression (Scheufele and Tewksbury, 2007). This work specifically focuses on framing "bias" that exists as a form of text in the news. More specifically, we focus on different writing factors such as word choices and the commission of extra information that sway individuals' perception of certain events.

Media Bias Detection In natural language processing (NLP), computational approaches for detecting media bias often consider linguistic cues that induce bias in political text (Recasens et al., 2013; Yano et al., 2010; Lee et al., 2019; Morstatter et al., 2018; Lee et al., 2019; Hamborg et al., 2019b; Lee et al., 2021b; Bang et al., 2021). For instance, Gentzkow and Shapiro count the frequency of slanted words within articles. These methods mainly focus on the stylistic ("how to cover") aspect of framing bias. However, relatively fewer efforts have been made toward the informational ("what to cover") aspect of framing bias (Park et al., 2011; Fan et al., 2019). Majority of literature doing informational detection are focused on more general factual domain (non-political information) under the name of "fact-checking" (Thorne et al., 2018; Lee et al., 2018, 2021a, 2020). However, these methods cannot be directly applied to media bias detection because there does not exist reliable source of gold-truth to fact-check biased text upon.

Media Bias Mitigation News aggregation, by displaying articles from different news outlets on a particular topic (e.g., Google News,[2] Yahoo News[3]), is the most common approach in NLP to mitigate media bias, but it still has limitations (Hamborg et al., 2019a). Other approaches have been proposed to provide additional information (Laban and Hearst, 2017), such as automatic classification of multiple viewpoints (Park et al., 2009), multinational perspectives (Hamborg et al., 2017), and detailed media profiles (Zhang et al., 2019b). However, these methods focus on providing a broader perspective to readers from an enlarged selection of articles, which still burdens the readers. Instead, we propose to automatically neutralize and summarize partisan headlines to produce a neutral headline summary.

Multi-document Summarization As a challenging subtask of automatic text summarization, multi-document summarization (MDS) aims to condense a set of documents to a short and informative summary (Lebanoff et al., 2018). Recently, researchers have applied deep neural models for the

[1]Priming happens when news tell readers in which context the event to be evaluated; Agenda-setting is news tell readers what to think about as most important problems

[2]https://news.google.com/

[3]https://news.yahoo.com/

MDS task thanks to the introduction of large-scale datasets (Liu et al., 2018; Fabbri et al., 2019). With the advent of large pre-trained language models (Lewis et al., 2019; Raffel et al., 2019), researchers have also applied them to improve the MDS models, performance (Jin et al., 2020; Pasunuru et al., 2021). In addition, many works have studied particular subtopics of the MDS task, such as agreement-oriented MDS (Pang et al., 2021), topic-guided MDS (Cui and Hu, 2021) and MDS of medical studies (DeYoung et al., 2021). However, few works have explored generating framing-bias-free summaries from multiple news articles. To add to this direction, we propose the NEUS task and creates a new benchmark.

3 Task and Dataset

3.1 Task Formulation

The main objective of NEUS is to generate a neutral headline summary H_{neu} given multiple news headlines $H_{0...N}$ with varying degrees and orientations of political bias. The neutral summary H_{neu} should (i) retain salient information and (ii) minimize as much framing bias as possible from the input headlines.

3.2 ALLSIDES Dataset

Allsides.com provides access to triplets of news, which comprise reports from left, right, and center American publishers on the same event, with an expert-written neutral summary of the headlines and its neutral title. The dataset language is English and mainly focuses on U.S. political topics that often result in media bias. The top-3 most frequent topics[4] are 'Elections', 'White House', and 'Politics'.

We crawl the headline triplets[5] to serve as the source inputs $\{H_L, H_R, H_C\}$, and the neutral headline summary to be the target output H_{neu} for our task. Note that "center" does not necessarily mean completely bias-free (all, 2021) as illustrated in Table 1. Although "center" media outlets are relatively less tied to a particular political ideology, their reports may still contain framing bias because editorial judgement naturally leads to human-induced biases. In addition, we also crawl the title triplets $\{T_L, T_R, T_C\}$ and the neutral issue title T_{neu} that are later used in our modeling.

[4]The full list is provided in the appendix.

[5]In some cases, Allsides does not provide all three reportings, and such cases are filtered out.

To make the dataset richer, we also crawled other meta-information such as date, topic tags, and media name. In total, we crawled $3,564$ triplets ($10,692$ headlines). We use 2/3 of the triplets, which is $2,276$, to be our training and validation set ($80:20$ ratio), and the remaining $1,188$ triples as our test set. We will publicly release this dataset for future research use.

4 Analysis of Framing Bias

The literature on media framing bias from the NLP community and social science studies provide the definition and types of framing bias (Goffman, 1974; Entman, 1993; Gentzkow et al., 2015; Fan et al., 2019) — *Informational framing bias* is the biased selection of information (tangential or speculative information) to sway the minds of readers. *Lexical framing bias* is a sensational writing style or linguistic attributes that may mislead readers. However, the definition is not enough to understand exactly how framing bias manifests in real examples such as, in our case, the ALLSIDES dataset. We conduct a case-study to obtain concrete insights to guide our design choices for defining the metrics and methodology.

4.1 Case-Study Observations

First, we identify and share the examples of framing bias in accordance with the literature (Table 1).

Informational Bias This bias exists dominantly in the form of "extra information" on top of the salient key information about an issue that changes the overall impression of it. For example, in Table 1, when reporting about the hold put on military aid to Ukraine (Issue A), the right-leaning media reports the speculative claim that there were "corruption concerns" and tangential information "decries media 'frenzy'" that amplifies the negative impression of the issue. Sometimes, media with different political leanings report additional information to convey a completely different focus on the issue. For Issue C, left-leaning media implies that Trump's statement about fake news has led to "CNN receiving another suspected bomb", whereas right-leaning media implies that the media is at fault by producing "biased reports".

