Faking Fake News for Real Fake News Detection: Propaganda-Loaded Training Data Generation

Kung-Hsiang Huang[♠] **Kathleen McKeown**[♣]
Preslav Nakov[◇] **Yejin Choi**[♡♦] **Heng Ji**[♠]
[♠] UIUC [♣] Columbia University [♡] University of Washington [◇] MBZUAI [♦] AI2
{khhuang3, hengji}@illinois.edu kathy@columbia.edu
preslav.nakov@mbzuai.ac.ae yejinc@allenai.org

Abstract

Despite recent advances in detecting fake news generated by neural models, their results are not readily applicable to effective detection of human-written disinformation. What limits the successful transfer between them is the sizable gap between machine-generated fake news and human-authored ones, including the notable differences in terms of style and underlying intent. With this in mind, we propose a novel framework for generating training examples that are informed by the known styles and strategies of human-authored propaganda. Specifically, we perform self-critical sequence training guided by natural language inference to ensure the validity of the generated articles, while also incorporating propaganda techniques, such as *appeal to authority* and *loaded language*. In particular, we create a new training dataset, PROPANEWS, with 2,256 examples, which we release for future use. Our experimental results show that fake news detectors trained on PROPANEWS are better at detecting human-written disinformation by 3.62–7.69% F1 score on two public datasets.[1]

1 Introduction

The dissemination of false information online can cause chaos, hatred, and trust issues, and can eventually hinder the development of society as a whole (Dewatana and Adillah, 2021; Wasserman and Madrid-Morales, 2019). In particular, human-written disinformation[2] is often used to manipulate certain populations and reportedly already had a catastrophic impact on multiple events, such as Brexit (Bastos and Mercea, 2019), the COVID-19 pandemic (van Der Linden et al., 2020), and the 2022 Russian assault on Ukraine.

Hence, there is an urgent need for a defense mechanism against human-written disinformation.[3] For this, we need a substantial amount of training data to build detectors. A naïve solution is to collect human-written news articles that contain inaccurate information by crawling untrustworthy news media. However, news articles published by suspicious sources do not necessarily contain false information, which means that annotators are required to fact-check every claim in each untrustworthy article. Moreover, articles containing false claims are often removed shortly after posting. While some work collected human-written fake news from fact-checking websites (Shu et al., 2018; Nguyen et al., 2020), the size of these datasets is limited. The curation process of these websites also requires a lot of manual efforts. Hence, such a solution is neither scalable nor reliable. Thus, an alternative direction complementing the existing efforts would be to generate training data automatically in a way that avoids these issues.

Our goal is to enhance disinformation detection by generating training examples that are better informed by the known styles and strategies of human-authored disinformation. We started by collecting human-written articles from untrustworthy sites[4], and we analyzed around 40 of them that spread false claims. Throughout our analysis, we found two characteristics of this human-written disinformation. First, about 33% of the articles used propaganda techniques to convince the audience that the fake information was actually authentic, and these techniques often involve the use of emotion-triggering language or logical fallacies (Da San Martino et al., 2019) to increase the impact on the reader. Statistics about the propaganda techniques are given in Appendix A.

[1] The code and data are released on GitHub: https://github.com/khuangaf/FakingFakeNews

[2] There are many types and definitions of *fake news*, but here we focus on text-only *disinformation*. Yet, we will also use the less accurate term *fake news* as it is more common.

[3] WARNING: This paper contains disinformation that may be sensitive or offensive in nature.

[4] These news sources are rated *low* for their factuality of reporting by mediabiasfactcheck.com.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 14571–14589
July 9-14, 2023

AJDABIYAH , Libya | Thu Apr 7 , 2011 6:34 pm EDT AJDABIYAH , Libya -LRB- Reuters -RRB- - Rebels fighting to overthrow Muammar Gaddafi said five of their fighters were killed ... "In rebel-held eastern Libya, wounded rebels being brought to a hospital Ajdabiyah said their trucks and tanks were hit on Thursday by a NATO air strike outside Brega. ~~NATO said it was investigating an attack by its aircraft on a tank column in the area along the Mediterranean coast on Thursday , saying the situation was " unclear and fluid . "~~ Rebels said at least five of their fighters were killed when NATO planes mistakenly bombed a rebel tank column near the contested port. "A number of vehicles were hit by a NATO strike ", officers from UN concluded. The fighting for Brega , the only active front , has dragged on for a week ...

Table 1: An example of our generated fake news. Given an authentic news article, our approach first identifies ~~a salient sentence~~, which it then replaces with a plausible but disinformative sentence that is coherent to the context. Finally, it generates a propaganda sentence to make the article resemble human-written fake news.

Second, more than 55% of the articles that we analyzed contained inaccurate information mixed with correct information: in fact, all claims, except for one or two, in these disinformation articles were factual, which makes the few false claims in these articles even more believable.

Prior work has made significant progress in generating fake news using large pre-trained sequence-to-sequence (seq2seq) models (Zellers et al., 2019; Fung et al., 2021; Shu et al., 2021). However, the articles generated by these approaches contain an overwhelmingly large proportion of false information and do not explicitly use propaganda.

To address these issues, here we propose a novel generation method. Given an authentic news article, we replace a salient sentence with a plausible but fake piece of information using a seq2seq model. As the generated texts can often be entailed by the original contexts, we incorporate a self-critical sequence training objective (Rennie et al., 2017) that incorporates a natural language inference (NLI) model into the loss function. Additionally, we use the NLI model to filter out generated sentences that can be inferred from the replaced ones. Then, we add propaganda techniques to mimic how humans craft disinformation. In particular, we automate two commonly used propaganda techniques, *appeal to authority* and *loaded language*, (Da San Martino et al., 2019) to add propaganda into the faked sentences.

Subsequently, we use the silver-standard training data generated from these two steps to train a detector. An example is shown in Table 1. We further recruited crowdsourcing workers to validate that some generated texts were indeed fake, so that we could construct a gold-standard training dataset.

Comparing our method to state-of-the-art fake news generation approaches, the evaluation results on two human-written fake news datasets show that detectors are substantially better at spotting human-written disinformation when trained on our generated fake news dataset.

Our ablation studies confirm the effectiveness of incorporating propaganda into the generated articles for producing better training data.

Our contributions can be summarized as follows:

- We propose an effective method to automatically generate more realistic disinformation compared to previous work.
- We develop the first automatic methods to generate specific propaganda techniques such that the generated articles are closer to disinformation written by humans.
- We demonstrate that detectors trained on our generated data, compared to generated articles using other methods, are better at detecting human-written disinformation.
- We release PROPANEWS, a dataset for disinformation detection containing 2.2K articles generated by our approach and validated by humans.

2 Training Data Generation

Our process of generating training data for propaganda-loaded disinformation consists of two main steps: disinformation generation (§2.1) and propaganda generation (§2.2). Below, we describe each of these steps in detail.

2.1 Disinformation Generation

Our disinformation generation approach aims at two sub-goals: (*i*) replacing a salient sentence in the given article with a sequence of generated coherent texts that looks plausible, and (*ii*) ensuring that the generated information cannot be entailed by the original masked-out sentence; otherwise, the generated texts will not be disinformative. To achieve the first sub-goal, we first identify salient sentences using extractive summarization, and we then perform mask-infilling with BART (Lewis et al., 2020). We achieve the second sub-goal using self-critical sequence training (Rennie et al., 2017) with an NLI component, which we use as a reward function for generation.

