Multi-Source Domain Adaptation with Weak Supervision for Early Fake News Detection

Yichuan Li*, Kyumin Lee*, Nima Kordzadeh*, Brenton Faber*, Cameron Fiddes*, Elaine Chen*, Kai Shu†
*Worcester Polytechnic Institute
Worcester, Massachusetts, USA
{yli29, kmlee, nkordzadeh, bdfaber, cmfiddes, echen2}@wpi.edu
†Illinois Institute of Technology
Chicago, Illinois, USA
kshu@iit.edu

***Abstract*—Recently, the massive and diverse fake news from politics to entertainment and health has amplified the social distrust problem and has become a big challenge for the society and research community. The existing fake news detection methods are mostly designed for either a specific domain or require huge labeled data from various domains. If there is not enough labeled data in a certain domain, existing models may not work well for detecting fake news from that domain. To overcome these limitations we propose a novel framework based on multi-source domain adaptation and weak supervision for early fake news detection. The framework transfers sufficient labeled source domains' knowledge into a target/new domain with limited or even no labeled data by the multi-source domain adaptation, and applies researchers' prior knowledge about fake news to the target domain by the weak supervision. The weak supervision assigns the weak labels to the unlabeled samples in the target domain through known heuristic rules. Our experimental results show that our approach outperforms 7 state-of-the-art methods in three real-world datasets. In particular, our model achieves, on average, 5.2% higher accuracy than the best baseline. Our model with a more advanced encoder can further boost the performance by 3.7%. The code is available at this clickable *link*.**

***Index Terms*—fake news detection, weak supervision, domain adaptation**

I. Introduction

Social media platforms provide users with a convenient and easily-accessed way to create, spread, and acquire diverse information. However, during the global pandemic of COVID-19, there has been an abundance of deliberate and domain variant disinformation[1] has been spread online. Such widespread fake news has already eroded the public trust in the government and professional journalism and has made a negative impact on both the online and offline worlds. Thus, it is important to identify fake news across different domains timely.

Due to the superior feature representation abilities, many deep learning based fake news detection methods [1], [2] have achieved promising results. However, the domain variance in fake news topics and word usages brings new challenges into the early fake news detection [3]. This is because the existing methods are specialized for one single news domain and may not be generalized well across multiple domains, especially unseen news domains [4], [5].

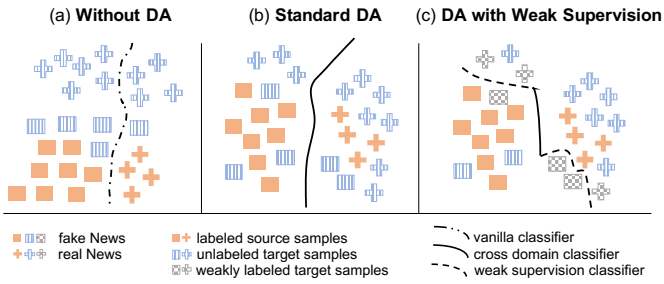


Fig. 1: (a) The vanilla classifier ignores the domain difference, (b) domain adaptation generalizes the classifier across domains, and (c) weak supervision utilizes the weakly labeled samples to adjust the decision boundary for the target domain. Best viewed in color.

Cross-domain early fake news detection is a non-trivial problem because of the following challenges. Firstly, fake news is generally complex in terms of linguistic styles and topics [6] and thus requires extensive training data for building an accurate detection models. However, newly emerging domains may not contain enough labeled data due to the high cost in time and labor of annotation. For example, journalists at PolitiFact[2] conduct thorough searches on online databases, consultation with experts, and a review of publications to rate and write each the fact-checking report. Thus, it is difficult to label enough newly emerging fake news timely. Secondly, the domain shifts or the distribution shifts across different news domains will exacerbate the generalization errors for the unseen news domain [4], [7]. It is not suitable to directly apply a model trained on high-resource domains (i.e., source domains with enough labeled data) to a low-resource domain (i.e., a target domain with very limited or even no labeled data) without any additional treatment, as shown in Figure 1(a). Thirdly, only limited information is available at the early stage

[1]https://archive.is/8Kopd

[2]https://archive.is/ORDpr

978-1-6654-3902-2/21/$31.00 ©2021 IEEE

of news dissemination. Although some fake news detection methods achieved strong performance by including additional information (such as information propagation paths over social media [8] and user credibility [9]), their approaches are relatively expensive as they require collecting additional data, and they require additional time to make predictions (e.g., the prediction can be done after fake news has been already propagated). Late fake news intervention/detection may be less useful as the social and knowledge-based consequences may have already been inflicted before the content is identified as fake. Agents are often unwilling to change or correct incorrect beliefs [10]. Therefore, better and timely methods would be to only require raw news content itself as input without additional information. Our goal is to develop a news content based fake news detection method to identify fake news early with limited labeled data.

Given these challenges, one straightforward approach is to employ the domain adaptation (DA) [11] on news content, where a model is trained on enough labeled source domains and limited labeled/unlabeled target domain. The benefit of the DA is that the model can align the representation space between the source and target domains, so the classifier's decision boundary can be generalized to the target domain. This approach is shown in Figure 1(b). Wang *et. al* [12] and Xu *et. al* [13] for example, applied adversarial training to learn an event, and then used domain invariant feature representation for the fake news detection, respectively. However, only using DA may fall short in fitting the target domain data. This is because DA learns the fake news classifier based on the generalized feature space and source domains' supervision signals without including supervision signals from the target domain. To provide additional supervision at the target domain, we also include fake news researchers' prior knowledge in the target domain to adjust the fake news classifier's decision boundary to the target domain. The addition is shown in Figure 1(c).

We propose a Multi-source Domain Adaptation with Weak Supervision (MDA-WS) for early fake news detection. Our approach integrates knowledge from multiple source domains and fake news researchers. To transfer the knowledge from multiple source domains, the model learns a domain invariant feature representation through adversarial training [14], and a set of fake news classifiers for multi-source domains. To incorporate fake news researchers' prior knowledge to the target domain and properly aggregate multiple source domains, we exploit the newly designed weight function, trained on weakly labeled target samples, to combine all the source classifiers' predictions. This approach is different from existing multi-source domain adaptation with weak supervision method CoDATS-WS [15], because we utilize the weakly labeled target samples instead of posterior distribution constraints and we treat each source differently, based on a weight model.

