Style-News: Incorporating Stylized News Generation and Adversarial Verification for Neural Fake News Detection

Wei-Yao Wang, Yu-Chieh Chang, Wen-Chih Peng
Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan
sf1638.cs05@nctu.edu.tw, jessie86915@gmail.com, wcpengcs@nycu.edu.tw

Abstract

With the improvements in generative models, the issues of producing hallucinations in various domains (e.g., law, writing) have been brought to people's attention due to concerns about misinformation. In this paper, we focus on neural fake news, which refers to content generated by neural networks aiming to mimic the style of real news to deceive people. To prevent harmful disinformation spreading fallaciously from malicious social media (e.g., content farms), we propose a novel verification framework, Style-News, using publisher metadata to imply a publisher's template with the corresponding text types, political stance, and credibility. Based on threat modeling aspects, a style-aware neural news generator is introduced as an adversary for generating news content conditioning for a specific publisher, and style and source discriminators are trained to defend against this attack by identifying which publisher the style corresponds with, and discriminating whether the source of the given news is human-written or machine-generated. To evaluate the quality of the generated content, we integrate various dimensional metrics (language fluency, content preservation, and style adherence) and demonstrate that Style-News significantly outperforms the previous approaches by a margin of 0.35 for fluency, 15.24 for content, and 0.38 for style at most. Moreover, our discriminative model outperforms state-of-the-art baselines in terms of publisher prediction (up to 4.64%) and neural fake news detection (+6.94% $\sim$ 31.72%).

1 Introduction

In recent years, social media have been used as platforms for people to share information due to non-distance on the Internet. However, the amount of deceptive news has also increased from vicious social media such as content farms by changing some words from their templates; this problem has been widely tackled by detecting the veracity of the news (Tseng et al., 2022; Wang and Peng, 2022; Du et al., 2023). With the advancement of generative pre-trained models (e.g., OpenAI (2023)), the issues of hallucinatory contents have been raised in various domains, e.g., scientific writing (Alkaissi and McFarlane, 2023), law (Forbes, 2023). In this paper, we focus on neural fake news, which has become an emerging societal crisis (Shu et al., 2021; Fung et al., 2021; Pegoraro et al., 2023; Reuters, 2023), aiming to produce human-like news via AI models at scale to defraud humans (Fung et al., 2021). Therefore, it is crucial to develop verification techniques for defending against neural fake news[1].

The recent progress of neural fake news lies primarily in synthetic news generation. Early research on synthetic news generation relied on hand-written rules (Van der Kaa and Krahmer, 2014) or templates (Leppänen et al., 2017). With the proposed controllable text generation (CTG), text generation can be applied based on given attributes (Keskar et al., 2019; Zhang et al., 2022). Grover (Zellers et al., 2019) produces CTG on multi-field documents to create synthetic news, including domain, date, authors, headline, and body. However, Grover neglects inherent factual discrepancies, which are tackled by retrieving external facts to enhance output consistency (Shu et al., 2021).

Despite the above progress, there are two limitations in the previous work. First, existing approaches to neural fake news detection fail to contemplate style information[2]. In this paper, we focus on an unexplored facet of the style of news in neural news generation: **publisher** (e.g., CNN or BBC), which can be adopted as a template for vicious social media (e.g., content farms) to produce fake news that can attract readers to read news

[1]We follow (Zellers et al., 2019) in using the term neural fake news to address machine-generated fake news.

[2]We note that authors in (Zellers et al., 2019) can be viewed as style information but are too sparse to learn the patterns.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1531–1541

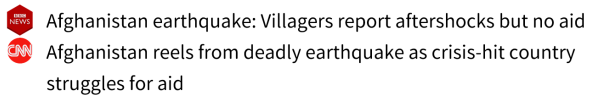

Figure 1: An example of news from different publishers.

from the corresponding publisher (Baptista and Gradim, 2020). For example, news content, political stance, and social engagements will be influenced by hyper-partisan publishers. Furthermore, different publishers are likely to describe an event with dissimilar content. As shown in Figure 1, we can observe that two publishers used different titles to describe the Afghanistan earthquake. The former states the event with the format *highlight: overview event*, whereas the latter uses a declarative sentence. These can be viewed as templates from specific publishers, where malicious groups are able to produce fake news based on the templates to deceive readers who often read specific news. Therefore, it becomes important to consider publisher information in synthetic neural news to detect it accurately before it is widely spread.

Second, previous work (e.g., Zellers et al. (2019); Shu et al. (2021)) evaluates the performance of defending neural fake news to classify the source of their generated news and the real news. We argue that the discriminators are trained to distinguish generated text from the corresponding generators, which makes the evaluation process unfair due to the fitting discriminators. It is essential to evaluate additional synthetic news that is not seen by models for fair comparisons.