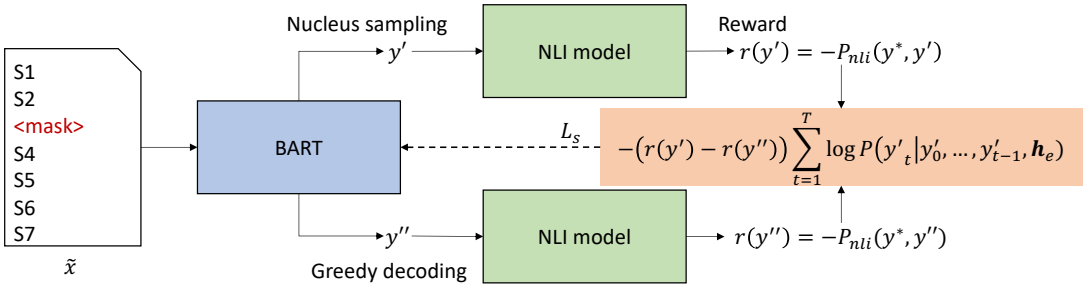

Figure 1: Illustration of our self-critical sequence training. Given a corrupted input article $\tilde{x}$, BART generates two sequences with nucleus sampling and greedy decoding, respectively. The reward for each sequence is computed as the negative entailment probability $-P_{ent}$ as output from the NLI model.

Salient Sentence Identification A salient sentence is critical for the overall semantics of the article. When it is manipulated or replaced, the complex events described in the article may be drastically changed. Yet, there is no salient sentence identification dataset publicly available. Motivated by the fact that sentences included in an extractive summary are often of higher importance, we take the scores computed by an extractive summarization model (Liu and Lapata, 2019), which predicts how likely each sentence is to belong to the summary, to estimate its saliency. We found that this yields reasonably good sentence saliency estimation. For each news outlet, we replaced one sentence that had the highest extractive summarization score with our generated disinformation.

Mask Infilling with BART To perform infilling, we took an approach similar to that of Donahue et al. (2020), but we used BART (Lewis et al., 2020). At training time, we randomly masked out a sentence y^* from a given article x. The bidirectional encoder first produces contextualized representations $\boldsymbol{h_e} = \text{Encoder}(\tilde{x})$ given the article with a masked-out sentence $\tilde{x} = x - y^*$. Then, the autoregressive decoder learns a maximum likelihood estimation that aims to maximize the probability of generating the next token y_t^* at time step t given all tokens in previous time steps $\{y_0^*, ..., y_{t-1}^*\}$ and the encoder hidden states $\boldsymbol{h_e}$ by minimizing the negative log probability of generating y_t^* as follows:

$$\mathcal{L}_m = -\sum_{t=1}^{T} \log P(y_t^* | y_0^*, ..., y_{t-1}^*, \boldsymbol{h_e}). \quad (1)$$

During inference time, rather than using random masking, $\tilde{x}$ is formed by masking out the sentence with the highest score computed by the extractive summarization model given the original document x, as discussed in the previous paragraph.

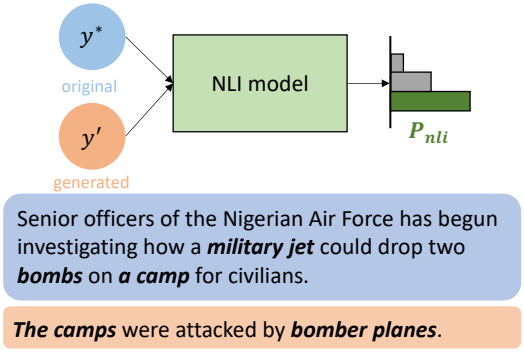

Figure 2: An example showing how the NLI model predicts an entailment from the masked out sentence y^* to the generated sentence y'.

Self-critical Sequence Training BART optimized via maximum likelihood estimation alone is capable of generating coherent texts. However, although the generated texts y' may be very different from the originally masked out sentence y^*, there is no guarantee that y' contains incorrect information. If the generated texts y' can be entailed by the masked out sentence y^*, then y' is actually not disinformative. An example is shown in Figure 2. Here, except for the lack of details, the generated sentence y' delivers the same message as the masked out sentence y^*. To reduce the probability that y' can be entailed by y^*, we leverage self-critical sequence training (Rennie et al., 2017; Bosselut et al., 2018) that rewards the model for generating sequences that cannot be entailed by the masked-out sentences. Self-critical sequence training (SCST) is a form of the REINFORCE algorithm (Williams, 1992) that allows direct optimization on non-differentiable functions. Using a baseline output y'' of the model to normalize the rewards, SCST avoids the challenge of directly estimating the reward signal or estimating normalization (Rennie et al., 2017). Since our goal is to avoid entailment from y^* to y', we define the reward as the negative entailment probability computed by a RoBERTa-based (Liu et al., 2019) NLI model

fine-tuned on Multi-NLI (Williams et al., 2018)[5],

$$r(y') = -P_{nli}(y^*, y'), \quad (2)$$

where $r(y')$ is the reward of the sequence sampled from the current policy y', and $P_{nli}(y^*, y')$ is the probability that y^* entails y'. To generate y', we use Nucleus Sampling (Holtzman et al., 2020) with $p = 0.96$, as this sampling method has shown advantages in *open-ended* generation (Holtzman et al., 2020; Zellers et al., 2019).

We generate the baseline output y'' using greedy decoding, then obtain the entailment probabilities between y' and y'' from the NLI model. We then compute the self-critical sequence training loss:

$$\mathcal{L}_s = -(r(y') - r(y'')) \sum_{t=1}^{T} \log P(y'_t | y'_0, .., y'_{t-1}, \boldsymbol{h_e}). \quad (3)$$

Here $r(y'')$ is a baseline reward, and $r(y') - r(y'')$ is a normalized reward. This loss function encourages BART to generate y' when $r(y') > r(y'')$, whereas it suppresses the probability of decoding y' when $r(y') < r(y'')$. An overview of SCST is shown in Figure 1.

The final objective function to minimize is a weighted sum of Equation (1) and Equation (3),

$$\mathcal{L}_{final} = \alpha \mathcal{L}_m + \beta \mathcal{L}_s, \quad (4)$$

where α and β are the weights for each loss.[6]

Post-processing To further ensure the quality of the disinformation generated, we reuse the NLI model discussed in the previous paragraph to filter out invalid outputs y' that can be entailed from the masked-out sentence y^*, as demonstrated in Figure 2. We found that incorporating the SCST loss (Equation (3)) into the training objective successfully reduces the invalid rate from 7.8% to 3.2%.

2.2 Propaganda Generation

After generating inaccurate information, we incorporate propaganda into each generated article. We chose two representative propaganda techniques of each type: emotional versus non-emotional. *Loaded language* is an emotional technique and it is also by far the most frequent propaganda technique as shown in Table 5 of (Da San Martino et al., 2019) and Table 2 of (Dimitrov et al., 2021).

[5]We use the fine-tuned NLI model from https://huggingface.co/roberta-large-mnli. Its accuracy is 90.2% on the dev set of MNLI, which is on par with state-of-the-art methods.