Lexical Bias This bias exists mainly as biased word choices that change the nuance of the information that is being delivered. For example, in Issue B, we can clearly observe that two media

Issue A: Trump Put Hold On Military Aid To Ukraine Days Before Call To Ukrainian President
Left: Trump ordered hold on military aid days before calling Ukrainian president, officials say
Right: Trump administration claims Ukraine aid was stalled over corruption concerns, decries media 'frenzy'
Center: Trump Put Hold on Military Aid Ahead of Phone Call With Ukraine's President
Issue B: Michael Reinoehl appeared to target right-wing demonstrator before fatal shooting in Portland, police say
Left: Suspect in killing of right-wing protester fatally shot during arrest
Right: Portland's Antifa-supporting gunman appeared to target victim, police say
Center: Suspect in Patriot Prayer Shooting Killed by Police
Issue C: Trump Says the 'Fake News Media' Are 'the true Enemy of the People'
Left: President Trump renews attacks on press as 'true enemy of the people' even as CNN receives another suspected bomb
Right: 'Great Anger' in America caused by 'fake news' — Trump rips media for biased reports'
Center: Trump blames 'fake news' for country's anger : 'the true enemy of the people'

Table 1: Illustration of differences in framing from Left/Right/Center media with examples from ALLSIDES dataset. We use titles for the analysis of bias, since they are simpler to compare and are representative of the framing bias that exists in the headline.

outlets change the framing of the issue by using different terms "suspect" and "gunman" to refer to the shooter, and "protester" and "victim" to refer to the person shot. Also, in Issue A, when one media outlet uses "(ordered) hold", another media uses "stalled", which has a more negative connotation.

4.2 Main Insights from Case-Study

Next, we share important insights from the case study observation that guide our metric and model design.

Relative Polarity Polarity is one of the commonly used attributes in identifying and analyzing framing bias (Fan et al., 2019; Recasens et al., 2013). Although informational and lexical bias is conceptually different, both are closely associated with polarity changes of concept, i.e., positive or negative, to induce strongly divergent emotional responses from the readers (Hamborg et al., 2019b). Thus, polarity can serve as a good indicator of framing bias.

However, we observe that the polarity of text must be utilized with care in the context of framing bias. *It is the relative polarity that is meaningful to indicate the framing bias, not the absolute polarity*. To elaborate, if the news issue itself is about tragic events such as "Terror Attack in Pakistan" or "Drone Strike That Killed 10 people", then the polarity of neutral reporting will also be negative.

Indicator of Framing We discover that the news title is very representative of the framing bias that exist in the associated headline and article – this makes sense because the title can be viewed as a succinct overview of the content that follows[6]. For instance, in Table 3 the source input example, the right-leaning media's title, and headline are mildly mocking of the "desperate" democrats' failed attempts to take down President Trump. In contrast, the left-leaning media's title and a headline show a completely different frame – implying that many investigations are happening and there is "possible obstruction of justice, public corruption, and other abuses of power."

[6]There can be exceptions for the non-main-stream media that use clickbait titles.

5 Metric

We use three metrics to evaluate summaries from different dimensions. For framing bias, we propose a polarity-based metric based on the careful design choices detailed in §5.1. For evaluating whether the summaries retain salient information, we adopt commonly used information recall metrics (§5.2). In addition, we use a hallucination metric to evaluate if the generations contain any unfaithful hallucinatory information because the existence of such hallucinatory generations can make the summary fake news (§5.3).

5.1 Framing Bias Metric

5.1.1 Design Consideration

Our framing bias metric is developed upon the insights we obtained from our case study in §4.

First of all, we propose to build our metric based on the fact that framing bias is closely associated with polarity. Both model-based and lexicon-based polarity detection approaches are options for our work, and we leverage the latter for the following reasons: 1) There is increasing demand for interpretability in the field of NLP (Belinkov et al., 2020; Sarker et al., 2019), and the lexicon-based approach is more interpretable (provides token-level

human interpretable annotation) compared to black-box neural models. 2) In the context of framing bias, distinguishing the subtle nuance of words between synonyms is crucial (e.g., dead vs. murdered). The lexicon-resource provides such token-level fine-grained scores and annotations, making it useful for our purpose.

Metric calibration is the second design consideration, and is motivated by our insight into the relativity of framing bias. The absolute polarity of the token itself does not necessarily indicate framing bias (i.e., the word "riot" has negative sentiment but does not always indicate bias), so it is essential to measure the relative degree of polarity. Therefore, calibration of the metric in reference to the neutral target is important. Any tokens existing in the neutral target will be ignored in bias measurement for the generated neutral summary. For instance, if "riot" exists in the neutral target, it will not be counted in bias measurement through calibration.

5.1.2 Framing Bias Metric Details

For our metric, we leverage Valence-Arousal-Dominance (VAD) (Mohammad, 2018) dataset which has a large list of lexicons annotated for valence (v), arousal (a) and dominance (d) scores. Valence, arousal, and dominance represent the direction of polarity (positive, negative), the strength of the polarity (active, passive), and the level of control (powerful, weak), respectively.

Given the neutral summary generated from the model $\hat{H}_{neu}$, our metric is calculated using the VAD lexicons in the following way:

1. Filter out all the tokens that appear in neutral target H_{neu} to obtain set of tokens *unique* to $\hat{H}_{neu}$. This ensures that we are measuring the *relative* polarity of $\hat{H}_{neu}$ in reference to the neutral target H_{neu} – results in calibration effect.
2. Select tokens with either positive valence ($v > 0.65$) or negative valence ($v < 0.35$) to eliminate neutral words (i.e., stopwords and non-emotion-provoking words) – this step excludes tokens that are unlikely to be associated with framing bias from the metric calculation.
3. Sum the arousal scores for the identified positive and negative tokens from Step 2 – positive arousal score ($Arousal_{+}$) and negative arousal score ($Arousal_{-}$). We intentionally separate the positive and negative scores for finer-grained interpretation. We also have the combined arousal score ($Arousal_{sum}$=$Arousal_{+}$+$Arousal_{-}$) for a coarse view.
4. Repeat for all $\{H_{neu}, \hat{H}_{neu}\}$ pairs in the test-set, and calculate the average scores to use as the final metric. We report these scores in our experimental results section (§7).

In essence, our metric approximates the existence of framing bias by quantifying how intensely aroused and sensational the generated summary is in reference to the target neutral reference. We publicly release our metric code for easy use by other researchers[7].

5.1.3 Human Evaluation

To ensure the quality of our metric, we evaluate the correlation between our framing bias metric and human judgement. We conduct A/B testing[8] where the annotators are given two generated headlines about an issue, one with a higher $Arousal_{sum}$ score and the other with a lower score. Then, annotators are asked to select the more biased headline summary. When asking which headline is more "biased", we adopt the question presented by Spinde et al. We also provide examples and the definition of framing bias for a better understanding of the task. We obtain three annotations each for 50 samples and select those with the majority of votes.

