Causal Understanding of Fake News Dissemination on Social Media

Lu Cheng[1], Ruocheng Guo[1], Kai Shu[2], Huan Liu[1]
[1] Computer Science and Engineering, Arizona State University, USA
[2] Department of Computer Science, Illinois Institute of Technology, USA
{lcheng35,rguo12,huanliu}@asu.edu,kshu@iit.edu

KDD'21

211005 Chia-Chun Ho

Outline

Introduction

Related Work

Problem Statement

Proposed Framework

Empirical Evaluation

Discussion

Comments

Introduction

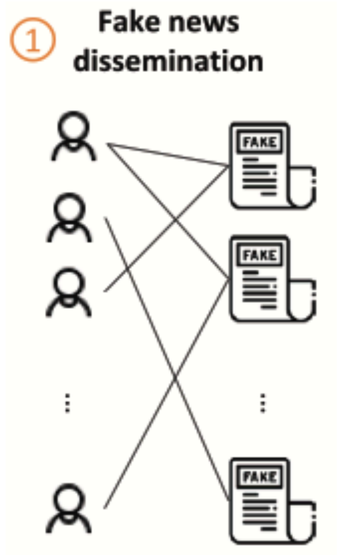
Fake news detection

- While great effort can be seen in computational fake news detection, less is known about what user attributes cause some users to share fake news.
- In contrast to the research focused on correlations between user profiles (e.g., age, gender) and fake news, this work seeks a more nuanced understanding of how user profile attributes are causally related to user susceptibility to share fake news.
- The key to identifying causal user attributes with observational data is to find *confounders* - variables that cause spurious associations between treatments (user profile attributes) and outcome (user susceptibility).

Introduction

Confounding bias in fake news

- Various studies in psychology and social science have shown the strong relationships of user behavior with user characteristics and activities such as information sharing, personality traits and trust.
- Characterizing user behavior has become a vital means to analyzing activities on networking sites.
- Informed by this, argue that fake news sharing behavior (i.e., the user-news dissemination relation characterized by a bipartite graph) is critical to address confounding in causal relations between user attributes and susceptibility.



Introduction

Learning fake news sharing behavior

- It's challenging because virtually all observational social media data is subject to selection bias due to self-selection and the actions of online news platforms.
- These biased data only partially describe how users share fake news.
- To alleviate the selection bias, one can leverage a technique commonly used in causal inference, particularly. Inverse Propensity Scoring (IPS) that creates a pseudo-population similar to data collected from a randomized experiment.
- Propensity describes the probability of a user being exposed to fake news pieces.
- By connecting fake news dissemination with causal inference, can derive an unbiased estimator for learning fake news sharing behavior under selection biases.

Introduction

Main contribution in three-fold

- Address a novel and important problem that complements earlier efforts on fake news detection.
  - In particular, this paper seek to answer why people share fake news by uncovering the causal relationships between user profiles and susceptibility.
- Show how learning fake news sharing behavior under selection biases can be approached with propensity-weighting techniques.
  - Designed 3 effective estimations of propensity score for fake news dissemination to learn unbiased embeddings of fake news sharing behavior.

Introduction

Main contribution in three-fold

- Under the multiple causal inference framework with mild assumptions, proposed to use the learned embeddings of fake news sharing behavior as the confounder, drawing from findings in social science.
    - This enable to learn a causal model that can identify causal user attributes and estimate their effects on user susceptibility.

Related Work
of fake news detection

- Content-based fake news detection
  - News content is typically represented by knowledge, style, or a latent representation.
  - Knowledge-guided methods seek to directly evaluate news authenticity by comparing its knowledge with that within a knowledge graph.
  - Style features can be word-level features such as TF-IDF / LIWC features.
  - Latent-representation-based methods have limited interpretability.

Related Work
of fake news detection

- Propagation-based fake news detection
  - Advocate the use of social context information (news cascade, stance, sentiments).
  - News cascade was used as multivariate time series to model the propagation path of each news story.
  - Stance graph built on user posts then detected by mining the stance correlations within a graph optimization framework.
  - Graphs explores relationship among article, publishers, users, and posts.
  - Underlying assumption is that the overall structure of fake news cascades differs from the true ones.

Related Work

of fake news detection

- Despite the remarkable progress in detecting fake news, comparatively fewer efforts seek to understand what user profile attributes cause users to spread fake news.
- Provide a novel *causal understanding* by learning unbiased fake news sharing behavior.
- Explicitly modeling fake news dissemination with a focus on discovering user attributes causally related to user susceptibility.

Problem Statement

Terminology Definition

- Users (who share fake news $\mathscr{C}$): $\mathscr{U} = \{1,2,\cdots, u, \cdots, U\}$
- Fake news $\mathscr{C} = \{1,2,\cdots, i, \cdots, N\}$
- Interaction between user u and fake news i: $Y_{ui} \in \mathscr{Y}$
  - $Y_{ui} = 1$, if u spread i.
  - $Y_{ui} = 0$, either u is not interested in i or u didn't observe i.
- Suppose users have m profile attributes denoted by matrix $A = (A_1, A_2, \cdots, A_m)$
- Each user u is also associated with an outcome $B \in (0,1]$, denoting u's susceptibility to spreading fake news.

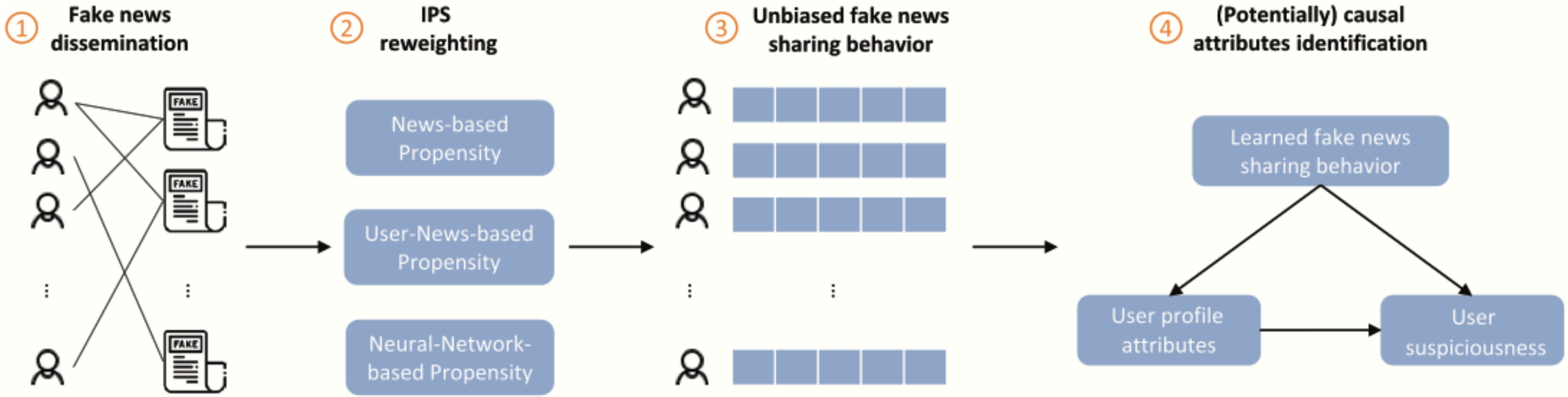
Problem Statement

Tasks Definition

- Fake News Sharing Behavior Learning
  - Given user group $\mathscr{U}$, the corpus of fake news $\mathscr{C}$, the set of user-fake news interactions $\mathscr{Y}$, aim to model the fake news dissemination process and learn fake news sharing behavior U under selection biases.
- Causal User Attributes Identification
  - Given user attributes A, the fake news sharing behavior U, the user susceptibility B, this task seeks to identify user attributes that potentially cause users to spread fake news and estimate the effect.

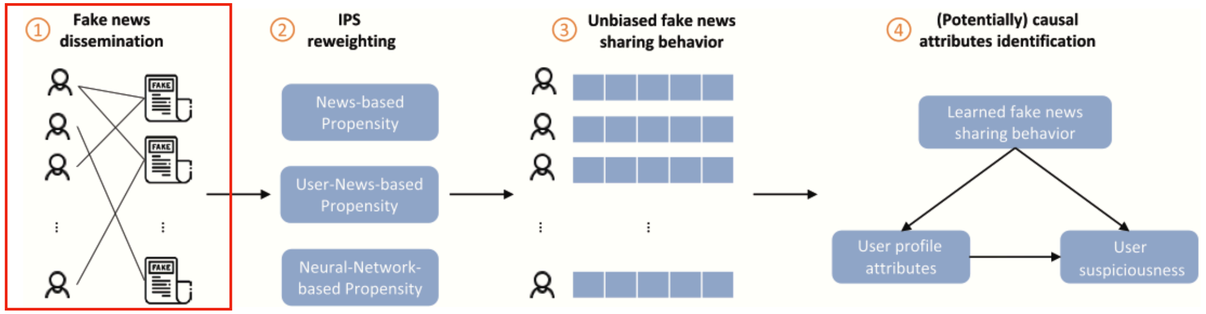
Proposed Framework

Framework Overview



Proposed Method

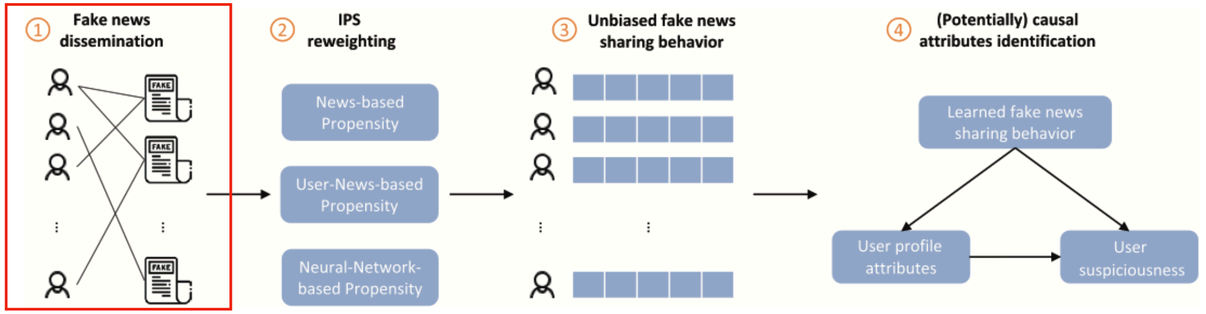
Modeling Fake News Dissemination



- Building a model that characterizes fake news dissemination.
- The key is the "implicit" feedback collect through natural behavior such as news reading or sharing of a user with unique profile attributes.
- By noting which fake news a user did and didn't share in the past, may infer fake news a user will be interested in sharing in the future.