In this work, we propose a novel framework, **Style-News**, with stylized news generation and two discriminators for publisher and neural fake news detection. Stylized news generation (SNG) is introduced to utilize publisher information as an explicit style for controllably generating human-like news content. To achieve fine-grained performance for SNG, the style discriminator is designed as a verifier for predicting the publisher of the generated content. In addition, neural fake news detection (NFND) is proposed to enhance the accuracy of distinguishing human-written and machine-generated news, which can be viewed as a news verifier. To tackle the second issue, we utilize the public dataset consisting of both synthetic and real news, VOA-KG2txt (Fung et al., 2021), which is generated by a separate model, to fairly verify the capability and robustness of our neural fake news detection and other baselines. The contributions of this paper are summarized as follows:

- We propose an adversarial framework with a threat modeling perspective to address the publisher-faceted issue of neural fake news. Meanwhile, the stylized news generation incorporates publisher information to produce style-adherence and human-like news content.
- To compare neural fake news detection fairly, we propose a fair evaluation pipeline by using an additional dataset instead of self-generated data to evaluate the robustness of our model and baselines. To the best of our knowledge, our work is the first to conduct comprehensive experiments for neural news generation and detection, which benefits future researchers with multi-dimensional performance aspects.
- Extensive experiments show that our generator significantly outperforms on multiple general news datasets in terms of fluency, content, and style qualities. Moreover, Style-News achieves a new state-of-the-art result on the neural fake news tasks, which demonstrates the effectiveness of our defense framework.

2 Related Work

Stylized text generation. Pre-trained language models (PLMs) have been widely adopted in various natural language tasks, which are trained on the large-scale corpus to have the ability to understand generic knowledge of text (Li et al., 2021). In recent years, generative PLMs have aimed to mimic the style of human beings to produce readable text from input prompts (Li et al., 2021). For instance, GPT-family (Radford et al., 2018, 2019; Brown et al., 2020; OpenAI, 2023) is a de facto generative model which achieves the robustness of text generation tasks. Accordingly, we adopted GPT-2 as the generation backbone following previous work.

Most of the research on stylized text generation puts efforts into the psycholinguistic aspect such as formal and casual with supervised settings (Wang et al., 2019; Verma and Srinivasan, 2019). However, supervised training requires a large amount of labeled data, which is difficult to generalize to practical tasks. Dathathri et al. (2020) tackled this issue by integrating a PLM with attribute classifiers to construct controlled text generation without training on the language model. StyleLM (Syed et al., 2020) pre-trains a Transformer-based masked language model and fine-tunes on an author-specific

corpus using DAE loss. However, the style of information has not been addressed in the neural news generation, which produces human-like news efficiently to deceive people based on existing publisher templates.

Neural fake news detection. The issues of fake news detection have been widely discussed since fake news covers a wide range of topics that may influence the public's views, political motives as well as social engagements (Shu et al., 2017). With the improvement of the generative PLM, Zellers et al. (2019) identified the problems of neural fake news and developed verification techniques by constructing controllable news generation as an adversary, and exploring potential defenses to mitigate the threats. To tackle the limitations of contradiction or missing details between the generated news and input prompt, FactGen (Shu et al., 2021) and InfoSurgeon (Fung et al., 2021) are introduced to improve the consistency of synthetic news by incorporating external knowledge. Nonetheless, previous work failed to explore style information to prevent neural fake news with specific templates, while we incorporate publisher information to generate style-aware news content to demonstrate the great potential of using style information and the awareness to defend against misinformation.

3 Approach

3.1 Problem Statement

In this paper, we address the neural news detection problem in an adversarial setting similar to (Zellers et al., 2019). We denote the **attack** phase as stylized neural news generation, and the **defense** phase as source discrimination.

In the attack phase, the input sequence contains news content, highlights, and publisher information. The highlights can be a news title or summary based on the different datasets. The goal of the generator G_{style} is to produce news content that mimics the style of real news conditioning on a specified publisher, which cannot be distinguished by the source discriminator D_{source}. The human-like news N_M generated by G_{style} can serve as potential threats that help the source discriminator learn to defend against neural fake news.

In the defense phase, the source discriminator D_{source} aims to learn to distinguish if input news N is human-written or machine-generated as:

$$D_{source}(N) \rightarrow y; y \in \{H, M\}, \quad (1)$$

where H and M denote human-written and machine-generated news, respectively.

3.2 Style-News Framework

Figure 2 presents the model architecture of StyleNews, where a stylized news generation module aims to generate synthetic news with a style-aware generator and style discriminator by taking news title, summary, content, and publisher as inputs. The neural fake news detection module classifies the source according to whether the input news content is human-written or machine-generated to enable the model to identify neural news. With the adversarial training on the generator and source discriminator, we are able to build up a stronger generator to produce style-aware news content; meanwhile, we have designed a robust verification mechanism to detect neural fake news. Detailed descriptions of the model are provided as follows.

3.3 Stylized News Generation

The stylized news generation module aims to produce expressive news based on writing style, which has not been utilized in the previous work. To generate stylized news, the style-aware generator is introduced by using publisher, title or summary, and content, and incorporates the style discriminator to reinforce the threat modeling aspect.

