Improving Generalizability of Fake News Detection Methods using Propensity Score Matching

Anonymous Author(s)

Abstract

Recently, due to the booming influence of online social networks, detecting fake news is drawing significant attention from both academic communities and general public. In this paper, we consider the existence of confounding variables in the features of fake news and use Propensity Score Matching (PSM) to select generalizable features in order to reduce the effects of the confounding variables. Experimental results show that the generalizability of fake news method is significantly better by using PSM than using raw frequency to select features. We investigate multiple types of fake news methods (classifiers) such as logistic regression, random forests, and support vector machines. We have consistent observations of performance improvement.

Introduction

In recent years, due to the rapid development of online social networks, more and more people tend to seek out and obtain news from social media than from traditional media. While it certainly makes people's life richer and easier, it gives fake news a lot of chance to spread. Compared with traditional suspicious information such as email spam and web spam, fake news has much worse societal impact. First, fake news spreads faster and broader. Traditional suspicious information often targets specific recipients and only produces a local impact. However, online fake news can disseminate exponentially, affecting more people. Second, there is no or very little cost of creating suspicious contents on social media, which makes malicious users easier to create fake news. Indeed, fake news has contributed to a wide range of social problems such as the polarization between political parties.

Due to the negative impact of fake news, fake news detection has aroused world-wide interest. However, statistical causal inference, which indicates generalizable features of causal information to infer the identity of fake news, has not been extensively investigated in this area, leaving a couple of related issues unaddressed. We identify three challenges that need to be addressed.

- First, there are various confounding variables in fake news corpora. Confounding variables are attributes that affect both dependent and independent variables. If they were left unnoticed when building machine learning algorithms, then we might not be able to learn generalizable causal features but only correlated features.
- Second, it is often difficult to extract useful features from text corpus. Traditional feature selection methods are susceptible to the existence of confounding variables. It is difficult to develop a feature selection method that can extract features of potential causal relationship.
- Third, since traditional methods did not take causality into consideration, machine learning models that were trained on one dataset might suffer from significant performance depreciation encoutering a new dataset that has a slightly different data distribution.

To address the above challenges, in this paper, we consider causal relations through a classical causality study method, Propensity Score Matching (Paul 2017). We obtain experimental data from open-source `FakeNewsNet` (Shu et al. 2018), which consists of data from two different sources, `politifact` and `gossipcop`. Then we select word features using Propensity Score Matching proposed. We adopted logistic regression in propensity score calculation. We evaluate feature selection methods using logistic regression and also extend them to other machine learning models. We conduct comparative analysis between datasets. Our method successfully improves the generalizability of the classifiers across multiple datasets.

Reproducibility

The `FakeNewsNet` (Shu et al. 2018) that we used in our paper is publicly available and can be found online.[1] Our code is publicly available.[2]

Related work

In this section, we discuss relevant topics to our proposed research work. The topics mainly include fake news detection and causal inference.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] https://github.com/KaiDMML/FakeNewsNet

[2] https://github.com/Arstanley/fakenews_pscore_match/

Fake News Detection

Online fake news detection has attracted a lot of attention from researchers. Most of them focus on applying machine learning classifiers to automatically identify fake news. We will summarize them into four categories based on the types of features they extracted and used.

- **Linguistic features extracted from news.** Castillo *et al.* used a series of linguistic features from news such as content length, emoticon, hashtag, etc. to access the credibility of a given set of tweets (Castillo, Mendoza, and Poblete 2011). Swear words, emotion words and pronouns are extracted to do credibility assessment (Gupta et al. 2014). Moreover, assertive verbs and factive verbs have been used (Potthast et al. 2017).
- **Linguistic features extracted from user comments.** User comments can reflect the authenticity of news. Zhao *et al.* detect fake news by inquiry phrases from users comments (Zhao, Resnick, and Mei 2015). Ma *et al.* and Chen *et al.* used RNN-based methods which captured linguistic features from users comments to detect rumors (Ma et al. 2016; Chen et al. 2018).
- **Structure features extracted from social networks.** Wu *et al.* proposed a graph kernel based hybrid SVM classifier which captured the high-order propagation patterns in addition to semantic features to do fake news detection (Wu, Yang, and Zhu 2015). Sampson *et al.* classified conversations through the discovery of implicit linkages between conversation fragments (Sampson et al. 2016).
- **Combine different types of features.** Castillo *et al.* usedfeatures from user's posting and re-posting behavior, from the text of the posts, and from citations to external sources (Castillo, Mendoza, and Poblete 2011). Yang *et al.* combined content-based, user-based, client-based, and location-based features (Yang et al. 2012). Kwon *et al.* examined a comprehensive set of user, structural, linguistic, and temporal features (Kwon, Cha, and Jung 2017).