A critical challenge of this evaluation is in controlling the potential involvement of the annotators' personal political bias. Although it is hard to eliminate such bias completely, we attempt to avoid it by collecting annotations from those indifferent to the issues in the test set. Specifically, given that our test set mainly covers US politics, we restrict the nationality of annotators to non-US nationals who view themselves bias-free towards any US political parties.

After obtaining the human annotations from A/B testing, we also obtain automatic annotation based on the proposed framing bias metric score, where the headline with a higher $Arousal_{sum}$ is chosen to be the more biased generation. The Spearman correlation coefficient between human-based and metric-based annotations is 0.63615 with a p-value < 0.001, and the agreement percentage 80%. These

[7] https://github.com/HLTCHKUST/framing-bias-metric

[8] Please refer the appendix for more detail of the A/B testing

values indicate that the association between the two annotations is statistically significant, suggesting that our metric provides a good approximation of the existence of framing bias.

5.2 Salient Info

The generation needs to retain essential/important information while reducing the framing bias. Thus, we also report ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) between the generated neutral summary, $\hat{H}_{neu}$, and human-written summary, H_{neu}. Note that ROUGE measures the recall (i.e., how often the n-grams in the human reference text appear in the machine-generated text) and BLEU measures the precision (i.e., how often the n-grams in the machine-generated text appear in the human reference text). The higher the BLEU and ROUGE1-R score, the better the essential information converges. In our results, we only report Rouge-1, but Rouge-2 and Rouge-L can be found in the appendix.

5.3 Hallucination Metric

Recent studies have shown that neural sequence models can suffer from the hallucination of additional content not supported by the input (Reiter, 2018; Wiseman et al., 2017; Nie et al., 2019; Maynez et al., 2020; Pagnoni et al., 2021; Ji et al., 2022), consequently adding factual inaccuracy to the generation of NLG models. Although not directly related to the goal of NEUS, we evaluate the hallucination level of the generations in our work. We choose a hallucination metric called FeQA (Durmus et al., 2020) because it is one of the publicly available metrics known to have a high correlation with human faithfulness scores. This is a question-answering-based metric built on the assumption that the same answers will be derived from hallucination-free generation and the source document when asked the same questions.

6 Models and Experiments[9]

6.1 Baseline Models

Since one common form of framing bias is the reporting of extra information (§4), summarization models, which extract commonly shared salient information, may already generate a neutral summaries to some extent. To test this, we conduct experiments using the following baselines.

[9]Experimental details are provided in the appendix for reproducibility.

- LEXRANK (Erkan and Radev, 2004): an extractive single-document summarization (SDS) model that extracts representative sentences that hold information common in both left- and right-leaning articles.
- BARTCNN: an abstractive SDS model that fine-tunes BART-large (Lewis et al., 2019), with 406M parameters, using the CNN/DailyMail (Hermann et al., 2015) dataset.
- BARTMULTI: a multi-document summarization (MDS) model that fine-tunes BART-large using Multi-News (Fabbri et al., 2019) dataset.
- PEGASUSMULTI: an MDS model that fine-tunes Pegasus-base (Zhang et al., 2019a), with 568M parameters, using the Multi-News dataset.

Since the summarization models are not trained with in-domain data, we provide another baseline model trained with in-domain data for a full picture.

- NEUSFT: a baseline that fine-tunes the BART-large model using ALLSIDES.

6.2 Our NEUS Models (NEUS-TITLE)

We design our models based on the second insight from the case study (§4) - the news title serves as an indicator of the framing bias in the corresponding headline. We hypothesize that it would be helpful to divide-and-conquer by neutralizing the the title first, then leveraging the "neutralized title" to guide the final neutral summary of the longer headlines. Multi-task learning (MTL) is a natural modeling choice because two sub-tasks are involved – title-level and headline-level neutral summarization. However, we also have to ensure a sequential relationship between the two tasks in our MTL training because headline-level neutral summarization leverages the generated neutral title as an additional resource.

We propose a simple yet elegant technique to preserve the sequential order in our MTL by adapting the idea of prompting, a method of reformatting NLP tasks in the format of a natural language response into a natural language input (Sanh et al., 2021). We train BART's autoregressive decoder to generate the target text Y formatted as follows:

$$\texttt{TITLE} \Rightarrow T_{neu}.\ \texttt{HEADLINE} \Rightarrow H_{neu},$$

where T_{neu} and H_{neu} denote the neutral title and neutral headline summary.

Models	Avg. Framing Bias Metric			Salient Info		Hallucination
	$Arousal_{+}\downarrow$	$Arousal_{-}\downarrow$	$Arousal_{sum}\downarrow$	BLEU↑	ROUGE1-R↑	FeQA↑
All Source input	6.76	3.64	10.40	8.27	56.57%	-
LEXRANK	3.02	1.74	4.76	**12.21**	39.08%	53.44%
BARTCNN	2.09	1.23	3.32	10.49	35.63%	58.03%
PEGASUSMULTI	5.12	2.39	7.51	6.12	**44.42%**	22.24%
BARTMULTI	5.94	2.66	8.61	4.24	35.76%	21.06%
NEUSFT	1.86	1.00	2.85	11.67	35.11%	**58.50%**
NEUS-TITLE	**1.69**	**0.83**	**2.53**	12.05	36.07%	45.95%

Table 2: 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 other scores, the *higher* number is the better (↑).

The input X to our BART encoder is formatted similarly to the target text Y:

$$\texttt{TITLE} \Rightarrow T_L.\ \texttt{HEADLINE} \Rightarrow H_L.[SEP]$$
$$\texttt{TITLE} \Rightarrow T_C.\ \texttt{HEADLINE} \Rightarrow H_C.[SEP]$$
$$\texttt{TITLE} \Rightarrow T_R.\ \texttt{HEADLINE} \Rightarrow H_R,$$

where $T_{L/C/R}$ and $H_{L/C/R}$ denote the title and headline from left-wing, center, and right-wing media, and [SEP] denotes the special token that separates different inputs. Note that the order of left, right, and center are randomly shuffled for each sample to discourage the model from learning spurious patterns from the input.

This technique allows us to optimize for both title and headline neutral summarization tasks easily by simply optimizing for the negative log-likelihood of the single target Y. The auto-regressive nature of the decoder also ensures the sequential relationship between the title and headline.