[6]Empirically, we set $\alpha = 1$ and $\beta = 0.01$.

Based on these two tables, we also see that *appeal to authority* is among the most frequent non-emotional techniques.

Appeal to Authority *Appeal to authority* is a propaganda technique that aims to strengthen or to invalidate an argument by referring to a statement made by authorities or experts (Da San Martino et al., 2019). We first collect experts from various domains, such as economics and immunology, from Wikidata.[7] In particular, we specify the *occupation* (P108) of each expert and we filter out entities that were born before 1940 to ensure recency. To consider only impactful entities, we rank all candidates based on the number of corresponding outcoming *statements* (i.e., connected concepts in Wikidata), inspired by PageRank (Page et al., 1999), and we add the top 100 entities for each occupation into the candidate list Z. Then, we include the person named entities extracted by a name tagger,[8] which are more relevant to the local context. This makes sense as we found that more than 73% of the news articles contain authorities. More details on how authority candidates Z are collected can be found in Appendix E.

Once we collect a candidate list Z, we then generate fake arguments made by each $z_i \in Z$ with the BART model that has already been fine-tuned in §2.1. In particular, a <mask> token is inserted right after the filled-in sentence y' in the input article to BART so that it knows where to perform infilling. To inform BART that it should generate a statement made by an authority, we prefix the decoder with the template *[z_i confirmed that "]*, where $z_i \in Z$ is the name of the authority.

The prefix ends with an opening quotation mark to indicate that it should be followed by a statement by authority z_i. To increase the diversity of the generated statements, we devise a variety of templates, as detailed in Appendix E. Finally, the best sequence s^* is selected with the lowest perplexity $s^* = \text{argmin}_{s_i} \text{Perplexity}(s_i)$, where s_i denotes the generated sequence using z_i as the authority.

Loaded Language *Loaded language* is another propaganda technique that uses emotion-triggering terms or phrases to influence the opinions of the audience (Da San Martino et al., 2019; Dimitrov et al., 2021). Often, *loaded language* uses sensational adverbs or adjectives to exaggerate a statement.

[7]https://query.wikidata.org/

[8]https://stanfordnlp.github.io/stanza

Technique	Generated Disinformation and Propaganda
Appeal to Authority	Cairo's Tahrir Square was the scene of clashes between protesters and police on Wednesday. " At least three people were killed and more than 600 were injured in the clashes," said Egypt's President.
Loaded Language	Cairo's Tahrir Square was the scene of deadly clashes between protesters and police on Wednesday.

Table 2: Examples of the two generated propaganda techniques, as shown by texts in blue. The first row shows how the argument is strengthened by appealing to an authority's statement, while the second row demonstrates how loaded language is introduced with an emotion-triggering term.

Based on this observation, we utilize the propaganda dataset released by Da San Martino et al. (2019) where propaganda techniques are annotated at the fragment level (i.e. span level). The dataset contains 2,547 *loaded language* instances. Yet, not every instance contains adjectives or adverbs that are emotion-triggering. To create valid training data for *loaded language* generation, we first use SpaCy to perform part-of-speech tagging and dependency parsing, and then keep the examples where there exists an adverb pointing to a verb or an adjective pointing to a noun through dependency parsing edges. This results in 1,017 samples of valid *loaded language* instances. Examples of the generated *appeal to authority* and *loaded language* are shown in Table 2.

Upon collecting the training data to generate *loaded language*, we fine-tune another BART on this dataset. Naïvely, we can take the articles with emotion-triggering adverbs or adjectives removed as input to BART and using the original article as the decoding target. However, we found that around 25% of the time BART does not exactly reproduce the unmasked texts due to hallucination. This observation is consistent with Donahue et al. (2020)'s findings. To this end, we propose a two-step generation approach. First, we train BART to insert a `<mask>` token into the target sentence in the input document marked with special tokens. Then, BART learns to infill the `<mask>` with an approach similar to what is discussed in §2.1 but without the SCST objective. Empirically, we found that this approach successfully reduces the chance of failure in generating the exact unmasked contexts to around 2%.

2.3 Intermediate Pre-training

As the size of TIMELINE17 (Tran et al., 2013) and the propaganda dataset (Da San Martino et al., 2019) are relatively small, we perform intermediate pre-training (IPT) on the news articles from CNN/DM, a large news summarization dataset (Hermann et al., 2015), for domain adaptation. Details of IPT can be found in Appendix F.

3 Our PROPANEWS Dataset

3.1 Data Source

When selecting the source of data, we considered two criteria. First, the news articles must have high trustworthiness. This ensures that, except for our manipulated sentences, the rest is genuine. Second, the news events described in the articles must be important. Motivated by these two criteria, we repurposed the TIMELINE17 dataset (Tran et al., 2013) as our source of data. It contains 17 timelines, each of which corresponds to a news event. Each timeline is associated with a series of news articles that span across a wide time span, implying the high importance and impact of these events. Moreover, the articles come from trustworthy media. In total, there are 4,535 news articles in TIMELINE17.

3.2 Crowdsourcing for Data Curation

We use Amazon's Mechanical Turk (AMT) to verify the quality and the correctness of the generated disinformation. In total, there are around 400 unique crowdsourcing workers contributing to approximately 2,000 Human Intelligence Tasks (HITs). For each HIT, the annotators were tasked to look for supporting evidence from trustworthy news media to determine whether the sentences generated are indeed *inaccurate*. Only those labeled as *inaccurate* were included in PROPANEWS, while the *accurate* counterparts were discarded. Appendix H gives additional details.

To measure the inter-annotator agreement (IAA), we use the Worker Agreement With Aggregate (WAWA) score (Ning et al., 2020; Sheng et al., 2021), which compares each annotator's answer to the aggregated answer obtained via majority votes and micro-averages the results across all samples.[9] The resulting WAWA precision, recall, and F_1 are 80.01%, 78.94%, and 79.47%, which indicates moderate to high agreement.

[9] We did not use other IAA metrics, such as Cohen's Kappa (Cohen, 1960), as we expect the vast majority of our generated disinformation to be inaccurate. WAWA provides a better approximation for inter-annotator agreement in our scenario.

4 Disinformation Detection

The disinformation detection task challenges detectors to determine whether a given input article contains inaccurate information or not. We experiment on four detectors, including HDSF (Karimi and Tang, 2019), GROVER (Zellers et al., 2019), BERT (Devlin et al., 2019), and ROBERTA (Liu et al., 2019). HDSF leverages the hierarchical structures of discourse-level features, such as dependency trees, to predict the veracity of a news article. GROVER is an unidirectional seq2seq model pre-trained on news documents. We use the discriminative version for detection which is adapted from its generative version by feeding the `[CLS]` token representations to a multi-layer perceptron. Similarly, BERT and ROBERTA take in the entire article as input and feed the representations of the first token to a classification head to determine the veracity of each article. In addition, all models are optimized using cross entropy. For fair comparison, we set the maximum sequence length to 512 and we use the LARGE variants for all models. More details can be found in Appendix J.

5 Experiments

In our experiments, we aim (1) to analyze the performance of different models on our new PROPANEWS dataset, (2) to examine the effect of various training data sets, and (3) to investigate how much silver-standard data is equivalent to gold-standard data.