Proposed Method

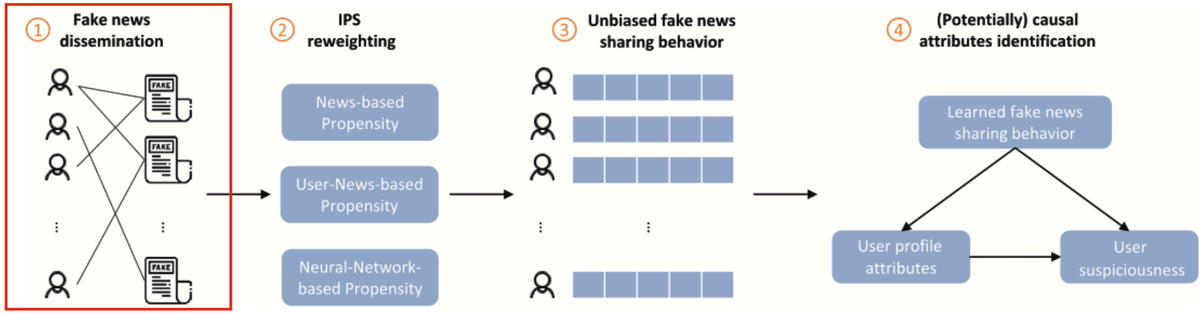
Modeling Fake News Dissemination



- To better formulate the process of fake news dissemination, introduce two binary variables highly related to this process:
  - Interestingness: $R_{ui} \in \{0,1\}$
    - $R_{ui} = 1(0)$ indicates u is interested (not interested) in i.
  - Exposure: $O_{ui} \in \{0,1\}$
    - $O_{ui} = 1$ denotes user u was exposed to fake news i and $O_{ui} = 0$, otherwise.
  - Assume that a user spreads fake news if s/he is both exposed to and interested in it.

Proposed Method

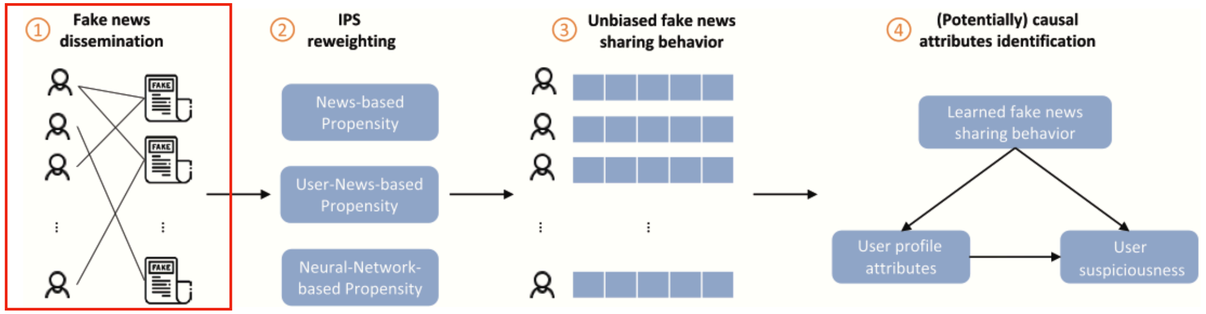
Modeling Fake News Dissemination



- $Y_{ui} = O_{ui} \cdot R_{ui}$
- $P(Y_{ui} = 1) = P(O_{ui} = 1) \cdot P(R_{ui} = 1) = \theta_{ui} \cdot \gamma_{ui}$
- As fake news dissemination is missing-not-at-random (MNAR), further assume that the probability of u spreading i is represented as the product of the exposure and interestingness parameters.

Proposed Method

Modeling Fake News Dissemination



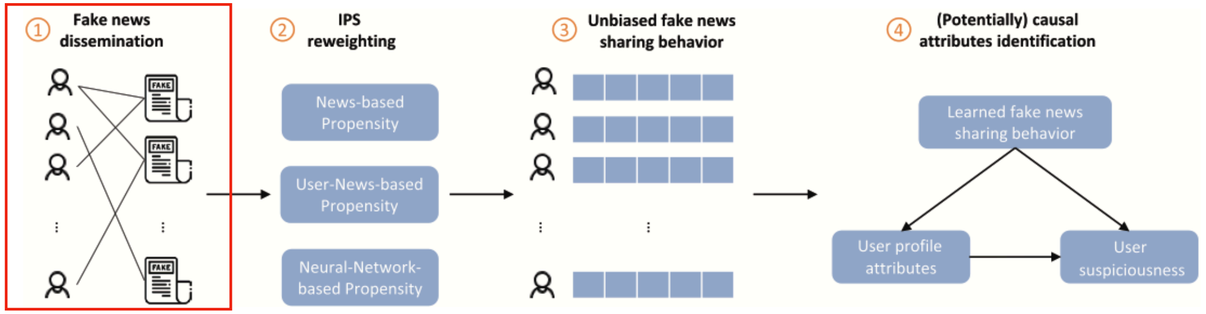
- Supposed have a pair of fake news (i, j) with $i \neq j$ and $\mathscr{D}_{pair} = \mathscr{U} \times \mathscr{C} \times \mathscr{C}$ is the set of all observed (positive) interactions (u, i) and unobserved (negative) interactions (u, j).

- As both R_{ui}, O_{ui} are unobserved, the model parameters are learned by pairwise BPR (Bayesian Personalized Ranking) loss that employs user-news interactions.

- Assume that observed user-news interactions better explain users' preferences than the unobserved ones, thereby, should be assigned higher prediction scores.

Proposed Method

Modeling Fake News Dissemination



- Define the ideal loss function of fake news dissemination as

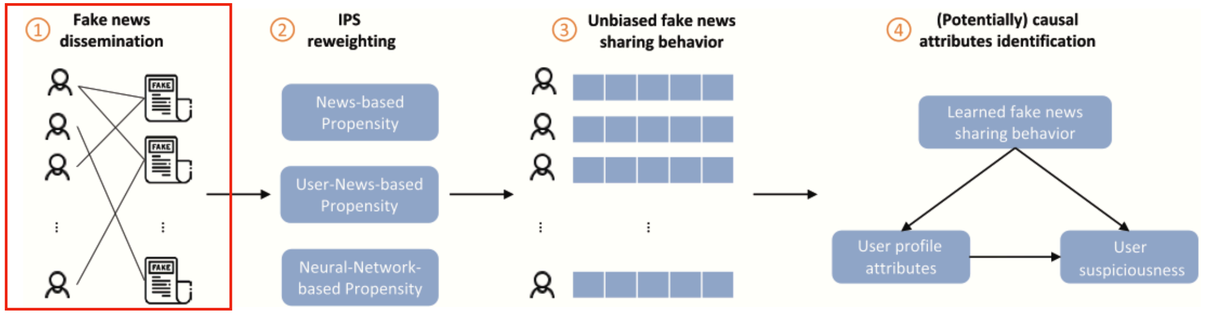
$$\mathscr{L}_{ideal}(\hat{S}) = \frac{1}{|\mathscr{D}_{pair}|} \sum_{(u,i,j)\in\mathscr{D}_{pair}} \gamma_{ui}(1-\gamma_{uj})\ell(\hat{S}_{uij})$$

  - $\hat{S}_{uij}$: difference between the predicted scores of fake news i and j, and $\ell = -\ln(\sigma(\cdot))$ represents the local loss for the triplet (u, i, j).

- To this end, modeling fake news dissemination is a statistical estimation problem where seek to estimate the ideal loss functions that returns news which users are most interested in using the observed user-news interaction.

Proposed Method

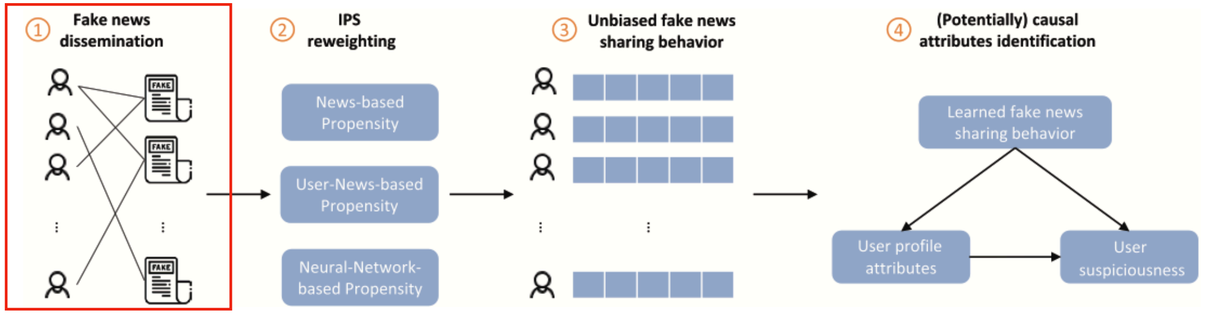
Learning Unbiased Sharing Behavior



- In previous model, leads to at least two major deficiencies:
  - Observational data only includes positive interactions between users and fake news whereas negative interactions are never observed.
  - So the above model cannot differentiate whether unshared fake news is uninteresting to the user or has yet to be exposed to the user.

Proposed Method

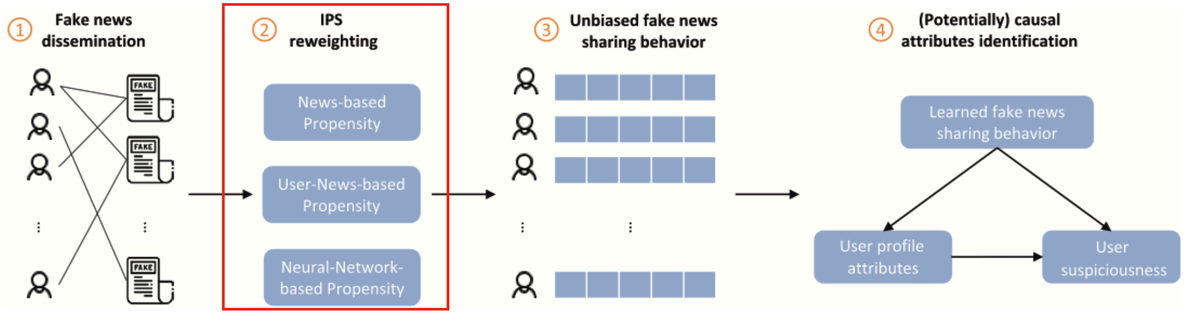
Learning Unbiased Sharing Behavior



- In previous model, leads to at least two major deficiencies:
  - Similar to preferential attachment theory in social network science, users are preferentially to interact with news that are already prevalent and online news platforms are also more likely to recommend popular news than tail ones.
- Models using these partially observed interactions will learn biased embedding of the fake news sharing behavior (or user embedding).

Proposed Method

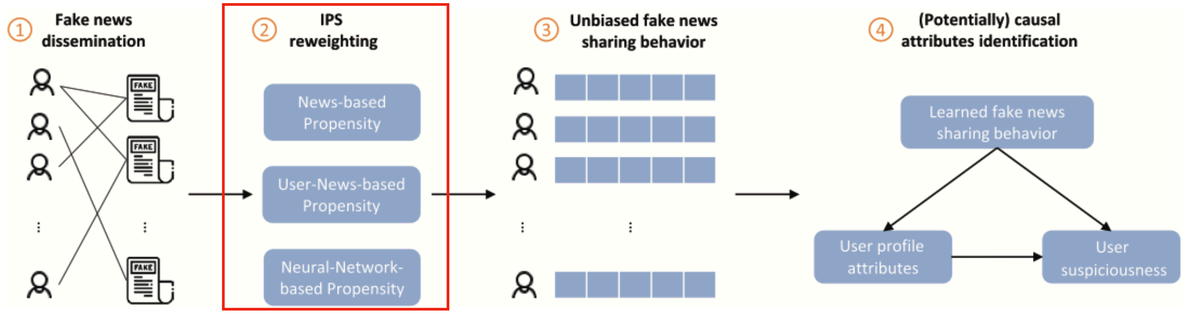
Learning Unbiased Sharing Behavior



- To handle selection bias, proposed to leverage IPS to learn unbiased fake news sharing behavior based on existing positive interactions between users and fake news.
- Propensity denotes the probability of exposing a user to fake news pieces.
- IPS works as a reweighing mechanism by assigning larger weights to news that is less likely to be observed.