The main contributions of this work are as follows:

- To the best of our knowledge, we are the first to study the problem of exploiting multi-source domain adaptation with weak supervision for early fake news detection.
- We propose a novel fake news detection framework, which integrates the cross-domain knowledge and prior knowledge about fake news from the literature. It does not require *clean* labels from the target domain during training, and only requires limited *clean* labeled target domain validation data for hyperparameter tuning.
- Extensive experimental results on three real-world fake news datasets show the effectiveness of MDA-WS over 7 state-of-the-art methods. The models achieves a 5.2% accuracy improvement over the best baseline, and an additional 3.7% accuracy improvement with a more advanced encoder.

II. Related Work

Domain Adaptation: Domain adaptation aims to map the target domain (test-dataset) into source domain (training-dataset), and then applies the classifier learned from the source domain to the target domain. Many works try to learn a domain-invariant feature representation, where the features' distribution is the same for samples sampled from target and source domains [11]. In multi-source domain adaptation, to integrate the information from different source domains, [15] considers all the source domains into one without considering the difference among these source domains. [16] weights the source domain based on the distance between the source domain and target domain. Different from existing works, our MDA-WS has a weight function to weight each source domain. Instead of weighting based on the domain distance, our weight function weights source domains based on their contributions on the weakly labeled target data, explicitly.

In addition, there are works in domain adaptation leveraging the weak supervision [15]. However, these weak supervisions are in different formats (e.g, providing the image-level label in segmentation tasks, posterior regularization, etc.), which cannot be applied to enforce and integrate different source domain knowledge in our work. Instead, our method learned an example-to-domain importance score for each source domain.

The most related work to ours is CoDATS-WS [15]. It also combined the multi-source domain adaptation and weak supervision. However, unlike ours, the source domains in CoDATS-WS were considered equally important and its weak supervision was prior distribution regularization which is hard to be obtained in advance. Instead, our method learned an example-to-domain importance score for each source domain.

Weak Supervision: To solve the limited labeled data problem in deep learning, weak supervision techniques have been developed. The weak supervision can provide an external but weak supervision signal to the model during the training. The weak supervision can be in a form of expected distribution constrain [17], weak labeling functions [18] and so on. Our work utilized the lexical characteristics of fake news as a weak labeling function to assign weak labels to unlabeled fake news contents in the target domain.

Fake News Detection: The existing fake news detection methods mainly focused on utilizing the news content and its social engagement [6]. Content based approaches learned

feature representations through the feature engineering or utilized deep learning to learn the content representation end-to-end [19]. Social engagements-based approaches extracted the auxiliary information from user profiles [20], and social discussion [21] and information propagation paths [7]. However, most of these methods are not specialized for cross-domain fake news detection and cannot be generalized for unseen news domains.

Several previous works aimed at cross-domain fake news detection. They learned the domain/topic-invariant features from the information propagation paths [3], [7] or news content [12]. [22] carefully did the feature engineering to identify the domain-invariant features. These works mainly exploited the supervision signals from the source domains without considering fake news researchers' prior knowledge in fitting decision boundary towards target news domain.

Two works utilized weak supervision in fake news detection [23], [24]. [23] leveraged the reinforcement learning to select high-quality weakly labeled samples from news' comments. [24] leveraged weak supervision from social engagements related to news. Compared with these approaches, our MDA-WS does not require any social information. In addition, our MDA-WS can integrate supervision signals from domain adaptation and weak supervision simultaneously for better fake news detection.

III. Problem Statement

Let $\{\mathcal{D}_{S_k}\}_{k=1}^{K}$ denote K-source domains' corresponding datasets (i.e., one fake news dataset from each domain), where $\mathcal{D}_{S_k} = \{(x_i^{S_k}, y_i^{S_k})\}_{i=1}^{|S_k|}$, and $x_i^{S_k}$ and $y_i^{S_k}$ represent news content and its *clean* label, respectively. In addition, we have unlabeled samples $X_T = \{x_i^T\}_{i=1}^{|X_T|}$ from a target domain (e.g., health), and a weak labeling function $g : x \rightarrow \hat{y}$ to weakly label a subset of X_T. The weakly labeled target domain data is denoted as $\mathcal{D}_{\widetilde{T}} = \{(x_i^{\widetilde{T}}, \hat{y}_i^{\widetilde{T}})\}_{i=1}^{|X_{\widetilde{T}}|}$. where $X_{\widetilde{T}} \subseteq X_T$. In this paper, we aim to learn a fake news classifier from $\{\mathcal{D}_{S_k}\}_{k=1}^{K}$, X_T and $\mathcal{D}_{\widetilde{T}}$, such that it will automatically predict whether an unseen news content in the target domain is fake or not. In the following sections, x, y and $\hat{y}$ represent news content, a *clean* label and a *weak* label, respectively.

IV. Our Proposed Framework

As shown in Figure 2, MDA-WS learns the supervision signals by: *(a)* multi-source domain adaptation (MDA) based on multiple labeled source domains, and *(b)* weak supervision based on fake news researchers' prior knowledge. Since existing MDA frameworks [16], [23], [25] did not exploit these two heterogeneous supervision signals at the same time, they cannot achieve optimal performance under this setting. In this section, our proposed model MDA-WS aims to better integrate the knowledge from multi-source domains and fake news researchers by answering three fundamental questions:

- How to exploit knowledge from different source domains?
- How to integrate weak supervision and domain adaptation without hurting a model's performance?

