Joint Estimation of User And Publisher Credibility for Fake News Detection

Rajdipa Chowdhury[§]
UC Santa Cruz
rachowdh@ucsc.edu

Sriram Srinivasan[§]
UC Santa Cruz
ssriniv9@ucsc.edu

Lise Getoor
UC Santa Cruz
getoor@ucsc.edu

ABSTRACT

Fast propagation, ease-of-access, and low cost have made social media an increasingly popular means for news consumption. However, this has also led to an increase in the preponderance of fake news. Widespread propagation of fake news can be detrimental to society, and this has created enormous interest in fake news detection on social media. Many approaches to fake news detection use the news content, social context, or both. In this work, we look at fake news detection as a problem of estimating the credibility of both the news publishers and users that propagate news articles. We introduce a new approach called the *credibility score-based model* that can jointly infer fake news and credibility scores for publishers and users. We use a state-of-the-art statistical relational learning framework called probabilistic soft logic to perform this joint inference effectively. We show that our approach is accurate at both fake news detection and inferring credibility scores. Further, our model can easily integrate any auxiliary information that can aid in fake news detection. Using the *FakeNewsNet*[1] dataset, we show that our approach significantly outperforms previous approaches at fake news detection by up to 10% in recall and 4% in accuracy. Furthermore, the credibility scores learned for both publishers and users are representative of their true behavior.

CCS CONCEPTS

• **Computing methodologies** → **Statistical relational learning**; • **Information systems** → **Social networks**;

KEYWORDS

Fake News Detection;Collective Classification;Social Network

ACM Reference Format:
Rajdipa Chowdhury[§], Sriram Srinivasan[§], and Lise Getoor. 2020. Joint Estimation of User And Publisher Credibility for Fake News Detection. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3412066

[1]github.com/KaiDMML/FakeNewsNet/tree/old-version/Data

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10.
https://doi.org/10.1145/3340531.3412066

1 INTRODUCTION

The number of users on social media and their increased engagement has led to a drastic shift in how users consume news. It has been reported that the number of users consuming news from social media increased from 49% to 62% from the years 2012 to 2016.[2] While social media is a convenient platform for users to consume and share news, it also makes it easy for users to publish inaccurate or intentionally misleading information (a.k.a., fake news). Content on social media can go viral, quickly misinforming millions of users, and can have severe consequences on the economy[3] or political stability of a nation [1]. Thus, it has become crucial to be able to detect fake news and mitigate the effects of spreading misinformation [9, 13].

Significant effort has been invested in fake news detection. Some of the initial works focused on exploiting the structure in news content in order to identify a news article as fake [4, 8, 17]. However, identifying fake news with text alone is challenging as, in many cases, these articles are written with the intention to misinform people, making it hard to distinguish from real news based on text alone. Alternatively, some approaches use the social context information available to effectively detect fake news [18, 19] . More recent methods use both the news content and social context to better detect fake news [11, 15, 16].

While the previous approaches have been effective, they tend to be complicated and focus only on detecting whether or not a news article is fake. In this work, we propose a simple yet effective approach for detecting fake news by inferring the credibility of the publisher publishing the news and the credibility of the users that share them; we refer to this approach as *credibility score-based model* (CSM). Some of the previous approaches that model user or publisher bias [5, 14] infer credibility indirectly by first learning the partisan bias of the user or publisher and then predicting their credibility. In this paper, we directly learn publisher credibility scores (PCSs) and user credibility scores (UCSs) by jointly inferring both the credibility scores (CSs) and the fake news labels. In order to do this, we make use of a powerful statistical relational learning (SRL) [6, 10] framework called probabilistic soft logic (PSL) [2]. Further, we show that any prior knowledge, such as a publisher's trustworthiness obtained from websites such as MBFC[4], can be easily incorporated in our model.

Our key contributions include: 1) we introduce a simple yet effective approach to identify fake news in social media; 2) we show how publisher and user credibility can be implicitly learned

[§]These authors contributed equally to this work.
[2]journalism.org/2016/05/26/news-use-across-social-media-platforms-2016
[3]cheq.ai/fakenews
[4]mediabiasfactcheck.com

by jointly inferring fake news labels and credibility scores; 3) we show that any external knowledge can be easily incorporated in our model; 4) through empirical evaluation on the FakeNewsNet dataset [12], we show that CSM can get up to 10% improvement in recall and 4% improvement in accuracy in the Politifact dataset; and 5) we show that the CSs learned for both the publishers and the users are accurate.