Proposed Method

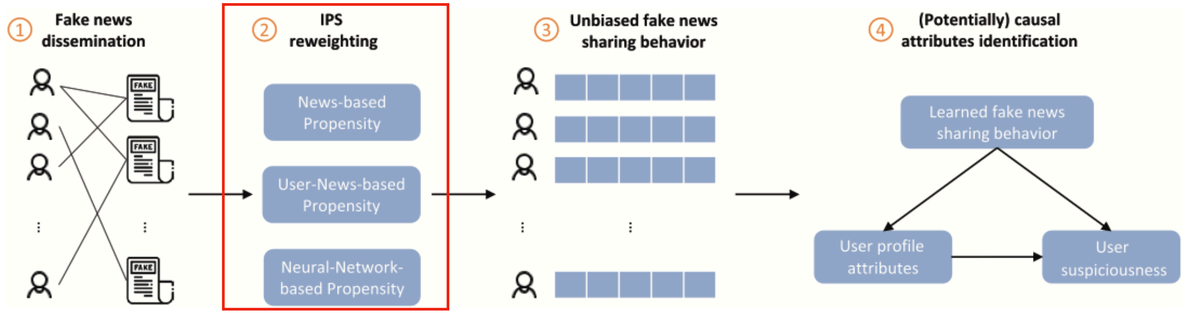
Learning Unbiased Sharing Behavior



- Define the propensity score in the fake news dissemination as follows:
  - $\theta_{ui} = P\left(O_{ui} = 1\right) = P\left(Y_{ui} = 1 \mid R_{ui} = 1\right)$
- Indicates that the propensity score is the probability of u spreading i given u is interested in i.
- This ensure that, in principle, there could be positive interaction between every pair of (u, i). Incorporating θ_{ui} into the ideal loss function of fake news dissemination, obtain the unbiased estimator:
  - $$\hat{\mathscr{L}}_{unbiased}(\hat{S}) = \frac{1}{|\mathscr{D}_{pair}|} \sum_{(u,i,j)\in\mathscr{D}_{pair}} \frac{Y_{ui}}{\theta_{ui}} \left(1 - \frac{Y_{uj}}{\theta_{uj}}\right) \ell\left(\hat{S}_{uij}\right)$$

Proposed Method

Learning Unbiased Sharing Behavior



- The loss function of $\hat{\mathscr{L}}_{unbiased}(\hat{S})$ is unbiased against the ideal loss of fake news dissemination in $\hat{\mathscr{L}}_{ideal}(\hat{S})$.

$$\mathbb{E}\left[\hat{\mathscr{L}}_{unbiased}(\hat{\mathbf{s}})\right] = \mathbb{E}\left[\frac{1}{|\mathscr{D}_{pair}|}\sum_{(u,i,j)\in\mathscr{D}_{pair}}\frac{Y_{u,i}}{\theta_{u,i}}\left(1-\frac{Y_{u,j}}{\theta_{u,j}}\right)\ell\left(\hat{s}_{uij}\right)\right]$$

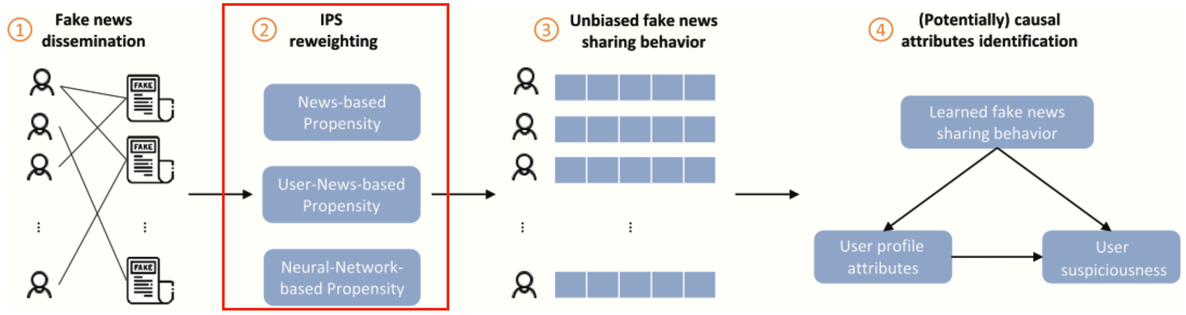
$$= \frac{1}{|\mathscr{D}_{pair}|}\sum_{(u,i,j)\in\mathscr{D}_{pair}}\frac{\mathbb{E}[Y_{u,i}]}{\theta_{u,i}}\left(1-\frac{\mathbb{E}[Y_{u,j}]}{\theta_{u,j}}\right)\ell\left(\hat{\mathbf{S}}_{uij}\right)$$

- $$= \frac{1}{|\mathscr{D}_{pair}|}\sum_{(u,i,j)\in\mathscr{D}_{pair}}\gamma_{u,i}(1-\gamma_{u,j})\ell\left(\hat{\mathbf{S}}_{uij}\right)$$

$$= \mathscr{L}_{ideal}(\hat{\mathbf{S}})$$

Proposed Method

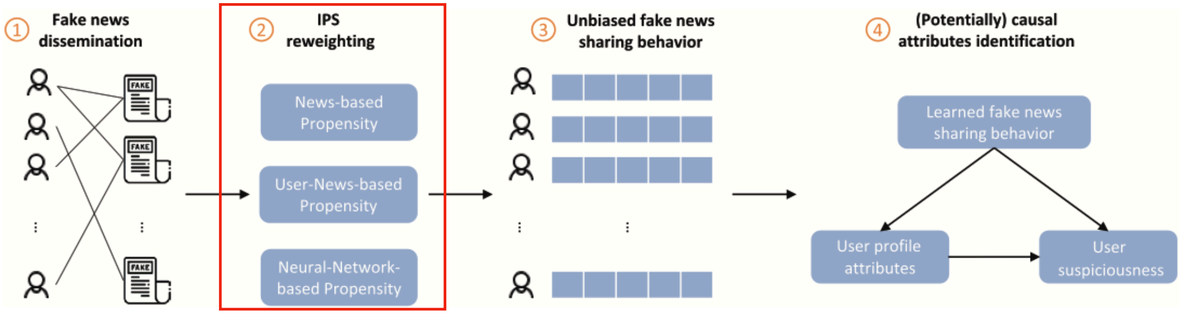
Propensity Score for Dissemination



- Here propose three estimations of propensity score based on user and news attributes.
  - News-based Propensity P_{news}
  - User-News-based Propensity P_{user}
  - Neural-Network-based Propensity P_{neural}

Proposed Method

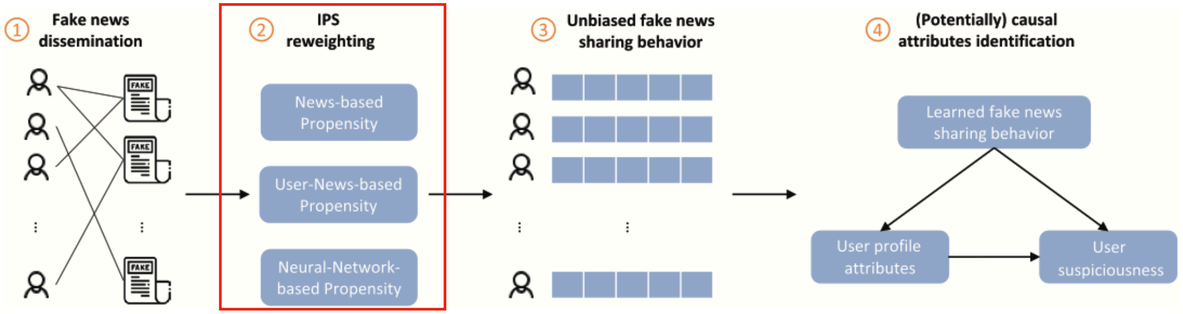
News-based Propensity



- *Definition*. Propensity using relative news popularity is defined as
- $$P_{news} = \hat{\theta}^{news}_{,i} = \left(\frac{\sum_{u \in \mathcal{U}} Y_{ui}}{\max_{i \in C} \sum_{u \in \mathcal{U}} Y_{ui}} \right)^{\eta}$$
- Popularity-related measures follow power law distribution, therefore, include the smoothing parameter $\eta \leq 1$ and set it to 0.5.
- Assume that the probability of a user observing fake news pieces is highly related to its popularity.

Proposed Method

User-News-based Propensity



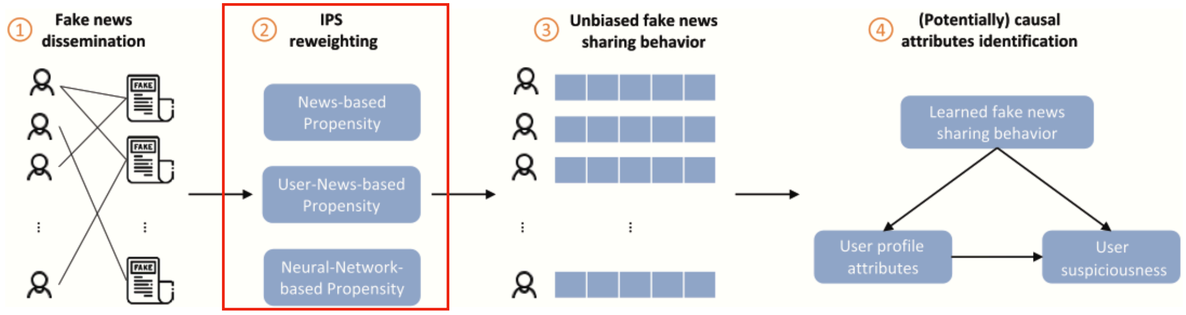
- *Definition*. Propensity using both relative news popularity and user popularity is defined as

- $$P_{user} = \hat{\theta}_{u,i}^{user} = \left(\frac{\sum_{u \in \mathcal{U}} Y_{ui} \cdot F_u}{\max_{i \in C} \sum_{u \in \mathcal{U}} Y_{ui} \cdot F_u} \right)^{\eta}$$

- P_{user} also considers the bias induced by the user popularity, that is users who are popular and active on social media are more likely to be exposed to fake news.