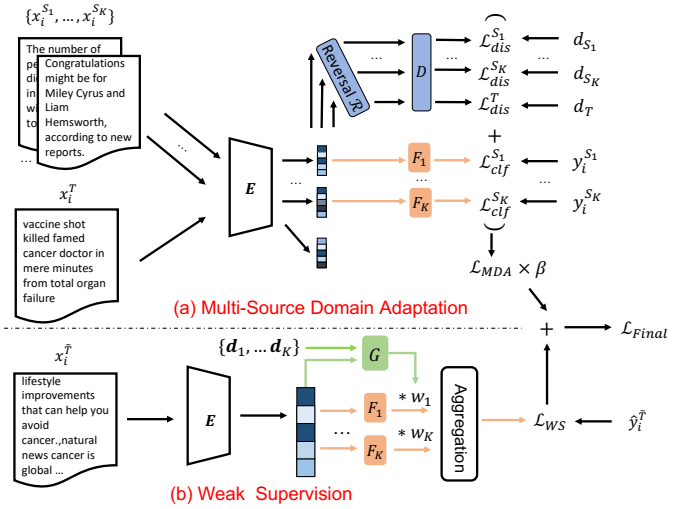


Fig. 2: The overall architecture of our MDA-WS. (a) Multi-source domain adaptation is trained with labeled multiple-source fake news domains $\{\mathcal{D}_{S_i}\}_{i=1}^{K}$ and one unlabeled target domain X_T. (b) Weak supervision utilizes weakly labeled target samples $\mathcal{D}_{\widetilde{T}}$ to enforce and integrate knowledge from multiple-source fake news domains. We jointly train them.

- How to provide high quality and enough weak supervision signals to prevent the model from fitting into the noise?

MDA-WS adapts multi-task learning framework [26] with a shared encoder E and a set of source-domain-specific classification heads $\{F_k\}_{k=1}^{K}$ and other auxiliary modules like domain discriminator D and weight function G. In section IV-A, we will introduce how to fully exploit the information from different source domains. In section IV-B and IV-C, we present our approach of integrating the weak supervision into the multi-source domain and converting the prior knowledge into high-quality weak labels. Lastly, in Section IV-D, we will formalize our final objective function.

A. Multi-Source Domain Adaptation

Inspired by the previous work [27] in multi-source domain adaptation, we focus on reducing both *(i)* the distribution distance between source domains and target domain, and *(ii)* the classification error on source domains to minimize the actual classification error at the target domain.

Domain Invariant Feature Representation Learning: To minimize the distribution distance and learn the domain invariant feature representation, we utilize an adversarial method [11], [14]. In particular, the encoder E in Figure 2 (a) tries to learn the domain invariant representations to fool the domain discriminator D, while D tries to accurately distinguish each sample's source domain. The formulation of the min-max game is as follows:

$$\min_{E} \max_{D} -\sum_{k=1}^{K} \mathbb{E}_{(x,y)\sim\mathcal{D}_{S_k}} l\left(D(E(x)), d_{S_k}\right) - \mathbb{E}_{x\sim X_T}\left[l\left(D(E(x)), d_T\right)\right] \quad (1)$$

where $d_{\{\cdot\}}$ is the domain indicator, the target domain is $d_T = 0$ and the source domain is $d_{S_k} = k$. l is the cross-entropy loss for multi-label classification.

To simplify the min-max game into one-step minimization, we adopt the gradient reversal layer represented as $\mathcal{R}(x)$ introduced in [11]. In the forward propagation, $\mathcal{R}(x)$ is the identity function. However, during the backward propagation, the gradients behind the $\mathcal{R}(x)$ and the gradients of encoder E are reversed by multiplying $-\lambda$, and the gradients of the preceding layers and the discriminator D are not changed. The λ controls the importance of distribution distance regularization in feature learning, and in this paper, we follow previous work [11], [12] set it to 1. The formulations of *reversal* forward- and backward-behaviors are:

$$\mathcal{R}(x) = x; \quad \frac{d\mathcal{R}}{dx} = -\lambda \mathbf{I} \tag{2}$$

where the $\mathbf{I}$ is the identity matrix. We can then convert the min-max game Eq. 1 into one-step optimization Eq. 3:

$$\begin{aligned} \mathcal{L}_{dis} = \min_{E,D} \textstyle\sum_{k=1}^{K} \mathbb{E}_{x \sim \mathcal{D}_{S_k}} l\left(D(\mathcal{R}(E(x))), d_{S_k}\right) \\ + \mathbb{E}_{x \sim X_T}\left[l\left(D(\mathcal{R}(E(x))), d_T\right)\right] \end{aligned} \tag{3}$$

Domain Specific Predictions: Since each news domain has its characteristics, only building a fake news classifier based on the domain invariant representations is sub-optimal. For example, celebrity fake news often mentions celebrities' divorce, while the politics fake news discusses election events. Therefore, the invariant representations would fail to capture these domain-specific characteristics and eventually restrict the capability of the fake news classifier [17].

To overcome this problem and minimize the classification error on the source domains, we build K different source classification heads $\{F_k\}_{k=1}^{K}$ to make the domain-specific predictions, as shown in Figure 2 (a).

Given the domain invariant representation of sample $x_i^{S_k}$ from source domain S_k, it will only go through the classification head F_k and back-propagate the gradient through head F_k towards encoder E. The objective function of the source domain's fake news classification is:

$$\mathcal{L}_{clf} = \textstyle\sum_{k=1}^{K} \mathbb{E}_{(x,y) \sim \mathcal{D}_{S_k}} l\left(F_k(E(x)), y\right) \tag{4}$$

By combining two objective functions in Eq. 3 and Eq. 4, the multi-source domain adaptation's objective function is:

$$\mathcal{L}_{MDA} = \mathcal{L}_{dis} + \mathcal{L}_{clf} \tag{5}$$

B. Weak Supervision

To better aggregate the contribution from different source domains, and include fake news researchers' prior knowledge into model training, we jointly train a weight function G and aforementioned multi-source domain adaptation modules on weakly labeled target samples $\mathcal{D}_{\widetilde{T}} = \{(x_i^{\widetilde{T}}, \hat{y}_i^{\widetilde{T}})\}_{i=1}^{|X_{\widetilde{T}}|}$ as shown in Figure 2 (b).