2 PROBLEM DEFINITION

Let $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$ be a set of N news articles, $\mathcal{P} = \{p_1, p_2, \ldots, p_P\}$ be a set of P publishers, and $\mathcal{U} = \{u_1, u_2, \ldots, u_U\}$ be a set of U users in a social network. We denote a user-news interaction matrix $S \in \{0,1\}^{U \times N}$, where $S_{u,n} = 1$ implies that user u shares news n in a social network at least once. Note that users may sometimes share fake news expressing their disagreement; we do not treat this differently as this is still fake news propagation. Next, we define a publisher-news matrix $I \in \{0,1\}^{P \times N}$, where $I_{p,n} = 1$ implies that news n was issued by publisher p. Further, we assume that, for a subset of publishers, a publisher trust score t_p can be obtained from external sources like MBFC. Given the information above our task is to label news articles $\mathcal{A}$ as fake ($l_a = 1$) or real ($l_a = 0$). Further, we assume that a subset of labels $L_o \subset L$ is observed and the rest $L_u = L - L_o$ are unobserved. Formally, the problem is defined as:

Definition 1. *Given news articles $\mathcal{A}$, users $\mathcal{U}$, publishers $\mathcal{P}$, user-news interaction matrix S, publisher-news matrix I, partially available publisher trust t_p, and partially observed news labels L_o, the task is to infer the rest of the labels L_u.*

In order to solve this problem, we propose to learn two latent factors, publisher and user credibility, which we use to infer the fake news labels jointly. To accomplish this, we use a powerful SRL framework called PSL, which we briefly describe in the next section.

3 PROBABILISTIC SOFT LOGIC

PSL is a probabilistic programming language that is effective at reasoning over structured data and output. A model in PSL is defined through a set of weighted first-order logical rules. These logical rules can be interpreted as a continuous relaxation of Boolean logic. A weighted logical rule is generally of the form:

$$w : \textsc{UserShare}(\mathtt{U}, \mathtt{N}) \wedge \textsc{UserCred}(\mathtt{U}) \rightarrow \neg \textsc{FakeNews}(\mathtt{N})$$

where $w \in \mathbb{R}^+$ is a learnable weight of the rule (also interpreted as importance of satisfying the rule), FakeNews, UserCred, and UserShare are predicates and N and U are placeholders for news articles and users. This rule when instantiated with data, i.e, $N = a \in \mathcal{A}$ and $U = u \in \mathcal{U}$ ($w : \textsc{UserShare}(\mathtt{u}, \mathtt{a}) \wedge \textsc{UserCred}(\mathtt{u}) \rightarrow \neg \textsc{FakeNews}(\mathtt{a})$), is referred to as ground rule and each predicate in a ground rule, such as FakeNews(a), is referred to as a ground predicate. We explain this rule in more detail in Section 4.2. Each ground predicate is represented as a continuous random variable in the range $[0, 1]$, and each ground rule represents a clique in a special type of Markov random field called a hinge-loss Markov random field (HL-MRF). Based on data, some random variables are observed X, and some are unobserved Y, and the task of inference in PSL is to estimate the value for Y given X. For example, for the above rule in our setting, all random variables generated by UserShare are observed while random variables generated from FakeNews are partially observed and random variables generated from UserCred are fully unobserved. The probability density of a HL-MRF is given by: $P(Y|X) \propto exp\Big(-\sum_{i=1}^{m} w_i \phi_i(Y,X)\Big)$ where, $\phi_i = \max\{0, \ell_i(Y,X)\}^{d_i}; d_i \in \{1,2\}$, m is the total number of cliques, ϕ_i is a potential function associated with each clique generated by a ground rule, ℓ_i is a linear function, d_i gives the flexibility to choose between linear and squared hinge loss (we only use squared in this paper), and w_i is the weight associated with the rule. The task of inference can be written as: $\arg\max_Y P(Y|X) = \text{argmin}_Y \sum_{i=1}^{m} w_i \phi_i(Y,X)$. The above expression is solved using alternating direction method of multipliers (ADMM) [3].

4 CREDIBILITY SCORE BASED MODEL

In this section, we describe our *credibility score-based model* (CSM), which jointly learns publisher and user credibility and infers news labels. We first discuss how the PCSs are inferred, followed by the UCSs.

