Unsupervised Domain Adaptation for COVID-19 Information Service with Contrastive Adversarial Domain Mixup

Huimin Zeng*, Zhenrui Yue*, Ziyi Kou*, Lanyu Shang*, Yang Zhang†, Dong Wang*

*School of Information Sciences
University of Illinois Urbana-Champaign, IL, USA
{huiminz3, zhenrui3, ziyikou2, lshang3, dwang24}@illinois.edu
†Department of Computer Science and Engineering
University of Notre Dame, IN, USA
yzhang42@nd.edu

Abstract—In the real-world application of COVID-19 misinformation detection, a fundamental challenge is the lack of the labeled COVID data to enable supervised end-to-end training of the models, especially at the early stage of the pandemic. To address this challenge, we propose an unsupervised domain adaptation framework using contrastive learning and adversarial domain mixup to transfer the knowledge from an existing source data domain to the target COVID-19 data domain. In particular, to bridge the gap between the source domain and the target domain, our method reduces a radial basis function (RBF) based discrepancy between these two domains. Moreover, we leverage the power of domain adversarial examples to establish an intermediate domain mixup, where the latent representations of the input text from both domains could be mixed during the training process. Extensive experiments on multiple real-world datasets suggest that our method can effectively adapt misinformation detection systems to the unseen COVID-19 target domain with significant improvements compared to the state-of-the-art baselines.

I. INTRODUCTION

In this work, we focus on COVID-19 misinformation detection, given its global impact of the ongoing pandemic and the "Infodemic"[1] it causes on social media [1]. Regarding COVID-19 misinformation detection, if the language models trained on non-COVID datasets without any fine-tuning on COVID-19 specific data, these models might suffer from a severe issue of generalization and perform poorly on the COVID-19 data, due to the domain shift between the non-COVID training data distribution and the test COVID-19 data distribution. Recently, the ongoing pandemic of COVID-19 inspires a variety of studies [2] to develop NLP models to provide reliable COVID-19 information services across various social media platforms (e.g., Twitter, Facebook). However, the supervised learning approaches often require a large-scale training dataset while collecting annotations for COVID training data is extremely expensive and time consuming due to the cost and complexity in recruiting the qualified annotators and keep the annotations update to date to accommodate the dynamics

of COVID-19 knowledge (e.g., different variants of the virus) [3]. Moreover, our unsupervised domain adaptation setting is motivated for a more general setting of any early-stage pandemic (not limited to COVID-19) where there is no ground-truth information about the novel disease at all, but the need for correct information is urgent. Therefore, it is critical to develop unsupervised domain adaptation frameworks to train COVID models so that knowledge from an existing data domain could be adapted and transferred to the unseen COVID data domain without requiring any ground-truth training labels.

In this paper, we explore an unsupervised domain adaptation problem for COVID-19 misinformation detection on social media. In particular, we propose an unsupervised domain adaptation framework **C**ontrastive **A**dversarial **D**omain **M**ixup (CADM), which uses adversarial domain mixup and contrastive learning to bridge the gap between the source training data domain and the target COVID data domain. The overview of our framework is shown in Figure 1. To demonstrate the effectiveness of the proposed CADM, we evaluate it on several real-world COVID-19 datasets. Our experimental results suggest that our CADM effectively adapts pre-trained language models to the target COVID domain, and consistently outperforms state-of-the-art baselines.

II. RELATED WORK

Misinformation Detection. Great efforts have been made to detect the misinformation from online platforms (e.g., social media). In [2], knowledge graphs are integrated into the misinformation detection framework to enhance the model's performance. The concurrent work [4] also proposed a domain adaptation framework to address the COVID-19 misinformation detection using label correction. However, such misinformation detection systems are built under a supervised or semi-supervised learning setting, but in practice labeled COVID-19 misinformation data is not always accessible. Therefore, this paper focuses on *unsupervised* domain adaptation of language models for COVID-19 misinformation detection, where the

