What Boosts Fake News Dissemination on Social Media? A Causal Inference View

Yichuan Li[1], Kyumin Lee[1], Nima Kordzadeh[1], and Ruocheng Guo[2]

[1] Worcester Polytechnic Institute, Worcester MA 01605, USA
{yli29,kmlee,nkordzadeh}@wpi.edu
[2] Bytedance Research, London, UK
rguo.asu@gmail.com

Abstract. There has been an upward trend of fake news propagation on social media. To solve the fake news propagation problem, it is crucial to understand which media posts (e.g., tweets) cause fake news to disseminate widely, and further what lexicons inside a tweet play essential roles for the propagation. However, only modeling the correlation between social media posts and dissemination will find a spurious relationship between them, provide imprecise dissemination prediction, and incorrect important lexicons identification because it did not eliminate the effect of the confounder variable. Additionally, existing causal inference models cannot handle numerical and textual covariates simultaneously. Thus, we propose a novel causal inference model that combines the textual and numerical covariates through soft-prompt learning, and removes irrelevant information from the covariates by conditional treatment generation toward learning effective confounder representation. Then, the model identifies critical lexicons through a post-hoc explanation method. Our model achieves the best performance against baseline methods on two fake news benchmark datasets in terms of dissemination prediction and important lexicon identification related to the dissemination. The code is available at `https://github.com/bigheiniu/CausalFakeNews`.

Keywords: Causal inference on text · Fake news propagation

1 Introduction

People often create various news related posts on social media platforms (e.g., sports, politics and finance), and the posts are shared by their friends or influencers, and are re-shared by other users as illustrated in Fig. 1(a). Some posts start a viral "chain reaction" [11], which amplifies the influence of the news. Fake news get the same benefit, and some users intentionally or unintentionally rephrase and summarize the fake news content to encourage other users to share them via social networks.

To increase the dissemination of a news post/tweet (e.g., number of retweets), posters may create the posts more clickbait [18], and use fewer jargon words [32] in them. However, these insights are mainly based on observations or statistical correlations between social media posts and corresponding quantity of engagements, and sometimes these insights may fail [18]. For example, news topics affect the posters' writing style and tendency to share [1]. Posts created by posters, who have many followers, intrinsically receive more share than ones who have fewer followers [33]. Thus, merely capturing the correlation between observed properties and dissemination of a given news post is a less robust estimation, and limits finding meaningful patterns/true causes

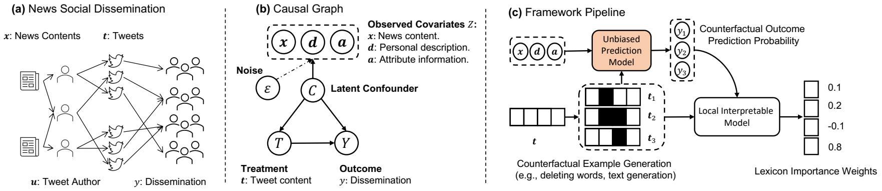

Fig. 1: (a) Overview of fake/real news dissemination on social media. A tweet author **u** imports news **x** from external websites and writes a tweet **t** to attract other Twitter users to disseminate **y**. (b) The causal graph contains hidden confounder C, noise ϵ, retweet status label **y**, and observed covariates Z, which consist of news content **x**, tweet author's profile information (**a**, **d**) and tweet content **t**. (c) The pipeline of identifying the important tweets' lexicon which causes fake/real news get disseminated.

To overcome the aforementioned limitations and understand news dissemination via social media (e.g. Twitter), we aim to answer the following research questions: **RQ1.** which news tweet[3] will receive more retweets? and **RQ2.** what textual features (lexicons) given a tweet corpus play decisive roles for the news dissemination via social networks? To answer **RQ1.**, we build a structural causal model (SCM) [19] as shown in Fig. 1(b) to model the causal relationship between news tweet and its dissemination (i.e., number of retweets). The SCM contains a hidden confounder C that influences both the probability of receiving treatment T (tweet) and outcome Y (a level of retweets). In particular, we go beyond correlation prediction $P(Y|T = \mathbf{t})$ and propose to use $P(Y|do(T = \mathbf{t}))$. By using *do*-operation, SCM can reduce the spurious correlations caused by confounder C (e.g. news topic) in news tweet dissemination prediction Y, leading to unbiased prediction. Since we do not have direct access to hidden confounder C, we adapt the idea of proximal variables [16, 15]. We assume an approximate measurement can model C, given the observed covariates Z, including news content and tweet poster's personal information. To answer **RQ2.**, we follow the previous work [7] by utilizing a post-hoc explanation method to interpret our model's prediction. The whole procedure is illustrated in Fig. 1(c).

