Energy-based Domain Adaption with Active Learning for Emerging Misinformation Detection

Kyumin Lee
Worcester Polytechnic Institute
100 Institute Rd, Worcester, MA, 01605
kmlee@wpi.edu

Guanyi Mou
Worcester Polytechnic Institute
100 Institute Rd, Worcester, MA, 01605
gmou@wpi.edu

Scott Sievert
Air Force Research Lab
26 Electronic Pkwy, Rome, NY 13441
scott.sievert.3@us.af.mil

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10021038

***Abstract*—Classifying whether collected information related to emerging topics and domains is fake/incorrect is not an easy task because we do not have enough labeled data in the domains. Given labeled data from source domains (e.g., gossip and health) and limited labeled data from a newly emerging target domain (e.g., COVID-19 and Ukraine war), simply applying knowledge learned from source domains to the target domain may not work well because of different data distribution. To solve the problem, in this paper, we propose an energy-based domain adaptation with active learning for early misinformation detection. Given three real world news datasets, we evaluate our proposed model against two baselines in both domain adaptation and the whole pipeline. Our model outperforms the baselines, improving at least 5% in the domain adaptation task and 10% in the whole pipeline, showing effectiveness of our proposed approach.**

***Index Terms*—domain adaptation, active learning, fake news, misinformation detection**

I. Introduction

The government (e.g., military units) and companies collect various information over time via online platforms such as news media and social media sites. Based on the collected information, stakeholders prepare planning and make important decisions. Sometimes, newly collected information from an emerging topic/domain may contain misinformation, and can easily fool the stakeholders because they are not familiar with the topic/domain. For example, military units may be interested in predicting veracity of information regarding recently occurred Ukraine war. How can we quickly and automatically determine the veracity? Especially, when we have limited labeled data from an emerging target domain, can we take advantage from other domains where we already have enough labeled data?

On one hand, transferring knowledge from source domain (i.e., other domains) to a target domain would be helpful in terms of overcoming limited labeled data issue in the target domain, since there would be common characteristics describing fake/misinformation across all domains. On the other hand, since data distribution of source and target domains would be different, simply applying the knowledge from the source domain to the target domain may not work well. For example, each domain may have the domain-specific terminology (e.g., words/phrases). If a model used it as a way to distinguish between misinformation and real information, the model may produce many false positives. What if we can get limited number of labeled data (i.e., a budget) in multiple-rounds from human experts for the target domain over time? What is the best way to select each subset of target domain data to be labeled?

The aforementioned problems motivate us to design and develop an intelligent machine learning model, which can automatically determine whether given textual information (e.g., news article) is fake or true. The prior work focused on either only domain adaptation [1; 2] for text and image data or active domain adaptation for image data given single source domain [3; 4]. However, researchers have rarely paid attention on active domain adaptation for textual data given multi-source domains.

Therefore, we propose an energy-based domain adaptation with active learning. Our proposed domain adaptation approach will focus on minimizing classification error of the source domain data and minimizing free energy difference between source and target domains called domain alignment. In active learning, our approach will focus on minimizing recently labeled target domain data's classification error and minimizing free energy difference between source and target domains. We explore a few active learning strategies to see which one produces the best result. In this paper, we limit the input as textual data without accessing other auxiliary/external data (e.g., social interaction information) for early detection.

II. Related Work

In this section, we summarize related work in fake news/misinformation detection, domain adaptation and active learning.

In fake news/misinformation detection, researchers focused on building a model based on content (e.g., news itself, news comments, social media posting), social engagement and consumer's preference. For example, [5] proposed content-based approach to learn content representation via feature learning and deep learning. [6] incorporated news comments to enhance the learning process and overcome labeling effort. Since confirmation bias plays an important role when a user determines whether he/she is going to consume content of news or social media posting, researchers developed a user preference-aware fake news detection model [7].