7 Results and Analysis

In this section, we point out noteworthy observations from the quantitative results in Table 2 along with insights obtained through qualitative analysis. Table 3 shows generation examples that are most representative of the insights we share.[10]

7.1 Main Results

Firstly, summarization models can reduce the framing bias to a certain degree (drop in $Arousal_{sum}$ score from 10.40 to 4.76 and 3.32 for LEXRANK and BARTCNN). This is because informational framing bias is addressed when summarization models extract the most salient sentences, which contain common information from the inputs. However, summarization models, especially LEXRANK cannot handle the lexical framing bias, as shown in Table 3. Moreover, if we further observe the results of LEXRANK, it is one of the best performing models in terms of ROUGE1-R (39.08%), the standard metric for summarization performance, but not in terms of the framing bias metric. This suggests that *having good summarization performance (*ROUGE1-R*) does not guarantee that the model is also neutral* – i.e., the requirement for summaries to be neutral adds an extra dimension to the summarization task.

Secondly, one interesting pattern that deserves attention is that only the single-document summarization models (BARTCNN and LEXRANK) reduced framing bias well, not the multi-document summarization models (PEGASUSMULTI and BARTMULTI). This is rather surprising because our task setup is more similar to MDS than SDS. One of the major contributors to high bias in the MDS models is probably the hallucination because MDS models portray drastically poor hallucination performance than all the other models (both the MDS models PEGASUSMULTI and BARTMULTI achieve 22.24% and 21.06%, when most of the other models achieve over 50%).[11] This suggests that the framing bias of MDS models may be related to the hallucination of politically biased content. We investigate into this in the next subsection (§7.2).

Thirdly, although summarization models help reduce the framing bias scores, we, unsurprisingly, observe a more considerable bias reduction when training with in-domain data. NEUSFT shows a further drop across all framing bias metrics without sacrificing the ability to keep salient information. However, we observe that NEUSFT often copies

[10]More examples are provided in the appendix.

[11]Note that 22.24% and 21.06% are already high FeQA scores, however, a comparatively low score in reference.

SOURCE: <Left> Title: Here Are The 81 People And Entities Close To Trump Democrats Are Investigating. **Headline:** Democrats on the House Judiciary Committee on Monday sent document requests to 81 agencies, entities and individuals close to President Donald Trump as part of a broad investigation into possible obstruction of justice, public corruption and other abuses of power. The list includes Trump's sons, Eric Trump and Donald Trump Jr., as well as his son-in-law, Jared Kushner. **<Center> Title:** House Panel Requests Documents From Associates of Trump. **Headline:** House Democrats intensified their investigations into President Trump and his associates Monday, demanding records from more than 80 people and organizations related to his business dealings, interactions with the Justice Department and communications with Russian President Vladimir Putin. **<Right> Title:** Dems Continue Their Assault on The Trump Administration By Launching Another Probe. **Headline:** Democrats are desperate to take down President Donald Trump. The Russia probe has proven to be ineffective and, quite frankly, a waste of time and taxpayer money. They didn't find what they wanted so now they're launching another probe.
TARGET: House Democrats launched a broad probe into President Trump on Monday, requesting documents from 81 agencies and individuals as they investigate his business dealings, interactions with Russia, and possible obstruction of justice.
Lexrank: Democrats are desperate to take down President Donald Trump. The Russia probe has proven to be ineffective and, quite frankly, a waste of time and taxpayer money.
NEUSFT: The Russia probe has proven to be ineffective and, quite frankly, a waste of time and taxpayer money.
NEUS-TITLE: TITLE=> House Panel Requests Documents. ARTICLE=> The House Select Committee on Intelligence has requested documents from 81 people and entities close to President Trump, including his sons Eric and Donald Trump Jr., as well as Jared Kushner.

Table 3: Generation examples for analysis purposes. Red highlights the tokens identified by VAD lexicons. Refer to the appendix for more examples.

SOURCE: ... President Trump on Saturday blasted what he called the "phony" BuzzFeed story and the mainstream media's coverage of it....
MDS Hallucination: president trump on sunday slammed what he called called a "phony" story by the "dishonest" and "fake news" news outlet in a series of tweets. ... "the fake news media is working overtime to make this story look like it is true," trump tweeted. "they are trying to make it look like the president is trying to hide something, but it is not true!"

Table 4: Illustration of hallucinatory framing bias from MDS models and the corresponding "most relevant source snippet" from the source input. Refer to the appendix for more examples with full context.

directly without any neutral re-writing – e.g., the NEUSFT example shown in Table 3 is a direct copy of the sentence from the input source.

Lastly, we can achieve slightly further improvement with NEUS-TITLE across all metrics except the FeQA score. This model demonstrates a stronger tendency to paraphrase rather than directly copy, and has comparatively more neutral framing of the issue. As shown in Table 3, when LEXRANK and NEUSFT are focused on the "ineffectiveness of Russia probe", the TARGET and NEUS-TITLE focus on the start of the investigation with the request for documents. NEUS-TITLE also generate a title with a similar neutral frame to the TARGET, suggesting this title generation guided the correctly framed generation.

7.2 Further Analysis and Discussion

Q: Is hallucination contributing to the high framing bias in MDS models? Through qualitative analysis, we discovered the MDS generations were hallucinating politically controversial or sensational content that did not exist in the input sources. This is probably originating from the memorization of either the training data or the LM-pretraining corpus. For instance, in Table 4, we can observe stylistic bias being injected – "the 'dishonest' and 'fake news' news outlet". Also, the excessive elaboration of the president's comment towards the news media, which does not appear the in source or target, can be considered informational bias – "they are trying to make it look like the president is trying to hide something, but it is not true!" *This analysis unveils the overlooked danger of hallucination, which is the risk of introducing political framing bias in summary generations.* Note that

this problem is not confined to MDS models only because other baseline models also have room for improvement in terms of the FeQA hallucination score.

Q: What are the remaining challenges and future directions? The experimental results of NEUS-TITLE suggest that there is room for improvement. We qualitatively checked some error cases and discovered that the title-generation is, unsurprisingly, not always accurate, and the error propagating from the title-generation step adversely affected the overall performance. Thus, one possible future direction would be to improve the neutral title generation, which would then improve the neutral summarization.

Another challenge is the subtle lexical bias involving nuanced word choices that manoeuvre readers to understand the events from biased frames. For example, "put on hold" and "stalled" both mean the same outcome, but the latter has a more negative connotations. Improving the model's awareness of such nuanced words or devising ways to incorporate style-transfer-based bias mitigation approaches (Liu et al., 2021) could be another helpful future direction.