Existing works of causal inference on text still cannot completely answer **RQ1.**. They can only handle either numerical [24, 29, 25] or textual [7, 23] covariates instead of both simultaneously. Other previous works [6, 18] extracted latent properties (e.g. sentiment) from the text and treated these properties as the treatment. Usually, the latent properties are binary or continuous scalar values. This approximation would not only propagate the error from the property extraction model to the causal inference model, but also cannot answer which specific post/tweet causes the fake/real news dissemination on social media. Although existing works [24, 25, 7] can model high-dimensional textual treatment, they mainly relied on the multi-layer perception (MLP) for modeling the causal relationship.

[3] A news tweet means a tweet mentions a certain news.

For both high-dimensional confounder and treatment, the MLP lacks the scalability, expressiveness and generalizability [5].

Thus, we propose a novel causal inference model based on the transformer model [34]. To represent *multimodal covariates*, our model adopts the soft prompt learning [14] to align the numerical and textual features. To represent *high-dimensional textual treatment*, the proposed model takes the treatment's raw text as input. The cross-attention inside the transformer naturally provides a way to capture the *complex relationship* between the high-dimensional confounder and treatment. Besides, we propose a two-stage training strategy to better model the dependency among covariates, confounder, treatment, and outcome. The two-stage training strategy consists of (i) *conditional treatment generation* and (ii) *outcome inference*. We evaluate effectiveness of our model in terms of robustness and explainability in two fake news benchmark datasets.

In short, this work has the following contributions: *i)* We propose a causal inference model that handles multimodal covariates and textual treatments, and estimates the outcome robustly; *ii)* We unbiasedly understand which lexicons inside tweets boost fake/real news dissemination on social media; *iii)* Our model achieves the best quantitative and qualitative results on the benchmark datasets. Experiments include the adjustment of different data distributions and interpreted lexicon explanation evaluation.

2 Problem Definition

Notation. Let boldface lowercase letter denote a vector or a sequence of words (e.g., $\mathbf{x}$), and boldface uppercase letter represent the matrix (e.g., $\mathbf{X}$), italic uppercase letter denote a causal inference variable (e.g., T), italic lowercase letter denote a word (e.g., w), and a calligraphic font represent a vocabulary set (e.g., $\mathcal{V}$). Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{|\mathbf{X}|}$ denote a corpus of news contents. For each news content $\mathbf{x}_i$, there are $\mathbf{T}_i = \{\mathbf{t}_j^i\}_{j=1}^{|\mathbf{T}_i|}$ tweets that mention the news $\mathbf{x}_i$. An author of the tweet $\mathbf{t}_j^i$ is $\mathbf{u}_j^i$, and her profile consists of numerical attributes $\mathbf{a}_j^i$ and personal textual description $\mathbf{d}_j^i$. It should be noticed that each of $\mathbf{x}_i$, $\mathbf{t}_j^i$ and $\mathbf{d}_j^i$ is a sequence of words $\{w_k\}_{k=1}^{||\cdot||}$. Each tweet $\mathbf{t}_j^i$ is associated with a discrete y_j^i retweet status (i.e., assigning each tweet to a class/bin based on the number of retweets). To answer the aforementioned two research questions in Sec. 1, we conduct the following studies:

- **Predict tweets causing fake/real news dissemination**. We aim to learn the causal relation $P(Y|do(T = \mathbf{t}))$ as shown in Fig. 1(b), where Y is the outcome, T is treatment. For more information about *do*-operation, please check Sec. 3.1.
- **Understand syntax playing decisive roles.** Then, we identify K important words $\{w_k\}_{k=1}^{K} \subseteq \mathcal{V}^{\mathbf{t}}$ that most significantly influence news tweets $\mathbf{T}$ dissemination. Here, $\mathcal{V}^{\mathbf{t}}$ is a vocabulary set of tweets.

3 Our Framework

Our proposed framework is designed based on a well-known sequence-to-sequence (seq2seq) model, `BART` [13], as shown in Fig. 2. In the following subsections, we will firstly introduce the preliminary knowledge. We then propose a two-stage training strategy

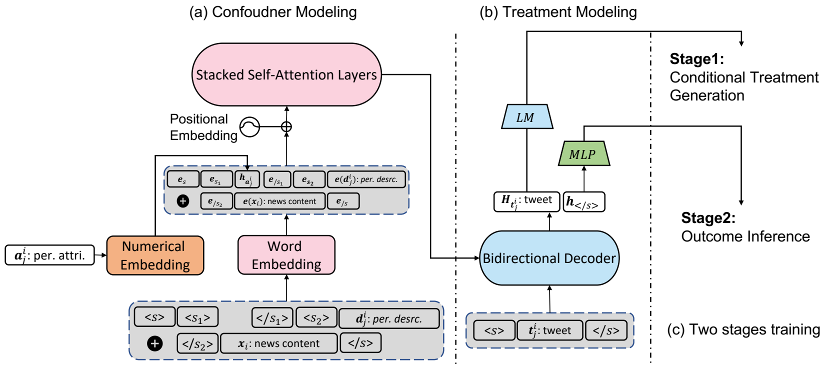

Fig. 2: The illustration of our model. The covariates (news content and the tweet author's profile) are input into the encoder and the treatment (tweet) is input into the decoder. The whole training process contains two stages. The first stage is to learn the hidden confounder's representation through conditional text generation. The second stage is to classify a level of news tweet dissemination by text classification.

to better capture the causal dependency showed in Fig 1(b). Lastly, we will discuss our approach for learning multimodal covariates embedding.