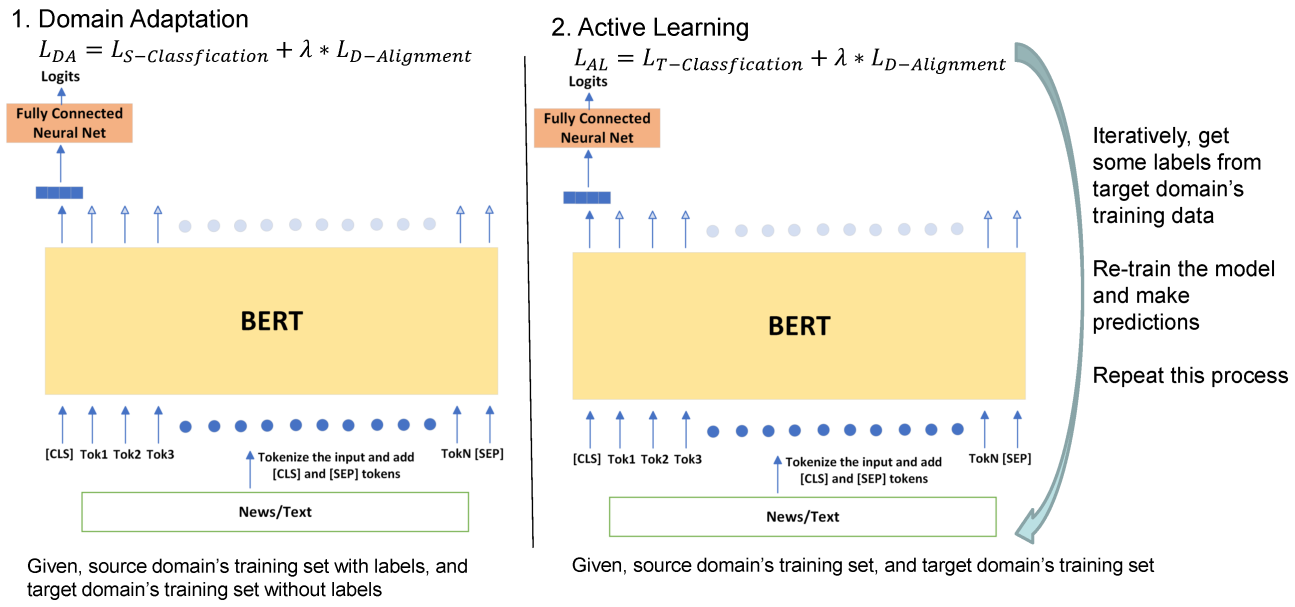

Fig. 1. Two-stage learning of our proposed framework.

Domain adaptation is a subcategory of transfer learning, aims to build a model from a single or multiple source domains and apply the model to a target domain. Researchers mostly focused on learning a domain-invariant feature representation learning [1; 8; 9; 2]. A weakly labeling approach for the target domain was proposed to further enhance the performance of a model [10].

Recently, researchers proposed methods to conduct domain adaptation with active learning [3; 4]. For example, Xie et al. [4] proposed an energy-based approach for an image classification task. This work is the most closely related to our work. However, the main differences between theirs and ours are our energy-based loss function is different from theirs, and a way to fine-tune a model given labeled data from the target domain by active learning is different. In addition, their two-stage active learning approach is different from our active learning strategies. In this work, we mainly focus on text data and combine multiple source domain data instead of one single source for one target domain.

III. Proposed Approach

We propose an energy-based domain adaptation (EDA) with active learning for fake news detection. Figure 1 shows a general view of our proposed approach which consists of two stage learning: (i) domain adaptation and (ii) active learning.

In the domain adaptation, given source domain's training data with labels and target domain's training data without labels, a pretrained language model's tokenizer tokenizes each document (i.e., a news article in this paper) into tokens. We chose BERT as our encoder, which processes input document and produces a vector representation. Note that we can replace the encoder with any other language models such as GPT-3 and T5. As illustrated in the figure, BERT tokenizer adds [CLS] and [SEP] tokens to tokens from the input document. In the end, [CLS]'s vector representation, which captures meaning of the input document, is fed to fully connected neural network, which produces logits. Since our fake news detection task is a binary classification task, the logits contain two scalar values.

Inspired from energy-based models [11; 4], we interpret each logit value as energy corresponding to each class. In the training process, we minimize true class's corresponding energy. When x is a high-dimensional variable and y is a discrete variable, a correctly trained model/energy function $E(x, y)$ should produce the lowest energy to correct class and the highest energy to incorrect class. For example, given a fake news article, our model should produce small energy corresponding to "fake news" class and high energy to "real news" class. Formally, in the prediction, the energy-based model should produce a class ID/value $\breve{y}$ which returns the smallest energy:

$$\breve{y} = argmin_{y \in Y} E(x, y) \quad (1)$$

In training the domain adaptation, we focus on minimizing two loss functions – source domain's supervision (i.e., fine-tuning BERT by the source domain's training set) and free energy difference between source and target domains called domain alignment (i.e., unsupervised learning).

$L_{s-classification}$ is measured as follows:

$$L_{s-classification} = E(x, y; \theta)^2 + exp^{(-E(x, \bar{y}; \theta))} \quad (2)$$

where y is the true class and $\bar{y}$ is the other/incorrect class given a news article from source domain's training set.

The intuition behind of Eq. 2 is that the model tries to minimize the loss by minimizing energy of the correct class and maximizing energy of the other/incorrect class.