Style-aware generator. Following (Zellers et al., 2019; Dathathri et al., 2020), we adopt GPT-2 (Radford et al., 2019) as the generator backbone to produce news content. However, GPT-2 cannot take news metadata (e.g., publisher) into account. Therefore, we convert the token sequence of the publisher, highlight, and content as text prompts with task-context tokens as shown in Figure 5. Formally, the prompt template of the input sentence is defined as:

$$S = \text{<|Start\_Publication|>}\ publisher \quad (2)$$
$$\text{<|End\_Publication|>}\ highlight\ \text{<sep>}\ content,$$

where <|Start_Publication|>, <|End_Publication|> and <sep> are denoted as special tokens for indicating the publisher information, and separator tokens, respectively. The special tokens <|Start_Publication|> and <|End_Publication|> enable the model to consider the importance of the publisher, which can also be controlled by different publishers during inferencing. We truncate the input to l tokens if the sequence length exceeds the maximum length.

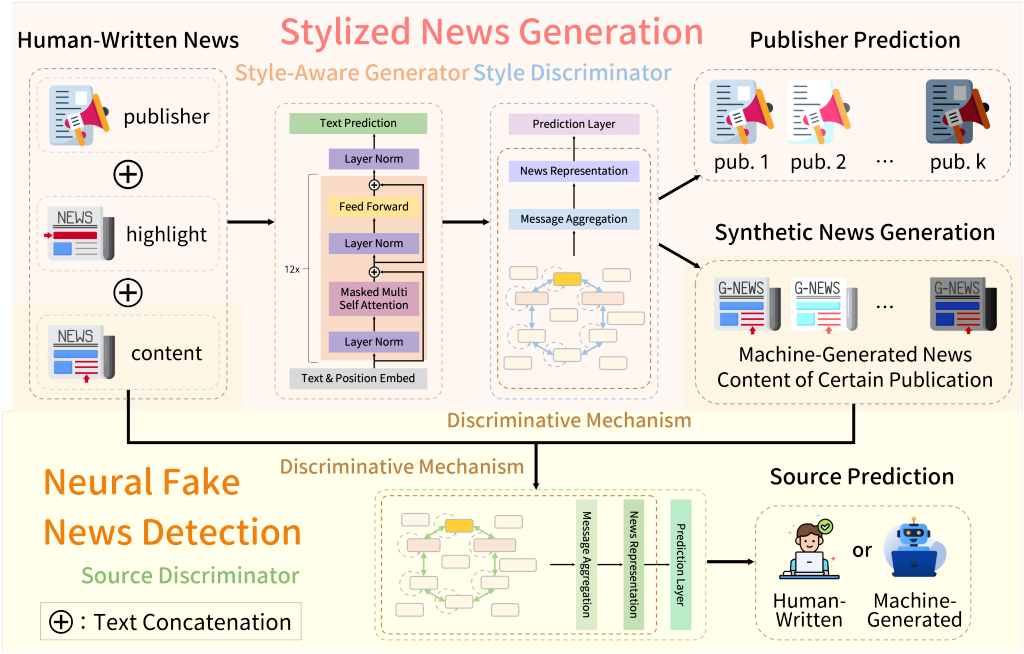

Figure 2: Illustration of the Style-News framework.

During the training stage, we randomly separate human-written news from the training set N_H into two groups for efficient training: the sampled group $N_H^{[sp]}$ and the unsampled group $N_H^{[usp]}$. $N_H^{[usp]}$ is used to train the parameters of the generative model and $N_H^{[sp]}$ is used to generate the synthetic news for training the source discriminator. Therefore, the input $N_H^{[usp]}$ contains news publisher, highlight, and content for the style-aware generator, and the objective function of G_{style} is defined as a language model problem:

$$L_{gen} = \sum_i (\log P(y_i|y_1, ..., y_{i-1})). \quad (3)$$

Discriminative mechanism (DM). The goal of the DM is to capture the representation of given news and distinguish between corresponding classes to reinforce the model quality. In the style discriminator D_{style}, the DM aims to identify which publisher the generated news belongs to. In the source discriminator D_{source}, the DM attempts to classify the source into human-written or machine-generated (this will be discussed in §3.4).

To capture the syntactic information, we propose a simple yet effective method by representing the input news (either human-written or machine-generated) in an inductive word graph as illustrated in Figure 3; this approach has been utilized in various text classification tasks (Zhang et al., 2020; Huang et al., 2022). Moreover, this design benefits our model generalizing to the unseen nodes compared with the common transductive graph models since the node embeddings in the word graph are initialized from the pre-trained word embeddings. Specifically, each token is represented as a node in the word graph, and each token embedding is converted from the GPT-2 pretrained model. Each node has two edges to connect with the former and latter tokens. This graph construction procedure enables the model to not only recognize the common tokens of the input sequence but also to capture the contextual information between tokens.

Formally, the i-th node w_i aggregate p hops neighbor information to encode the contextual representations as r'_{w_i}:

$$r'_{w_i} = (1-\alpha)\text{AGG}(\{r_{w_j}, w_j \in n(w_i)\}) + \alpha r_{w_i}, \quad (4)$$

where r_{w_j} is the node embedding, $n(w_i)$ is denoted as the p-hop neighborhood tokens of w_i, AGG is the message aggregation with max pooling, and $\alpha \in \mathbb{R}^1$ is a trainable weight for adjusting the importance between the node itself and the neighbor.