3.1 Preliminary

***Do*-Notation** ($P(Y|do(T=\mathbf{t}))$) [19] is different from conditional correlation $P(Y|T=\mathbf{t})$, which is based on the sub-population of the dataset. The *Do*-Notation will change the data distribution by making intervention [9], even if the sub-population is unseen in the collected data. For example, in the collected data, if users whose followers are greater than 10 never posting $\mathbf{t}$, the $P(Y|T=\mathbf{t})$ cannot answer *what-if* this group of users have tweeted t, how many retweets they will receive. But $P(Y|do(T=\mathbf{t}))$ will model all groups of users including follower greater than 10 have posted $\mathbf{t}$. The observation/collected data often lacks this intervention. To estimate $P(Y|do(T=\mathbf{t}))$, it often requires the model to block all the incoming paths to the treatment ($C \rightarrow T$ as shown in Fig. 1(b)). This incoming path blocking can be modeled as conditioning on the confounder C: $P(Y|do(T=\mathbf{t})) = \int_C P(Y|T=\mathbf{t}, C=\mathbf{c})P(C=\mathbf{c})d\mathbf{c}$.

BART [13] is a denoising transformer-based language model. Both encoder and decoder of BART are stacked with the Transformer blocks. Each block contains a self-attention layer to interact with and aggregate the information from either the encoder or decoder: $Self-Attention(\mathbf{Q},\mathbf{K},\mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\mathbf{V})$. But the decoder has an additional cross attention layer, where $\mathbf{Q}=\mathbf{Q}_{dec}$, $\mathbf{K}=\mathbf{K}_{enc}$, $\mathbf{V}=\mathbf{V}_{enc}$. This difference allows us to use different inputs (see Sec. 3.3) for the encoder (multimodal covariates as input) and decoder (treatment as input) for better feature interaction modeling.

3.2 Causal Feature Representation Learning

We aim to learn the causal relation $P(Y|do(T=\mathbf{t}))$ to predict how dissemination (Y) would be changed if a tweet (T) had been written differently about the same news by the same author (Z). Based on the backdoor criterion [19] and the causal graph Fig. 1(b), we can predict each tweet's dissemination status by $P(Y|T = \mathbf{t}, Z) = \int_C P(Y|T, C)P(C|Z)dC$. As the encoder is a deterministic function, given covariates $Z = \mathbf{z}$, we have $P(C = \mathbf{c}|Z = \mathbf{z}) = 1$ if $\mathbf{c} = enc(\mathbf{z})$, and $P(C = \mathbf{c}|Z = \mathbf{z}) = 0$ if $\mathbf{c}$ takes other values. This implies $P(Y|T, Z = \mathbf{z}) = P(Y|T, C = enc(\mathbf{z}))$. We separate the modeling of $P(Y|T, C = enc(\mathbf{z}))$ into two stages and optimize them sequentially.

Stage 1: Conditional Treatment Generation. As the approximation of confounder C, the covariates Z may contain irrelevant noise to the treatment T. To partial out the noise and learn better confounder representation, we design a new task - conditional text generation, generating the treatment tweets T conditioned on covariates Z. We consider the following conditional text generation model $P(T|Z) = \int_C P(T|C)P(C|Z)dC$. Based on the previous discussion, we can have $P(T|Z = \mathbf{z}) = P(T|C = enc(\mathbf{z}))$. Our model takes the embedding of multimodal covariates formalized in Sec. 3.3 and outputs the treatment words T autogressively as follows:

$$P(w_1, \cdots, w_{|\mathbf{t}_j^i|} \mid z = (\mathbf{x}, \mathbf{d}, \mathbf{a})) = \prod_{k=1}^{|\mathbf{t}_j^i|} P(w_k|C = enc(z), \mathbf{w}_{<k}) \tag{1}$$

where $w_{<k}$ is a sequence of previously generated $k - 1$ tokens.

Stage 2: Outcome Inference. Inspired by previous works [29, 23, 25, 7], to avoid the underrepresentation of treatment, we separate the location of input of covariates and treatment. The previous works do the outcome inference by assigning confounder's hidden representation into different MLPs [29, 23] or boosting the outcome prediction from confounder to treatment [25, 7]. Instead, in this work, the encoder takes the embedding of covariates as input, while the decoder takes treatment/tweet embedding as input. These isolated inputs for the encoder and decoder bring well-discriminated representation learning for confounder and treatment [36].