We started the neutral summarization task assuming that framing bias originates from the source inputs. However, our results and analysis suggest that hallucination is another contributor to framing bias. Leveraging hallucination mitigation techniques would be a valuable future direction for the NEUS task. We believe it will help to reduce informational framing bias, although it may be less effective to lexical framing biases. Moreover, our work can also be used to facilitate hallucination research as well. We believe the proposed framing bias metric will help researchers evaluate hallucinatory phenomena from different angles other than "factuality". The proposed framing bias metric could also be adapted to the hallucination problem without a "neutral" reference. The source input can substitute the "neutral" reference to measure if the generated summary is more politically biased than the source – a potential indication of political hallucination.

8 Conclusion

We introduce a new task of Neutral Multi-News Summarization (NEUS) to mitigate media framing bias by providing a neutral summary of headlines, along with the dataset ALLSIDES and a set of metrics. Throughout the work, we share insights to understand the challenges and future directions in the task. We show the relationships among polarity, extra information, and framing bias, which guides us to the metric design, while the insight that the title serves as an indicator of framing bias leads us to the model design. Our qualitative analysis reveals that hallucinatory content generated by models may also contribute to framing bias. We hope our work stimulates researchers to actively tackle political framing bias in both human-written and machine-generated texts.

Ethical Considerations

If we could automatically generate a neutralized version of a media report, it would be a meaningful solution to framing bias. However, the idea of unbiased journalism has been challenged [12] because journalists will make their own editorial judgements that can never be guaranteed to be completely bias-free. Therefore, we aim to provide NEUS which provides a comprehensive summary of headlines from different media, instead of trying to neutralize an article.

One of the considerations is the bias induced by the computational approach. Automatic approaches replace a known source bias with another bias caused by human-annotated data or the machine learning models. Understanding the risk of uncontrolled adoption of such automatic tools, careful guidance should be provided in how to adopt them. For instance, an automatically generated neutral summary should be provided with reference to the original source instead of standing alone.

We use news from English-language sources only and largely American news outlets throughout this paper. Partisanship from this data refers to domestic American politics. We note that this work does not cover media bias at the international-level or in other languages. In future work, we will explore the application of our methodology to different cultures or languages. However, we hope the paradigm of NEUS, providing multiple sides to neutralize the view of an issue, can encourage future research in mitigating framing bias in other languages or cultures.

[12] https://www.allsides.com/blog/does-unbiased-news-really-exist

References

2021. Center – what does a "center" media bias rating mean?

Yejin Bang, Nayeon Lee, Etsuko Ishii, Andrea Madotto, and Pascale Fung. 2021. Assessing political prudence of open-domain chatbots. *arXiv preprint arXiv:2106.06157*.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5.

Peng Cui and Le Hu. 2021. Topic-guided abstractive multi-document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1463–1472.

Claes De Vreese. 2004. The effects of strategic news on political cynicism, issue evaluations, and policy support: A two-wave experiment. *Mass Communication & Society*, 7(2):191–214.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms2: Multi-document summarization of medical studies. *arXiv preprint arXiv:2104.06486*.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, pages 390–397.

Robert M Entman. 2002. Framing: Towards clarification of a fractured paradigm. *McQuail's Reader in Mass Communication Theory. London, California and New Delhi: Sage*.

Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. *arXiv preprint arXiv:1909.02670*.

Matthew Gentzkow and Jesse M Shapiro. 2006. Media bias and reputation. *Journal of political Economy*, 114(2):280–316.

Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Matthew Gentzkow, Jesse M Shapiro, and Daniel F Stone. 2015. Media bias in the marketplace: Theory. In *Handbook of media economics*, volume 1, pages 623–645. Elsevier.

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

Tim Groeling. 2013. Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science*, 16:129–151.

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

Felix Hamborg, Norman Meuschke, and Bela Gipp. 2017. Matrix-based news aggregation: exploring different news perspectives. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–10. IEEE.

Felix Hamborg, Anastasia Zhukova, and Bela Gipp. 2019b. Illegal aliens or undocumented immigrants? towards the automated identification of bias by word choice and labeling. In *International Conference on Information*, pages 179–187. Springer.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6244–6254.

Philippe Laban and Marti A Hearst. 2017. newslens: building and visualizing long-ranging news stories. In *Proceedings of the Events and Stories in the News Workshop*, pages 1–9.

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.

Nayeon Lee, Yejin Bang, Andrea Madotto, Madian Khabsa, and Pascale Fung. 2021a. Towards few-shot fact-checking via perplexity. *arXiv preprint arXiv:2103.09535*.

Nayeon Lee, Belinda Z Li, Sinong Wang, Pascale Fung, Hao Ma, Wen-tau Yih, and Madian Khabsa. 2021b. On unifying misinformation detection. *arXiv preprint arXiv:2104.05243*.

Nayeon Lee, Belinda Z Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? *arXiv preprint arXiv:2006.04102*.

Nayeon Lee, Zihan Liu, and Pascale Fung. 2019. Team yeon-zi at semeval-2019 task 4: Hyperpartisan news detection by de-noising weakly-labeled data. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1052–1056.

Nayeon Lee, Chien-Sheng Wu, and Pascale Fung. 2018. Improving large-scale fact-checking using decomposable attention models and lexical tagging. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1133–1138.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating political bias in language models through reinforced calibration.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Maxwell McCombs and Amy Reynolds. 2009. How the news shapes our civic agenda. In *Media effects*, pages 17–32. Routledge.

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.

Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R Corman, and Huan Liu. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing*, 1(2):1–18.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829.

Richard Yuanzhe Pang, Adam D Lelkes, Vinh Q Tran, and Cong Yu. 2021. Agreesum: Agreement-oriented multi-document summarization. *arXiv preprint arXiv:2106.02278*.

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.

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*, pages 443–452.

Souneil Park, Kyung-Soon Lee, and Junehwa Song. 2011. Contrasting opposing views of news articles on contentious issues. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 340–349.

Ramakanth Pasunuru, Mengwen Liu, Mohit Bansal, Sujith Ravi, and Markus Dreyer. 2021. Efficiently summarizing text and graph encodings of multi-document clusters. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4768–4779.

Elizabeth M Perse and Jennifer Lambe. 2016. *Media effects and society*. Routledge.

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. *arXiv preprint arXiv:1910.10683*.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Abeed Sarker, Ari Z Klein, Janet Mee, Polina Harik, and Graciela Gonzalez-Hernandez. 2019. An interpretable natural language processing system for written medical examination assessment. *Journal of biomedical informatics*, 98:103268.

Dietram A Scheufele. 2000. Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. *Mass communication & 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.

All Sides. 2018. Media bias ratings. *Allsides.com*.