After updating each node embedding, the news representation r'_N is computed by aggregating node embeddings of news:

$$r'_N = \sum_{w_i \in N} r'_{w_i}, \quad (5)$$

Finally, the news representation r'_N is then fed into

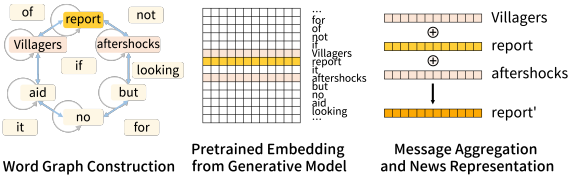

Figure 3: An example of the discriminative mechanism for *Villagers report aftershocks but no aid*.

a linear layer to predict the label:

$$\hat{y} = W' r'_N + b', \tag{6}$$

where $W' \in \mathbb{R}^{d_r \times d_c}$ is a matrix that maps the news representation into the number of classes (i.e., publisher or news source) and $b' \in \mathbb{R}^{d_c}$ is the bias.

To train the style discriminator, we minimize the cross-entropy loss $\mathcal{L}_{style}$:

$$\mathcal{L}_{style} = - \sum_{i \in \{pub1, \ldots, pubk\}} y_i \ln(\text{softmax}(\hat{y}_i)). \tag{7}$$

3.4 Neural Fake News Detection

As the generator is capable of creating various types of news content based on the publisher, a source discriminator is introduced to prevent high-quality synthetic news from maliciously spreading and further misleading the public.

Specifically, the source discriminator D_{source} adopts the same architecture as the DM. The input is randomly sampled from either human-written H or machine-generated M (by Style-News) news content for the D_{source}. We utilize the pretrained embeddings from G_{style} to build the word graph and train D_{source} by minimizing the cross-entropy loss for the class of news content:

$$\mathcal{L}_{source} = - \sum_{i \in \{H, M\}} y_i \ln(\text{softmax}(\hat{y}_i)). \tag{8}$$

3.5 Training Schedule

To construct an adversarial structure, we establish Style-News in a nested loop and jointly train the style-aware generator with the style discriminator and the source discriminator respectively. The training procedure is illustrated in Algorithm 1, where we train G_{style}, D_{style}, and D_{source} in order. Generally, the inner loop is stylized news generation (Line 4-13), and the outer loop is neural fake news detection (Line 3-16), which is able to align the latent space of these modules and thus meet the goal of threat modeling. We note that in the phase

Algorithm 1 Training procedure of Style-News.

Input: The human-written news with publisher, title, and content N_H; first and second stage epoch number $epochs_{style}$ and $epochs_{source}$

Output: Style-Aware Generator G_{style}, Style Discriminator D_{style}, and Source Discriminator D_{source}

1: Initialize the G_{style} from pretrained GPT-2 and add the special tokens; Initialize the parameters in D_{style} and D_{source} with Glorot uniform initializer
2: Randomly separate N_H into the sampled and unsampled groups $N_H = \{N_H^{[sp]}, N_H^{[usp]}\}$.
3: **for** $epoch_2$ = 1 to $epochs_{source}$ **do**
4: Train G_{style} by $N_H^{[usp]}$ via maximizing L_{gen} (Eq.3)
5: **for** $epoch_1$ = 1 to $epochs_{style}$ **do**
6: **if** $epoch_2$ == 1 **then**
7: Train D_{style} by $N_H^{[usp]}$ via minimizing $\mathcal{L}_{style}$ (Eq.7)
8: **else**
9: Train D_{style} by N_M via minimizing $\mathcal{L}_{style}$ (Eq.7)
10: **end if**
11: **end for**
12: Generate synthetic news N_M from $N_H^{[sp]}$
13: ▷ Stylized News Generation
14: Concatenate N_H and N_M and randomly shuffle them as the input to the D_{source}
15: Train D_{source} via minimizing $\mathcal{L}_{source}$ (Eq.8)
16: ▷ Neural Fake News Detection
17: **end for**

of training D_{style} (Line 5-11), the input is human-written news in the first epoch to equip the ability for understanding news content of real publishers (Line 7). Afterwards, the input is the synthetic news generated by G_{style} to detect the publisher of the generated news content (Line 9).

4 Experiments and Analysis

4.1 Dataset

We performed experiments on two news datasets that contain publisher metadata: CNN/DailyMail (Hermann et al., 2015; See et al., 2017) and All the News (Thompson, 2018). The CNN/DailyMail dataset is written by journalists at CNN and the Daily Mail, and contains over 300,000 unique news