After that, we add a MLP f to the seq2seq model to predict outcome based on the interacted representation of treatment and covariates. Specifically, we take the hidden representation of $\langle/s\rangle$ token, $\mathbf{h}_{\langle/s\rangle}$, from treatment to predict outcome (i.e., a level of tweet dissemination). Usually, the $\langle/s\rangle$ is appended to the end of treatment. Overall, the objective function for outcome inference is:

$$L_{outcome} = \mathbb{E}\left[CrossEntropy\left(y_j^i, f(\mathbf{h}_{\langle/s\rangle})\right)\right] \tag{2}$$

3.3 Multimodal Covariates Embedding

To better learn representation of the hidden confounder C, we consider as many observed covariates as possible. In particular, besides the news content, we also consider a tweet author's personal information. Previous researches [33, 21] found that the number of followers and friends, status, lists, and user's verification status influenced the retweetability. Thus, our covariates consist of textual news content $\mathbf{x}$, personal description $\mathbf{d}$, and numerical attribute information $\mathbf{a}$ (i.e., *verified status*, and *number of tweets, followers count, friends count* and *number of lists*).

However, BART [13] is pre-trained on text only but not the tweet author's numerical data. To fill the gap between the pre-training and our downstream fine-tuning, in our framework, we utilize the soft-prompt [14] to align the modality. It wraps the input with unique tokens. These unique tokens do not need to be included in the vocabulary, and their embeddings are learned from scratch. During the training process, they will gradually capture the modality information. This setting will provide more flexibility and better coverage than actual tokens [26, 12]. Specifically, the input of our encoder is:

$$\left[\mathbf{e}_{</s>}, \mathbf{e}_{</s_1>}, \mathbf{h}_{\mathbf{a}^i_j}, \mathbf{e}_{</s_1>}, \mathbf{e}_{<s_2>}, \mathbf{e}(\mathbf{d}^i_j), \mathbf{e}_{</s_2>}, \mathbf{e}(\mathbf{x}_i), \mathbf{e}_{</s>}\right] \tag{3}$$

where $\{\mathbf{e}_{\cdot}\}$ are the soft-prompts to wrap different inputs; $\mathbf{h}_{\mathbf{a}^i_j}$ is the hidden vector of numerical features. $\mathbf{e}(\mathbf{d}^i_j)$ and $\mathbf{e}(\mathbf{x}_i)$ are the word embeddings for the tweet author's personal description, and news content, respectively.

4 Experiment

In this section, we answer the following research questions. **RQ 1.1.** How accurately can our model estimate a level of news tweet dissemination (i.e., predict the class/bin of the number of retweet given the news tweet)? **RQ 1.2.** What contribution does each observed covariate make for learning representation of the hidden confounder? **RQ 2.** What is the effectiveness of the identified words $\{w_k\}_{k=1}^{K}$ from the tweets $\mathbf{T}$ in terms of boosting the news dissemination?

4.1 Evaluation Datasets

We use two fake news benchmark datasets from FakeNewsNet [30]: (1) PolitiFact and (2) GossipCop. They contain news content, veracity labels (fake vs. real), and social context tweets (tweets mention the news). The statistical information of the tweets is listed in Table. 1. Since most of the tweets received zero or one retweet, we group the tweets based on their corresponding $\#\ of\ retweets$ into binary ($|reweets| = 0$ and $|reweets| > 0$), ternary ($|reweets| = 0$, $|reweets| = 1$ and $|reweets| > 1$), and quaternary ($|reweets| = 0$, $|reweets| = 1$, $1 < |reweets| <= 10$ and $|reweets| > 10$) classes as the ground truth. For example, in the ternary setting, class 0 is a tweet receiving 0 retweet, class 1 is a tweet receiving one retweet, and class 2, the most viral class, is a tweet receiving more than one retweet.

These three different settings can help evaluate the robustness of the proposed model. Because fake news inherently received more retweets than real news [35], we train and evaluate two different models – one for fake news and the other one for real news. We split each dataset into training, validation, and test sets by a ratio of 70%:10%:20% based on the number of tweets. To evaluate the model's robustness, we follow the previous work [4] by creating biased data distributions for the training and validation sets, and unbiased distribution for test set. Because of transportability and omitted spurious dependency from the confounder [20, 4] a causal model is expected to be more robust to distribution shift than non-causal model. We repeat the experiment for five times and report the averaged results.

Dataset	Veracity	News	Tweets	Retweets
PolitiFact	Real	335	16,376	35,586
	Fake	247	11,975	27,627
GossipCop	Real	13,601	563,056	218,760
	Fake	4,111	101,910	225,447

Table 1: Basic statistics of FakeNewsNet [30].