Proposed Method

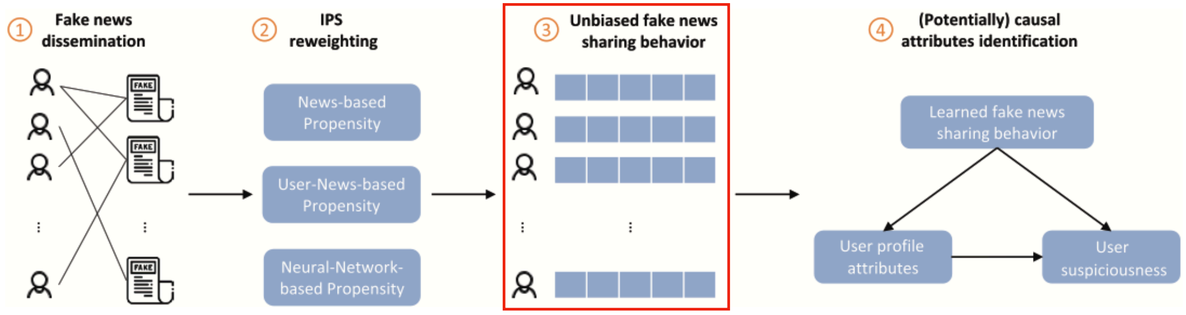
Neural-Network-based Propensity



- *Definition*. Propensity encoded by neural networks is defined as
- $P_{neural} = \hat{\theta}_{,i}^{neural} = \sigma\left(\boldsymbol{e}_i\right)$
- e_i: latent representation of news content
- $\sigma(\cdot)$: sigmoid function
- Implicitly encode the popularity of fake news in the latent space based on the news content.

Proposed Method

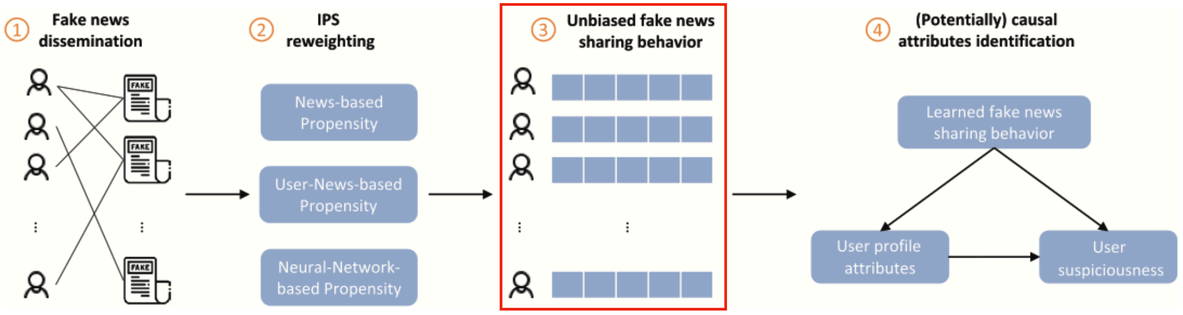
Variance Reduction



- IPS-based approaches often suffer from large variance as the propensity score can be extremely small.
- Unpopular fake news has low exposure probability.
- To reduce to variance, employ the following non-negative loss:
- $$\hat{\mathscr{L}}_{non-neg}(\hat{S}) = \frac{1}{|\mathscr{D}_{pair}|} \sum_{(u,i,j)\in\mathscr{D}_{pair}} \max\left\{ \ell_{unbiased}\left(\hat{S}_{uij}\right), 0 \right\}$$

Proposed Method

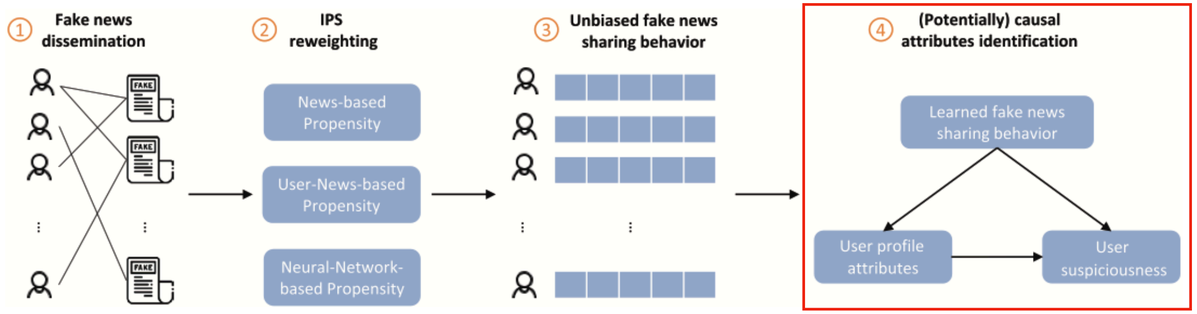
Variance Reduction



- The final loss function for modeling the unbiased fake news dissemination is formulated as
  - $\arg\min_{\boldsymbol{U},\boldsymbol{V}} \hat{\mathscr{L}}_{non-neg}(\hat{\boldsymbol{S}}) + \lambda \left(\|\boldsymbol{U}\|_2^2 + \|\boldsymbol{V}\|_2^2 \right)$
  - $\boldsymbol{U}, \boldsymbol{V}$: user (fake news sharing behavior), news embeddings
  - λ: hyper-parameter that controls the weight of l2-regularization

Proposed Method

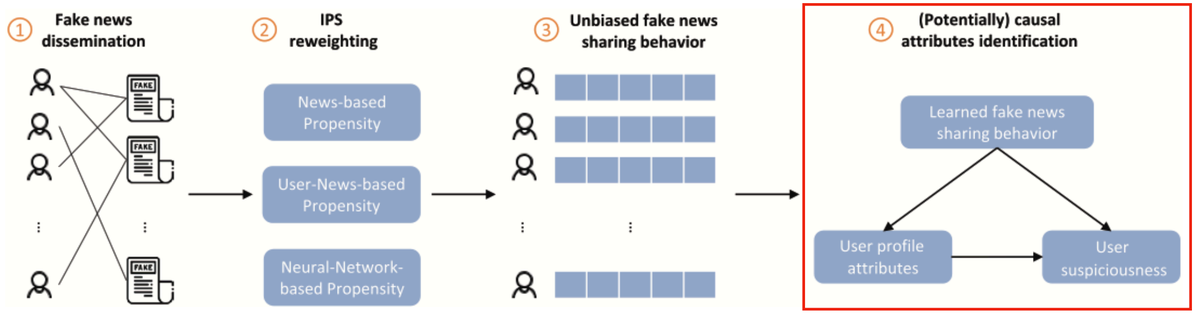
Identifying Causal User Attributes



- Discusses how to simultaneously identify multiple user attributes that potentially cause user susceptibility and estimate the effects.
- Causal inference is the anchor of knowledge to understand the underlying mechanism that drives people to spread fake news.
- Tackling a multiple causal inference task where user attributes represent the multiple treatments and user susceptibility denotes the outcome.
- The goal is to estimate simultaneously the effects of individual user attributes on how likely a user spread fake news pieces.

Proposed Method

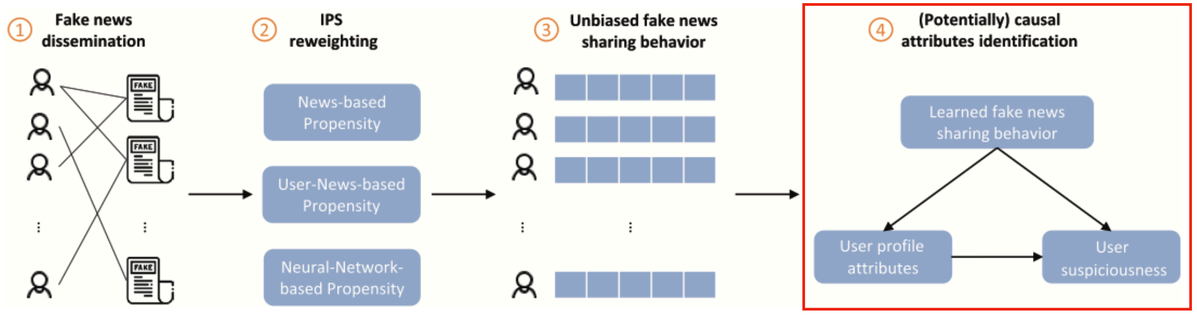
Identifying Causal User Attributes



- Suppose u's attributes are encoded in a vector $\boldsymbol{a} = (a_1, a_2, \cdots, a_m), \boldsymbol{a} \in \boldsymbol{A}$.
- For each user u, there is a potential outcome function that maps configurations of the attributes to user susceptibility $B_u \in (0,1]$ defined as
  - $B_u = n^u_{fake} / \left(n^u_{fake} + n^u_{true} \right)$, n^u_{fake}: number of fake news u has shared.
- Assume that a large portion of news a user has shared is fake, more susceptible s/he is to share fake news.

Proposed Method

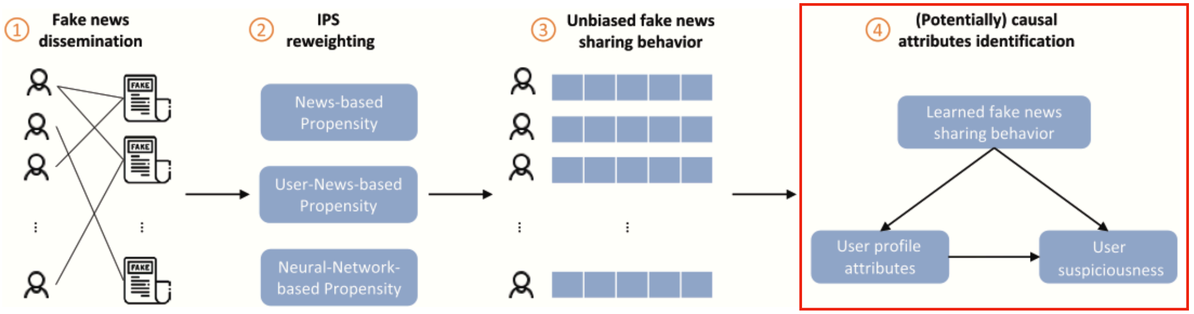
Identifying Causal User Attributes



- Multiple causal inference seeks to identify the sampling distribution of the potential outcomes $B_u(\boldsymbol{a})$ for each configuration of attribute $\boldsymbol{a}$.
- However in observational studies, can only observe one potential outcome of a user under one configuration of $\boldsymbol{a}$, a.k.a. the "fundamental problem of causal inference".

Proposed Method

Identifying Causal User Attributes



- Given user profile attributes $\boldsymbol{A}$, confounder $\boldsymbol{U}$ and the user susceptibility $\boldsymbol{B}$, build the causal model to identify the causal user attributes and estimate their effects:
  - $B_u = \boldsymbol{\beta}^\top \boldsymbol{a}_u + \boldsymbol{\gamma}^\top \boldsymbol{U}_u$
  - β, γ: coefficients
  - β: denote how user attributes affect individual decision to share fake news.

Empirical Evaluation

Datasets and settings

Dataset	# Real	# Fake	# Total	# Users
PolitiFact	624	432	1,056	110,127
GossipCop	16,817	5,323	22,140	194,788

- PolitiFact, GossipCop
- Train / Test = 80/20
- The training data is randomly selected from the original data (thus biased) whilst from the rest data.
- Create the test data such that expose each user to each fake news as uniformly as possible (i.e., with equal probability, thus less biased).

Empirical Evaluation

Baselines

- The author not aware of any similar work in the literature of fake news that learns the embeddings of fake news sharing behavior and identifies causal user profile attributes.
- Problem setting is closely related to recommender system.
- Bayesian personalized ranking for matrix factorization (BPRMF)
- neural collaborative filtering model (NCF)
- For each baseline has three different variants corresponding to the three estimated propensity scores.

Empirical Evaluation

Research Questions

- RQ1: How does the proposed model fare against standard recommendation model w.r.t. the performance of predicting fake news that user will share?
- RQ2: How is the fake news sharing behavior different from the true news sharing behavior in the latent space?