Criteria	Metric	CopyTransformer	GPT-2	PPLM$_{gen}$	Grover$_{gen}$	FactGen	**Style-News**
Fluency	Mauve (↑)	0.7836	0.8050	<u>0.8827</u>	0.8314	0.7836	**0.8832**
	Frontier (↓)	0.9999	0.9300	<u>0.6634</u>	0.7299	0.9999	**0.6734**
Content	SacreBLEU (↑)	5.5527	8.1374	<u>14.7936</u>	0.3084	13.1285	**18.1064**
	MoverScore (↑)	0.5166	0.5369	0.5217	0.5010	<u>0.5434</u>	**0.5523**
Style	Accuracy (↑)	0.8918	<u>0.9273</u>	0.5949	0.8378	0.7392	**0.9609**
	F1 (↑)	0.5205	0.6898	0.5303	<u>0.8379</u>	0.5000	**0.8792**
Avg. Rank		5.0	<u>3.3</u>	<u>3.3</u>	4.0	4.3	**1.0**

Table 1: Performance of synthetic news generation on CNN/DailyMail. The best result in each row is in boldface and the second best result is underlined.

articles and highlight sentences. The training, validation, and testing sets are used as the official split. The All the News dataset encompasses 143,000 articles and essays from 15 American publishers. We picked the data with the common top 5 publishers (NPR, New York Post, Reuters, Washington Post, and Breitbart[3]) to ensure that the news of publishers has sufficient news to show the corresponding template patterns.

To defend against neural fake news, we follow (Zellers et al., 2019; Shu et al., 2021) to test our source discriminator on machine-generated data. However, previous evaluations only measured the effectiveness on their own self-generated datasets, which failed to measure the robustness of their discriminators. Therefore, we utilized a public dataset containing both human-written and machine-generated news, VOA-KG2txt (Fung et al., 2021), to fairly examine the discriminative performance of our models and the baselines. VOA-KG2txt includes 15,000 real news articles from Voice of America and 15,000 neural fake news articles produced by the KG-to-text approach (Fu et al., 2020). The testing set of these datasets is balanced. The statistics of the datasets are described in Table 6. All the results are the average of 5 random seeds.

4.2 Implementation Details

The word representation dimension d_r is set to 768. For training the style-aware generator, we set the learning rate to 5×10^{-5}, warmup steps to 1000, and weight decay to 0.01. The batch size in the training phase and generation phase was set to 2 and 32 respectively. For training the style discriminator and source discriminator, we used the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 10^{-3}, and weight decay was set to 10^{-4}. Dropout with a probability of 0.1 was applied after the linear layer. The max length of the token sequence l was restricted to 1024. The token would be converted to <UNK> special token if it does not match the dictionary of the pretrained model. The number of hop p is set to 1. The $epochs_{style}$ and $epochs_{source}$ in Algorithm 1 were set to 10 and 5 respectively. For the baseline models, we used default parameter settings as in their official implementations. All the training and evaluation phases were conducted with Pytorch 1.7 on a machine with Ubuntu 20.04, Intel(R) Xeon(R) Silver 4110 CPU, and Nvidia GeForce RTX 2080 Ti GPU.

4.3 Results of the Generative Models

Generative baselines. We selected 5 neural fake news generative baselines in this experiment to compare the generation quality of our Style-News. Specifically, we compared CopyTransformer (See et al., 2017), GPT-2 (Radford et al., 2019), PPLM$_{gen}$ (Dathathri et al., 2020), Grover$_{gen}$ (Zellers et al., 2019), and FactGen (Shu et al., 2021) for all the generative settings.

Evaluation metrics. Since there is no existing work considering different facets of generation quality[4], we introduce three evaluation facets to assess the quality of generated news content: **language fluency:** Mauve score (Pillutla et al., 2021) and Frontier Integral (Liu et al., 2021), **content preservation:** SacreBLEU (Post, 2018) and MoverScore (Zhao et al., 2019), and **style adherence:** RoBERTa-large with the training sets of CNN/DailyMail and All the News. Detailed descriptions are introduced in Appendix A.2.2.

[3]Breitbart is known for publishing conspiracy theories, which can be further examined for the generation quality of the fake news publisher (Higdon, 2020).

[4]We note that Shu et al. (2021) failed to consider the style aspect as evaluation, and the BLEU score is more suitable for content preservation instead of language fluency since repeated patterns have a larger score.

Criteria	Metric	CopyTransformer	GPT-2	PPLM_{gen}	$\text{Grover}_{\text{gen}}$	FactGen	**Style-News**
Fluency	Mauve (↑)	0.7849	0.8508	0.8707	<u>0.8847</u>	0.7756	**0.8881**
	Frontier (↓)	0.9966	0.7764	0.6865	<u>0.6642</u>	1.0000	**0.6467**
Content	SacreBLEU (↑)	0.3338	2.7831	<u>11.3883</u>	1.0472	6.2606	**14.5186**
	MoverScore (↑)	0.4942	0.5189	**0.5519**	0.5218	0.5304	<u>0.5448</u>
Style	Accuracy (↑)	0.2984	0.4721	0.5571	<u>0.5793</u>	0.4828	**0.5937**
	F1 (↑)	0.1478	0.3158	<u>0.4822</u>	0.4733	0.4136	**0.4921**
Avg. Rank		5.7	4.3	<u>2.3</u>	2.8	4.2	**1.2**

Table 2: Performance of synthetic news generation on All the News. The best result in each row is in boldface and the second best result is underlined.