4.2 Experiment Setting

Baseline Methods. Since there is no prior work that involves learning the hidden confounder from textual and numerical data and taking the textual data as the treatment, we mainly focus on four variations of our model and one causal inference work as baseline methods. To prove the effectiveness of the SCM mentioned in Fig 1(b), we discard covariates (`w/o` Z) and treatment (`w/o` T) of our model. Secondly, to understand the effectiveness of the treatment generation in the hidden confounder representation learning, we consider the following two variations without the first stage. The *Con. w/o seq2seq* concatenates the observed covariates and treatment as input for both encoder and decoder while *w/o seq2seq* isolates the inputs for encoder and decoder like our model. All the variation-based baseline methods will only minimize the objective function of the outcome inference which is the main task. Besides, we customize an existing *causal inference model*, `DeepResidual` (`DR`) [7, 22], to handle the textual and numerical data. Different from our model, `DR` models the causal dependency through boosting. It firstly estimates the outcome relying on only the confounder and then takes the concatenation of the predicted probability and treatment's representation to conduct the outcome prediction. In Tables 2 and 3, each cell contains two performance numbers, for fake and real news data respectively. For example, 75.02/74.78 of our model under the binary setting in Table 2 mean the accuracy (ACC) for fake news and real news test sets on PolitiFact, respectively. The best performance is **bold** and the second-best is underlined.
Implementation Details. We utilize the `BART-base` [13] as our main module for feature representation learning and outcome inference. For a fair comparison, the DR takes the confounder as input for the encoder and treatment for the decoder like our model and `w/o seq2seq`. The learning rate for conditional text generation is set to $1e-5$, the learning rate for outcome prediction's seq2seq module is selected from $[1e-5, 1e-6, 0]$, and the learning rate for the prediction head is selected from $[1e-4, 1e-5, 1e-3]$ by the grid search. Because the size of the GossipCop dataset is larger than the PolitiFact, the number of training epochs for the first stage is set to 3 and 5 in GossipCop and PolitiFact, respectively. We report the test results based on the best validation results.

4.3 Main Results

RQ1.1 : Unbiased Dissemination Estimation. As the results showed on Table. 2, we firstly observe that incomplete causal dependency-based baselines (i.e., `w/o` T and `w/o` Z) achieve worse performance than the complete causal dependency-based baselines (i.e., `Con. w/o seq2seq`, `w/o seq2seq` and `DR`). Secondly, our model after the conditional treatment generation shows better performance than `w/o seq2seq` and `Con. w/o seq2seq`. This indicates the importance of learning hidden confounder by capturing the relationship between the covariates and treatment. Thirdly, isolating the input for encoder and decoder only contributes when combined with the Treatment Generation. This is observed because `w/o seq2seq` shows competitive average rank compared with `Con w/o seq2seq`. Fourthly, overall our model shows the best performance compared with all the baselines across three experiment settings. This indicates our model can learn better confounder feature representation and provide robust outcome prediction by overcoming the challenge of data distribution shifting between the training and test sets.

Models	Binary		Ternary		Quaternary	
	ACC(%)	micro-AUC(%)	ACC(%)	micro-AUC(%)	ACC(%)	micro-AUC(%)
PolitiFact						
`DR`	70.59/71.15	82.99/80.78	57.62/65.71	73.69/**83.18**	60.98/60.99	**85.59**/84.96
`w/o` *T*	70.97/73.61	79.79/81.58	66.62/63.92	83.12/77.42	60.93/60.09	85.46/84.82
`w/o` *Z*	70.55/72.12	80.06/81.58	68.72/65.08	83.33/79.59	60.25/59.40	84.77/84.37
`Con. w/o seq2seq`	71.30/73.82	79.87/82.59	69.42/67.16	83.22/80.88	60.19/60.79	84.91/85.01
`w/o seq2seq`	74.46/70.38	82.37/80.29	69.76/63.59	84.13/79.37	60.59/60.35	84.86/84.91
Our model	**75.02/74.78**	**83.17/82.61**	**72.08/67.49**	**85.24**/82.18	**63.35/62.69**	85.52/**86.02**
GossipCop						
`DR`	75.89/89.28	74.88/94.10	**74.66**/86.12	85.89/**97.25**	74.60/84.38	91.91/96.41
`w/o T`	74.49/85.98	74.76/94.08	72.10/84.10	82.64/94.04	67.07/84.05	88.97/94.48
`w/o C`	73.80/84.83	76.88/92.76	71.70/84.20	84.32/96.59	71.03/84.59	90.35/96.07
`Con. w/o seq2seq`	74.65/87.56	73.95/93.23	73.10/83.99	83.82/94.57	72.14/85.34	90.86/95.64
`w/o seq2seq`	74.11/88.02	74.55/93.26	73.17/84.22	86.27/88.23	72.77/85.60	91.43/96.29
Our model	**76.01/92.44**	**79.22/94.83**	74.17/**87.73**	**86.30**/96.83	**74.83/86.46**	**92.06/96.94**

Table 2: News tweet dissemination prediction on PolitiFact and GossipCop datasets with handling the distribution difference between the training and test sets.