5.1 Data

PROPANEWS The PROPANEWS dataset consists of 2,256 distinct articles, with a balanced number of fake and real documents. Within the fake articles, 30% use *appeal to authority*, another 30% include *loaded language*, and the remaining 40% simply contains inaccurate information. We split the data into 1,256:500:500 for training, validation, and testing.

Evaluation Data We use two sets of human-written articles released by Nguyen et al. (2020) and Shu et al. (2018) to evaluate the effectiveness of our approach. The articles in each dataset are collected from two fact-checking websites, SNOPES and POLITIFACT, respectively. Articles no longer accessible via the given URL are removed. Statistics about both datasets are shown in Appendix I.

Other generated training data We compare PROPANEWS to the following approaches. **GROVER-GEN** (Zellers et al., 2019) generates headlines which condition on the original body texts, followed by body text generation conditioning on the generated headlines. **FACTGEN** (Shu et al., 2021) enhances the factual consistency of the generated article with a fact retriever that fetches supporting information from external corpora. **FAKEEVENT** (Wu et al., 2022) generates sentences sequentially with condition on the manipulated knowledge elements of each sentence. Also, we form the **PN-SILVER** dataset by resampling our generated data but disregarding the annotator validation. Furthermore, we construct additional training sets by replacing the salient sentence in each article with one sentence generated by each baseline method, as indicated by **-1SENT**. To ensure fair comparisons, all generators take in the same set of authentic articles as inputs.

5.2 Results and Discussion

Human-written disinformation detection To study the effectiveness of human-written disinformation detection, we train GROVER-LARGE and ROBERTA-LARGE on different training datasets and evaluate them on the SNOPES and POLITIFACT datasets, as shown in Table 3. Both models perform best when trained on PROPANEWS, compared to training on other datasets. Consider ablating human validation, detectors trained on PN-SILVER still outperform their counterparts trained on other datasets. This shows that our generative method produces articles that are more similar to human-written disinformation. To further verify this finding, we measure the similarity between articles generated by different approaches and disinformative articles in the POLITIFACT dataset using the MAUVE metric (Pillutla et al., 2021). MAUVE computes the similarity between two text distributions by adding the areas under a divergence curve, and has been shown to produce better approximations than other metrics such as JS divergence (Martins et al., 2020). We find that the MAUVE score with POLITIFACT for PROPANEWS and GROVER-GEN is 17.1% and 13.7%, respectively, suggesting that the generated documents in PROPANEWS are closer to human-written disinformation. These results confirm that the advantage of our generated articles in defending against human-written disinformation is resulted from the closer gap between them.

Test Data →	POLITIFACT				SNOPES			
Detectors → Training Data ↓	ROBERTA-LARGE		GROVER-LARGE		ROBERTA-LARGE		GROVER-LARGE	
Without human validation (silver)								
GROVER-GEN	57.65	(±7.6)	52.77	(±2.1)	48.42	(±2.2)	49.53	(±0.1)
GROVER-GEN-1SENT	49.65	(±5.2)	47.48	(±1.8)	44.44	(±3.2)	50.10	(±2.1)
FAKEEVENT	46.33	(±2.6)	50.27	(±5.9)	45.36	(±1.2)	47.40	(±1.3)
FAKEEVENT-1SENT	47.32	(±3.2)	50.12	(±3.2)	46.62	(±2.9)	47.29	(±2.7)
FACTGEN	48.46	(±2.2)	51.79	(±3.6)	41.98	(±5.4)	50.47	(±4.9)
FACTGEN-1SENT	41.19	(±3.5)	40.92	(±4.1)	40.01	(±3.8)	45.52	(±3.7)
PN-SILVER	**60.39***	(±3.9)	**55.23***	(±5.8)	**51.52****	(±3.4)	**52.39****	(±4.1)
With human validation (gold)								
PROPANEWS	**65.34****	(±4.5)	**60.43****	(±6.2)	**53.03****	(±3.7)	**54.09****	(±2.8)
w/o AA	63.21**	(±3.2)	58.28**	(±4.2)	50.78*	(±1.8)	53.22**	(±3.7)
w/o LL	64.65**	(±1.8)	56.93**	(±5.3)	51.92**	(±3.4)	51.68*	(±1.4)
w/o AA & LL	61.83*	(±4.9)	52.82	(±3.3)	52.77**	(±2.7)	50.93	(±2.7)

Table 3: AUC (in %) for different models on the SNOPES and POLITIFACT datasets when trained on various data sets. The bottom rows show different variants of PROPANEWS. AA denotes *appeal to authority*, whereas LL refers to *loaded language*. We report the mean and the standard deviation of four runs. Statistical significance over previous best approaches computed using the paired bootstrap procedure (Berg-Kirkpatrick et al., 2012) is indicated with $^{**}(p < .01)$ and $^{*}(p < .05)$.

Comparing each baseline method and its counterpart that only generates one sentence to be substituted for the salient sentence (i.e., **-1SENT**), we found significant performance drops on GROVER-GEN and FACTGEN when only generating one sentence. This is likely caused by the incoherence between the right context and the sentence generated by these approaches due to the left-to-right fashion of text generation. While FAKEEVENT does not see the right context, it additionally conditions on knowledge elements corresponding to the sentence, which discourages it from producing topically irrelevant content and thus does not lead to huge performance drop.

In Table 4, we show two examples of disinformative articles from POLITIFACT where ROBERTA is able to classify them as inaccurate when trained on PN-SILVER, but fails when trained on GROVER-GEN. Both articles contain propaganda, which is incorporated into PN-SILVER but not into GROVER-GEN. This demonstrates that detectors trained on our generated data perform better at detecting human-written disinformation that has such properties.

Is propaganda generation helpful for disinformation detection? We further conduct an ablation study to analyze the contributions of each propaganda technique. As shown in the bottom of Table 3, both *appeal to authority* and *loaded language* prove beneficial in enhancing models' abilities to detect human-written disinformation.

We can further see in Table 3, when comparing PROPANEWS WITHOUT AA& LL to other generation approaches, that both models trained on our generated data, even without the incorporation of propaganda techniques, still outperform their counterparts trained on other datasets. This illustrates that our generated disinformation texts are closer to news articles written by humans.

How good is the generation quality? To evaluate the quality of our generation approach, we asked Amazon Mechanical Turk (AMT) workers to rate the plausibility of 100 generated articles from PROPANEWS and to determine the degree by which their answer to this question is influenced by the generated propaganda. Each article was rated by three different AMT workers. For comparison, we also asked the AMT workers to rate the plausibility of 100 generated articles from GROVER-GEN. The average plausibility scores for PROPANEWS and GROVER-GEN were 2.25 and 2.15 (out of 3), respectively. indicating that our generation approach has a slight advantage over GROVER-GEN in terms of plausibility. Moreover, among the articles in PROPANEWS that are rated highly plausible, 29.2% of the workers think that the generated propaganda highly affects their response (i.e. rated 3 out of 3) that the generated article is plausible. This demonstrates the effectiveness of our propaganda techniques in increasing the plausibility of generated articles. Survey details and score distributions are discussed in Appendix K.