The second loss $L_{D-Alignment}$ is measured as follows:

$$L_{D-Alignment} = max(0, \mathbb{E}_{x \sim \mathcal{D}_S} F(x; \theta) - \mathbb{E}_{x \sim \mathcal{D}_T} F(x; \theta)) \quad (3)$$

where free energy is measured by $F(x;\theta) = -log\sum_{y\in Y} exp^{(-E(x,y;\theta))}$. The main purpose of the domain alignment loss is to minimize free energy difference between source domain and target domain.

Finally, we combine the two loss functions as follows:

$$L_{DA} = L_{S-classification} + \lambda * L_{D-Alignment} \quad (4)$$

where λ is a hyper-parameter to tune importance of each loss in the domain adaptation process.

As shown in Figure 1, once the model is trained by the domain adaptation, we run active learning, which iteratively subsamples some instances from the target domain's training set based on a predefined budget and gets true labels from human experts/the oracle. Then, we use labeled target instances to further train our model. We conduct the active learning multiple rounds to see whether which active learning strategy contributes the most positively, improving performance of the model.

In each round, once we got labeled sample from the target domain, we further re-train the model, and then measure the following loss function. The training process is to minimize the following loss:

$$L_{AL} = L_{T-classification} + \lambda * L_{D-Alignment} \quad (5)$$

where again, λ is a hyper-parameter to tune importance of each loss in the active learning process. As the name $L_{T-classification}$ indicates, we measure classification loss of labeled target samples. $L_{D-Alignment}$ tries to keep similar free energy between source domain and target domain.

We consider two types of active learning strategies: (i) random selection (EDA-random) and (ii) uncertainty-based selection (EDA-uncertainty). The uncertainty-based selection approach estimates each target domain training instance's energy of being fake news and being real news, and selects top-k most uncertain instances, which have the smallest energy difference.

IV. Experiments

We answer the following research questions:

- RQ1: What is the effectiveness of our approach in domain adaptation?
- RQ2: What is the effectiveness of our approach with active learning?

First of all, we describe datasets and experiment setting. Then, experiment results are presented.

TABLE I
THE STATISTICAL INFORMATION OF THE DATASETS.

Datasets	Fake	Real
GossipCop	4,252	4,252
PolitiFact	260	260
Health	1,997	1,997

A. Datasets and Experiment Setting

Dataset. Three fake news benchmark datasets were used for experiments: GossipCop (gossip), PolitiFact [12] and Health DETERRENT (health) [13]. Each dataset contains news content and labels (i.e., fake news or real news). Table I presents each dataset's statistical information. We evaluated our model and baselines under the standard balance setting. We utilized the BERT-base tokenizer to tokenize each news article and used BERT-base as the encoder.

Experiment Setting. Each dataset was split to training, validation, and test sets with a ratio of 70%, 10% and 20%, respectively. We selected one of the datasets as a target domain dataset, and the remaining two datasets were used as a source domain dataset by combining them together. We did this process three times so that we could have three pairs of source datasets and target dataset. Given a pair, we trained our model by the labeled source domain training set and unlabeled target domain training set in the domain adaptation, and then trained our model by the labeled target domain training data sampled by an active learning strategy, and source and target domains' training sets for domain alignment. The target domain's validation set was used for hyperparameter tuning. Finally, the optimized model was evaluated over the target domain's test set. We applied the same setting to all baselines as well. We report the average results over the three pairs of source and target domains.

We compare our model with two baselines: EADA [4] and BERT with cross entropy (our own baseline). EADA utilizes a two-stage active learning method in which first selects top-k samples from target domain training set by free energy, and then further sample instances from the top-k samples based on uncertainty. BERT with cross entropy is only compared for the domain adaptation task.

For a fair comparison, We used the same experiment setting. For example, BERT-base was used as encoder for our model and the baselines. We conducted a grid search for the learning rate in a range of 0.00008, 0.00005 and 0.00001. λ was set as 0.1 in Eq. 4 and Eq. 5, following the prior work [4]. RAdam optimizer was applied. In the domain adaptation, all the models were trained for 15 epochs. We saved the best checkpoint at the end of each epoch and report the test result for checkpoint with the best validation accuracy score. The saved checkpoint was used in the active learning stage. In the active learning, a budget was 15 (i.e., sample 15 instances from unlabeled target domain training set) and four rounds were run. We repeated experiments three times with different seeds, and the average result was reported. We implemented our model with PyTorch[1] (version 1.11.0) and used Hugging Face BERT-base-uncased as the encoder and tokenizer.

Since the datasets were balanced, we utilized accuracy as an evaluation metric to represent the performance of our approach and the baseline methods.

B. Results

Table II shows experiment results of our EDA and two baselines (EADA and BERT) in the domain adaptation task. Our approach was better than two baselines in the first two pairs

[1]https://pytorch.org/

TABLE II
PREDICTION RESULTS OF OUR MODEL AND BASELINES IN THE DOMAIN ADAPTATION WITHOUT ACTIVE LEARNING (S: SOURCE, T: TARGET).