RQ1.2: Contribution of Different Covariates. This ablation study ablates several components of the tweet author's profile, such as personal text description (named as `w/o pers. descr.`), personal numerical attributes (`w/o pers. ATT`), and all the profile information (`w/o pers.`). As shown in Fig. 3, our model achieves the best performance in the average rank of accuracy and micro-AUC. This indicates that learning the hidden confounder's representation from the tweet author's profile has positive contributions. Besides, we can observe that the user's both personal textual description and numerical attributes have positive contributions to the social engagements prediction (Avg. Rank `w/o pers. descr.` and `w/o pers. descr.` < `w/o pers.`).

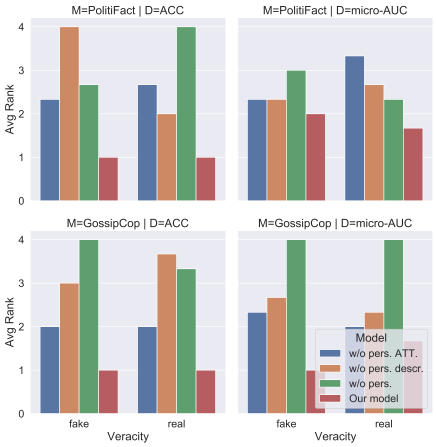

Fig. 3: Ablation study of covariates for fake and real news on PolitiFact and GossipCop datasets. A lower value ("rank") is better.

So far, we have discussed three different label settings. From now on, we will focus on the binary and quaternary settings due to limited space.

4.4 Lexicons Boosting Dissemination

To understand the key syntax of tweets causing news dissemination (**RQ2.**), we need to interpret the causal inference model. Since our model itself is not interpretable[4], we utilize the model-agnostic explanation method - `LIME` [27] to do the model interpretation. LIME will create many counterfactual tweets and fit a locally-linear model based on the

[4] We leave the syntax self-interpreted causal inference model as our future work.

pairs of counterfactual tweets and their unbiased retweet status prediction. It should be noticed that the unbiased prediction is from our model's output. In this experiment, we only generate counterfactual examples on the tweet content (treatment) and keep the news content and tweet author's profile (covariates) static. Since the baseline method `w/o T` did not utilize the treatment, it is impossible to make an intervention on the treatment tweet. Therefore, we discard `w/o T` in the following evaluations. To measure the effectiveness of our model with `LIME` in identifying words for the news dissemination, we provide quantitative and qualitative analysis. Due to the high computation costs for `LIME`, we follow the previous work [3], randomly sampling 500 instances from test sets.

Quantitative Evaluation. A well-trained model can capture the important information, and thus provide a more meaningful explanation compared with a poorly trained model under the same explanation method [28]. We use area over the perturbation (AOPC) score [28, 17] as an evaluation metric. It will calculate the average prediction probability change when deleting the top-L words from `LIME` [3]. A higher AOPC score is preferred. Table. 3 shows the results under binary and quaternary settings. Our model shows significant AOPC score improvement over all baseline methods. This indicates our model can provide more important/interpretable words.

Qualitative Analysis. In Table. 4, we list the Top-30 salient words. These words are ranked by average LIME weights and filtered by their TF-IDF scores. Due to different retweet count distribution between fake and real news, given two different causal inference models for fake and real news, we report these Top-30 salient words for both of them, separately. Since our goal is to understand what syntax boosts news dissemination the most, we only focus the $|reweets| > 10$ under the quaternary setting. In *fake news tweets*, there exists announcement words such as "*revealed*" and "*claim*". PolitiFact contains the sequential connectives like "*until*" and "*then*". It could indicate that the fake news tweets have similar writing style with real news to cause more dissemination [31]. Besides, we observe the emotional words like "*hating*", "*bullshit*" in PolitiFact and "*trouble*", "*death*", "*jealous*", "*happy*" in GossipCop. These words drive audiences' emotions and encourage them to engage [10]. In *real news tweets*, we observe referenced words (e.g., "*call*", "*posted*", "*approve*", "*info*") play important roles in both datasets. These words will evoke users' curiosity to share [10]. Besides, these real news tweets contain many time-relevant words such as "*year*" and "*week*". This finding indicates the use of more precise instruction strategy in real news tweets. Lastly, in GossipCop, there are many degree adverbs (e.g., words like "*very*", "*really*" and "*super*"). These degree words make the speakers' utterance expressive [8], which may receive more retweets.