4.1 Publisher credibility

Fake news is often written by publishers with strong partisan bias which affects the credibility of the published news [5]. Websites like MBFC analyze and provide such biases and also generate trustworthiness scores t_p for publishers. These scores are between zero to five, with five being most credible and zero being least. We create a predicate called $\textsc{MBFC}(\mathtt{P}) = \frac{t_p}{5}$ which represents the MBFC score and is fully observed. While this is a good source of information for publisher credibility, it is not complete and can be biased. Therefore, we treat it as a prior for publisher credibility. We introduce a new predicate called $\textsc{PubCred}(\mathtt{P})$, which represents the latent PCS. As this predicate models a latent variable, this value is unobserved for all publishers and needs to be inferred. We first incorporate the prior information using the following rules:

$$w_1 : \textsc{MBFC}(\mathtt{P}) \rightarrow \textsc{PubCred}(\mathtt{P}) \quad (1)$$

$$w_2 : \neg\textsc{MBFC}(\mathtt{P}) \rightarrow \neg\textsc{PubCred}(\mathtt{P}) \quad (2)$$

Next we learn the PCS from data by jointly inferring the value for unobserved labels of news articles Y_u with the rules below:

$$w_3 : \textsc{PubCred}(\mathtt{P}) \wedge \textsc{NewsPub}(\mathtt{N}, \mathtt{P}) \rightarrow \neg\textsc{FakeNews}(\mathtt{N}) \quad (3)$$

$$w_4 : \textsc{FakeNews}(\mathtt{N}) \wedge \textsc{NewsPub}(\mathtt{N}, \mathtt{P}) \rightarrow \neg\textsc{PubCred}(\mathtt{P}) \quad (4)$$

where $\textsc{NewsPub}(\mathtt{N}, \mathtt{P})$ is fully observed and is true when $I_{P,N} = 1$. The above rules encode the intuition that a publisher that is not credible will publish fake news, and a credible publisher will not. Performing inference with the above set of rules generates predictions for $Y_u \in [0,1]^{|Y_u|}$ and a PCS in range $[0, 1]$ for all publishers. We refer to a model that uses the above rules only as *publisher CSM* (PCSM).

4.2 User credibility

Unlike publisher credibility, there is no explicit information available for user credibility. Previous approaches [7, 14] generally exploit user attributes or behavior or learn partisan bias to estimate user credibility. In our approach, we learn user credibility by jointly reasoning about fake news and user's news sharing behavior. In order to accomplish this, we introduce a latent predicate $\textsc{UserCred}(\mathtt{U})$, which is fully unobserved and represents a UCS in the range $[0, 1]$.

Table 1: Performance of different approaches at fake news detection. Numbers in bold are significant with $p < 0.05$

Datasets	Metrics	LR-Unigram	LR-Bigram	TriFN	PCSM	UCSM	CSM
Politifact	Accuracy	0.801 (0.059)	0.852 (0.047)	0.878 (0.017)	0.875 (0.047)	**0.890 (0.048)**	**0.913 (0.040)**
	Precision	0.778 (0.089)	0.807 (0.076)	0.867 (0.034)	**0.957 (0.050)**	0.846 (0.074)	0.879 (0.069)
	Recall	0.876 (0.084)	0.934 (0.055)	0.893 (0.023)	0.791 (0.092)	**0.957 (0.038)**	**0.961 (0.032)**
	F1	0.818 (0.059)	0.863 (0.050)	0.88 (0.015)	0.862 (0.058)	**0.896 (0.048)**	**0.917 (0.040)**
Buzzfeed	Accuracy	0.713 (0.067)	0.734 (0.076)	**0.864 (0.026)**	0.827 (0.053)	**0.858 (0.051)**	**0.858 (0.050)**
	Precision	0.697 (0.094)	0.704 (0.086)	**0.849 (0.040)**	0.787 (0.070)	0.779 (0.077)	0.787 (0.069)
	Recall	0.770 (0.141)	0.803 (0.154)	0.893 (0.013)	0.888 (0.106)	**0.993 (0.018)**	**0.979 (0.035)**
	F1	0.717 (0.084)	0.740 (0.085)	**0.87 (0.019)**	0.829 (0.063)	**0.871 (0.047)**	**0.870 (0.043)**

A value of one implies the user is entirely credible, while a value of zero implies the opposite. We introduce the following rules to incorporate and learn UCS:

$$w_5 : \text{UserShare}(U, N) \wedge \neg\text{FakeNews}(N) \rightarrow \text{UserCred}(U) \quad (5)$$

$$w_6 : \text{UserShare}(U, N) \wedge \text{UserCred}(U) \rightarrow \neg\text{FakeNews}(N) \quad (6)$$

where $\text{UserShare}(U, P)$ is fully observed and is true when $S_{U,N} = 1$. The above rules encode the intuition that a credible user will often share real news while a user that is not credible will share more fake news. Finally, performing inference with the above set of rules leads to the prediction for $Y_u \in [0, 1]^{|Y_u|}$ and a latent UCS for all users. We refer to a model that uses the above rules only as *user CSM* (UCSM). Our final model combines PCSM and UCSM to generate CSM.