---

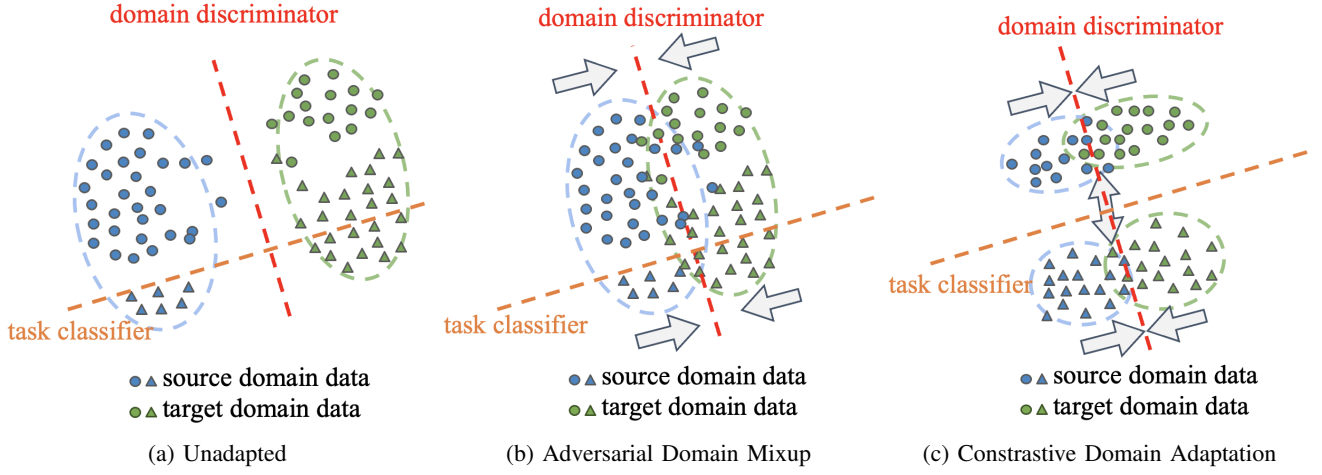[1] https://www.who.int/health-topics/infodemic#tab=tab_1

Fig. 1: The Overview of Our Contrastive Adversarial Domain Mixup (CADM): firstly, in (a), a pre-trained models will generate labels for target domain examples via pseudo labeling, where the green triangles belong to one class (e.g., true information in the misinformation detection task) and the green dots belong to another class (e.g., false information in the misinformation detection task). In addition, a domain discriminator is trained. Then, in (b), the well-trained domain discriminator will establish an intermediate domain mixup by perturbing latent representations of input text from both domains towards each other. At the same time, in (c), we also compute a contrastive adaptation loss over the perturbed adversarial representations, and optimize it to further reduce the domain discrepancy and increase the models' performance. Note that (b) and (c) are executed alternately.

models are trained to adapt knowledge from a source domain to the unknown COVID-19 domain.

Domain Adaptation. Only limited domain adaptation methods are developed for misinformation detection. In [5], domain adversarial training is implemented for misinformation task, so that the models are trained to learn domain-invariant features. Utilizing contrastive methods, [6] propose to quantify and reduce the domain discrepancy using explicit distance measures (e.g., maximum mean discrepancy) to bridge the gap between source domain and target domain. However, the unsupervised domain adaptation methods have not been systematically studied in the COVID-19 domain. In this work, inspired by the idea of adversarial examples [7], [8] and domain mixup [9], we propose to establish a smoothed intermediate training domain by perturbing the latent representations of the input from both source domain and target domain towards each other with a domain discriminator and perform contrastive training on the smoothed domain to transfer knowledge from the source training domain to the target COVID-19 domain.

III. PROBLEM STATEMENT

Regarding misinformation detection, we aim at training a model f, which takes an input text x (a COVID-19 claim or a piece of news) to predict whether the information contained in x is valid or not (i.e., a binary classification task). Moreover, in our domain adaptation problem, we use $\mathcal{P}$ to denote source domain data distribution and $\mathcal{Q}$ for the target domain data distribution. Each data point (x, y) contains an input segment of COVID-19 claim or news (x) and a label $y \in \{0, 1\}$ ($y = 1$ for true information and $y = 0$ for false information). To differentiate the notations of the data sampled from the

source distribution $\mathcal{P}$ and the target distribution $\mathcal{Q}$, we further introduce two definitions of the domain data:

- **Source domain**: The subscript s is used to denote the source domain data: $\mathcal{X}_s = \{(x_s, y_s) | (x_s, y_s) \sim \mathcal{P}\}$.
- **Target domain**: Similarly, the subscript t is used to denote the target domain data: $\mathcal{X}_t = \{x_t | x_t \sim \mathcal{P}\}$. Note that in our unsupervised setting, the ground truth labels of target domain data y_t are not used during training.

Our goal is to adapt a classifier f trained on $\mathcal{P}$ to $\mathcal{Q}$. For a given target domain input x_t, a well-adapted model aims at making predictions as correctly as possible.

IV. SOLUTION

Domain Discriminator. The first step of our framework is to train a domain discriminator f_D to classify whether the input data belongs to the source or target domain. The domain discriminator shares the same BERT Encoder f_e with the COVID model and has a different binary classification module f_D. The domain discriminator takes the token [CLS] representation from the BERT encoder as input to predict the domain of the input data as in [10]:

$$\hat{y} = f_D(z), \tag{1}$$

where z is the representation of token [CLS].