Empirical Evaluation

Research Questions

- RQ1: How does the proposed model fare against standard recommendation model w.r.t. the performance of predicting fake news that user will share?
- RQ2: How is the fake news sharing behavior different from the true news sharing behavior in the latent space?

Experiments

Results

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	13.31	16.38	18.77	20.8
BPRMF-N	14.92$^{\uparrow 12.2\%}$	17.61$^{\uparrow 7.5\%}$	19.70$^{\uparrow 5.0\%}$	21.52$^{\uparrow 3.5\%}$
BPRMF-U	14.97$^{\uparrow 12.6\%}$	17.70$^{\uparrow 8.1\%}$	19.73$^{\uparrow 5.1\%}$	21.58$^{\uparrow 3.8\%}$
BPRMF-Neu	**15.72**$^{\uparrow 18.2\%}$	**18.76**$^{\uparrow 14.5\%}$	**21.03**$^{\uparrow 12.0\%}$	**22.96**$^{\uparrow 10.4\%}$

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	10.52	11.32	11.86	12.30
BPRMF-N	12.38$^{\uparrow 17.7\%}$	13.11$^{\uparrow 15.8\%}$	13.60$^{\uparrow 14.7\%}$	13.97$^{\uparrow 13.6\%}$
BPRMF-U	12.22$^{\uparrow 16.2\%}$	12.95$^{\uparrow 14.4\%}$	13.42$^{\uparrow 13.2\%}$	13.81$^{\uparrow 12.3\%}$
BPRMF-Neu	**12.74**$^{\uparrow 21.1\%}$	**13.56**$^{\uparrow 19.8\%}$	**14.08**$^{\uparrow 18.7\%}$	**14.49**$^{\uparrow 17.8\%}$

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
NCF	5.87	8.01	9.72	11.63
NCF-N	7.59$^{\uparrow 29.3\%}$	9.50$^{\uparrow 18.6\%}$	11.22$^{\uparrow 15.4\%}$	12.74$^{\uparrow 9.5\%}$
NCF-U	**8.99**$^{\uparrow 53.2\%}$	**10.93**$^{\uparrow 36.5\%}$	**12.73**$^{\uparrow 31.0\%}$	**14.42**$^{\uparrow 24.0\%}$
NCF-Neu	8.36$^{\uparrow 42.4\%}$	10.53$^{\uparrow 31.5\%}$	12.39$^{\uparrow 27.5\%}$	13.97$^{\uparrow 20.1\%}$

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
NCF	4.41	4.97	5.37	5.77
NCF-N	5.96$^{\uparrow 35.1\%}$	6.50$^{\uparrow 30.8\%}$	6.91$^{\uparrow 28.7\%}$	7.23$^{\uparrow 25.3\%}$
NCF-U	**7.36**$^{\uparrow 66.9\%}$	**7.91**$^{\uparrow 59.2\%}$	**8.33**$^{\uparrow 55.1\%}$	**8.68**$^{\uparrow 50.4\%}$
NCF-Neu	6.53$^{\uparrow 48.1\%}$	7.14$^{\uparrow 43.7\%}$	7.57$^{\uparrow 41.0\%}$	7.91$^{\uparrow 37.1\%}$

L: GossipCop R: PolitiFact

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	12.36	22.18	31.10	39.51
BPRMF-N	14.45$^{\uparrow 16.9\%}$	25.11$^{\uparrow 13.2\%}$	34.34$^{\uparrow 10.4\%}$	42.72$^{\uparrow 8.1\%}$
BPRMF-U	14.78$^{\uparrow 19.6\%}$	25.65$^{\uparrow 15.6\%}$	34.91$^{\uparrow 12.2\%}$	**43.63**$^{\uparrow 10.4\%}$
BPRMF-Neu	**14.90**$^{\uparrow 20.6\%}$	**25.83**$^{\uparrow 16.5\%}$	**35.13**$^{\uparrow 13.0\%}$	43.55$^{\uparrow 10.2\%}$

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	5.33	7.51	9.22	10.71
BPRMF-N	6.39$^{\uparrow 19.9\%}$	8.73$^{\uparrow 16.2\%}$	10.49$^{\uparrow 13.8\%}$	11.97$^{\uparrow 11.8\%}$
BPRMF-U	**6.54**$^{\uparrow 22.7\%}$	8.92$^{\uparrow 18.8\%}$	10.69$^{\uparrow 15.9\%}$	**12.21**$^{\uparrow 14.0\%}$
BPRMF-Neu	6.53$^{\uparrow 22.5\%}$	**8.93**$^{\uparrow 18.9\%}$	**10.71**$^{\uparrow 16.2\%}$	12.19$^{\uparrow 13.8\%}$

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
NCF	9.59	18.45	27.30	36.33
NCF-N	**10.42**$^{\uparrow 8.7\%}$	**19.34**$^{\uparrow 4.8\%}$	28.58$^{\uparrow 4.7\%}$	37.07$^{\uparrow 2.0\%}$
NCF-U	10.29$^{\uparrow 7.3\%}$	19.29$^{\uparrow 4.6\%}$	27.34$^{\uparrow 0.1\%}$	34.87$^{\downarrow 4.0\%}$
NCF-Neu	10.20$^{\uparrow 6.4\%}$	19.11$^{\uparrow 3.6\%}$	**28.74**$^{\uparrow 5.3\%}$	**38.39**$^{\uparrow 5.7\%}$

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
NCF	3.72	5.66	7.35	8.94
NCF-N	4.13$^{\uparrow 11.2\%}$	6.09$^{\uparrow 7.6\%}$	**7.85**$^{\uparrow 6.8\%}$	9.36$^{\uparrow 4.7\%}$
NCF-U	**4.19**$^{\uparrow 12.6\%}$	**6.18**$^{\uparrow 9.2\%}$	7.75$^{\uparrow 5.4\%}$	9.10$^{\uparrow 1.8\%}$
NCF-Neu	4.04$^{\uparrow 8.6\%}$	5.99$^{\uparrow 5.8\%}$	7.82$^{\uparrow 6.4\%}$	**9.52**$^{\uparrow 6.5\%}$

- Observing that indeed the imposed IPS re-weighting confers an advantage to alleviating the selection bias in fake news dissemination.
- The improvement is most significant when K is small, e.g., $K = 20$.

Experiments

Results

L: GossipCop R: PolitiFact

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	13.31	16.38	18.77	20.8
BPRMF-N	14.92↑12.2%	17.61↑7.5%	19.70↑5.0%	21.52↑3.5%
BPRMF-U	14.97↑12.6%	17.70↑8.1%	19.73↑5.1%	21.58↑3.8%
BPRMF-Neu	**15.72**↑18.2%	**18.76**↑14.5%	**21.03**↑12.0%	**22.96**↑10.4%

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	10.52	11.32	11.86	12.30
BPRMF-N	12.38↑17.7%	13.11↑15.8%	13.60↑14.7%	13.97↑13.6%
BPRMF-U	12.22↑16.2%	12.95↑14.4%	13.42↑13.2%	13.81↑12.3%
BPRMF-Neu	**12.74**↑21.1%	**13.56**↑19.8%	**14.08**↑18.7%	**14.49**↑17.8%

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
NCF	5.87	8.01	9.72	11.63
NCF-N	7.59↑29.3%	9.50↑18.6%	11.22↑15.4%	12.74↑9.5%
NCF-U	**8.99**↑53.2%	**10.93**↑36.5%	**12.73**↑31.0%	**14.42**↑24.0%
NCF-Neu	8.36↑42.4%	10.53↑31.5%	12.39↑27.5%	13.97↑20.1%

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
NCF	4.41	4.97	5.37	5.77
NCF-N	5.96↑35.1%	6.50↑30.8%	6.91↑28.7%	7.23↑25.3%
NCF-U	**7.36**↑66.9%	**7.91**↑59.2%	**8.33**↑55.1%	**8.68**↑50.4%
NCF-Neu	6.53↑48.1%	7.14↑43.7%	7.57↑41.0%	7.91↑37.1%

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	12.36	22.18	31.10	39.51
BPRMF-N	14.45↑16.9%	25.11↑13.2%	34.34↑10.4%	42.72↑8.1%
BPRMF-U	14.78↑19.6%	25.65↑15.6%	34.91↑12.2%	**43.63**↑10.4%
BPRMF-Neu	**14.90**↑20.6%	**25.83**↑16.5%	**35.13**↑13.0%	43.55↑10.2%

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	5.33	7.51	9.22	10.71
BPRMF-N	6.39↑19.9%	8.73↑16.2%	10.49↑13.8%	11.97↑11.8%
BPRMF-U	**6.54**↑22.7%	8.92↑18.8%	10.69↑15.9%	**12.21**↑14.0%
BPRMF-Neu	6.53↑22.5%	**8.93**↑18.9%	**10.71**↑16.2%	12.19↑13.8%

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
NCF	9.59	18.45	27.30	36.33
NCF-N	**10.42**↑8.7%	**19.34**↑4.8%	28.58↑4.7%	37.07↑2.0%
NCF-U	10.29↑7.3%	19.29↑4.6%	27.34↑0.1%	34.87↓4.0%
NCF-Neu	10.20↑6.4%	19.11↑3.6%	**28.74**↑5.3%	**38.39**↑5.7%

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
NCF	3.72	5.66	7.35	8.94
NCF-N	4.13↑11.2%	6.09↑7.6%	**7.85**↑6.8%	9.36↑4.7%
NCF-U	**4.19**↑12.6%	**6.18**↑9.2%	7.75↑5.4%	9.10↑1.8%
NCF-Neu	4.04↑8.6%	5.99↑5.8%	7.82↑6.4%	**9.52**↑6.5%

- This indicates that the IPS-reweighting strategy is more effective when predicting fake news that is highly likely to be shared.

Experiments

Results

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	13.31	16.38	18.77	20.8
BPRMF-N	14.92↑12.2%	17.61↑7.5%	19.70↑5.0%	21.52↑3.5%
BPRMF-U	14.97↑12.6%	17.70↑8.1%	19.73↑5.1%	21.58↑3.8%
BPRMF-Neu	**15.72**↑18.2%	**18.76**↑14.5%	**21.03**↑12.0%	**22.96**↑10.4%

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	10.52	11.32	11.86	12.30
BPRMF-N	12.38↑17.7%	13.11↑15.8%	13.60↑14.7%	13.97↑13.6%
BPRMF-U	12.22↑16.2%	12.95↑14.4%	13.42↑13.2%	13.81↑12.3%
BPRMF-Neu	**12.74**↑21.1%	**13.56**↑19.8%	**14.08**↑18.7%	**14.49**↑17.8%

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
NCF	5.87	8.01	9.72	11.63
NCF-N	7.59↑29.3%	9.50↑18.6%	11.22↑15.4%	12.74↑9.5%
NCF-U	**8.99**↑53.2%	**10.93**↑36.5%	**12.73**↑31.0%	**14.42**↑24.0%
NCF-Neu	8.36↑42.4%	10.53↑31.5%	12.39↑27.5%	13.97↑20.1%

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
NCF	4.41	4.97	5.37	5.77
NCF-N	5.96↑35.1%	6.50↑30.8%	6.91↑28.7%	7.23↑25.3%
NCF-U	**7.36**↑66.9%	**7.91**↑59.2%	**8.33**↑55.1%	**8.68**↑50.4%
NCF-Neu	6.53↑48.1%	7.14↑43.7%	7.57↑41.0%	7.91↑37.1%

L: GossipCop R: PolitiFact

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	12.36	22.18	31.10	39.51
BPRMF-N	14.45↑16.9%	25.11↑13.2%	34.34↑10.4%	42.72↑8.1%
BPRMF-U	14.78↑19.6%	25.65↑15.6%	34.91↑12.2%	**43.63**↑10.4%
BPRMF-Neu	**14.90**↑20.6%	**25.83**↑16.5%	**35.13**↑13.0%	43.55↑10.2%

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
BPRMF	5.33	7.51	9.22	10.71
BPRMF-N	6.39↑19.9%	8.73↑16.2%	10.49↑13.8%	11.97↑11.8%
BPRMF-U	**6.54**↑22.7%	8.92↑18.8%	10.69↑15.9%	**12.21**↑14.0%
BPRMF-Neu	6.53↑22.5%	**8.93**↑18.9%	**10.71**↑16.2%	12.19↑13.8%

(a) Recall@K with K=20,40,60,80.