Generation performance. Table 1 and Table 2 present the quality of the generation results in terms of language fluency (fluency), content presentation (content), and style adherence (style)[5]. We can observe that Style-News consistently outperforms the generative baselines by a margin of 0.35 for fluency, 15.24 for content, and 0.38 for style at most for both datasets, which demonstrates the realistic-looking performance of our generated news content. We summarize the observations as follows.

1) Using pre-trained models (i.e., GPT-2, PPLM_{gen}, $\text{Grover}_{\text{gen}}$, Style-News) to generate news content improves generation performance in terms of language fluency, which indicates the significance of incorporating prior knowledge from the pre-trained data. 2) We observed that the controllable text generative models (i.e., PPLM_{gen}, $\text{Grover}_{\text{gen}}$, Style-News) perform better on the style adherence aspect since non-controllable models fail to take the publisher information into account. Therefore, our style-aware generator integrating publishers into the prompt to manipulate the style of news is superior to these baselines. 3) It is worth noting that all baselines are biased to generate human-like news content for only some facets, which indicates that they often focus on only specific aspects. With the threat modeling design for the style-aware generator and style discriminator, our Style-News is capable of getting a human-like text with all criteria from the better-detected discriminator.

4.4 Results of the Discriminative Models

Discriminative baselines. To validate the performance of our proposed discriminators (including publisher and neural fake news detection), we further conducted experiments on publisher prediction and neural fake news detection with strong discriminative baselines: RoBERTa (Liu et al., 2019), PPLM_{def} (Dathathri et al., 2020), $\text{Grover}_{\text{def}}$ (Zellers et al., 2019), GET (Xu et al., 2022), as well as CoCo (Liu et al., 2022). To investigate the effectiveness of feature-based methods, we follow the setting as (Aich et al., 2022) to add Linear Regression (LR), SVM, Ridge Regression (RR), KNN, and Random Forecast (RF) as machine learning baselines.

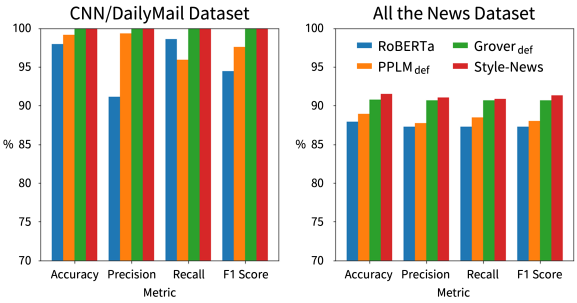

Figure 4: Performance of publisher prediction on CNN/DailyMail and All the News.

Evaluation metrics. We adopt the common classification metric, macro F1 score, for measuring both publisher and neural fake news classifications. We set the training epoch to 5 and selected the best model evaluating the validation set for all discriminative experiments.

Publisher prediction. To examine the capability of distinguishing publishers of news, we carried out experiments with the discriminators to classify the publishers of real news. Figure 4 demonstrates the correctness of predicting the corresponding publisher given the news content[6]. Our Style-News outperforms the baselines by up to 4.64% on All the News, and classifies perfectly on CNN/DailyMail, which demonstrates that jointly training the style-aware generator and style discriminator enables the

[5]The generation samples are discussed in Appendix B.3.

[6]GET is neglected due to the gradient explosion while training on CNN/DailyMail and All the News.

LR	SVM	RR	KNN	RF	RoBERTa	PPLM_{def}	$\text{Grover}_{\text{def}}$	GET	CoCo	Style-News
72.35	75.84	68.07	81.10	80.67	92.15	89.83	74.82	82.45	93.82	**98.55**

Table 3: The F1 scores of neural fake news detection.

Criteria	Metric	w/o Style w/o Source	w/o Style	w/o Source	**Style-News**
Fluency	Mauve (↑)	0.8045	0.8047	0.8164	**0.8832**
	Frontier (↓)	0.9317	0.9310	0.8924	**0.6734**
Content	SacreBLEU (↑)	8.1374	8.1405	7.9399	**18.1064**
	MoverScore (↑)	0.5369	0.5371	0.5345	**0.5523**
Style	Accuracy (↑)	0.9266	0.9253	0.9284	**0.9609**
	F1 Score (↑)	0.6842	0.6822	0.7127	**0.8792**

Table 4: Ablation study on CNN/DailyMail.

model to understand the publisher template. The observations are summarized as follows:

1) All models exhibit almost perfect performance on CNN/DailyMail since there are only two publishers, while the prediction gap becomes large on All the News. 2) RoBERTa hinders the performance compared with controllable generative models (i.e., PPLM_{def}, $\text{Grover}_{\text{def}}$, and Style-News), which indicates that additional information during training generators helps the corresponding discriminators to distinguish style information (e.g., publisher in this paper). 3) Both $\text{Grover}_{\text{def}}$ and Style-News achieve perfect performance on CNN/DailyMail but Style-News is superior to $\text{Grover}_{\text{def}}$ on All the News. This comparison reveals the importance of considering publishers in the generator as well as using the discriminative mechanism.