PolitiFact	Binary	Quaternary
`DR`	0.67/-0.60	0.94/1.21
`w/o` *Z*	0.26/0.10	2.11/1.48
`Con. w/o seq2seq`	0.13/0.99	3.19/2.28
`w/o seq2seq`	0.32/-0.60	6.57/6.11
Our model	**5.10/3.53**	**9.88/9.97**
GossipCop	Binary	Quaternary
`DR`	0.35/-0.60	0.98/0.50
`Con. w/o seq2seq`	-1.25/1.04	0.66/2.92
`w/o seq2seq`	0.37/1.84	0.43/0.01
`w/o seq2seq`	0.58/2.03	0.98/4.12
Our model	**1.04/2.03**	**1.45/8.47**

Table 3: AOPC [28] score(‰) for LIME in interpreting counterfactual outcome inference.

We also provide a fake news tweet[5] (shown in Table. 5) to understand the context information of the salient word. The less retweeted tweet were very similar to the most

[5] Due to space limit, we only report the fake news' tweets.

	PolitiFact	*GossipCop*
Fake	revealed, emoji, hashtag, engages, doing, center, stands, goals, joining, flag, claims, until, loyalty, claim, really, then, crap, that, video, aside, betrayal, pleas, hating, normal, bullshit, protesters, honor, private, has, right	these, thanks, children, director, very, hired, paid, separate, year, contributed, scale, fifth, with, said, happy, confusing, between, picture, hashtag, announced, ice, jealous, until, detective, mystery, situation, trouble,complicated, death, bus
Real	emoji, station, attack, info, hashtag, programme, approve, lapse, posted, audio, space, advisor, climate, globe, hideaway, cancel, nite, cannabis, week, quot, dedicada, uncooperative, blaring, call, via, morphing, easily, here, soon, year	peaceful, design, why, reports, emoji, toughness, pic, already, hashtag, tribute, premiere, disagree, fight, knows, reluctantly, pieces, had, kill, myself, confirms, moment, are, been, star, victory, match, super, really, that, his

Table 4: Top-30 salient words for the tweets' dissemination prediction.

retweeted tweet except the missing of identified salient word. This confirms that the importance of these salient words to cause fake/real news tweets get disseminated.

Retweet	Fake-PolitiFact, Keyword: "*claim*"	Fake-GossipCop, Keyword: "*reveal*"
Most	Dying 78 year old cia agent admits to killing marilyn monroe *claim*s he carried out 37.	AQUAMAN Movie *reveal* First Look at Nicole Kidman as Atlanna
Less	Dying 78 year old cia agent admits to killing marilyn monroe	Nicole Kidman was pelted with rocks while filming new flick Aquaman.

Table 5: Examples of the most and least retweeted tweets of fake news. The most retweeted tweets contain salient words identified via LIME from our model.

5 Related Work

User Engagements on News. Social media platforms provide a new way for news organizations to distribute content and receive feedback from the users through user engagement [1]. Many works try to predict the user engagements based on the news content and users' reactions [18]. Although they have achieved excellent performance in the prediction, they did not tackle the causal relationships between the social media posts and user engagements. The most similar work is [18], which estimated the causal effects of tweets' editing styles on boosting users' engagements. However, identifying the treatment (editing styles) is dependent on an out-of-domain style classifier, which will inevitably bring the measurement error for the downstream causal effect estimation. [35] revealed that fake news spreads faster, deeper, and more broadly than the truth. The authors found that fake news was more novel than the truth and inspired the replies' fear, disgust and surprise. [4] identified several tweet authors' attributes in causing the fake news spread on social media. Different from these works, we revealed which tweet caused news disseminated and what types of lexicons in these tweets played the decisive roles.

Causal Inference on Text. Text data provides a new perspective for researchers to understand the causal effects of the treatment's intervention. In this paper, the textual data appeared in both the proxy of confounder and treatment. To learn the hidden confounder from observed covariates, existing works map the covariates into a low-dimensional vector through methods like Latent Dirichlet Allocation (LDA) [2] and auto-encoder [15]. As for textual treatment, most works extract text properties from the text. They utilize a classifier to predict the text properties such as sentiment [23] and clickbait [18], which will inevitably generate measurement errors and require additional efforts to label the dataset for classifier training [6]. Others map the treatment text to latent vectors [24, 7]. However, conventional methods cannot handle the multimodal covariates and cannot exploit the pre-trained language model in outcome inference. In addition, these methods ignore the dependency between the textual treatment and multimodal covariates. Thus, they cannot provide the correct outcome estimation.

6 Conclusion

In this paper, we propose a causal inference model to unbiasedly know which tweets cause fake/real news disseminated on social media and what lexicons play a critical role inside the tweet. our model successfully represents the multimodal covariates (news content and user personal attributes) and textual treatments (tweet). The comprehensive experiment results indicate the robustness and effectiveness of our model in resolving the confounding bias. In our qualitative analysis, we identify salient words from two fake news benchmark datasets. These salient words can not only be used as the fake news detection features, but also help further research in better understanding of behind mechanism of fake news dissemination on social media.