K	20	40	60	80
NCF	9.59	18.45	27.30	36.33
NCF-N	**10.42**↑8.7%	**19.34**↑4.8%	28.58↑4.7%	37.07↑2.0%
NCF-U	10.29↑7.3%	19.29↑4.6%	27.34↑0.1%	34.87↓4.0%
NCF-Neu	10.20↑6.4%	19.11↑3.6%	**28.74**↑5.3%	**38.39**↑5.7%

(b) NDCG@K with K=20,40,60,80.

K	20	40	60	80
NCF	3.72	5.66	7.35	8.94
NCF-N	4.13↑11.2%	6.09↑7.6%	**7.85**↑6.8%	9.36↑4.7%
NCF-U	**4.19**↑12.6%	**6.18**↑9.2%	7.75↑5.4%	9.10↑1.8%
NCF-Neu	4.04↑8.6%	5.99↑5.8%	7.82↑6.4%	**9.52**↑6.5%

- No evidence showing that which IPS strategy is most superior.
- Estimating propensity using user popularity and news content may be more effective than using news popularity alone.

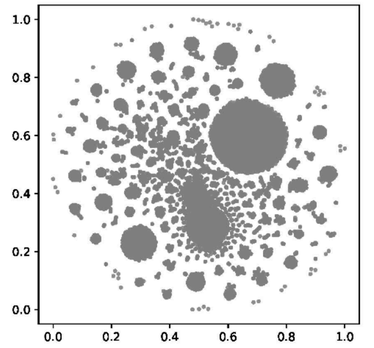
Empirical Evaluation

Research Questions

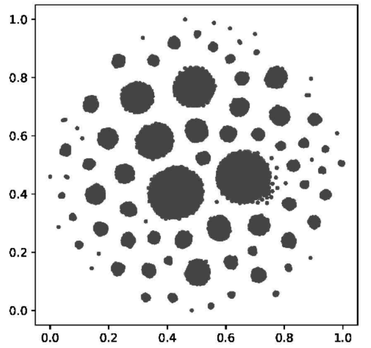
- RQ1: How does the proposed model fare against standard recommendation model w.r.t. the performance of predicting fake news that user will share?
- RQ2: How is the fake news sharing behavior different from the true news sharing behavior in the latent space?

Empirical Evaluation

Comparing News Sharing Behavior



(a) Fake news sharing behavior. Silhouette Coefficient=-0.124

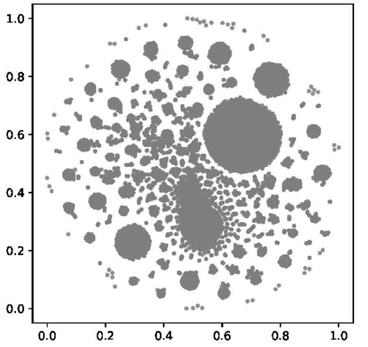


(b) True news sharing behavior. Silhouette Coefficient=0.903

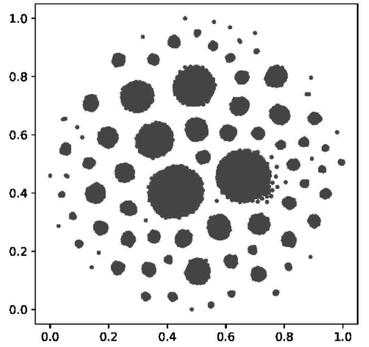
- Extract embeddings of users who shared fake news and who only shared true news.
- Run BPRMF-N on PolitiFact as working example and visualize in 2D space using t-SNE.
- Select users who only spread fake/true news and further conduct random sampling to make the number of analysis.
- Further performed DBSCAN clustering and compute Silhouette Coefficient of the inferred clusters.

Empirical Evaluation

Comparing News Sharing Behavior



(a) Fake news sharing behavior. Silhouette Coefficient=-0.124

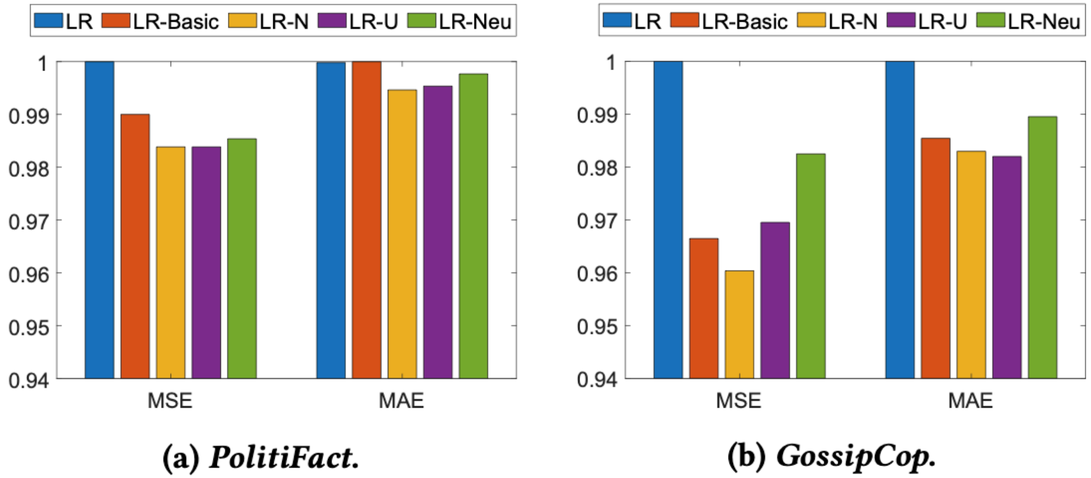


(b) True news sharing behavior. Silhouette Coefficient=0.903

- Fake news sharing behavior are more concentrated on a single primary cluster.
- True news sharing behavior are better separated into multiple and smaller clusters.
- Also evidenced by the results of Silhouette Coefficient, value of which ranges from -1 to 1. A larger value denotes that a sample is further away from its neighboring clusters.
- The Silhouette Coefficient of true news sharing behavior is close to 1, indicating that the samples are well matched to their own clusters.

Empirical Evaluation

Effect on Predicting User Susceptibility

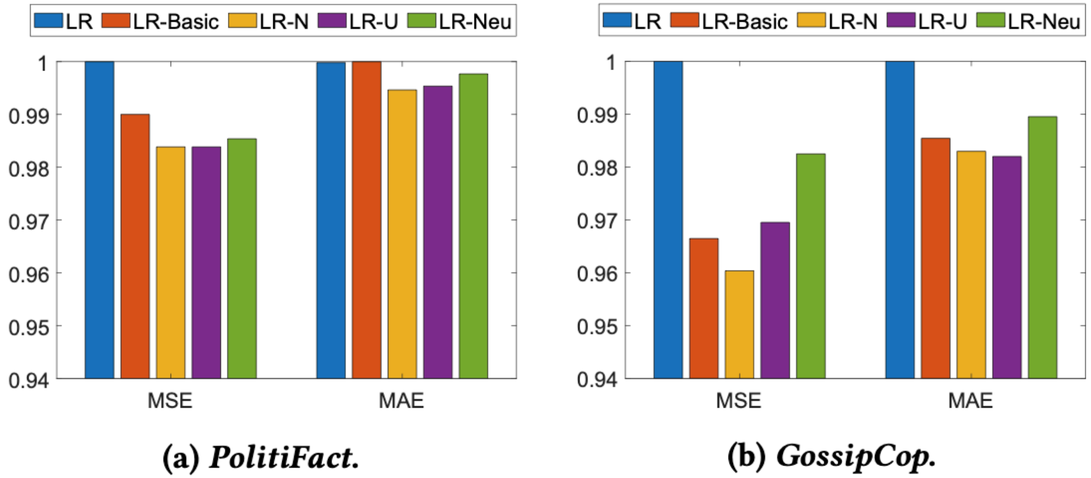


(a) *PolitiFact.* (b) *GossipCop.*

- Feed B_u along with a_u and U_u into equation to predicting user susceptibility.
- Take the simple Linear Regression (LR) as the basic model and compare the performance:
  - LR: solely consists of the user attributes.
  - LR-Basic: both user attributes and embeddings of user sharing behavior learned via BPRMF.
  - LR-N: both user attributes and embeddings of user sharing behavior learned via BPRMF-N.
  - LR-U: both user attributes and embeddings of user sharing behavior learned via BPRMF-U.
  - LR-Neu: both user attributes and embeddings of user sharing behavior learned via BPRMF-Neu.