Regarding the training of the f_D, we explicitly define the domain label y_D of the source domain data as $y_D = 0$ and the domain label of the target domain data as $y_D = 1$. Note that the domain labels y_D are not the same as data label y (true information or false information). Therefore, the training of the domain discriminator is formulated as:

$$\min_{f_D} \mathbb{E}_{(x, y_D) \sim \mathcal{X}'} \Big[l(f_D(f_e(x)), y_D) \Big], \tag{2}$$

where $\mathcal{X}'$ represents the merged datasets of both source domain and target domain training data with domain labels.

Adversarial Domain Mixup. After the domain discriminator is trained, we propose to directly perturb the latent representations of the input data from both source domain and target domain towards the decision boundary of the domain discriminator as shown in Figure 1b. To this end, the perturbed representations (i.e., domain adversarial representations) from both domains could become closer to each other, indicating a reduced domain gap. Herein, the generated domain adversarial representations from both domains form a smoothed intermediate domain mixup in the latent feature space of the model. Mathematically, the optimal perturbation $\boldsymbol{\delta}^*$ to perturb the latent representation $\boldsymbol{z}$ of a training sample $\boldsymbol{x}$ could be found by solving an optimization problem:

$$\mathcal{A}(f_e, f_D, \boldsymbol{x}, y_D, \epsilon) = \max_{\boldsymbol{\delta}} \left[l(f_D(\boldsymbol{z} + \boldsymbol{\delta}), y_D) \right] \quad (3)$$
$$s.t. \quad \|\boldsymbol{\delta}\| \leq \epsilon, \quad \boldsymbol{z} = f_e(\boldsymbol{x}).$$

Note that in the above equation, we introduce a hyperparameter ϵ to bound the norm of the perturbation δ, so that the infinity solution could be avoided. Eventually, after applying Equation 3 to all training samples in the merged training set $\mathcal{X}'$, we obtain the adversarial domain mixup $\mathcal{Z}'$:

$$\mathcal{Z}' = \{ \boldsymbol{z}' | \boldsymbol{z}' = \boldsymbol{z} + \mathcal{A}(f_e, f_D, \boldsymbol{x}, y_D, \epsilon), (\boldsymbol{x}, y_D) \in \mathcal{X}' \}$$
$$:= \mathcal{Z}'_s \cup \mathcal{Z}'_t, \quad (4)$$

where $\mathcal{Z}'_s$ are perturbed source features and $\mathcal{Z}'_t$ are perturbed target features. We use the projected gradient descent (PGD) to approximate the solution of Equation 3 as in [7], [8].

Contrastive Domain Adaptation. Next, inspired by [6], we propose a two-fold contrastive adaptation loss over $\mathcal{Z}_{adv}$ to further adapt knowledge from the source data domain to the target data domain. Firstly, we reduce the domain discrepancy among intra-class representations. That is, if a representation from the source data domain has a label of being true (or false) and a representation from the target data domain has a pseudo label of being true (or false), then these two representations are considered as intra-class representations and we reduce the domain discrepancy between them. The second level of our contrastive adaptation loss is defined for inter-class representations. As shown in Figure 1c, the discrepancy between the representations of true information and false information will be enlarged.

To compute the our proposed contrastive adaptation loss, we propose to measure the discrepancy among token classes using radial basis functions (RBF). In [11], RBF is proved to be an efficient tool to quantify uncertainty in deep neural networks. Since our pseudo labeling process is designed to automatically filter out low-confident labels for the target domain data, using RBF to measure the discrepancy among token classes could efficiently improve the quality of the pseudo labels and ultimately contribute the domain adaptation of the model. Formally, with the definition of the RBF kernel $k(z_1, z_2) = \exp[-\frac{\|z_1 - z_2\|^2}{2\sigma^2}]$, we define the class-aware loss for the misinformation detection task as follows:

$$\mathcal{L}_{con}(\mathcal{Z}') = -\sum_{i=1}^{|\mathcal{Z}'_s|}\sum_{j=1}^{|\mathcal{Z}'_t|} \frac{\mathbb{1}(y_s^{(i)} = 0, \hat{y}_t^{(j)} = 0)k(z_s^{(i)}, z_t^{(j)})}{\sum_{l=1}^{|\mathcal{Z}'_s|}\sum_{m=1}^{|\mathcal{Z}'_t|}\mathbb{1}(y_s^{(l)} = 0, \hat{y}_t^{(m)} = 0)}$$
$$-\sum_{i=1}^{|\mathcal{Z}'_s|}\sum_{j=1}^{|\mathcal{Z}'_t|} \frac{\mathbb{1}(y_s^{(i)} = 1, \hat{y}_t^{(j)} = 1)k(z_s^{(i)}, z_t^{(j)})}{\sum_{l=1}^{|\mathcal{Z}'_s|}\sum_{m=1}^{|\mathcal{Z}'_t|}\mathbb{1}(y_s^{(l)} = 1, \hat{y}_t^{(m)} = 1)}$$
$$+\sum_{i=1}^{|\mathcal{Z}'_s|}\sum_{j=1}^{|\mathcal{Z}'_s|} \frac{\mathbb{1}(y_s^{(i)} = 1, y_s^{(j)} = 0)k(z_s^{(i)}, z_s^{(j)})}{\sum_{l=1}^{|\mathcal{Z}'_s|}\sum_{m=1}^{|\mathcal{Z}'_s|}\mathbb{1}(y_s^{(l)} = 1, y_s^{(m)} = 0)}$$
$$+\sum_{i=1}^{|\mathcal{Z}'_t|}\sum_{j=1}^{|\mathcal{Z}'_t|} \frac{\mathbb{1}(\hat{y}_t^{(i)} = 1, \hat{y}_t^{(j)} = 0)k(z_t^{(i)}, z_t^{(j)})}{\sum_{l=1}^{|\mathcal{Z}'_t|}\sum_{m=1}^{|\mathcal{Z}'_t|}\mathbb{1}(\hat{y}_t^{(l)} = 1, \hat{y}_t^{(m)} = 0)},$$
$$(5)$$

where $\hat{y}_t$ is the pseudo label of the target domain samples and $\boldsymbol{z}$ denotes the representation of token CLS.

Overall Contrastive Adaptation Loss. Now, we merge the cross-entropy loss of the task classification problem and the above contrastive adaptation loss into a single optimization objective for the COVID model:

$$\mathcal{L}_{all} = \mathcal{L}_{ce}(\mathcal{X}) + \lambda\mathcal{L}_{con}(\mathcal{Z}'), \quad (6)$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss over the training data with ground-truth label or the pseudo label, and λ is the hyperparameter to adjust the domain adaptation strength. Moreover, we compute the RBF kernel with multiple bandwidths for $\mathcal{L}_{con}$, since multiple bandwidths of the RBF kernel encourage the model to learn a smoothed and generalized feature space [6].

V. EVALUATION

A. Experimental Design

In our experiments, we use three source misinformation datasets (GossipCop [12], LIAR [13] and PHEME [14]) released *before* the COVID outbreak and two COVID misinformation datasets (Constraint [15] and ANTiVax [16]) collected *after* the outbreak as target datasets. Following [5], the commonly used RoBERTa [17] was selected as the misinformation detection model. Moreover, regarding the baseline methods, we select DAAT [18], where the misinformation detection model is post-trained to improve the domain-adversarial adaptation, and EADA [5], where energy-based domain adversarial training is performed using autoencoder. Finally, in terms of the evaluation metrics, we use the balanced accuracy, accuracy and F1 score to evaluate the models' performance as in [4].

B. Evaluation Results

The results are reported in Table I. We observe that our method consistently outperforms all baseline algorithms in terms of adapting the models to the unseen COVID-19 domain. For instance, consider the setting where the model is adapted from source domain PHEME to target domain Constraint. This setting is the most challenging one: the model can only make random guess on the target Constraint dataset (BA=0.4889) without any domain adaptation. Under this setting, EADA failed to adapt the model by achieving a balanced accuracy of 0.4944. As for the other baseline DAAT, the adapted model only performs slightly better, namely BA=0.5227. In contrast,

TABLE I: Results of domain adaptation for COVID-19 misinformation detection.

Target Dataset	Source Dataset	LIAR			GossipCop			PHEME		
	Metric	BA ↑	Acc. ↑	F1 ↑	BA ↑	Acc. ↑	F1 ↑	BA ↑	Acc. ↑	F1 ↑
Constraint	No Adaptation	0.7231	0.7322	0.7822	0.5638	0.5832	0.7110	0.4889	0.5047	0.6360
	DAAT	0.7606	0.7626	0.7795	0.7178	0.7276	0.7806	0.5227	0.5411	0.6763
	EADA	0.7776	0.7794	0.7950	0.5210	0.5430	0.6944	0.4944	0.4969	0.6391
	CADM (Ours)	**0.7858**	**0.7944**	**0.8317**	**0.7787**	**0.7780**	**0.7828**	**0.6430**	**0.6547**	**0.7301**
ANTiVax	No Adaptation	0.5444	0.4929	0.4162	0.5695	0.6501	0.7673	0.5294	0.6196	0.7531
	DAAT	0.6228	0.5778	0.5393	0.6692	0.7161	0.7918	0.5895	0.6498	0.7518
	EADA	0.6184	0.6642	0.7490	0.5509	0.6434	0.7709	0.5411	0.6328	0.7632
	CADM (Ours)	**0.7123**	**0.7158**	**0.7652**	**0.7522**	**0.7701**	**0.8152**	**0.6752**	**0.7323**	**0.8107**