Neural fake news detection. To defend against synthetic fake news, we conducted experiments to examine the robustness of our Style-News. We utilized VOA-KG2txt as the evaluated dataset to draw a fair comparison between our model and the baselines. Table 3 shows the performance of discriminative models on detecting the source of the input news content. Specifically, our model surpasses all the baselines from 6.94% to 31.72%. We conclude the observations as follows:

1) The models with the word graph (i.e., GET and Style-News) are superior to $\text{Grover}_{\text{def}}$, which verifies that the word graph can capture the syntactic meanings as structural information. 2) GET has substantially worse performance on neural fake news detection tasks since it takes claim-evidence interactions while there is no precise evidence of neural fake news in the real world. In addition, $\text{Grover}_{\text{def}}$ suffers from degradation performance due to the sparsity learning from author information. Attributed to capturing publisher style from the news, our model is thus able to effectively distinguish neural news. 3) All baselines degrade their performance in detecting neural news on the additional dataset, while our model consistently detects almost perfectly. This suggests that evaluating the classification only on their generated news fails to measure the robustness of unseen news since the discriminators did not train on them. Our model, in contrast, is still capable of classifying the news as either machine-generated or human-written, which can be used to not only defend against self-generated news but also against existing neural fake news.

4.5 Result Analysis

Ablation study. To quantify the contributions of different discriminators of Style-News, we further conducted ablation experiments on CNN/DailyMail. As shown in Table 4, it is obvious that removing any discriminator (w/o Style and w/o Source) results in a significant performance drop in terms of all generated aspects. Also, as expected, only using the generator leads to inferior performance in all metrics. These results illustrate the reasonable and effective design of our model. In addition, without the assistance of the style discriminator (w/o Style), the performance drops significantly in style adherence in terms of an F1 score of 0.21, indicating that the style discriminator can help enhance the ability to capture the writing style of the corresponding publisher.

Human evaluation. We randomly sampled 3 generated news articles of each model from both

Criteria	CopyTransformer	GPT-2	PPLM_{gen}	$\text{Grover}_{\text{gen}}$	FactGen	**Style-News**
Language (↑)	1.00	1.67	1.67	**2.33**	1.67	**2.33**
Content (↑)	1.00	1.67	1.89	2.11	2.00	**2.33**
Style (↑)	1.11	1.67	1.56	2.00	1.78	**2.44**

Table 5: The human evaluations of generated samples in both the CNN/DailyMail and the All the News datasets.

CNN/DailyMail and All the News, which were annotated by 9 annotators without advanced knowledge of the source of the generated content to reflect real-world reader scenarios. They were asked to evaluate the generated news in terms of language fluency, content preservation, and style adherence. We provided some sample news from the corresponding publisher to let annotators evaluate style adherence. The details of human evaluation questions were designed similarly to (Shu et al., 2021), i.e., the annotators should evaluate each question with a score of 1 (the worst) to 3 (the best).

Table 5 lists the human evaluation results, which illustrate that our Style-News significantly outperforms all generative baselines in terms of all three aspects. Quantitatively, our approach achieves 17% and 37% performance improvement over the best baseline in content preservation and style adherence, respectively. This again reveals the enhancement of considering publisher information for stylized news generation.

5 Conclusion

This paper presents Style-News, a novel adversarial framework to defend against the urgent neural fake news problem. Distinct from existing generative models, our style-aware generator produces news with text prompts not only from news highlights and content, but also from publisher information, allowing the integration of additional metadata in the realm of text-metadata compositions. Meanwhile, our neural fake news detection captures syntactic information by constructing the input as a graph for distinguishing the human-like news content. Style-News sets new state-of-the-art results on both neural news generation and detection benchmarks with our comprehensive metrics. We believe Style-News serves as a flexible framework for neural fake news detection, and multiple interesting directions could be further explored within the framework, such as prompt design, few-shot examples, etc.

6 Ethics Considerations

We discuss the potential usage and the potential risks of Style-News for ethical considerations.

Journalism assistants Inspired by (Shu et al., 2021) who discussed helping journalists with claim generation using their fact retrieval mechanism, our method can provide alternative perspectives and inspire journalists to enrich their news content. However, the results generated by Style-News can only serve as a reference and cannot be used directly.

Veracity of machine-generated news Following (Zellers et al., 2019), one of the goals of Style-News is to detect machine-generated news. This task is necessary based on a strong assumption that machine-generated news is fake and can be harmful to the public. Nonetheless, as we mention above, machine-generated news can also be regarded as a template or an inspiration for journalists. Therefore, we suggest that future work verify the factual claims of machine-generated news and release open-source datasets generated by different algorithms or researchers to construct stronger detectors.

7 Limitations

The major limitation of Style-News is the machine-generated news with further human modifications, i.e., multi-hop modifications. The manual rewriting can be regarded as another various style, which increases the difficulty of neural fake news detection. In addition, Style-News focuses on effective performance to mitigate the spread of neural fake news, but does not take the computation resource into account, which may be more efficient by introducing adapters into the model.

Acknowledgments

We thank the anonymous reviewers for their insightful comments and feedback. This work was partially supported by the Ministry of Science and Technology of Taiwan under Grants 112-2917-I-A49-007.