Empirical Evaluation

Effect on Predicting User Susceptibility

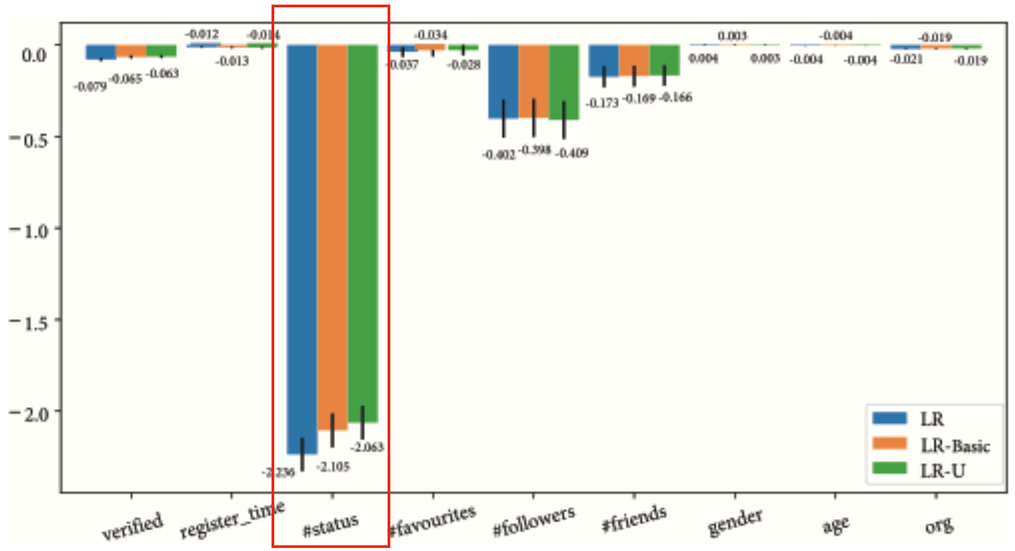


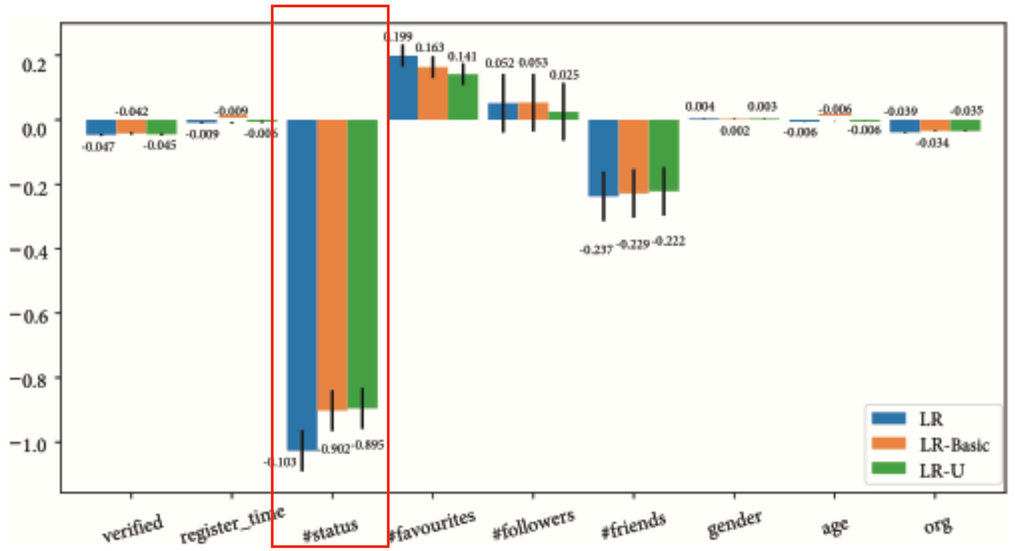
(a) *PolitiFact.* (b) *GossipCop.*

- Observing that the learned embeddings of user sharing behavior can improve the accuracy of predicting user susceptibility.
- Further, when taking the input of unbiased embeddings, can achieve the best results.
  - Especially for LR-N and LR-U.
- Conclude that when predicting user susceptibility, incorporating the unbiased embeddings of fake news sharing behavior as the confounder has more positive influence on standard predictive models compared to biased embeddings.

Empirical Evaluation

Identification of Causal User Profile Attributes



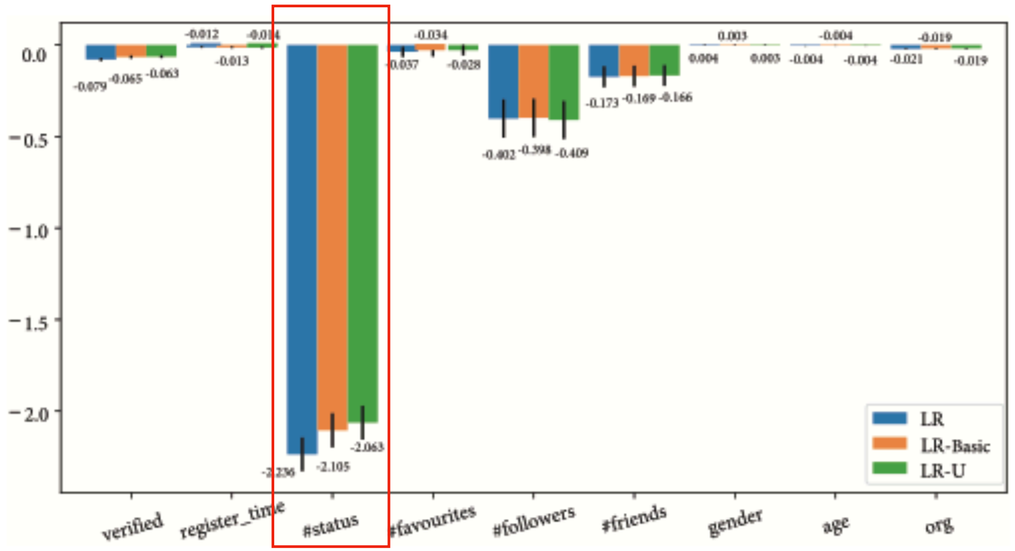


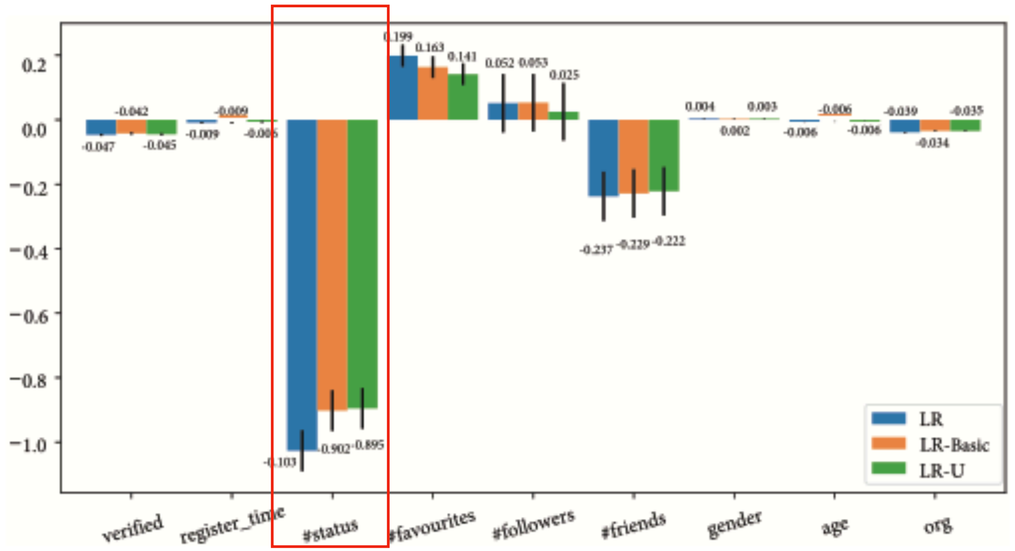
- Observe that LR-U presents more conservative estimations of the effects, see, e.g., #status – the number of Tweets (including retweets) issued by the user – for both datasets.

- This is partly because the unbiased embeddings can better alleviate the influence of confounding bias on the outcome.

Empirical Evaluation

Identification of Causal User Profile Attributes

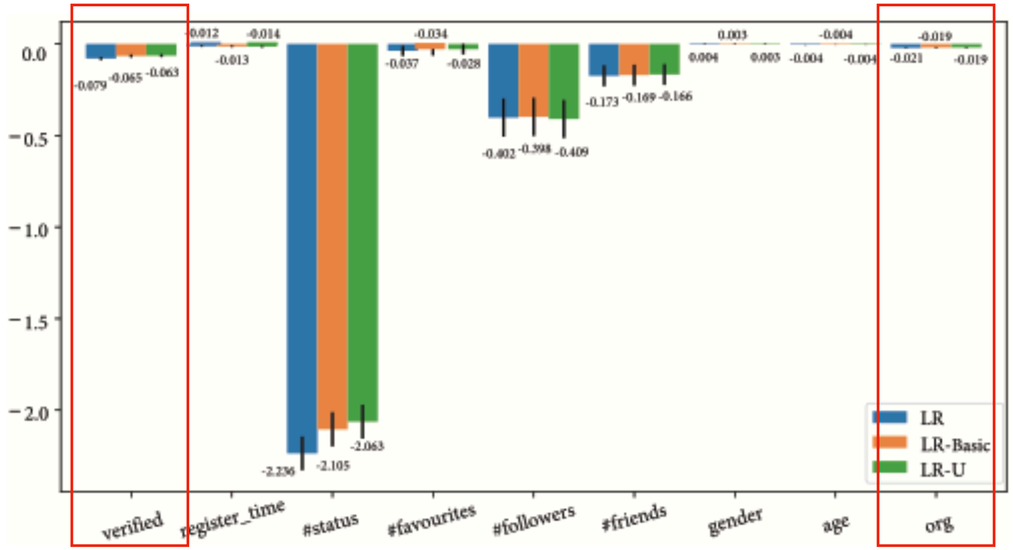


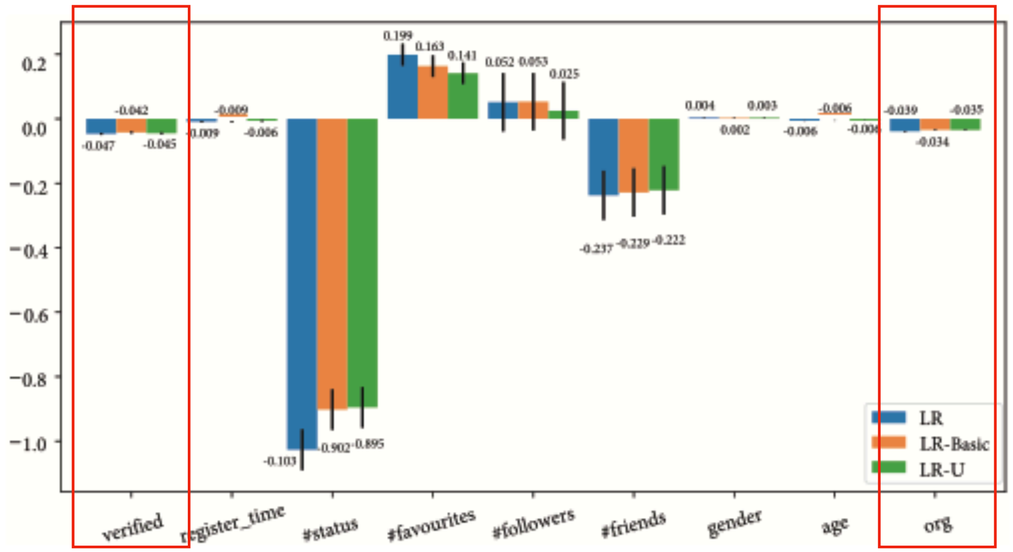


- #status has the largest effect on identifying a susceptible user, and the causal effect is negative.
- May infer that users who have historically issued more tweets (regardless they are fake or not) are less susceptible to spread fake news.