5 EMPIRICAL EVALUATION

In this section, we demonstrate the performance of CSM using the FakeNewsNet dataset. Using our empirical evaluation, we answer three research questions: RQ1) is CSM effective at fake news detection? RQ2) is UCSM better than PCSM? RQ3) are the PCSs and the UCSs learned representative of data?

5.1 Experimental Setup and Methods

The FakeNewsNet dataset from Shu et al. [15] contains data from two sources, Politifact and Buzzfeed. The social context for these datasets is mined from Twitter. Politifact dataset contains 23,865 users, 88 publishers, 37,259 social engagements, 120 fake news, and 120 real news. Similarly, the Buzzfeed dataset contains 15,257 users, 27 publishers, 25,240 social engagements, 91 fake news, and 91 real news. Similar to previous work [15], we use accuracy, precision, recall, and F1 to evaluate the performance of different approaches. For a fair comparison, we use the same approach as TriFN [15] and randomly choose 80% data for training and 20% for testing, we repeat this process 30 times (instead of 10 times in TriFN), and report the mean performance and their standard deviations. We learn all the hyperparameters using the training data only. We perform independent T-test to ensure significance with $p < 0.05$. In this work, we evaluate on three baselines and three CSMs:

LR-Unigram: a logistic regression (LR) model which uses the unigrams of the news content as features.

LR-Bigram: a LR model that uses bigrams of the news content as features. Both LR-Unigram and LR-Bigram are common ways of performing classification based on text.[5]

TriFN: work by Shu et al. [15] that makes use of news features, user features, user-news interaction, user-user interaction, and publisher-news interaction to perform fake news detection. They show that their approach outperforms many previous approaches [4, 8]. Note, as their code is not publicly available, we report the evaluation metrics provided in their paper.

PCSM: model defined by the rules in Section 4.1.

UCSM: model defined by the rules in Section 4.2.

CSM[6]: combine rules from PCSM and UCSM and the weights for rules are learned using continuous random grid search.

5.2 Performance Analysis

In this section, we answer RQ1 and RQ2. First, to answer RQ1, we compare CSM with the LR-Unigram, LR-Bigram, and TriFN. Table 1 shows accuracy, precision, recall, and F1 obtained on the Politifact and the Buzzfeed datasets. Overall we observe that LR-Bigram tends to be better than LR-Unigram on all metrics and for both datasets. Next, we observe that TriFN is better than both LR models on almost all metrics except recall in Politifact. Finally, when we compare our approach CSM with TriFN which, to the best of our knowledge, is the previous state-of-the-art, we observe that CSM is significantly better than TriFN in the Politifact dataset on all metrics except precision where they are similar. In the Buzzfeed dataset, we observe that there is no significant difference between TriFN and CSM in terms of accuracy and F1, but TriFN is significantly better than CSM at precision while CSM is better than TriFN at recall. This indicates that CSM tends to be conservative and labels more news as fake compared to TriFN. Overall, we observe that CSM is effective at fake news detection and mostly outperforms TriFN. We believe that the effectiveness of CSM is due to the joint inference performed using PSL, as both CSM and TriFN use the same information sources to perform fake news detection.

To answer RQ2, we evaluate PCSM and UCSM models on both the Politifact and Buzzfeed datasets and report the four metrics in Table 1. We observe here that UCSM outperforms PCSM on almost all metrics in both datasets. We believe UCSM outperforms PCSM here because the signal obtained from users sharing news is stronger and more reliable than the PCSs obtained using a few news articles. Publisher's credibility can be seen as a natural prior when not enough user shares are available. However, with sufficient social

[5] https://www.kaggle.com/mdepak/fakenewsnet

[6] https://github.com/linqs/chowdhury-cikm20

interaction on a news article, the signal obtained from user shares becomes more informative than the publisher's credibility and, therefore, UCSM outperforms PCSM.

5.3 Credibility Score Analysis

In this section, we answer RQ3 by analyzing the PCSs and the UCSs learned. In order to understand the PCSs, we define the *true news publish rate* (TNPR). TNPR is the fraction of real news issued by a publisher, i.e., $TNPR(p) = \frac{\sum_{n=1}^{N} I_{p,n}(1-\hat{l}_n)}{\sum_{n=1}^{N} I_{p,n}}$ where $\hat{l}_n$ is the true label for news n. In Fig. 1 we show the TNPR, MBFC (t_p) and the PCSs learned using the CSM on the Buzzfeed dataset (results are similar for Politifact). To visualize, we choose only those publishers for which MBFC provided a trustworthiness score. We observe that PCS is meaningful and acts as a posterior computed using MBFC as prior and fake news data as likelihood. In cases such as "opposingviews.com" the MBFC score is significantly lower than the TNPR observed from data and the PCS has an updated value.