References

Ankit Aich, Souvik Bhattacharya, and Natalie Parde. 2022. Demystifying neural fake news via linguistic feature-based interpretation. In COLING, pages 6586–6599. International Committee on Computational Linguistics.

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. Cureus, 15(2).

João Pedro Baptista and Anabela Gradim. 2020. Understanding fake news consumption: A review. Social Sciences, 9(10):185.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In NeurIPS.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In ICLR. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT (1), pages 4171–4186. Association for Computational Linguistics.

Wei-Wei Du, Hong-Wei Wu, Wei-Yao Wang, and Wen-Chih Peng. 2023. Team triple-check at factify 2: Parameter-efficient large foundation models with feature representations for multi-modal fact verification. In DE-FACTIFY@AAAI, volume 3555 of CEUR Workshop Proceedings. CEUR-WS.org.

Forbes. 2023. Lawyer used chatgpt in court—and cited fake cases. a judge is considering sanctions. https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/?sh=3308159a7c7f.

Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. 2020. MAGNN: metapath aggregated graph neural network for heterogeneous graph embedding. In WWW, pages 2331–2341. ACM / IW3C2.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen R. McKeown, Mohit Bansal, and Avi Sil. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In ACL/IJCNLP (1), pages 1683–1698. 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 NIPS, pages 1693–1701.

Nolan Higdon. 2020. The anatomy of fake news. https://www.ucpress.edu/book/9780520347878/the-anatomy-of-fake-news.

Yen-Hao Huang, Yi-Hsin Chen, and Yi-Shin Chen. 2022. Contexting: Granting document-wise contextual embeddings to graph neural networks for inductive text classification. In COLING, pages 1163–1168. International Committee on Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. CoRR, abs/1909.05858.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In ICLR (Poster).

Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. 2017. Data-driven news generation for automated journalism. In INLG, pages 188–197. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Pretrained language models for text generation: A survey. CoRR, abs/2105.10311.

Lang Liu, Krishna Pillutla, Sean Welleck, Sewoong Oh, Yejin Choi, and Zaïd Harchaoui. 2021. Divergence frontiers for generative models: Sample complexity, quantization effects, and frontier integrals. In NeurIPS, pages 12930–12942.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Yu Lan, and Chao Shen. 2022. Coco: Coherence-enhanced machine-generated text detection under data limitation with contrastive learning. CoRR, abs/2212.10341.

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. CoRR, abs/1907.11692.

OpenAI. 2023. GPT-4 technical report. CoRR, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In ACL, pages 311–318. ACL.

Alessandro Pegoraro, Kavita Kumari, Hossein Fereidooni, and Ahmad-Reza Sadeghi. 2023. To chatgpt, or not to chatgpt: That is the question! CoRR, abs/2304.01487.

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 NeurIPS, pages 4816–4828.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In WMT, pages 186–191. Association for Computational Linguistics.

Piotr Przybyla. 2020. Capturing the style of fake news. In AAAI, pages 490–497. AAAI Press.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

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.

Reuters. 2023. Google cautions against 'hallucinating' chatbots, report says. https://www.reuters.com/technology/google-cautions-against-hallucinating-chatbots-report-2023-02-11/.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In ACL (1), pages 1073–1083. Association for Computational Linguistics.

Kai Shu, Yichuan Li, Kaize Ding, and Huan Liu. 2021. Fact-enhanced synthetic news generation. In AAAI, pages 13825–13833. AAAI Press.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. SIGKDD Explor., 19(1):22–36.

Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. 2020. Adapting language models for non-parallel author-stylized rewriting. In AAAI, pages 9008–9015. AAAI Press.

Andrew Thompson. 2018. All the news dataset. https://www.kaggle.com/datasets/snapcrack/all-the-news.

Yu-Wun Tseng, Hui-Kuo Yang, Wei-Yao Wang, and Wen-Chih Peng. 2022. KAHAN: knowledge-aware hierarchical attention network for fake news detection on social media. In WWW (Companion Volume), pages 868–875. ACM.

Hille Van der Kaa and Emiel Krahmer. 2014. Journalist versus news consumer: The perceived credibility of machine written news. In Proceedings of the computation+ journalism conference, Columbia university, New York, volume 24, page 25.

Gaurav Verma and Balaji Vasan Srinivasan. 2019. A lexical, syntactic, and semantic perspective for understanding style in text. CoRR, abs/1909.08349.

Wei-Yao Wang and Wen-Chih Peng. 2022. Team yao at factify 2022: Utilizing pre-trained models and co-attention networks for multi-modal fact verification (short paper). In DE-FACTIFY@AAAI, volume 3199 of CEUR Workshop Proceedings. CEUR-WS.org.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. In EMNLP/IJCNLP (1), pages 3571–3576. Association for Computational Linguistics.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In WWW, pages 2501–2510. ACM.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In NeurIPS, pages 9051–9062.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. CoRR, abs/2201.05337.

Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. 2020. Every document owns its structure: Inductive text classification via graph neural networks. In ACL, pages 334–339. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In EMNLP/IJCNLP (1), pages 563–578. Association for Computational Linguistics.