Despite traditional methods of fake news detection, recently researchers focus on more specific yet challenging problems in this domain. Zhang *et al.* detected fauxtography (misleading images) on social media using directed acyclic graphs produced by user interactions (Daniel (Yue) Zhang and Wang 2018). Wang *et al.* adopted adversarial neural networks for early stage fake news detection (Yaqing Wang and Gao 2018). Shu *et al.* employed a co-attention neural network to detect fake news in an explainable framework using both user-based features and content-based features (Kai Shu and Liu 2019).

In this paper, our work will focus on the generalizing ability of machine learning models. The main objective of our work is not to improve the absolute performance of fake news detection. Here we employ content-based features only but our proposed methods can be generalized to other types of features if causal relationship exists.

Causal Machine Learning

Most of existing popular machine learning algorithms hold the assumption of independent and identically distributed (IID) data. Indeed they have reached impressive results in various big data problems (Y. LeCun and Hinton 2015). However, this is a strong assumption and not suitable for a lot of real world situations (Scholkopf 2019). Recently, various researchers have achieved great progress in causal machine learning. Pearl *et al.* introduced the causal graphs and structural causal models, incorporating the notion of intervention into statistical machine learning models (Pearl 2009). In addition, confounding variables have been discussed in different domains. Lu *et al.* addressed the presence of confounding variable in the setting of reinforcement learning by extending an actor-critic reinforcement learning algorithm to its de-confounding variant (C. Lu and Hernánde 2018). In the realm of text classification, Landeiro studied the problem by explicitly indicating the specific confounding variables that might misguide the classifier (V.Landeiro and Culotta 2016).

Propensity score matching, the technique that we use in this work, was proposed by Rosenbaum *et al.* to address the presence of confounding variables in statistical experiments (Rosenbaum and Rubin 1985). It has been extended to user-generated data and sentiment classification (N.A. Rehman and Chunara 2016; Dos Reis and Culotta 2015; Paul 2017). Paul *et al.* generalized the propensity score matching to a feature selection technique that takes confounding variables into consideration (Paul 2017). Causal methods have not been fully investigated in the realm of fake news detection.

Problem Definition

In this section, we formally define *confounding variables* and the task of *de-counfounding fake news detection*.

Definition 1 (Confounding Variables) *Let X be some independent variable, Y some dependent variable. We say Z is a confounding variable that confounds X and Y if Z is an unobserved variable that influence both X and Y (wik 2019).*

In this paper, we focus on mitigating the effects of confounding variables in fake news detection. First, we explain why causal techniques are necessary, and we justify the presence of confounding variables in fake news detection. Indeed, a great amount of fake news have political purpose. As a result, one classifier might take word "trump" as a useful feature, but intuitively, "trump" does not *causally* indicate a piece of news being fake. To make predictions based on such features will inevitably result in weak robustness. When encountering another news dataset that might be less political, it would perform worse. In order to address this issue, We formulate the problem as follows:

Let $N = (T, C)$ be a news entity that consists of a title T and contents C, $L = \{Fake, Real\}$ a binary label, we define the problem of de-counfounding fake news detection as follows: Assume there exists a confounding variable Z that confounds N and L. Given a dataset $S = \{(N_i, L_i) \mid 1 <= i <= n\}$ which consists of all the news and their corresponding labels, we aim at learning a map from the input space N to the label space L.

Proposed Approach

Overview

In this work, we use propensity score matching to select *decounfounded* features for fake news detection. Following the methods proposed in (Paul 2017), for each feature, we first calculate the propensity score of every sample regarding the specific feature. Then, we employ a one-to-one matching based on the propensity score. Finally, chi-square test statistics are used to rank the causal relevance of the features. More details are provided in the following sections.

Propensity Score

In statistics, causality analysis is often conducted with control-treatment pairs. However, in most of the real-word situations, it is impossible to obtain control-treatment pairs. Propensity score intends to solve the problem. As initially proposed in (Rosenbaum and Rubin 1985), it is defined as the probability of a subject to receive a certain treatment. In the realm of fake news detection, we regard each word feature as a treatment and each news sample as a subject. Then we formally define the propensity score as below.

Definition 2 (Propensity Score) *Let w be a word feature, X a news corpus. We say $psm(w, X)$ is the propensity score of w regarding X and*

$$psm(w, X) = P(w|X - \{w\})$$

The propensity score can be estimated in many different ways (2011a 2011). In this work, we conduct experiment with logistic regression and random forest regression. And we compare them in the experiment section.