In this paper, the weak supervision is in the form of a weak label, where we implement a weak labeling function $g: \ x \rightarrow \hat{y}$ based on psychology or computation research findings in the literature, interpreting them as fake news researchers' prior knowledge. Note that the detail of our weak labeling function is described in Section IV-C. A weak labeling function assigns a *weak* label for each sample from a subset of unlabeled samples X_T in the target domain. The weakly labeled target dataset $\widetilde{D_T}$ is utilized to learn the weight function G and Encoder E and classification heads $\{F_k\}_{k=1}^{K}$. The objective function of weak supervision is:

$$\begin{aligned} \mathcal{L}_{WS} \quad &= \mathbb{E}_{(x,\hat{y}) \ \sim \mathcal{D}_{\widetilde{T}}} l\left(F(x), \hat{y}\right) \\ &= \mathbb{E}_{(x,\hat{y}) \ \sim \mathcal{D}_{\widetilde{T}}} l\left(\textstyle\sum_{k=1}^{K} w_k F_k(E(x)), \hat{y}\right) \end{aligned} \tag{6}$$

where $\{w_k\}_{k=1}^{K}$ are float scores under the interval $[0, 1]$ generated by weight function G for every weakly labeled target sample, to be used for computing averaged sum of the logits from $\{F_k\}_{k=1}^{K}$.

Training the model by these weakly labeled target samples brings two advantages: *(i)* the weak supervision helps the model fits well into the target domain by adjusting the decision boundary as shown in Figure 1(c); and *(ii)* these weakly labeled samples do not require manual annotations, which resolves the limited labeled data issue at a new/an emerging target domain such as health. This also makes our work different from semi-supervised domain adaptation [28], where the label at the target domain is from manual annotation instead of labeling functions.

The steps of outputting $\{w_k\}_{k=1}^{K}$ from G are as follows: given hidden representation of the weakly labeled target samples $E(x)$ and learnable source domain embedding vector $\mathbf{d}_k \in \mathbb{R}^Z$ for each source classification head, our weight function G will output $\{w_k\}_{k=1}^{K}$ through the shallow multi-layer perceptions C with non-linearity:

$$w_k = G(x)_k = \sigma(C([E(x); \mathbf{d}_k])) \tag{7}$$

where "$[\ \cdot\ ;\ \cdot\ ]$" is the concatenate operation and σ is the *sigmoid* activation function. The additional domain embedding vector $\{\mathbf{d}_k\}_{k=1}^{K}$ helps us to capture the global information inside each domain. In our ablation study on Section V-D, we have shown the effectiveness of the domain embedding vector.

Different from learning domain importance weights $\{w_k\}_{k=1}^{K}$ from the distribution distance between the source and target domains [29], [30], our approach considers the contribution towards the weakly labeled samples $\mathcal{D}_{\widetilde{T}}$ from the target domain as the importance weight. In this way, the weak supervision can directly enforce and reformulate the knowledge from different source domains. In addition, by concatenating the domain embedding vector $\{\mathbf{d}_k\}_{k=1}^{K}$ and weakly labeled samples hidden representation $E(x)$, the weight model can better integrate the local and global information towards target sample. We will present this in our ablation study in Section V-D.

C. Weak Labeling Function

Inspired by the setting of weak labeling functions in [31], our weak labeling function contains a feature transformation

function f and a threshold N as shown in Figure 3. As a way to apply fake news researchers' prior knowledge, we analyzed the previous work in fake news content and found the following characteristics: *(i)* There are more *second-person pronouns* in fake news than real news [32], [33] because real news editors prefer removing personal languages and such pronouns were usually an indication of imaginative writing which is close to the fake news; *(ii)* Researchers [19] found that *swear words* are more often used in fake news content because the fake news writers pay less attention to informal words, and *(iii)* The *number of adverbs* is also an indicator of the news veracity. Fake news used more adverbs to exaggerate [19], [32]. Although other weak labeling functions utilize social media engagements [23], [24] and agencies' trustworthiness [34], they require not only the news content but also additional contextual information which is not suitable for early fake news detection.

Based on these characteristics, we came up three transformation functions f s, each of which measures news content's $LIWC_{score}$ according to one of the categories: *you*, *swear* and *adverb* [35]. We consider choosing a transformation function as a hyperparameter. Each $LIWC$ category contains multiple pre-selected words and the calculation of $LIWC_{score}$ is as follows:

$$LIWC_{score} = \frac{\#\ of\ matched\ words}{total\ \#\ of\ words\ in\ news} * 100\% \quad (8)$$

To properly label the target training samples and reduce the noise from the weak labeling function, we only keep and label the largest N samples as fake news and the smallest N samples as real news (refer to Figure 3). These $|X_{\widetilde{T}}| = 2N$ samples with the weak labels are the weakly labeled dataset $\mathcal{D}_{\widetilde{T}}$. The target domain's validation set is used for the hyperparameter tuning of the weak labeling function (i.e., the best f and N)[3]. For these $2N$ weakly labeled samples, the accuracy scores for the best combinations are (N=75 and f = "*you*" in Health), ($N = 25$ and f = "*swear*" in PolitiFact) and ($N = 50$ and f = "*adverb*" in GossipCop) were 0.81, 0.70, and 0.58, respectively. It should be noticed that these combinations are selected based on hyperparameter analysis shown in Figure 5. These accuracy scores are better than the random guess, 0.5, which proves the effectiveness of our weak labeling function. The detailed dataset information is described in Section V-A.

D. Final Objective Function

Our final objective function consists of **two parts**: (i) training multi-source domain adaptation and (ii) training weak supervision. In the first part, we input labeled source domain datasets $\{\mathcal{D}_{S_k}\}_{k=1}^{K}$ and unlabeled target domain samples X_T into the framework. In the second part, we input the same $\{\mathcal{D}_{S_k}\}_{k=1}^{K}$ and X_T with the weakly labeled target sample data $\mathcal{D}_{\widetilde{T}}$. Overall, the objective function of MDA-WS is:

$$\mathcal{L}_{final} = \beta \mathcal{L}_{MDA} + \mathcal{L}_{WS} \quad (9)$$

[3]We also tried combining outcomes of three weak labeling functions/transformation functions and selected weak labels based on the majority vote. But, this approach produced more noise and resulted in worse performance. Therefore, we selected the best one for each target domain.