Empirical Evaluation

Identification of Causal User Profile Attributes

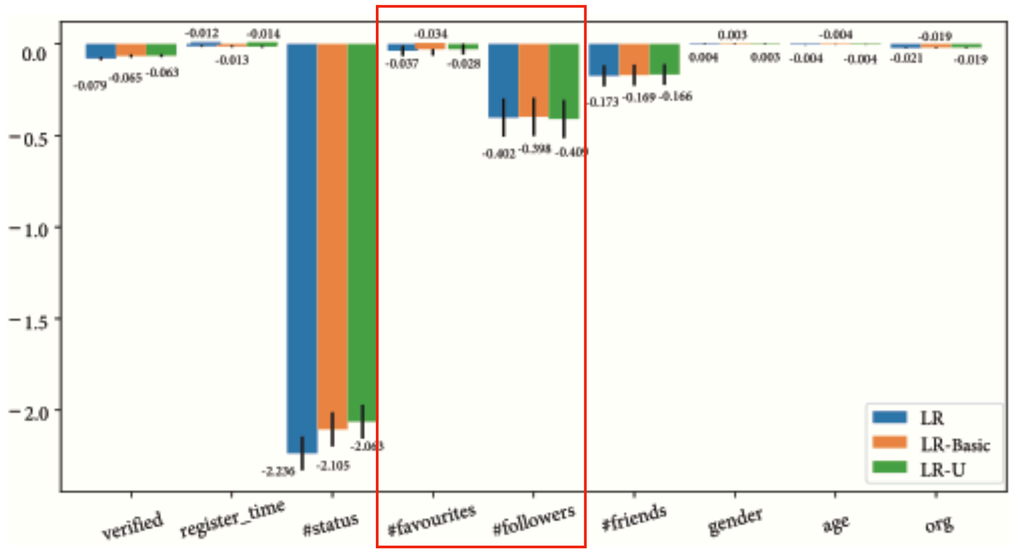


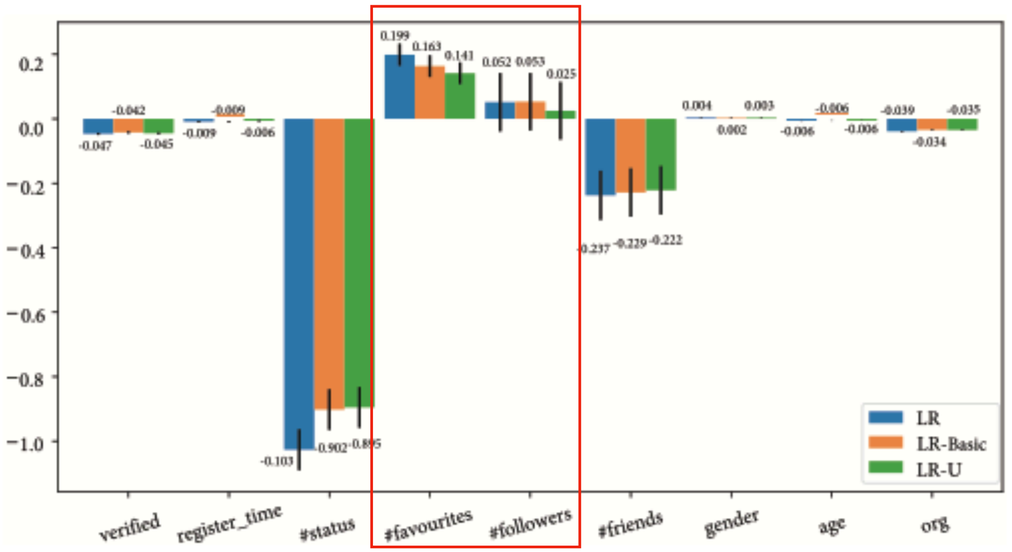


- Intuitively, verified users and organizations are less susceptible to share fake news.
- While lacking ground truth for causal user attributes, by identifying profile attributes that are intuitively causes, causal models might be applied to discovering more intrinsic user attributes that describe why people share fake news.

Empirical Evaluation

Identification of Causal User Profile Attributes

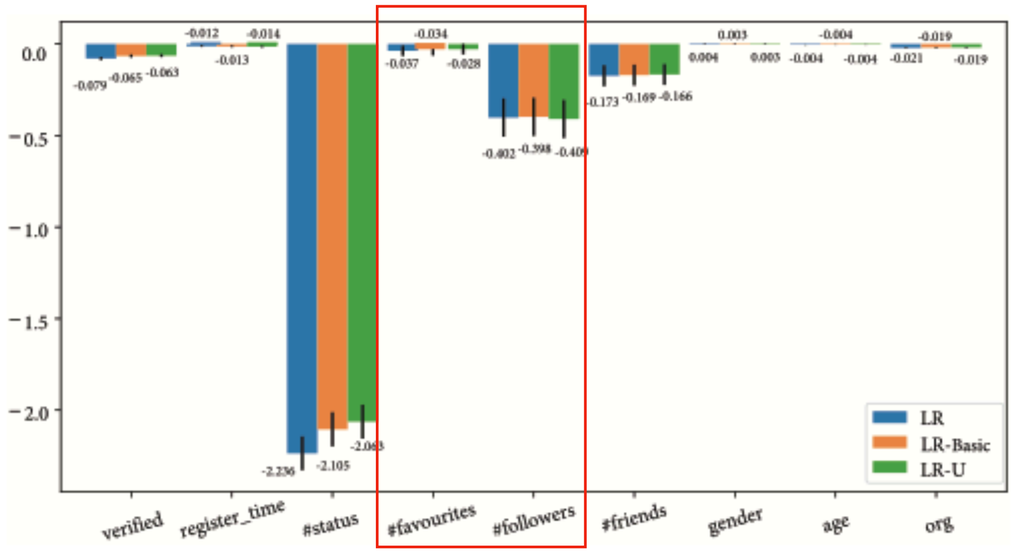


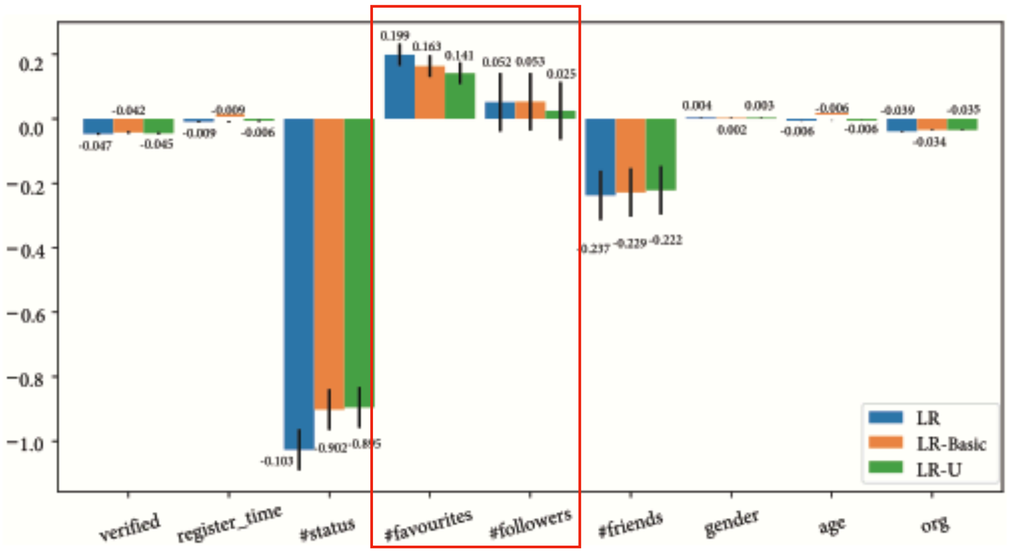


- Of particular interest is that there are two contradictory results across the two datasets: effects of both #favorites and #followers are negative in PolitiFact but become positive in GossipCop.

Empirical Evaluation

Identification of Causal User Profile Attributes

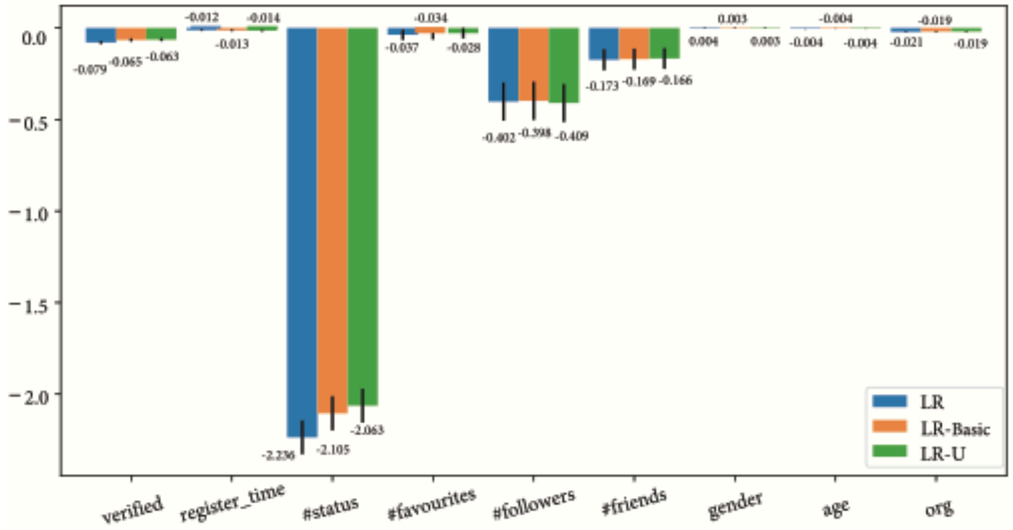


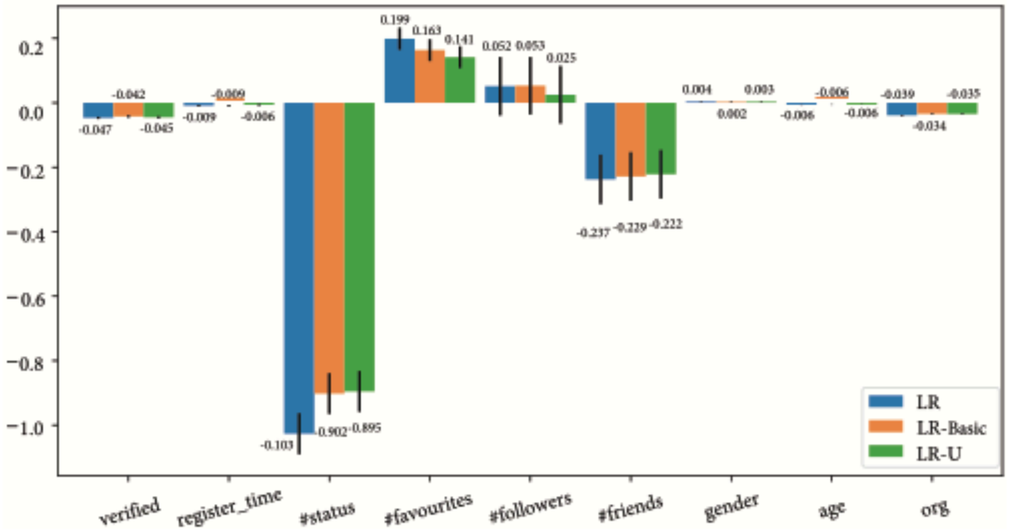


- Based on the causal transportability theory, these two attributes are less likely to be the causes of user susceptibility to share fake news.
- The category of online news platforms is a possible confounder that is left out by the surrogate confounder.

Empirical Evaluation

Identification of Causal User Profile Attributes





- Shows that proposed framework can learn unbiased embeddings of fake news sharing behavior that lead to more accurate predictions of fake news that users will share and user susceptibility.

Discussion

- Discuss the importance of understanding the causal relationships between user profile attributes and user susceptibility in combating the growing concerns about fake news.
- The results shown in this work demonstrate the efficacy of IPS-weighted news sharing models for learning unbiased fake news sharing behavior and the causal regression models for identifying user attributes potentially causing user susceptibility.
- Do not consider other important information sources such as social networks and comments of each news.
- The news content and attributes have yet to be fully explored.

Comments

of Causal Understanding of Fake News Dissemination on Social Media

- Focus on find sharing behavior causal understanding
- Not fully explore the dissemination graph
  - Like News Publisher, comments, user-user following
- The identified causal user attribute is simple and cannot give robust proof the relation with user susceptibility.
- The IPS-reweighting can used in other works as feature attribute.