Timo Spinde, Christina Kreuter, Wolfgang Gaissmaier, Felix Hamborg, Bela Gipp, and Helge Giese. 2021. Do you think it's biased? how to ask for the perception of media bias. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 61–69. IEEE.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Yifan Zhang, Giovanni Da San Martino, Alberto Barrón-Cedeno, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, et al. 2019b. Tanbih: Get to know what you are reading. *EMNLP-IJCNLP 2019*, page 223.

Appendix

A Topics covered in ALLSIDESdataset

The ALLSIDESdataset language is English and mainly focuses on U.S. political topics that often result in media bias. The top-5 most frequent topics are 'Elections', 'White House', 'Politics', 'Coronavirus', 'Immigration'.

The full list is as follow (in a descending order of frequency): ['Elections', 'White House', 'Politics', 'Coronavirus', 'Immigration', 'Violence in America', 'Economy and Jobs', 'Supreme Court', 'Middle East', 'US House', 'Healthcare', 'World', 'US Senate', 'National Security', 'Gun Control and Gun Rights', 'Media Bias', 'Federal Budget', 'Terrorism', 'US Congress', 'Foreign Policy', 'Criminal Justice', 'Justice Department', 'Trade', 'Impeachment', 'Donald Trump', 'North Korea', 'Russia', 'Education', 'Environment', 'Free Speech', 'FBI', nan, 'Abortion', 'General News', 'Disaster', 'US Military', 'Technology', 'LGBT Rights', 'Sexual Misconduct', 'Voting Rights and Voter Fraud', 'Joe Biden', 'Race and Racism', 'Economic Policy', 'Justice', 'Holidays', 'Taxes', 'China', 'Polarization', 'Democratic Party', 'Religion and Faith', 'Sports', 'Homeland Security', 'Culture', 'Cybersecurity', 'National Defense', 'Public Health', 'Civil Rights', 'Europe', 'Great Britain', 'Banking and Finance', 'Republican Party', 'NSA', 'Business', 'State Department', 'Facts and Fact Checking', 'Media Industry', 'Labor', 'Veterans Affairs',

'Campaign Finance', 'Life During COVID-19', 'Transportation', 'Marijuana Legalization', 'Agriculture', 'Arts and Entertainment', 'Fake News', 'Campaign Rhetoric', 'Nuclear Weapons', 'Israel', 'Asia', 'CIA', 'Role of Government', 'George Floyd Protests', "Women's Issues", 'Safety and Sanity During COVID-19', 'Animal Welfare', 'Treasury', 'Science', 'Climate Change', 'Domestic Policy', 'Energy', 'Housing and Homelessness', 'Bridging Divides', 'Mexico', 'Inequality', 'COVID-19 Misinformation', 'ISIS', 'Palestine', 'Bernie Sanders', 'Tulsi Gabbard', 'Sustainability', 'Family and Marriage', 'Pete Buttigieg', 'Welfare', 'Opioid Crisis', 'Amy Klobuchar', 'Food', 'EPA', 'South Korea', 'Alaska: US Senate 2014', 'Social Security', 'US Constitution', 'Tom Steyer', 'Andrew Yang', 'Africa']

B Additional Salient Information Score Results

We report additional Salient information F1 (Table 5) and Recall (Table 6) scores for ROUGE1, ROUGE2 and ROUGEL.

	ROUGE1 F1	ROUGE2 F1	ROUGEL F1
LEXRANK	33.60%	13.60%	29.77%
BARTCNN	33.76%	13.67%	30.57%
PEGASUSMULTI	30.03%	10.28%	26.70%
BARTMULTI	23.01%	6.84%	20.55%
NEUSFT	36.76%	16.27%	32.86%
NEUS-TITLE	35.49%	15.69%	32.05%

Table 5: Additional Salient Info Scores. F1 scores for ROUGE1, ROUGE2 and ROUGEL for ALLSIDES testset. For the scores, the *higher* number is the better.

	ROUGE1 RECALL	ROUGE2 RECALL	ROUGEL RECALL
LEXRANK	39.08%	17.66%	34.69%
BARTCNN	35.63%	15.32%	32.22%
PEGASUSMULTI	44.42%	16.99%	39.45%
BARTMULTI	35.76%	12.48%	32.08%
NEUSFT	35.11%	15.74%	31.43%
NEUS-TITLE	36.07%	16.47%	32.63%

Table 6: Additional Salient Info Scores. Recall scores for ROUGE1, ROUGE2 and ROUGEL for ALLSIDES testset. For the scores, the *higher* number is the better.

C Details for Human Evaluation (A/B testing)

We first presented the participants with the definition of framing bias from our paper, and also showed examples in Table 1 to ensure they understand what framing bias is. Then we asked the following question: "Which one of the articles do you believe to be more biased toward one side or the other side in the reporting of news?" This is modified to serve as a question for AB testing based on "To what extent do you believe that the article is biased toward one side or the other side in the reporting of news?" The original question is one of the 21 questions which are suitable and reliable for measuring the perception of media bias, designed by Spinde et al. (2021).

The participants (research graudate students) have different nationalities including Canada, China, Indonesia, Iran, Italy, Japan, Poland and South Korea (ordered in an alphabetical order). All of participants answered to be not having political leaning towards U.S. politics. All participants are fully explained on the usage of collected data in this particular work and agreed on it.

D Experimental Setup Details

All our experimental codes are based on the HuggingFace (Wolf et al., 2020). We used the following hyperparameters during training and across models: 10 epoch size, $3e-5$ learning rate, and a batch size of 16. We did not do hyper-parameters tuning since our objective is to provide various baselines and analysis. Training run-time for all of our experiments are fast ($<$ 6hr). We ran all experiments with one NVIDIA 2080Ti GPU with 16 GB of memory. The experiment was a single-run.

E Generation Examples from Different Models

To help better understand performances of each models, we provide more examples of generation from all baseline models and our proposed models along with the target neutral headline summary. The examples can be found in Table 7, 8, 9.

F Illustration of hallucinatory framing bias

In Table 10, we provide more examples of hallucinatory framing bias that was discussed in analysis.