Acknowledgement. This work was supported in part by NSF grant CNS-1755536 and WPI TRIAD. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

References

1. Aldous, K.K., An, J., Jansen, B.J.: View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations. ICWSM (2019)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (mar 2003)
3. Chen, H., Ji, Y.: Learning variational word masks to improve the interpretability of neural text classifiers. In: EMNLP (2020)
4. Cheng, L., Guo, R., Shu, K., Liu, H.: Causal understanding of fake news dissemination on social media. KDD (2021)
5. Deng, Z., Zheng, X., Tian, H., Zeng, D.D.: Deep causal learning: Representation, discovery and inference. arXiv preprint arXiv:2211.03374 (2022)
6. Egami, N., Fong, C.J., Grimmer, J., Roberts, M.E., Stewart, B.M.: How to make causal inferences using texts. CoRR **abs/1802.02163** (2018)
7. Fytas, P., Rizos, G., Specia, L.: What makes a scientific paper be accepted for publication? (2021)
8. Indhiarti, T.R., Chaerunnisa, E.R.: A corpus-driven collocation analysis of degree adverb very, really, quite, and pretty (2020)
9. Keith, K.A., Jensen, D., O'Connor, B.: Text and causal inference: A review of using text to remove confounding from causal estimates. In: ACL (2020)
10. Kilgo, D.K., Sinta, V.: Six things you didn't know about headline writing: Sensationalistic form in viral news content from traditional and digitally native news organizations. In: ISOJ. vol. 6, pp. 111–130 (2016)
11. Kushin, M.J., Yamamoto, M.: Did social media really matter? college students' use of online media and political decision making in the 2008 election. Mass Communication and Society **13**(5), 608–630 (2010)

12. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: EMNLP (2021)
13. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL (2020)
14. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J.: Gpt understands, too (2021)
15. Louizos, C., Shalit, U., Mooij, J.M., Sontag, D., Zemel, R., Welling, M.: Causal effect inference with deep latent-variable models. NeurIPS (2017)
16. Miao, W., Geng, Z., Tchetgen Tchetgen, E.J.: Identifying causal effects with proxy variables of an unmeasured confounder. Biometrika **105**(4), 987–993 (2018)
17. Nguyen, D.: Comparing automatic and human evaluation of local explanations for text classification. In: NAACL HLT (2018)
18. Park, K., Kwak, H., An, J., Chawla, S.: How-to present news on social media: A causal analysis of editing news headlines for boosting user engagement. In: ICWSM (2021)
19. Pearl, J.: Causality. Cambridge university press (2009)
20. Pearl, J., Bareinboim, E.: Transportability of causal and statistical relations: A formal approach. In: AAAI (2011)
21. Petrovic, S., Osborne, M., Lavrenko, V.: RT to win! predicting message propagation in twitter. In: ICWSM. The AAAI Press (2011)
22. Pryzant, R., Basu, S., Sone, K.: Interpretable neural architectures for attributing an ad's performance to its writing style. In: EMNLP Workshop BlackboxNLP (2018)
23. Pryzant, R., Card, D., Jurafsky, D., Veitch, V., Sridhar, D.: Causal effects of linguistic properties. In: NAACL HLT (2021)
24. Pryzant, R., joo Chung, Y., Jurafsky, D.: Predicting sales from the language of product descriptions. In: eCOM@SIGIR (2017)
25. Pryzant, R., Shen, K., Jurafsky, D., Wagner, S.: Deconfounded lexicon induction for interpretable social science (2018)
26. Qin, G., Eisner, J.: Learning how to ask: Querying lms with mixtures of soft prompts. In: NAACL-HLT (2021)
27. Ribeiro, M.T., Singh, S., Guestrin, C.: ”why should I trust you?”: Explaining the predictions of any classifier. In: KDD (2016)
28. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.: Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Netw. Learn. Syst **28**(11), 2660–2673 (2017)
29. Shi, C., Blei, D.M., Veitch, V.: Adapting neural networks for the estimation of treatment effects. In: NeurIPS (2019)
30. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. arXiv preprint arXiv:1809.01286 (2018)
31. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter **19**(1), 22–36 (2017)
32. Strekalova, Y.A., Krieger, J.L.: Beyond words: Amplification of cancer risk communication on social media. Journal of Health Communication **22**(10), 849–857 (2017)
33. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: SocialCom (2010)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. NeurIPS (2017)
35. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
36. Zhang, Y.F., Zhang, H., Lipton, Z.C., Li, L.E., Xing, E.P.: Can transformers be strong treatment effect estimators? arXiv preprint arXiv:2202.01336 (2022)