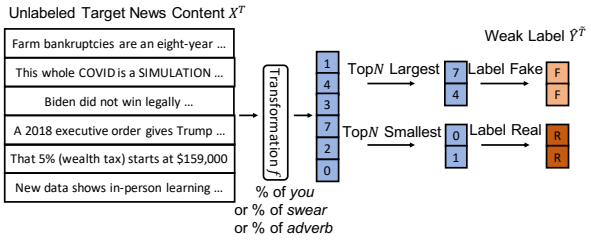


Fig. 3: The procedure of a weak labeling function.

where β is the hyperparameter controlling multi-source domain adaptation and weak supervision loss.

V. Experiments

Through the experiments, we answer the following research questions: ***RQ1***: How accurately can MDA-WS detect fake news in the target domain? ***RQ2***: Does each component of MDA-WS positively contribute? ***RQ3***: How does hyperparameter values affect the MDA-WS's performance? ***RQ4***: How many clean labeled samples can be saved by MDA-WS? ***RQ5***: Can the weight function G properly identify the contribution of each source domain? ***RQ6***: Do baselines achieve better performance by utilizing our weakly labeled samples $\mathcal{D}_{\widetilde{T}}$? Does our MDA-WS still outperform the baselines?

A. Evaluation Datasets

We selected fake news datasets from three different news domains: GossipCop (GC), PolitiFact (PF) [36] and Health DETERRENT (HD) [37]. These datasets contain news contents and veracity labels annotated by professional journalists. For HD, the authors split it into cancer and diabetes. In our study, we combined them into one health dataset. Datasets' statistical information is listed in Table I. Since some hyperlinks of the news content are inaccessible because of the deletion, crawling news content will inevitably shrink the size of the datasets. We evaluate our model and baselines under the standard balance setting. We truncated news content into 300 tokens and utilized the RoBERTa-base's tokenizer[4] to encode text.

Dataset Setting: We split each dataset into training, validation, and test sets with a ratio of 7:1:2. We picked up a dataset as the target domain dataset in the round-robin, and the remaining two datasets were the source domain datasets. In other words, there are three pairs of two source domain datasets and a target domain dataset. Given a pair, we train our MDA-WS by the *labeled* source domain training sets ($\{D_{S_k}\}_{k=1}^{K}$ where K=2) and unlabeled target domain training set (X_T) with weakly labeled target domain data $\mathcal{D}_{\widetilde{T}}$, a subset of the target domain training set with weak labels as described in Section IV-D. Then, the target domain's validation is used for hyperparameter tuning. Finally, the optimized model will be evaluated over the target domain's test set. We apply the same setting to all the baselines described in Section V-B. This

[4]https://github.com/huggingface/transformers

TABLE I: The statistical information of the datasets.

Datasets	Fake	Real
GossipCop	4,252	4,252
PolitiFact	260	260
Health	1,992	1,992

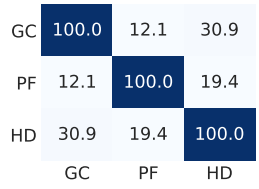


Fig. 4: Vocabulary overlap (%) between different domains.

setting can be considered as the upper bound on how well these unsupervised domain adaptation methods can perform [38].

Dataset Analysis: To understand cross-domain relevance of the datasets, we measured the vocabulary overlap in Figure 4. We excluded less frequent words (<3) and stop words from NLTK[5] in each dataset. We can observe that GossipCop and Health had the highest vocabulary overlap (30.9%), while PolitiFact was less vocabulary overlap with GossipCop (12.1%) and Health (19.4%). In our further analysis in Figure 7, our weight function G can capture the domain relevance, and assigns a large averaged importance weight to more relevant source domain. Since the cross-domain similarity is low, it is important to apply multi-source domain adaptation with weak supervision.

B. Experimental Setting

Baseline Methods. We compare our MDA-WS with 7 baselines: Text-CNN [39], EANN [12], DANN [11], CNN-DDS [38], MoE-A [17], MDAN [27] and CoDATS-WS [15].

Implementation Details. For a fair comparison, Text-CNN [39] was used as the encoder of all the baselines and ours. The WordPiece embedding of Text-CNN was initialized from the RoBERTa-base[6], and was frozen during training. Text-CNN has three 1D convolutional layers with kernel size 3, 4, and 5. Each layer has 100 filters. We trained no-domain-adaptation *(NDA)* methods like Text-CNN and EANN by using a single source domain dataset or combination of two source domain datasets to report both results. For *NDA* and DANN trained on a single-source dataset, we report the best single-source performance. We use the target validation set to tune the hyperparameters and report each model's performance in the target test set. For model optimization, we use the optimizer Adam with a learning rate of 0.001 and weight decay of 0. All the models are trained for 50 epochs. We save the best checkpoint at the end of each epoch and report the test result for checkpoint with the best validation accuracy score. The training is repeated 5 times and the average result is reported. We implemented MDA-WS with PyTorch[7] (version 1.7.0) and utilized the Ray[Tune][8] for the hyperparameter search. All the codes are available at[9].

[5]https://www.nltk.org/book/ch02.html
[6]https://huggingface.co/roberta-base
[7]https://pytorch.org/
[8]https://docs.ray.io/en/master/tune/
[9]https://github.com/bigheiniu/BigData-MDA-WS

Evaluation Metrics. Since these datasets are balanced, we utilize evaluation metrics such as accuracy, and macro- precision, recall, and F1 to represent the performance of our approach and these baseline methods. These evaluation tools are from scikit-learn [40].