Matching

By pairing the subjects that have similar propensity scores, we can eliminate the bias caused by confounding variables. Among different matching strategies (2011a 2011), we use one-to-one matching for its efficiency. We rank the subjects by their propensity score and greedily find the matched subjects. Each pair of matched subjects consists of (1) a treatment unit, (2) a text corpus that contains the word feature and a control unit, and (3) a text corpus that does not contain the word feature but has a similar propensity score. We finally calculate the chi-square test statistics for each feature with the paired subjects with

$$X^2 = \frac{(TN - CP)^2}{TN + CP}, \tag{1}$$

where TN stands for treatment-negative and CP stands for control-positive.

Experiments

Dataset

We conduct experiments to evaluate our propensity score matching-based approach with a widely used fake news dataset, `FakeNewsNet` (Shu et al. 2018). The dataset consists of news corpora from two primary sources:

- `PolitiFact`: In `PolitiFact`, political news collected from various sources are fact-checked by experts and journalists. Specifically, we use the sample data provided by (Shu et al. 2018). It consists of $1,056$ data points for the `PolitiFact` section. And it includes 624 real news and 432 fake news documents.
- `GossipCop`: `GossipCop` is a website that collects entertainment stories from various sources with fact-check score that ranges from 1 to 10. However, since the website intends to showcase mostly fake stories, the majority of the stories have scores less than five. Real stories are collected from *E! Online*, a widely-accepted reliable entertainment website. The samples that we use in this work consists of 16,817 real stories and 5,323 fake stories.

Since for both `PolitiFact` and `GossipCop`, real samples are fewer than fake samples, in order to balance the label distribution, we randomly sample 432 and 5323 real samples from the two datasets, respectively.

Experiment Settings and Results

We use document frequency as the baseline feature selection method and we intend to compare the generalization ability. Also, we want to observe how well the model performance evolves when adding more features. For simplicity, we consider a standard logistic classifier with default parameters provided by scikit-learn (Pedregosa et al. 2011). We train two classifiers on both datasets separately with different percentage of features and then evaluate on each other. We visualize our results in Figure 1 and Figure 2. The curve starts at the point where 1% of the top features are selected. As we can see, PSM consistently outperforms the baseline feature selection method based on document frequency in the task of fake news prediction across data regardless of the percentage of features used. We notice that the gap is smaller in Figure 2 than that in Figure 1. It is due to the fact that entertainment stories have more diverse themes which makes feature selected by document frequency more reliable. In general, as illustrated by both graphs, by applying propensity score matching to feature selection, we obtain models with better generalization ability. To summarize the graph to a more concise metric, we calculate the AUROC in Table 1.

It is noticeable that when using a relatively small percentage of selected features, our PSM-based method significantly outperforms the baseline method. It is reasonable since the top features selected from PSM should be more accurately representative of what fake news might look like and less subject to specific topics. To illustrate this point, we provide an empirical analysis in the next section.

Empirical Analysis

To better understand how propensity score matching improves the generalization ability, we showcase some specific examples in this section. We only consider the model trained on `PolitiFact` in this section for a clearer representation. We show top five features from the baseline method and also top five features from propensity score matching in Table 2. Clearly, when using document frequency for feature selection, most of the top features turn to be political figures. It

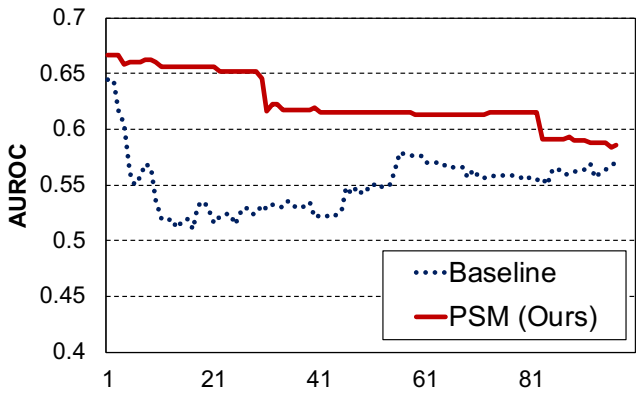

Figure 1: We developed a fake news classification model and trained it on PolitiFact. We evaluated the model on GossipCop. Clearly, our PSM-based method performs better than the baseline method (higher AUROC).