Article and Analysis
Article: ... Statement from FDA Commissioner Scott Gottlieb, M.D., on FDA's ongoing efforts to help improve effectiveness of influenza vaccinesFor Immediate Release: ... **Analysis:** *Appealing to authority* is common in human-written fake news.
Article: ... Regardless of how much we hate Nacy Pelosi, she represents a Congressional District that saw a million fraudulent votes from illegal immigrants... **Analysis:** The use of *loaded language* often indicates disinformation.

Table 4: Examples from POLITIFACT where ROBERTA-LARGE successfully predicts the veracity when trained on PN-SILVER, but classifies incorrectly when trained on GROVER-GEN.

6 Related Work

Fake News Generation and Detection There has been a focus in prior research on using neural networks to automatically generate fake news as a means of defending against the proliferation of machine-generated fake news. Zellers et al. (2019) pre-trained a generator with the GPT-2 architecture (Radford et al., 2019) on a large-scale news corpus and demonstrated that it was effective in detecting neural fake news. More recently, Fung et al. (2021) improved the controllability of the generated fake news by conditioning the generator on knowledge elements, such as entities, relations and events, extracted from the original news article. Shu et al. (2021) enhanced the factuality of the generated article by introducing a fact retriever that fetches relevant information from external corpora. Mosallanezhad et al. (2021) used adversarial reinforcement learning to generate topic-preserving articles. These studies developed methods for generating fake news that is hard to distinguish from real news to humans. Nevertheless, due to the overwhelming amount of inaccurate information introduced and the lack of propaganda techniques in the generated texts, these approaches are suboptimal for detecting human-written fake news, as shown in §5.2. In contrast, we generate fake news by incorporating propaganda techniques and preserving the majority of the correct information. Hence, our approach is more suitable for studying defense against human-written fake news. Also, since our dataset is annotated with the exact offset of the disinformative passages, it enables research on interpretable detection of fake news.

Propaganda Generation and Detection There is little previous work on propaganda generation. Zellers et al. (2019) is the only relevant work, and it studied the generation of propaganda to communicate targeted disinformation. In contrast, we generate propaganda techniques to bring the generated articles closer to human-written fake news.

To the best of our knowledge, we are the first to study the incorporation of specific propaganda techniques into generated articles. Prior work on propaganda detection mainly focused on document-level detection. Early work collected propaganda datasets using distant supervision (Rashkin et al., 2017) by assigning the same propaganda label to each news outlet under the same source based on the news-media-level label of corresponding news source listed on trustworthy sites. However, classifiers trained on such datasets may only learn to recognize the bias of each news source instead of propaganda (Martino et al., 2020). Our dataset avoids such issues by explicitly incorporating propaganda into each generated article. Furthermore, Da San Martino et al. (2019) presented a fragment-level propaganda detection dataset, where specific propaganda techniques were labeled onto spans of text instead of each document. Recent approaches for detecting these propaganda techniques rely on pre-trained transformers (Morishita et al., 2020; Feng et al., 2021). In contrast, we focus on detecting disinformative articles with propaganda signals.

7 Conclusions and Future Work

We have proposed a novel method for generating disinformation that is closer to human-written fake news. Evaluation on two human-written fake news datasets, POLITIFACT and SNOPES, demonstrated the effectiveness of our generated data PROPANEWS in enabling better detection performance on human-written fake news. We hope that the dataset presented in this work, PROPANEWS, can serve as an enabling resource for detecting human-written fake news and encouraging future research in this direction.

In future work, we plan to extend our approach to other languages and to cover more propaganda techniques. We are also interested in studying other aspects of fake news generation, such as novelty and elaboration, as well as engaging linguistic style.

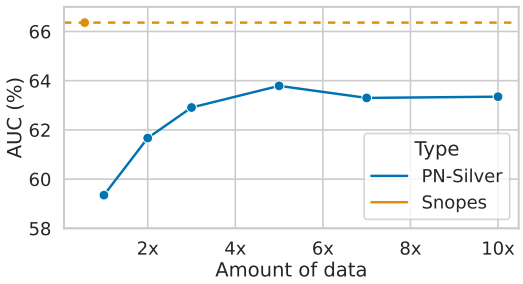

Figure 3: Performance comparison of RoBERTa-Large on the PolitiFact dataset when trained on Snopes and different size of PN-silver.

8 Limitations

To understand the gap between our automatic data generation method and fake news written by humans, we expanded PN-silver to different sizes and compared the performance of RoBERTa-Large when trained on these generated datasets and the human-written fake news dataset, Snopes. Note that since the Timeline17 dataset only contains around 4K samples, we additionally crawled New York Times news articles as an input to our generator for the "5 times" to "10 times" experiments. The results are shown in Figure 3. Although the detector performance at first improves as we add more *silver* training data, it reaches a plateau after the size is increased five-fold. This illustrates that while our approach is more effective compared to baseline generation methods, there is still a clear gap between our generated articles and human-crafted fake news, likely in aspects such as style (as discussed in §5.2), intent (i.e., limited modeling of propaganda techniques), and falsehood (i.e., the generated content is 100% false).

Despite the advantages of our generation approach, as compared to previous methods, it is uncapable of generating other propaganda techniques covered in (Da San Martino et al., 2019), such as *straw man*. Thus, our method is not generic enough to handle all types of propaganda techniques within a unified framework. Moreover, our approach is limited to generating English-only news articles, and cannot be applied to other languages.

9 Ethical Statement and Broader Impact

Our objective for developing a generative approach that produces more realistic news articles is to advance the field of disinformation detection and to bring awareness that the current approaches for generating training data for fake news detection are sub-optimal.

We acknowledge that our generator may produce toxic text as it was fine-tuned on propagandistic datasets. We also understand the dual-use concerns for such a generation framework. One potential concern is the possibility of using the generator to produce fake news for political gain or to sow social discord. Another concern is the potential for the generator to be used to generate fake news that could cause harm, such as false medical information or misleading financial advice. Additionally, the generator might be used to create false evidence or to fabricate information to support false allegations in legal or regulatory proceedings.

Therefore, to contribute to future studies on human-written disinformation detection, we decided to release the codebase for only the detectors used in the experiments as well as the generated data but not the generator.

We highlight some scenarios that illustrate appropriate and inappropriate uses of our generator:

- **Appropriate:** Researchers can use our framework to produce more challenging training data for learning stronger detectors.
- **Inappropriate:** The method should not be used to intentionally create or propagate false information.
- **Inappropriate:** The propaganda generation technique should not be used for political campaigns or any malicious purposes.

Both inappropriate uses could lead to harmful consequences, such as undermining trust in the media and causing social unrest.

Acknowledgement

This research is based upon work supported by U.S. DARPA SemaFor Program No. HR001120C0123 and DARPA MIPs Program No. HR00112290105. The views and the conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and to distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Marco T Bastos and Dan Mercea. 2019. The Brexit botnet and user-generated hyperpartisan news. *Social science computer review*, 37(1):38–54.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 173–184, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. 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)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hernawan Dewatana and Siti Ummu Adillah. 2021. The effectiveness of criminal eradication on hoax information and fake news. *Law Development Journal*, 3(3):513–520.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Zhida Feng, Jiji Tang, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, and Li Chen. 2021. Alpha at SemEval-2021 task 6: Transformer based propaganda classification. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 99–104, Online. Association for Computational Linguistics.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pages 1693–1701, Montreal, Quebec, Canada.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the 8th International Conference on Learning Representations*, ICLR '20, Addis Ababa, Ethiopia. OpenReview.net.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3432–3442, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR '15, San Diego, CA, USA.