C. Effectiveness of our MDA-WS (RQ1)

To answer *RQ1*, we compared our MDA-WS with the baselines and experimental results are shown in Table II. We have several observations:

- Overall, our proposed method MDA-WS achieved the best performance compared with all the baselines over three different target domain datasets, improving 5.2% accuracy on average compared with the best baseline. When we closely compare our MDA-WS with CoDATS-WS, the result demonstrated that our weak labeling function provided high-quality supervision signals compared with the posterior regularization of CoDATS-WS. This also demonstrated the importance of treating each source domain differently. In addition, compared with MoE-A, which weights each source by the distance between the target and source domain, our model achieved considerable performance improvement. This indicates the combination of our weak supervision and weight function G contributes to cross-domain fake news detection.
- Single-Source Domain Adaptation *SDA* methods (i.e., DANN and CNN-DDS) outperformed *NDA* (i.e., Text-CNN and EANN) with one source domain dataset. In general, *MDA* (i.e., MDAN and MoE-A) achieved better performance than *SDA* (i.e., DANN, CNN-DDS). *NDA* with multiple source training data achieved better performance than *NDA* with a single source. However, in Health dataset, *Text-CNN* with multiple sources showed a performance drop. This may suggest the importance of domain adaptation for better supervision signals with omitting the domain conflict.

D. Ablation Study (RQ2)

Effects of MDA and WS. Our proposed method exploits two kinds of supervision signals for cross-domain fake news detection. To understand the contribution of each supervision signal and prove that weak supervision can boost the performance of multi-source domain adaptation and vise versa, we eliminate one of them (*w/o MDA* or *w/o WS*) in model training. We can observe that both WS and MDA positively contributed to the model's performance. Its result confirms the importance of combining knowledge from multiple source domains and fake news researchers. In addition, we observe that *w/o WS* shows relative performance drop ($2\% \sim 6\%$) compared with *w/o MDA*. This indicates that limited prior knowledge/*weak supervision* makes more contribution than additional training data from source domains/*MDA*.

Effects of Weak Supervision Components. To understand the contribution of the weight function G and explicit domain embedding vector $\{\mathbf{d}_k\}_{k=1}^{K}$, we consider two variants of MDA-WS: *(i)* Weight function G generates uniform weight for the

TABLE II: The experiment results on three fake news domains. Our MDA-WS significantly outperformed all baselines (t-test with $p < 0.05$). The best performance is bold and the second best is underlined.

Methods	GossipCop				PolitiFact				Health				Avg. Rank
	ACC	Precision	Recall	F1	ACC	Precision	Recall	F1	ACC	Precision	Recall	F1	Acc
Single Source													
Text-CNN	56.54	57.75	56.59	54.92	60.76	60.48	60.38	60.28	64.41	64.70	64.42	64.27	9.00
EANN	58.86	59.06	58.90	58.69	61.86	68.09	61.92	58.45	60.85	60.49	60.90	60.28	8.00
DANN	59.37	59.75	59.40	59.02	63.76	64.59	63.84	63.87	69.85	69.91	69.74	70.34	6.33
CNN-DDS	56.69	56.84	56.72	56.51	66.48	67.09	66.92	66.83	71.25	71.51	71.33	71.28	7.00
Multiple Sources													
Text-CNN	57.02	57.06	57.05	57.04	67.72	67.89	67.69	67.61	50.96	51.20	51.06	49.11	7.67
EANN	58.73	58.82	58.76	58.68	67.26	67.68	67.12	66.84	72.09	75.51	72.14	71.25	5.67
MoE-A	60.18	60.27	60.20	60.13	69.47	70.28	69.03	68.53	61.96	69.40	62.07	58.12	5.00
MDAN	61.63	63.05	61.66	60.90	70.67	70.88	70.58	70.47	72.34	75.31	72.37	71.57	2.67
Multiple Sources + Weak Supervision													
CoDATS-WS	61.91	62.49	61.97	61.55	69.21	69.24	69.04	68.96	82.46	83.32	82.53	82.42	2.67
MDA-WS	**66.18**	**66.44**	**66.21**	**66.09**	**75.80**	**77.40**	**75.96**	**75.69**	**88.78**	**89.36**	**88.70**	**88.65**	**1.00**

source domains, termed as *w/o* G. In our case, the weights for two source domains will be both set to 0.5. *(ii)* Weight function generates weights without domain embedding $\{\mathbf{d}_k\}_{i=1}^{K}$. We observe that our original weight function performed better than the two variants over all of the target domain datasets, indicating the effectiveness of our weight function G and the domain embedding vector $\{\mathbf{d}_k\}_{k=1}^{K}$.

Effects of Basic Encoder. To understand the influence of the encoder, we replaced Text-CNN with RoBERTa-base. MDA-WS$_{RoBERTa-base}$ achieved 3.7% accuracy improvement compared with our original MDA-WS. The result of MDA-WS$_{RoBERTa-base}$ confirms that an advanced encoder can produce better performance, and also shows the flexibility of our framework.

TABLE III: Ablation study ***w.r.t*** accuracy ($p < 0.05$).

Model	GossipCop	PolitiFact	Health
w/o MDA	61.32	72.11	82.14
w/o WS	58.43	70.19	76.12
w/o Gate	57.38	54.81	85.48
w/o $\{\mathbf{d}_k\}_{i=1}^{K}$s	58.45	65.38	86.48
MDA-WS	66.18	75.80	88.78
MDA-WS$_{RoBERTa-base}$	**69.96**	**80.77**	**90.99**

E. Further Analysis

Hyperparameter Analysis *(RQ3).* To conduct hyperparameter (*HP*) analysis, we varied values of the four hyperparameters β, f, N and Z. Figure 5 shows how accuracy was changed when we changed each hyperparameter. We first observed that in GossipCop and Health, mid-range values of β achieved the best performance while in PolitiFact, small β is preferred. This is consistent with the observation from Figure 4, in which PolitiFact has relatively low similarities with GossipCop and Health. A large β score in PolitiFact would make the model overfitting with the source domains. Secondly, the performance gaps across different transformation function f were due to domain shift. Thirdly, in GossipCop and Health, Mid-range N achieved the best performance because large N would bring much noise in the training, while small N would constrain the researchers' prior knowledge. However, in PolitiFact, small N achieves better performance. This is due to the small size of the dataset itself that $N = 25$ already covers enough part of the dataset ($(25 \times 2)/520 = 9.6\%$). Lastly, for a domain embedding size Z, we observe that the best performance was achieved in all three datasets when $Z = 128$, because it can overcome the Underfitting and overfitting problems.