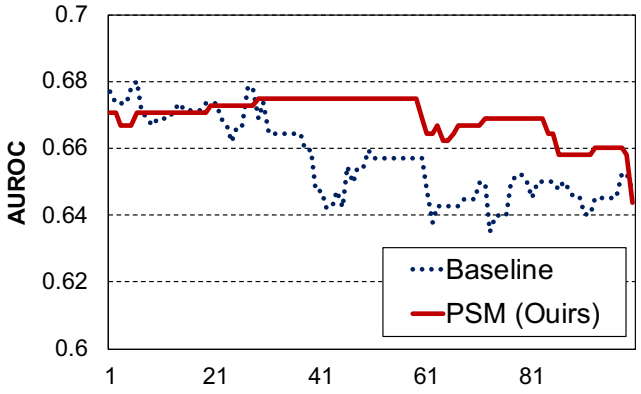

Figure 2: We developed a fake news classification model and trained it on GossipCop. We evaluated the model on PolitiFact. Clearly, our PSM-based method performs better than the baseline method (higher AUROC).

is self-evident that features selected by document frequency only reflects the distribution of the PolitiFact dataset, so it cannot be generalized well when there is a change of distribution. In contrast, top features acquired from PSM are more likely to be patterns of fake stories in general, and words like "confirmed" and "inside" could be reliable features that generalize to other datasets.

Future Work

Although we obtain improvements in the generalization of fake-news detection, there still remains a couple of challenges in this area. One thing worth noticing is that PSM only accounts for biases caused by observed variables. Researchers could focus on mitigating the biases caused by latent variables. One approach could be extending PSM to latent representations learned by deep neural networks. Another direction of improvements could be causal fake-news detection with Bayesian Networks and Structural Equation Models proposed by Pearl *et al.* (Pearl 2009) which ensures there will be no hidden confounding variables.

	Baseline	PSM (Ours)
PolitiFact	0.56	**0.68**
GossiCop	0.63	**0.67**

Table 1: Experimental results (AUROC Score) show that PSM-based method performs better than baseline method.

Baseline	PSM (ours)
"trump"	"makes"
"obama"	"leaves"
"senator"	"confirmed"
"donald"	"nightmare"
"action"	"inside"

Table 2: Top features that the baseline method and our PSM-based method discovered in the PolitiFact dataset.

Conclusions

In this work, we conducted a data-driven study on the generalization ability of fake news detection models. We approached the task by introducing propensity score matching into the feature selection process. We study the generalization ability of fake news detection models. In conclusion, our experimentation shows significant improvement of using propensity score matching as feature selection compared with baseline model on the generalizability.

References

2011a, P. A. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*.

C. Lu, B. S., and Hernánde, J. M. 2018. Deconfounding reinforcement learning in observational settings. *arxiv preprint arXiv:1812.10576*.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, 675–684. ACM.

Chen, T.; Li, X.; Yin, H.; and Zhang, J. 2018. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 40–52. Springer.

Daniel (Yue) Zhang, Lanyu Shang, B. G. S. L. K. L. H. Z. M. T. A., and Wang, D. 2018. Fauxbuster: A content-free fauxtography detector using social media comments. *IEEE International Conference on Big Data (BigData)*.

Dos Reis, V. L., and Culotta, A. 2015. Using matched samples to estimate the effects of exercise on mental health via twitter. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Gupta, A.; Kumaraguru, P.; Castillo, C.; and Meier, P. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*, 228–243. Springer.

Kai Shu, Limeng Cui, S. W. D. L., and Liu, H. 2019. defend: Explainable fake news detection. *Proceedings of 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Kwon, S.; Cha, M.; and Jung, K. 2017. Rumor detection over varying time windows. *PloS one* 12(1):e0168344.

Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B. J.; Wong, K.-F.; and Cha, M. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Ijcai*, 3818–3824.

N.A. Rehman, J. L., and Chunara, R. 2016. Using propensity score matching to understand the relationship between online health information sources and vaccination sentiment. *AAAI Spring Symposium on Observational Studies through Social Media and Other Human-Generated Content*.

Paul, M. J. 2017. Feature selection as causal inference: Experiments with text classification. *Conference on Compuatational Natural Language Learning*.

Pearl, J. 2009. Causality.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and Stein, B. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Rosenbaum, P., and Rubin, D. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician 39:33-38*.

Sampson, J.; Morstatter, F.; Wu, L.; and Liu, H. 2016. Leveraging the implicit structure within social media for emergent rumor detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2377–2382. ACM.

Scholkopf, B. 2019. Causality for machine learning. *arxiv preprint arXiv:1911.10500*.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

V.Landeiro, and Culotta, A. 2016. Robust text classification in the presence of confounding bias. *AAAI*.

2019. Confounding variables. *Wikipedia*.

Wu, K.; Yang, S.; and Zhu, K. Q. 2015. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*, 651–662. IEEE.

Y. LeCun, Y. B., and Hinton, G. 2015. Deep learning. *Nature*.

Yang, F.; Liu, Y.; Yu, X.; and Yang, M. 2012. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, 13. ACM.

Yaqing Wang, Fenglong Ma, Z. J. Y. Y. G. X. K. J. L. S., and Gao, J. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. *ACM SigKDD*.

Zhao, Z.; Resnick, P.; and Mei, Q. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, 1395–1405. International World Wide Web Conferences Steering Committee.