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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*, ICLR '19, New Orleans, LA, USA. OpenReview.net.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4826–4832. ijcai.org.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2020. Sparse text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4252–4273, Online. Association for Computational Linguistics.

Terufumi Morishita, Gaku Morio, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at SemEval-2020 task 3: Exploring the representation spaces of transformers for human sense word similarity. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 286–291, Barcelona (online). International Committee for Computational Linguistics.

Ahmadreza Mosallanezhad, Kai Shu, and Huan Liu. 2021. Generating topic-preserving synthetic news. In *Proceedings of the 2021 IEEE International Conference on Big Data*, Big Data '21, pages 490–499.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20, pages 1165–1174, Ireland (Virtual Event). ACM.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaïd Harchaoui. 2021. MAUVE: measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, NeurIPS '21, pages 4816–4828.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '17, pages 1179–1195, Honolulu, HI, USA. IEEE Computer Society.

Niloufar Salehi, Lilly Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. 2015. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1621–1630, Seoul, Republic of Korea. ACM.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. "nice try, kiddo": Investigating ad hominems in dialogue responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 750–767, Online. Association for Computational Linguistics.

Kai Shu, Yichuan Li, Kaize Ding, and Huan Liu. 2021. Fact-enhanced synthetic news generation. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, AAAI '21', pages 13825–13833. AAAI Press.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media. *ArXiv:1809.01286*.

Giang Binh Tran, Tuan Tran, Nam Khanh Tran, Mohammad Alrifai, and Nattiya Kanhabua. 2013.

Leveraging learning to rank in an optimization framework for timeline summarization. In *Proceedings of the SIGIR 2013 Workshop on Time-aware Information Access*, TAIA '13.

Sander van Der Linden, Jon Roozenbeek, and Josh Compton. 2020. Inoculating against fake news about COVID-19. *Frontiers in psychology*, 11:2928.

Herman Wasserman and Dani Madrid-Morales. 2019. An exploratory study of "fake news" and media trust in Kenya, Nigeria and South Africa. *African Journalism Studies*, 40(1):107–123.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, United States. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, NeurIPS '19, pages 9051–9062, Vancouver, BC, Canada.

A Distribution of Propaganda

Figure 4 shows the distribution of the propaganda techniques used in the human-written fake news we collected and analyzed in §1. Note that one article may contain multiple propaganda techniques.

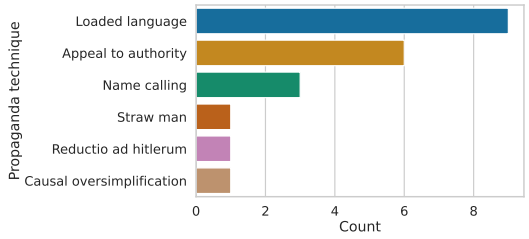

Figure 4: Total number of propaganda techniques used in the human-written fake news we analyzed.

B Additional Research Questions

Q1: Is the detector learning to distinguish between fake/real news articles or simply learning to detect the use of propaganda techniques? In Table 3, **PROPANEWS w/o AA & LL** is the variant of our proposed dataset with both propaganda techniques removed. By training detectors on this version of the proposed dataset, the model is still effective in identifying human-written articles containing false information. Therefore, the detectors trained on our generated data have learned to distinguish between fake and real articles instead of exploiting propaganda information only. On the other hand, comparing the detectors trained on **PROPANEWS** and their counterparts trained on **PROPANEWS w/o AA & LL** in Table 3, we see that propaganda can help improve the detection of real human-written fake news. We further want to emphasize that fake news detection is an extremely challenging task that requires both factual and stylistic analysis as demonstrated by our experiments and by the relatively low performance of prior SOTA models.

Q2: Do real articles make use of propaganda techniques, such as *appeal to authority* and *loaded language*? The similarity between our generated text and the real articles in PolitiFact is 7.3% as per the MAUVE measure, which is much lower than the similarity between the generated text and the fake news articles, as discussed in §5.2. It is possible that some real news articles can contain propaganda. However, according to MAUVE, the real articles in POLITIFACT do not contain much loaded language or appeal to authority.

C Further Analysis

C.1 Remaining Challenges

To better understand the remaining disinformative articles that the detectors failed to identify, we conducted additional analysis by comparing the ROBERTA predictions and the labels. As a result, we identified the following three major modeling capabilities required for successful detection:

Static knowledge enrichment About 30% of misclassification is due to the lack of static knowledge that can be found in public databases, such as law dictionaries. For example, in this article,[10] Alexandria Ocasio-Cortez falsely states that the U.S. Immigration Customs Enforcement (ICE) is required to fill 34,000 beds every day. According to the Appropriations Act of 2016,[11], ICE is only required to detain 34,000 available beds. Therefore, to detect such kind of misinformation, the detector needs to be enriched with static knowledge bases.

Dynamic knowledge acquisition Around 48% of the misclassified human-written disinformation is due to the inability to acquire dynamic knowledge from new news sources. For instance, COVID19-related articles are usually published after 2020, while ROBERTA was pre-trained on news articles released before 2019. It is very challenging for ROBERTA to detect disinformation of such topics unless the detector is equipped with the capability to acquire dynamic knowledge from news articles. Particularly, ROBERTA achieves an accuracy of 69.0% on detecting fake articles published before 2019, but its accuracy drops to 51.9% when testing on articles published after 2019.

Multi-document reasoning The rest of the incorrect detection is caused by the lack of multi-document reasoning ability. For instance, a news article[12] wrongly associates Hillary Clinton with a flawed immigration policy of the former government, and strengthens such a statement by referring to a Senate report and relevant news articles. However, the cited report does not mention Clinton, and the other news articles contain disinformation. To correctly detect this piece of disinformation, detectors should reason across multiple documents.

[10] https://tinyurl.com/static-knowledge
[11] https://www.congress.gov/114/bills/hr2029/BILLS-114hr2029enr.pdf
[12] https://tinyurl.com/multi-doc

D Qualitative Examples of Generated Articles

In Table 8, we show a comparison of generated articles given the same input data across different generative methods. Our approach produces articles with a small fraction of inaccurate information, which matches a property of human-written fake news discussed in §1.

E Appeal to Authority Details

To recap, we first gather a list of authorities Z for each article from Wikidata and the corresponding context. The best *appeal to authority* sequence s^* is selected, i.e., the one with the lowest perplexity $s^* = \text{argmin}_{s_i} \text{PPL}(s_i)$, where s_i denotes the generated sequence using z_i as the authority. However, this process results in every sequence s^* containing the substring "confirms that", which makes it trivial for detectors to classify these generated documents as fake by simply detecting such substrings. Therefore, we devise an algorithm to diversify the templates so that these generated articles are not easily detectable.

First, we define a set of verbs V that can be swapped with "confirms": V = {*said*, *concluded*, *confirmed*, *emphasized*, *stated*, *argued*}. Then, we diversify the generated structure of the generated sentence s^* by reordering the subject, the verb, and the object. Next, we swap the verb with another verb from V. Finally, in order to diversify the context, we append a preposition from the preposition set PP = {*on*, *at*, *in*} to the output of the previous step, and then we feed the sequence to BART to generate the context. An example of this process is given in Table 6.