Target: *The House Ways and Means Committee has requested that the IRS turn over the last six years of tax returns from President Trump.*
LEXRANK: Representative Richard Neal invoked law that allows House Ways and Means Committee chair to obtain any person's tax information. The request by House Democrats for six years of Donald Trump's tax returns opens a serious new political, legal, constitutional and personal front in the party's battle to use its new majority to check a defiant President.
BARTCNN: Rep. Richard Neal invoked law that allows House Ways and Means Committee chair to obtain any person's tax information. The request by House Democrats for six years of Donald Trump's tax returns opens a serious new political, legal, constitutional and personal front in the party's battle to use its new majority to check a defiant President.
PEGASUSMULTI: The battle over President Trump's tax returns is officially on. Rep. Richard Neal, chairman of the House Ways and Means Committee, on Wednesday formally asked the IRS for six years of Trump's personal and business tax returns, as well as the returns of some of his companies, reports the Washington Post. The panel has the power to subpoena someone's tax information, and Neal has signaled that he plans to do just that. "The American people have a right to know who the president of the United States is, and we have a responsibility to do our part to ensure that the tax laws are administered in an efficient and thorough manner," he said in a statement, per the Hill. The White House has signaled that it will fight the request, with press secretary Sarah Huckabee Sanders calling it "an abuse of the tax-writing committee's authority." The letter to the IRS was signed by Neal, a Massachusetts Democrat, and was obtained by the Wall Street Journal. The Journal notes that Neal has the authority under a law passed in 1969 that allows the House Ways and Means Committee to obtain tax information from anyone.
BARTMULTI: the head of the powerful house tax-writing committee has asked the government for six years of president trump's tax returns, a move that is expected to lead to a long court battle with the white house, the washington post reports. in a letter to the irs on wednesday, rep. richard h. Neal, chairman of the house ways and means committee, invoked a law that allows any person to obtain any person's personal tax information, the wall street journal reports. according to the hill, the move is a " serious new political, legal, constitutional, and personal front in the party's battle to use its new majority to check a defiant president. " it is a bid to solve one of the most tantalizing and enduring mysteries of the trump campaign and his presidency — what is it in the president's unseen financial and business life that he doesn't want americans to see?
NEUSFT: House Ways and Means Committee Chair Richard Neal (D-MA) has formally requested President Trump's tax returns from the IRS.
NEUS-TITLE: TITLE=> House Committee Seeks Trump Tax Returns. ARTICLE=> The House Ways and Means Committee formally requested President Trump's tax returns on Wednesday, kicking off what could be a protracted legal fight between Congress and the Trump administration.

Table 7: Neutral Summary Generation Examples from baseline models and NEUS-TITLE.

Target: *The State Department has blocked Gordon Sondland, the U.S. Ambassador to the European Union, from testifying at a scheduled deposition on the Ukraine controversy. Sondland is a key figure as he had texted with Bill Taylor, a top U.S. diplomat in Ukraine, arguing over allegations of a quid pro quo. This move by the State Department has triggered a new clash with House Democrats as part of their impeachment inquiry, and President Trump has dismissed the deposition as a "a kangaroo court" as he defended the State Department decision.*
LEXRANK: The Trump administration on Tuesday blocked a planned deposition from Gordon Sondland, the U.S. ambassador to the European Union and a central figure in the Democrats' impeachment inquiry, according to a statement by his lawyer. Democrats swiftly called the administration's move an act of obstruction and accused President Trump of hiding information about his efforts to pressure the Ukrainian government to investigate his political rivals.
BARTCNN: The State Department instructed Ambassador to the European Union Gordon Sondland not to appear for the deposition, according to his counsel. Democrats swiftly called the administration's move an act of obstruction and accused President Trump of hiding information about his efforts to pressure the Ukrainian government to investigate his political rivals.
PEGASUSMULTI: The Trump administration has blocked US ambassador to the European Union Gordon Sondland from testifying before a House panel investigating possible obstruction of justice by the president, the New York Times reports. Sondland is a key witness in the probe into whether Trump obstructed justice when he fired FBI chief James Comey. Sondland had been scheduled to testify in front of the House Judiciary Committee on Tuesday, but the State Department ordered him not to appear, according to the Washington Post. Sondland is a former British ambassador to the US. In a Tuesday tweet, President Trump said he would love to send Sondland to testify, "but unfortunately he would be testifying before a totally compromised kangaroo court, where Republican's rights have been taken away, and true facts are not allowed out for the public to see." He added: "I believe you are incorrect about President Trump's intentions."
BARTMULTI: a key witness in the house impeachment inquiry was ordered not to appear at a scheduled deposition on tuesday, ramping up tensions between congressional democrats and the trump administration. the state department instructed ambassador to the eu and key figure in the impeachment inquiry, Gordon sondland, not to show up for the deposition, according to his counsel. democrats called the move an act of obstruction and accused president trump of hiding information about his efforts to pressure the ukraine government to investigate his political rivals, the new york times reports. " this is a clear attempt to block testimony from key witnesses in the ongoing impeachment investigation, " said house minority leader nancy pelosi, per the washington post. " it's an impeachable act. " in a tweet tuesday night, trump said he " would love to send ambassador [sundland], a really good man and great american, to testify, but unfortunately he would be testifying before a totally compromised kangaroo court, where republican ' s rights have been taken away, and true facts are not allowed out for the public to see. "
NEUSFT: The Trump administration blocked a planned deposition from Gordon Sondland, the U.S. ambassador to the European Union and a central figure in the Democrats' impeachment inquiry, according to a statement from his lawyer. President Trump responded via Twitter, saying that he "would love to send Ambassador Sonderland, a really good man and great American, to testify, but unfortunately he would be testifying before a totally compromised kangaroo court, where Republican's rights have been taken away, and true facts are not allowed out for the public to see."
NEUS-TITLE: TITLE=> State Dept Orders Ambassador Sondland Not to Testify Before House. ARTICLE=> The State Department has ordered Ambassador Gordon Stondland not to testify before the House of Representatives in the impeachment inquiry of President Donald Trump. Democrats have accused the Trump administration of blocking statements from witnesses.

Table 8: Continued from Previous Page: Neutral Summary Generation Examples from baseline models and NEUS-TITLE.