Method	S: health&gossip T: politifact	S: gossip&politifact T: health	S: politifact&health T: gossip	Avg. Accuracy
EADA	51.2%	63.4%	55.5%	56.7%
BERT	58.3%	61.7%	**57.2%**	59.0%
Our EDA	**63.5%**	**65.9%**	56.8%	**62.0%**

TABLE III
PREDICTION RESULTS OF OUR MODEL AND A BASELINE WITH ACTIVE LEARNING AT 60 LABELED TARGET TRAINING DATA (S: SOURCE, T: TARGET).

Method	S: health&gossip T: politifact	S: gossip&politifact T: health	S: politifact&health T: gossip	Avg. Accuracy
EADA	72.4%	**88.0%**	54.0%	71.4%
EDA-Random	81.7%	**87.9%**	65.0%	78.2%
EDA-Uncertainty	**86.8%**	85.0%	**65.3%**	**79.0%**

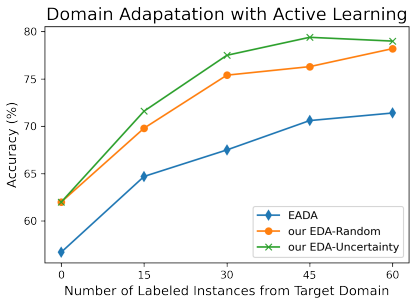

Fig. 2. Prediction results with active learning over four rounds, each of which received true labels of 15 instances sampled from the target domain training set.

and was competitive in the third pair. EADA performed poorly compared with the other baseline and our model. It means that their loss function, which tried to minimize correct class' energy and maximize overall energy, did not learn the decision boundary properly. Overall, our approach outperformed the baselines, and achieved at least 5% improvement compared with the best baseline (BERT).

Next, we measure the effectiveness of our model with active learning against EADA. Figure 2 shows how prediction performance has changed as we get more labels from unlabeled target domain training set. Our EDA with two active learning strategies outperformed the baseline (EADA), achieving about 10% improvement. Table III shows more detailed results with 60 labels. Limited labeled data (i.e., 60 labels) already significantly improved our model's performance compared with one without active learning (i.e., 62% vs. 79% accuracy of EDA-uncertainty), indicating the effectiveness of active learning.

V. CONCLUSION

In this paper, we proposed an energy-based domain adaptation with active learning for fake news detection. Our model outperformed baselines in both domain adaptation and the whole pipeline, achieving at least 5% and 10% improvements, respectively. In the future, we will investigate a more advanced domain adaptation approach, which learns each source domain's importance/similarity to a target domain instead of simply combining two source domain data. We are also interested in studying more advanced active learning strategies.

ACKNOWLEDGMENT

This work was supported by the Air Force Research Lab Summer Faculty Fellowship Program, and NSF grants CNS-1755536 and IIS-2039951. We thank Daniel Carpenter and Lee Seversky for their support and advice.

REFERENCES

[1] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015.
[2] 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," in *KDD*, 2018.
[3] V. Prabhu, A. Chandrasekaran, K. Saenko, and J. Hoffman, "Active domain adaptation via clustering uncertainty-weighted embeddings," in *ICCV*, 2021.
[4] B. Xie, L. Yuan, S. Li, C. H. Liu, X. Cheng, and G. Wang, "Active learning for domain adaptation: An energy-based approach," in *AAAI*, 2022.
[5] 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," in *EMNLP*, 2017.
[6] Y. Wang, W. Yang, F. Ma, J. Xu, B. Zhong, Q. Deng, and J. Gao, "Weak supervision for fake news detection via reinforcement learning," in *AAAI*, 2020.
[7] Y. Dou, K. Shu, C. Xia, P. S. Yu, and L. Sun, "User preference-aware fake news detection," in *SIGIR*, 2021.
[8] Y. Han, S. Karunasekera, and C. Leckie, "Graph neural networks with continual learning for fake news detection from social media," *arXiv preprint arXiv:2007.03316*, 2020.
[9] A. Silva, L. Luo, S. Karunasekera, and C. Leckie, "Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data," in *AAAI*, 2021.
[10] Y. Li, K. Lee, N. Kordzadeh, B. Faber, C. Fiddes, E. Chen, and K. Shu, "Multi-source domain adaptation with weak supervision for early fake news detection," in *Big Data*, 2021.
[11] Y. LeCun and F. J. Huang, "Loss functions for discriminative training of energy-based models," in *International workshop on artificial intelligence and statistics*, 2005.
[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] L. Cui, H. Seo, M. Tabar, F. Ma, S. Wang, and D. Lee, "Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation," in *KDD*, 2020.