F Intermediate Pre-training Details

For domain adaptation, we perform intermediate pre-training (IPT) on the CNN/DM dataset, a large summarization corpus containing more than 280K news articles from CNN and Daily Mail. The IPT objectives for disinformation generation and propaganda generation are mostly the same as described in the previous sections, but with some minor changes due to different goals in the IPT phase. When performing IPT for disinformation generation, we removed $\mathcal{L}_s$ from the final loss function (Equation (4)) as the goal for IPT is only to learn to generate coherent sentences, and thus IPT is not needed.

Detector	Dev Acc. (%)	Test Acc. (%)
HDSF	52.4 (±0.6)	50.6 (±2.4)
BERT	57.7 (±1.0)	58.0 (±1.2)
GROVER	60.3 (±5.8)	63.3 (±5.0)
ROBERTA	**70.5** (±0.3)	**69.8** (±1.1)

Table 5: Evaluation of various detectors on the PROPANEWS development and test set. We report the mean and the standard deviation over four runs.

Moreover, in order to create training samples for *loaded language* IPT, we gather all the appearances of adjectives pointing to a noun or adverbs pointing to a verb via dependency parsing graphs without considering whether the samples actually contain *loaded* terms since the goal here is to enable BART to identify where properly to insert which adjectives or adverbs.

G Benchmarking Detectors

The performance of various detectors on the PROPANEWS dataset is shown in Table 5. We find that ROBERTA and GROVER demonstrate advantages over BERT. This could be explained by the fact that ROBERTA and GROVER are pre-trained on news domain corpora, whereas BERT has no access to such domains during pre-training. In addition, we find that HDSF performs much worse than the other three models. This reflects that large-scale pre-training of language models brings more benefit to detection performance than explicit modeling of discourse-level features.

H Human Validation Details

Next, we describe the details of human validation, where AMT workers were tasked to validate whether the generated sentences contained inaccurate information. We recruited AMT workers from USA and Canada. To ensure the annotation quality, only workers who had an acceptance rate greater than 95% and more than 100 accepted HITs in the past were allowed to work on our annotation task. This greatly reduced the chances of collecting annotations from scammers. Each HIT was designed such that the annotators were rewarded $12-$15 per hour, which complies with the ethical research standards outlined by AMT (Salehi et al., 2015). In each HIT, the annotators were presented an article with the generated part marked in boldface. The questions and the guidelines are given below. (Note that we only use the annotators' response for Q1 to validate our generated data. The annotations for the other questions will be used for future research.)

Step	Generated Sequence
1	Panmure Gordon analyst Peter Hitchens confirmed that “ the US government is likely to agree to reduce its estimate of the size of the spill, which would cut BP fines ”.
2	“ The US government is likely to agree to reduce its estimate of the size of the spill, which would cut BP fines, ” Panmure Gordon analyst Peter Hitchens confirmed.
3	“ The US government is likely to agree to reduce its estimate of the size of the spill, which would cut BP fines, ” Panmure Gordon analyst Peter Hitchens said.
4	“ The US government is likely to agree to reduce its estimate of the size of the spill, which would cut BP fines, ” Panmure Gordon analyst Peter Hitchens said in a conference.

Table 6: An illustration of how appeal to authority is performed. In step 1, we generate a statement using BART with the prefix [*Panmure Gordon analyst Peter Hitchens confirmed that "*]. In step 2, we move the subject and the verb to the back of the sentence to diversify the sentence structure. In step 3, we swap the verb with another verb from the verb set V. In step 4, we append a preposition *in* to the sequence in step 3 and we use the resulting sequence as a prefix to BART's decoder to generate the rest of the context. For steps 1 and 4, we mark the prefix sequence to the decoder in yellow, and the generated sequence in blue. To increase the diversity of the generated sequences, step 2 to 4 are each performed 50% of the time.

Q1: Is the generated text in boldface **Accurate** or **Inaccurate**? (If you cannot find any supporting evidence, please select **Inaccurate**.) Note that a statement (in quotation marks) made by a person is only accurate if this person actually made the exact same statement. If the statement in quotation marks is just a paraphrase of what the person actually said, then the statement is inaccurate.

- **Inaccurate**: Any false information presented in the generated text makes it inaccurate.

- **Accurate**: All the information in the generated text must be accurate.

Q2: Enter the URL of the news article you found that supports your decision in the previous response in the below box. Put down “from context" if the evidence can be found in the context.

Q3: Does the generated text in boldface deliver the same sentiment as the rest of the article?

- **False**: The sentiment of the generated text is NOT the same as the rest of the article.

- **True**: The sentiment of the generated text is the same as the rest of the article.

Q4: Is the discourse of the generated text in boldface consistent with the rest of the article?

- **False**: The discourse of the generated text is NOT consistent with the rest of the article.

- **True**: The discourse of the generated text is consistent with the rest of the article.

Q5: If there is any grammatical error or inconsistent discourse, please rewrite and correct generated text and put it in the below box. Just put down the corrected generated text in bold is enough. For example, “Harry is a boy. He likes go to school.” Please put in “He likes to go to school.” in the box below.

I Statistics about the Evaluation Datasets

In Table 7, we give some statistics about the two evaluation datasets used in our experiments. The reported numbers are not the same as those in the original papers (Nguyen et al., 2020; Shu et al., 2018) since some of the articles were no longer accessible via the provided URLs.

Dataset	# Real	# Fake
SNOPES	430	280
POLITIFACT	517	369

Table 7: Statistics about the two evaluation datasets, SNOPES and POLITIFACT.

J Detector Implementation Details

For our experiments with BERT and ROBERTA, we used AdamW (Loshchilov and Hutter, 2019) with a batch size of 2 and gradient accumulation steps of 8. We set the learning rate and the weight decay to 5e-5 and 1e-5 for the parameters that have been pre-trained, and 1e-3 and 1e-3 for the other parameters. For the GROVER detector, we follow the original detection setting. GROVER is trained using Adam (Kingma and Ba, 2015) with a learning rate of 2e-5 and a batch size of 64.

Similarly, we follow the original recipe to train HDSF, which is optimized with Adam with a learning rate of 1e-2. All detectors are fine-tuned for at most 20 epochs where the best model is determined by the accuracy on the development set.

All experiments are conducted on an Ubuntu 18.04 machine with NVIDIA Tesla V100. We use PyTorch 1.10.0 and Transformers 4.3.0 for constructing all models and loading pre-trained weights, except for GROVER, which operates on Tensorflow 1.13.1. The training time for BERT and ROBERTA, each of which has 340M parameters, is around 2-3 hours, while for GROVER, which contains 355M parameters, it is about 1 hour.

K Human Evaluation Details

In this section, we describe the survey we did with AMT workers for evaluating the quality of the generated articles. The annotators were presented with a generated article and were asked to answer a few questions regarding its quality. **Q2** is only applicable for evaluating generated articles from PROPANEWS, in which we show the sentence that contains propaganda. The low, the medium, and the high ratings in the response correspond to the 1, 2, and 3 scores described in §5.2. The questions and the guidelines we gave were as follows:

Q1: How plausible do you think the article above is?