Target: *Ukrainian police have opened an investigation into whether or not U.S. Ambassador Marie Yovanovitch came under surveillance before she was recalled from her post in Ukraine last May. Democrats have released documents that show Lev Parnas, an associate of Rudy Giuliani, communicating about Yovanovitch's removal.*
LEXRANK: Ukraine's government announced Thursday that police are investigating whether ousted U.S. ambassador Marie Yovanovitch was subject to illegal surveillance, in response to new documents released ahead of President Trump's impeachment trial. Those documents, released by Democratic lawmakers, showed Lev Parnas – an associate of Trump lawyer Rudy Giuliani – communicating about the removal of Marie Yovanovitch as the ambassador to Ukraine.
BARTCNN: Police in Ukraine have opened a criminal investigation into whether allies of President Trump had the U.S. ambassador to the country under surveillance. Newly released documents suggest that Ms. Yovanovitch was being watched in Kyiv. The Ministry of Internal Affairs of Ukraine says it's asking the FBI to help its inquiry.
PEGASUSMULTI: The Ukraine has opened a criminal investigation into allegations that President Trump's allies had the US ambassador to the country under surveillance while she was stationed in the country, the New York Times reports. According to the Wall Street Journal, the allegations surfaced in newly released documents released by Democrats ahead of Trump's impeachment trial. The documents show that Rudy Giuliani's associate, Lev Parnas, communicated about the removal of Marie Yovanovitch as ambassador to Ukraine with Yovanovitch's former boss at the State Department, Sergey Kislyak. Yovanovitch was removed from her post in May of last year. Ukraine's Ministry of Internal Affairs says it's asking the FBI to help with the investigation. "Ukraine cannot ignore such illegal activities on the territory of its own state," the ministry says in a statement
BARTMULTI: new documents released ahead of president trump's impeachment trial suggest that the us ambassador to the country was under surveillance while she was stationed in the country, the new york times reports. according to the wall street journal, the allegations were made in a text message between a us politician and an associate of trump lawyer rudy giuliani. in the messages, the politician says he'd like to see former ambassador to ukraine and current ambassador to russia, former deputy foreign minister oleksandr turchynov, removed from his post. " i would like to remove him from his position, " the politician wrote in the message, which was sent in april 2017. the messages were sent in response to newly released documents from the house intelligence committee. those documents show that former trump lawyer paul ryan had been trying to get rid of yovanovitch, who was ambassador at the time, for more than a year, the journal reports. the messages also show that ryan's chief of staff, george w. bush, had been in touch with ryan about removing her from her post.
NEUSFT: Ukraine's national police are investigating allegations that U.S. Ambassador Marie Yovanovitch was subjected to illegal surveillance while she was stationed in Kyiv.
NEUS-TITLE: TITLE=> Ukraine Opens Investigation Into Claims U.S. Ambassador Yovanovitch Was Under Surveillance. ARTICLE=> Ukraine's national police are investigating allegations that former U. S. ambassador Marie Yovanovich was subjected to surveillance while she was stationed in Kyiv, following the release of documents by Democratic lawmakers.

Table 9: Continued from Previous Page: Neutral Summary Generation Examples from baseline models and NEUS-TITLE.

[1] Source: Newly released documents suggest that Ms. Yovanovitch was being watched in Kyiv while she was the U.S. ambassador there. The police in Ukraine have opened a criminal investigation into whether allies of President Trump had the United States ambassador to the country under surveillance while she was stationed in Kyiv, the Ukrainian government said on Thursday. [SEP] Ukraine's government announced Thursday that police are investigating whether ousted U.S. ambassador Marie Yovanovitch was subject to illegal surveillance, in response to new documents released ahead of President Trump's impeachment trial. Those documents, released by Democratic lawmakers, showed Lev Parnas – an associate of Trump lawyer Rudy Giuliani – communicating about the removal of Marie Yovanovitch as the ambassador to Ukraine. [SEP] Ukraine's national police are investigating allegations that U.S. Ambassador Marie Yovanovitch was under surveillance in Kyiv last spring — an allegation made in WhatsApp messages between a U.S. politician and Lev Parnas. "Ukraine cannot ignore such illegal activities on the territory of its own state," the Ministry of Internal Affairs of Ukraine said in a statement Thursday. The ministry says it's asking the FBI to help its inquiry.
[1] MDS Hallucination: new documents released ahead of president trump's impeachment trial suggest that the us ambassador to the country was under surveillance while she was stationed in the country, the new york times reports. according to the wall street journal, the allegations were made in a text message between a us politician and an associate of trump lawyer rudy giuliani. in the messages, the politician says he'd like to see former ambassador to ukraine and current ambassador to russia, former deputy foreign minister oleksandr turchynov, removed from his post. "i would like to remove him from his position," the politician wrote in the message, which was sent in april 2017. the messages were sent in response to newly released documents from the house intelligence committee. those documents show that former trump lawyer paul ryan had been trying to get rid of yovanovitch, who was ambassador at the time, for more than a year, the journal reports. the messages also show that ryan's chief of staff, george w. bush, had been in touch with ryan about removing her from her post.
[2] Source: "If I threaten someone, you'll know it," White House counselor Kellyanne Conway told reporters Friday morning in a bizarre hours-long effort to discredit an audio recording that seemingly showed her threatening to use West Wing resources to investigate the personal life of a reporter a day earlier. "I never threatened anyone," Conway told a gaggle of journalists at the White House. "Don't use those words. I didn't threaten. It's not a threat. I never threatened her." [SEP] A tense phone conversation between a reporter for the Washington Examiner and White House senior counselor Kellyanne Conway was published by the newspaper on Thursday. In the conversation, Conway objected that a story written by the reporter, Caitlin Yilek, mentioned that her husband George Conway is a fierce critic of President Trump on Twitter. Yilek was writing a story on Conway possibly becoming President Trump's next White House chief of staff if Trump decides to move on from the official now in the position, Mick Mulvaney. [SEP] Kellyanne Conway defended her phone call with a Washington Examiner reporter during an appearance on Fox News. Fox News host Sandra Smith described the phone call Friday between reporter Caitlin Yilek and the counselor to the president as a "heated clash" during which things got "very personal." Conway claimed, "Everything I said in that phone call, I've said publicly before. Every single thing. I said it again." Smith asked Conway if there was anything she regretted about how she handled her phone call, which she denied.
[2] MDS Hallucination: a recording of a tense phone call between senior white house adviser and reporter catherine yilek was published by the washington examiner on friday, and it's not going over well. in the recording, yalek says she was told by the white house that she would be fired if she didn't back off from her reporting on president trump's controversial son-in-law, and that she'd have to go to court if she did not. "i'm not going to do that," she says in the recorded conversation, per the new york times. "you're going to get fired if you don't shut the f — up." in the call, she also says she'll use the office of management and budget to investigate the personal life of the reporter. "if i threaten someone, you'll know it," the caller can be heard saying in the audio recording, per politico. "don ' t use those words. it ' s not a threat. i never threatened anyone." but on monday, white house counselor to the president katie holmes told fox news that she had never threatened the reporter.

Table 10: Examples of hallucinatory framing bias from MDS models and the corresponding the source input.