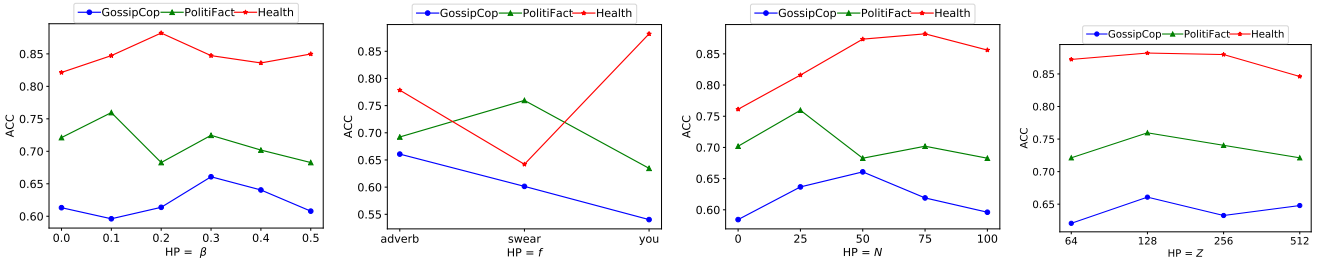
Effects of Clean samples *(RQ4).* To understand the superiority of our approach in the quantity of clean labeled samples, we compare MDA-WS with supervised classification and semi-supervised domain adaptation (SSDA). Specifically, the supervised classification is the basic encoder trained on labeled target samples, while the semi-supervised learning is based on MDA-WS but replaces weakly labeled samples to manually labeled target domain samples $D_T = \{(x_i^T, y_i^T)\}_{i=1}^{M}$, where $M < |X_T|$. As the result shows in Figure 6a and 6b, MDA-WS can achieve compatible performance without manually labeled samples. These flat lines are because MDA-WS did not use any manually labeled samples in this setting. Specifically, in the Health domain, MDA-WS can reduce at least 40 labeled samples in supervised classification and SSDA. This result indicates MDA-WS can solve the limited label problem in newly emerged news domains.

Weight Function Analysis *(RQ5).* To answer *RQ5*, we visualized the normalized average weights associated with the classification heads in Figure 7. The visualization weights $\{\hat{w}_k\}_{k=1}^{K}$ are calculated as follows:

$$\hat{w}_k = \frac{\frac{1}{|\mathcal{D}_{S_k}|}\sum_{i=1}^{|\mathcal{D}_{S_k}|} w_k^i}{\sum_{k'=1}^{K}\frac{1}{|\mathcal{D}_{S_{k'}}|}\sum_{i=1}^{|\mathcal{D}_{S_{k'}}|} w_{k'}^i} \tag{10}$$

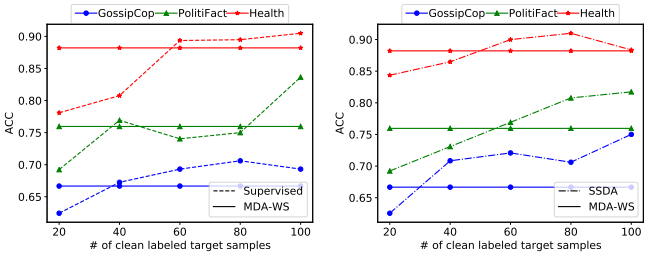
The figure shows that each source's weight was different, meaning each source contributed unequally. Specifically, when a target domain dataset was GossipCop, Health DETERRENT's classification head was higher or more important than PolitiFact's classification head. This pattern was also observed in Figure 4. This result indicates that the weight function can properly identify the instructive domain.

Baseline Methods with Weakly Labeled Target Samples *(RQ6).* To understand the contribution of weakly labeled samples $\mathcal{D}_{\widetilde{T}}$ and effectiveness of MDA-WS, we also consider providing these weakly labeled samples to two best baseline



(a) MDA importance score β. (b) *Trans* Function f in weak supervision. (c) Threshold N in weak supervision. (d) Domain embedding size Z in weak supervision.

Fig. 5: Hyperparameter (HP) analysis ***w.r.t*** accuracy (ACC).



(a) Supervised classification *vs.* MDA-WS (b) SSDA *vs.* MDA-WS

Fig. 6: The efficiency of MDA-WS in saving *clean* labeled target domain samples $\mathcal{D}_T$, compared with supervised classification and semi-supervised domain adaptation (SSDA).

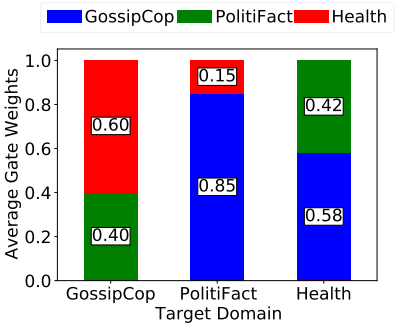


Fig. 7: Average normalized expert weights for each domain.

methods MDAN and CoDATS-WS. It should be noticed that the weak labeling function f and the number of weakly labeled samples N are consistent with MDA-WS. From the result shown in Table IV, we can observe that both baseline methods got improved performance, but are still worse than our method MDA-WS. This not only indicates the effectiveness of the weak supervision, but also the advance information aggregation mechanism of our method MDA-WS.

VI. Conclusion and Future Work

In this paper, we have studied a cross-domain early fake news detection problem, especially focusing on limited or even no labeled data at new/emerging news domains. It is a challenging problem because a newly emerging domain may have only limited annotated data for early fake news detection. To perform early fake news detection effectively under the limitation, we proposed a novel framework based on multi-source domain adaptation and weak supervision. Our proposed MDA-WS successfully integrated knowledge from multiple news source domains and fake news researchers' prior knowledge. Specifically, MDA-WS learned the domain-invariant feature representation through the adversarial training and utilized the weakly labeled samples to train a weight function in order to aggregate the output from source-specific fake news classifiers/classification heads. The comprehensive experiments conducted on three different target domains showed that our proposed model outperformed 7 baselines, improving 5.2% accuracy compared with the best baseline. In addition, we further improved our model's performance by 3.7% accuracy, using a more advanced encoder.