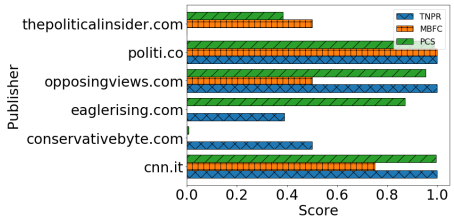
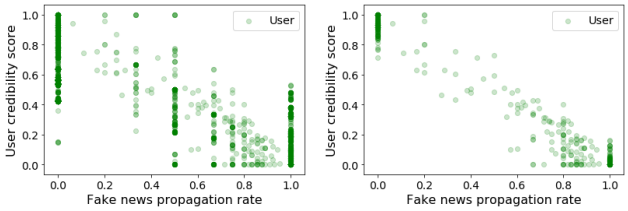


Figure 1: The PCS, MBFC score, and true news publishing rate in Buzzfeed dataset.

Next, we analyze the UCSs learned. In Fig. 2, we show a scatter plot of the UCSs learned using CSM and the *fake news propagation rate* (FNPR) of the user in the Buzzfeed dataset (we observe a similar graph for the Politifact dataset). FNPR is the ratio of fake news shared by the user, i.e., $FNPR(u) = \frac{\sum_{n=1}^{N} \hat{l}_n \cdot S_{u,n}}{\sum_{n=1}^{N} S_{u,n}}$. We observe in Fig. 2a that at $FNPR = 0$, the UCSs are concentrated close to one indicating credible users tend to propagate less fake news. As the FNPR increases to one, we see a gradual decline in the UCSs, and eventually, when $FNPR = 1$, the UCSs of users is concentrated at zero. Further, in Fig. 2b, we observe that when we plot the UCSs for only users with at least five shares, we see that even the variance in scores goes down indicating higher confidence. This shows that the UCSs learned using the CSM is indicative of the true credibility of users in social media.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we introduce a novel approach to identify fake news in social media. Our approach jointly identifies fake news and learns publisher and user credibility scores, which are meaningful and representative of true credibility. We show that prior knowledge can be easily incorporated in our model to improve credibility scores and fake news detection. Through a series of empirical evaluations, we show the effectiveness of our approach at fake news detection. Further, in our experiments, we show the usefulness of the credibility scores learned. This work can be further extended in many interesting ways, such as making use of other social information like



(a) All users **(b) Users with at least 5 shares**

Figure 2: Figure showing the CS learned and fake news propagation rate of users on Buzzfeed dataset. Darker color indicates more users.

friendships to improve fake news detection and including temporal information for early fake news detection.

7 ACKNOWLEDGEMENTS

This work was partially supported by the National Science Foundation grants CCF-1740850 and IIS-1703331, AFRL and the Defense Advanced Research Projects Agency.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. *Social Media and Fake News in the 2016 Election.* Technical Report.
[2] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. 2017. Hinge-Loss Markov Random Fields and Probabilistic Soft Logic. *JMLR* 18 (2017), 3846–3912.
[3] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. 2011. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *FTML* 3 (2011), 1–122.
[4] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information Credibility on Twitter. In *WWW*.
[5] Robert M. Entman. 2007. Framing Bias: Media in the Distribution of Power. *J. Commun* 57 (2007), 163–173.
[6] Lise Getoor and Ben Taskar. 2007. *Introduction to Statistical Relational Learning.* The MIT Press.
[7] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *CSCW*.
[8] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015.* Technical Report.
[9] Francesco Pierri and Stefano Ceri. 2019. False News On Social Media: A Data-Driven Survey. *SIGMOD* 48 (2019), 18–27.
[10] Luc De Raedt and Kristian Kersting. 2011. *Statistical Relational Learning.* Springer.
[11] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A Hybrid Deep Model for Fake News Detection. In *CIKM*.
[12] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *ArXiv* abs/1809.01286 (2018).
[13] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD* 19 (2017), 22–36.
[14] Kai Shu, Suhang Wang, and Huan Liu. 2018. Understanding user profiles on social media for fake news detection. In *MIPR*.
[15] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *WSDM*.
[16] Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking url recommendation to combat fake news. In *SIGIR*.
[17] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359 (2018), 1146–1151.
[18] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL*.
[19] Samuel C Woolley and Philip N Howard. 2018. *Computational propaganda: political parties, politicians, and political manipulation on social media.* Oxford University Press.