the performance of the model trained using our framework could be significantly improved (e.g., BA=0.6430). Similar trends could be observed on other adaptation settings with different source and target domain combinations. We observe that our adapted models achieve a better performance than the baselines on all metrics. Moreover, we also observe that our CADM is more robust than the other two baseline methods. For instance, regardless of the source domain dataset and target domain dataset, our method could consistently adapt the misinformation detection model to the target domain. However, on some datasets, the baseline methods successfully adapt the model but on other datasets the baseline methods simply fail. For instance, from GossipCop to Constraint, the model was adapted towards a worse direction, when EADA is deployed.

VI. CONCLUSION

In this paper, we present a novel unsupervised domain adaptation framework for COVID-19 misinformation detection on social media data. Our unsupervised framework is motivated by the fact that the ground-truth labels of the COVID-19 data are not always available but the need for high-quality information services is always persistent and urgent. In addition to COVID-19, our method has the potential to provide efficient solutions to many other information services on social media platforms, when the training labels of target domain data are missing. Extensive experimental results suggest that our method could effectively adapt the misinformation detection models from source domain to the target COVID-19 domain without requiring labels of COVID-19 data.

REFERENCES

[1] S. Malla and P. Alphonse, "Covid-19 outbreak: an ensemble pre-trained deep learning model for detecting informative tweets," *Applied Soft Computing*, vol. 107, p. 107495, 2021.

[2] Z. Kou, L. Shang, Y. Zhang, Z. Yue, H. Zeng, and D. Wang, "Crowd, expert & ai: A human-ai interactive approach towards natural language explanation based covid-19 misinformation detection," in *Proc. Int. Joint Conf. Artif. Intell.(IJCAI)*, 2022, pp. 5087–5093.

[3] D. Oniani and Y. Wang, "A qualitative evaluation of language models on automatic question-answering for covid-19," in *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2020, pp. 1–9.

[4] Z. Yue, H. Zeng, Z. Kou, L. Shang, and D. Wang, "Contrastive domain adaptation for early misinformation detection: A case study on covid-19," *arXiv preprint arXiv:2208.09578*, 2022.

[5] H. Zou, J. Yang, and X. Wu, "Unsupervised energy-based adversarial domain adaptation for cross-domain text classification," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 1208–1218.

[6] Z. Yue, B. Kratzwald, and S. Feuerriegel, "Contrastive domain adaptation for question answering using limited text corpora," *arXiv preprint arXiv:2108.13854*, 2021.

[7] H. Zeng, C. Zhu, T. Goldstein, and F. Huang, "Are adversarial examples created equal? a learnable weighted minimax risk for robustness under non-uniform attacks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 10 815–10 823.

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[9] M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Q. Tian, and W. Zhang, "Adversarial domain adaptation with domain mixup," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6502–6509.

[10] S. Lee, D. Kim, and J. Park, "Domain-agnostic question-answering with adversarial training," *arXiv preprint arXiv:1910.09342*, 2019.

[11] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *International conference on machine learning*. PMLR, 2020, pp. 9690–9700.

[12] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big data*, vol. 8, no. 3, pp. 171–188, 2020.

[13] W. Y. Wang, "" liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint arXiv:1705.00648*, 2017.

[14] C. Buntain and J. Golbeck, "Automatically identifying fake news in popular twitter threads," in *2017 IEEE International Conference on Smart Cloud (SmartCloud)*. IEEE, 2017, pp. 208–215.

[15] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "Fighting an infodemic: Covid-19 fake news dataset," in *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Springer, 2021, pp. 21–29.

[16] K. Hayawi, S. Shahriar, M. A. Serhani, I. Taleb, and S. S. Mathew, "Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection," *Public health*, vol. 203, pp. 23–30, 2022.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[18] C. Du, H. Sun, J. Wang, Q. Qi, and J. Liao, "Adversarial and domain-aware bert for cross-domain sentiment analysis," in *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 2020, pp. 4019–4028.