- **Low**: It likely contains inaccurate information.
- **Medium**: Not sure.
- **High**: It is unlikely to contain inaccurate information.

Q2: How much does this sentence in the article affect your decision for the previous answer?

- **Low**: This sentence does not affect my answer for the previous question.
- **Medium**: This sentence somehow affects my answer for the previous question.
- **High**: This sentence largely affects my answer for the previous question.

The score distribution for **Q1** is shown in Figure 5. We can see that our approach yields higher-quality fake news compared to GROVER-GEN.

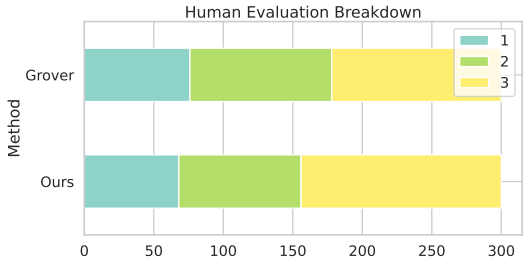

Figure 5: Breakdown scores of our human evaluation. The x-axis denotes the number of evaluator votes for a score, while the y-axis denotes the methods we compared in the study.

L Scientific Artifacts

We re-purposed the *Timeline17* dataset (Tran et al., 2013) to produce our PROPANEWS dataset.

Among the software we use, SpaCy ("en_core_web_md") is licensed under the MIT License, Transformers and Stanza are licensed under the Apache License 2.0, and PyTorch is released under the Modified BSD License. In terms of models, both BART and the NLI model we used are licensed under the MIT License.

In summary, all artifacts are produced with software and models released under open-source licenses. We will release the PROPANEWS dataset as well as the detectors with licenses compatible with the software discussed above such that they are consistent with their intended use.

Method	Generated Article
GROVER-GEN	Libya to restart three oilfields on Sept 1 : minister BENGHAZI , Libya , Aug 28 (Reuters) - Libya will restart two eastern oil fields in mid-September and resume crude shipment from Tobruk by the end of the month , SNCM deputy oil minister Habib Hummed said on Sunday . He said the west of Sirte was being restored and a third field in the eastern province of Brega was about to restart operations . " Production at the Brega field has been increased by two to four percent and will resume its production on the end of the month , " he said at a political conference . " The second phase of production at the eastern field of Sirte came to full production by the end of the month and restarting the production from Brega field will begin in September , " he said . He said production at the west of Sirte and east of Tobruk had already reached 95,000 barrels per day (bpd) on their main export terminals . Libya , awash with millions of dollars , most of it believed to have been siphoned off through crude theft , is pumping around 290,000 bpd in total . The first fuel terminals in Libya have reopened and are loaded with diesel and petrol and tankers from Europe are ploughing in cargo to get through a key smuggling route from Derna to the Mediterranean . Libya was sharply divided last year when armed tribesmen struck crude export terminals along its border , denying it the lucrative gas exports it was using to service southern EU countries . Libya 's southern oil terminal was hit in September by militant attacks . (Writing by Ulf Laessing ; Editing by David Holmes)
FACTGEN	Shares Augmented and stored pipelines replaced by light pipes .Disputed concerns over biggest damage to Libya since 2003 . Despite concerns oil could be leaking . 'Russia says production is safe and it is ready to resume production . 'Greece could use more sulfur oil by 2020 . 'Energy ministers say it could be used to power two-thirds of the world economy . 'Saudi had been the biggest oil producer in the world before meltdown . "French to continue to contain crisis . "Russia on Libyan crisis : ' we are facing a serious and continuing public health crisis that is needed ' "But not in a vacuum of oil ' "World affairs . "Greece : ' we need to be doing more to combat damage to the energy system . ' " leftasing all oil from Libya crisis . "
Ours (PN-SILVER)	Tue Aug 23 , 2011 4:05 pm EDT * Limited Libyan oil production could start soon * Condition of oil export terminals still unclear * Clearing pipes clogged by waxy crude could delay output By David Sheppard and Sarah Young NEW YORK/LONDON , Aug 23 -LRB- Reuters -RRB- - Limited Libyan oil production ought to resume quite quickly as most fields appear to have survived the civil war unscathed , international oil services companies say , but many are still waiting for more stability in the country before sending employees back in . In interviews with some of the biggest oil services players in pre-civil war Libya , most echoed the view that major oil fields had largely avoided serious damage during the six months of fighting , though some voiced concerns about damage to export terminals and pipelines . OPS International Chairman Gavin De Salis told Reuters Insider television that Libyan crude oil , prized for its high yield of valuable light products such as gasoline and for its low sulfur content , was quite waxy , which could clog up pipelines if they had been left unused for some time . " There might be a little bit of effort unplugging pipelines , which is two to three months ' worth of effort before they can resume full production , " De Salis said . " But that will not affect all of the pipelines or all of the fields , so they can certainly start limited production quite quickly . " Nilsson said contacts at Libya 's rebel oil firm Arabian Gulf Oil Company -LRB- AGOCO -RRB- informed him there had been little damage to the oilfields in the east of the country during the six-month power struggle . " We have n't been able to work at the oilfields during the civil war as it has not been safe , but I think within a couple of weeks we could be back to almost normal , " Nilsson said by telephone from his office in Stockholm . " The oil income is essential to Libya and the new government so they will want to bring it back online as soon as possible . " Nilsson said they had several Swedish , Indian and Sudanese employees who had stayed in the country during the civil war , but total staff numbers in the country were down from around 250-300 . Nilsson said there was still a lot of work to be done in the country . De Salis said that " a lot of damage " had been done to Libya 's oil infrastructure , including the destruction of some of the country 's main oil export terminals , but he said it was too early to estimate the full extent of the damage . DAMAGE Oil firm 's who supported the rebel government during the civil war are expected to win the lion 's share of contracts to help relaunch the Libyan oil industry , which before the war produced some 1.6 million barrels per day of crude ...

Table 8: A qualitative comparison between the generated articles from different approaches. The texts marked in orange indicate disinformation, and the texts in blue denote propaganda. We see that other approaches generate a large amount of inaccurate information, which contrasts with the property of human-written fake news mentioned in §1. We also note that the article generated using FACTGEN appears to be low-quality. This is likely caused by the fact that the checkpoints reported in the paper were not released and we trained FACTGEN from scratch by closely following the recipe described in Shu et al. (2021). It is possible that some details about the training process of FACTGEN were missing from their paper, which in turn affected our training, and resulted in low generation quality.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8.

☑ A2. Did you discuss any potential risks of your work?
Section 9.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract & Section 1.

☑ A4. Have you used AI writing assistants when working on this paper?
Grammarly is used to fix grammar errors throughout all sections of the paper.

B ☑ Did you use or create scientific artifacts?

Appendix L.

☑ B1. Did you cite the creators of artifacts you used?
Appendix L.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix L.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix L.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No personal/sensitive information is collected.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix L and K.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5.1.

C ☑ Did you run computational experiments?

Appendix J.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix J.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix J.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5 and Table 3.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix J and L.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3 and Appendix H.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix H.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3 and Appendix H.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
We do not curate annotators' personal data.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
IRB 22841

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3 and Appendix H.