In the future, we will investigate an alternative weak labeling function which can potentially work well for any target domain. In particular, we plan to study a universal labeling function based on fake news' general attributes like text perplexity [41], logic reasoning [42] and etc. Another possible future improvement is to reduce the human efforts at constructing these weak labeling functions to catch up quickly evolving fake news formats like a short-form, video sharing, etc. We are also interested in automatically generating weak labeling functions from natural language description or constrains entailed in a dataset.

TABLE IV: Baseline methods' accuracy score with additional weakly labeled samples $\mathcal{D}_{\widetilde{T}}$.

Model	GossipCop	PolitiFact	Health
MDAN	61.63	70.67	72.34
MDAN *w/* $\mathcal{D}_{\widetilde{T}}$	61.67	72.12	85.73
CoDATS-WS	61.91	69.21	82.46
CoDATS-WS *w/* $\mathcal{D}_{\widetilde{T}}$	62.21	72.12	86.86
MDA-WS	66.18	75.80	88.78

VII. Acknowledgment

This work was supported in part by NSF grant CNS-1755536 and WPI TRIAD. We also thank the anonymous reviews for helpful comments on this work.

REFERENCES

[1] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," ser. CIKM '17, 2017. [Online]. Available: https://doi.org/10.1145/3132847.3132877

[2] J. Zhang, B. Dong, and P. S. Yu, "Fakedetector: Effective fake news detection with deep diffusive neural network," 2019.

[3] A. Silva, L. Luo, S. Karunasekera, and C. Leckie, "Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data," 2021.

[4] M. Janicka, M. Pszona, and A. Wawer, "Cross-domain failures of fake news detection," *Computación y Sistemas*, vol. 23, 10 2019.

[5] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," ser. ACL 2018, 2018, pp. 3391–3401.

[6] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," 2017.

[7] Y. Han, S. Karunasekera, and C. Leckie, "Graph neural networks with continual learning for fake news detection from social media," 2020.

[8] J. Zhang, B. Dong, and P. S. Yu, "Deep diffusive neural network based fake news detection from heterogeneous social networks," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 1259–1266.

[9] A. Dhiman and D. Toshniwal, "An unsupervised misinformation detection framework to analyze the users using covid-19 twitter data," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 679–688.

[10] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake news early detection: An interdisciplinary study," 2020.

[11] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," ser. ICML'15, 2015.

[12] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "Eann: Event adversarial neural networks for multi-modal fake news detection," ser. KDD '18, 2018.

[13] B. Xu, M. Mohtarami, and J. Glass, "Adversarial domain adaptation for stance detection," 2019.

[14] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[15] G. Wilson, J. R. Doppa, and D. J. Cook, "Multi-source deep domain adaptation with weak supervision for time-series sensor data," 2020.

[16] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton, "Adaptive mixture of local expert," *Neural Computation*, 1991.

[17] J. Guo, D. J. Shah, and R. Barzilay, "Multi-source domain adaptation with mixture of experts," 2018.

[18] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel," *Proc. VLDB Endow.*, 2017.

[19] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," ser. ACL '17, 2017.

[20] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," ser. WWW '11, 2011.

[21] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "Defend: Explainable fake news detection," ser. KDD '19, 2019.

[22] S. Castelo, T. Almeida, A. Elghafari, A. Santos, K. Pham, E. Nakamura, and J. Freire, "A topic-agnostic approach for identifying fake news pages," *Companion Proceedings of The 2019 World Wide Web Conference*, May 2019.

[23] Y. Wang, W. Yang, F. Ma, J. Xu, B. Zhong, Q. Deng, and J. Gao, "Weak supervision for fake news detection via reinforcement learning," 2020.

[24] K. Shu, G. Zheng, Y. Li, S. Mukherjee, A. H. Awadallah, S. Ruston, and H. Liu, "Leveraging multi-source weak social supervision for early detection of fake news," 2020.

[25] S. Paul, Y.-H. Tsai, S. Schulter, A. K. Roy-Chowdhury, and M. Chandraker, "Domain adaptive semantic segmentation using weak labels," 2020.

[26] R. Caruana, "Multitask learning," Tech. Rep., 1997.

[27] Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig, "Controllable invariance through adversarial feature learning," 2018.

[28] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," 2019.

[29] H. Zhao, S. Zhang, G. Wu, J. P. Costeira, J. M. F. Moura, and G. J. Gordon, "Multiple source domain adaptation with adversarial training of neural networks," 2017.

[30] S. Zhao, G. Wang, S. Zhang, Y. Gu, Y. Li, Z. Song, P. Xu, R. Hu, H. Chai, and K. Keutzer, "Multi-source distilling domain adaptation," 2020.

[31] P. Varma and C. Ré, "Snuba: Automating weak supervision to label training data," *Proc. VLDB Endow.*, 2018.

[32] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," ser. ACL '11, 2011.

[33] P. Rayson, A. Wilson, and G. Leech, *Grammatical word class variation within the British National Corpus Sampler*, 2002.

[34] S. Helmstetter and H. Paulheim, "Weakly supervised learning for fake news detection on twitter," in *2018 ASONAM*, 2018, pp. 274–277.

[35] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," 2010.

[36] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media," 2019.

[37] L. Cui, H. Seo, M. Tabar, F. Ma, S. Wang, and D. Lee, "Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation," ser. KDD '20, 2020.

[38] X. Ma, P. Xu, Z. Wang, R. Nallapati, and B. Xiang, "Domain adaptation with BERT-based domain classification and data selection," ser. ACL 2019, 2019.

[39] Y. Kim, "Convolutional neural networks for sentence classification," 2014.

[40] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[41] N. Lee, Y. Bang, A. Madotto, and P. Fung, "Misinformation has high perplexity," 2020.

[42] A. Groza, "Detecting fake news for the new coronavirus by reasoning on the covid-19 ontology," 2020.