Emotion-guided Cross-domain Fake News Detection using Adversarial Domain Adaptation

Arjun Choudhry[†*], Inder Khatri[†*], Arkajyoti Chakraborty[†], Dinesh Kumar Vishwakarma[†], Mukesh Prasad[‡]
[†]Biometric Research Laboratory, Delhi Technological University, New Delhi, India
[‡]School of Computer Science, University of Technology Sydney, Australia

Abstract

Recent works on fake news detection have shown the efficacy of using emotions as a feature or emotions-based features for improved performance. However, the impact of these emotion-guided features for fake news detection in cross-domain settings, where we face the problem of domain shift, is still largely unexplored. In this work, we evaluate the impact of emotion-guided features for cross-domain fake news detection, and further propose an emotion-guided, domain-adaptive approach using adversarial learning. We prove the efficacy of emotion-guided models in cross-domain settings for various combinations of source and target datasets from FakeNewsAMT, Celeb, Politifact and Gossipcop datasets.

1 Introduction

In recent years, our reliance on social media as a source of information has increased multi-fold, bringing along exponential increase in the spread of *fake news*. To counter this, researchers have proposed various approaches for fake news detection (Shu et al., 2019; Sheng et al., 2022). However, models trained on one domain are often brittle and vulnerable to incorrect predictions for the samples of another domain (Saikh et al., 2019; Pérez-Rosas et al., 2018). This is primarily due to the shift between the two domains, as depicted in Figure 1(1). To handle this, some domain-adaptive frameworks (Zhang et al., 2020; Huang et al., 2021; Li et al., 2021) have been proposed which help align the source and target domains in the feature space to ameliorate domain shift across different problems. These frameworks guide the feature extractors to extract domain-invariant features by aligning the source and target domains in the feature space, thus generalizing well across domains. However, due to the absence of labels in the target-domain data, the adaptation is often prone to negative transfer, which can disturb the class-wise distribution and affect the discriminability of the final model, as shown in Figure 1(2).

Some recent studies have observed a correlation between the veracity of a text and its emotions. There exists a prominent affiliation for certain emotions with fake news, and for other emotions with real news (Vosoughi et al., 2018), as illustrated in Figure 1(3). Further, some works have successfully utilized emotions as features, or emotion-guided features to aid in fake news detection (Guo et al., 2019; Zhang et al., 2021; Choudhry et al., 2022). However, we observe that these works only consider the in-domain setting for evaluation, without analyzing the robustness of these frameworks to domain shift in cross-domain settings. This is another important direction that needs to be explored.

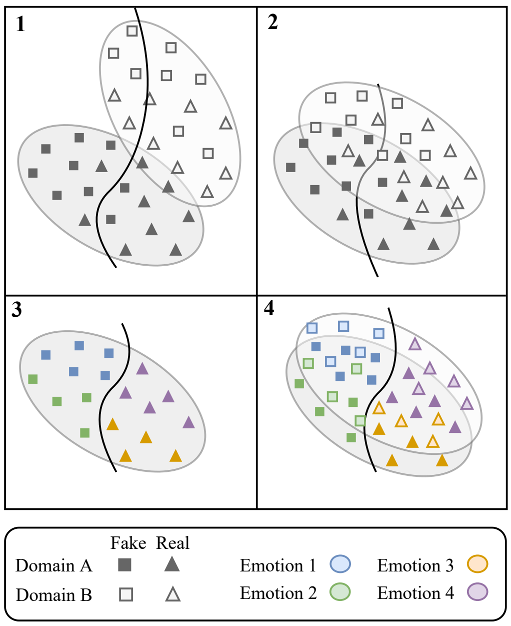

Figure 1: (1) Cross-domain texts not aligned. (2) Domain adaptation leads to some alignment. (3) Emotion-guided classification in one domain. (4) Emotion-guided domain adaptation leads to improved alignment of the two domains.

*Equal contribution

Proceedings of the 19th International Conference on Natural Language Processing (ICON), pages 75 - 79
December 15-18, 2022

In this paper, we study the efficacy of emotion-aided models in capturing better generalizable features for cross-domain fake news detection. Table 1 shows the improvements observed in various cross-domain settings when our emotion-guided models were evaluated in cross-domain settings. We observe that emotion-guided frameworks show improved performance in cross-domain settings, as compared to their baseline models without the said emotion-aided features, thus underscoring the generalized feature extraction in emotion-aided models. We further propose an emotion-guided unsupervised domain adaptation framework, which utilizes emotion labels in a multi-task adversarial setting for better feature alignment across domains. The emotion labels for emotion classification, trained parallel to the fake news detection branch in the multi-task learning setup, help provide additional supervision for improved alignment during domain adaptation, mitigating the issue of incorrect alignment of domains. This is illustrated in Figure 1(4)). This leads to better discriminability. We experimentally prove the efficacy of our approach across a variety of datasets in cross-domain settings for various combinations of single-task or multi-task, domain-adaptive or non-adaptive, emotion-guided or unguided settings on the accuracy of the models.

Our contributions can be summarized as follows:

- We suggest the use of emotion classification as an auxiliary task for improved fake news detection in cross-domain settings, indicating the applicability of emotion-guided features across domains.
- We compare how Ekman's and Plutchik's base emotion classes individually affect the performance of our multi-task domain-adaptive framework, and if there are meaningful differences between them.
- We propose an emotion-guided domain-adaptive framework for fake news detection across domains. We show that domain-adaptive fake news detection models better align the two domains with the help of supervised learning using emotion-aided features.
- We evaluate our approach on a variety of source and target combinations from four datasets. Our results prove the efficacy of our approach.

2 Related Works

Several studies over the last few years have explored the correlation of fake news detection with emotions. K et al. (2020) *emotionized* text representations using explicit emotion intensity lexicons. Guo et al. (2019) utilized the discrepancies between publisher's emotion and the thread's comments' emotions to detect fake news. However, most of these methods relied upon some additional inputs during evaluation. Choudhry et al. (2022) proposed an emotion-aided multi-task learning approach, where emotion classification was the auxiliary task implicitly aligning fake news features according to emotion labels.

Inspired by Ganin et al. (2015), Zhang et al. (2020) proposed the first fake news detection work to tackle domain shifts between different datasets. They proposed a multi-modal framework with a Gradient Reversal Layer (GRL) to learn domain-invariant features across different domains and used a joint fake news detector on the extracted features. Huang et al. (2021) proposed a robust and generalized fake news detection framework adaptable to a new target domain using adversarial training to make the model robust to outliers and Maximum Mean Difference (MMD)-based loss to align the features of source and target. Li et al. (2021) extended the problem by treating it as a multi-source domain adaptation task, using the labeled samples from multiple source domains to improve the performance on unlabeled target domains. They also utilized weak labels for weak supervision on target samples to improve performance.

However, no previous work has aligned features between different domains using emotion-guided features and domain adaptation using adversarial training. We show that applying both of these approaches leads to improved performance due to better alignment of inter-domain features.

3 Proposed Methodology

3.1 Datasets, Emotion Annotation & Preprocessing

We use the FakeNewsAMT (Pérez-Rosas et al., 2018), Celeb (Pérez-Rosas et al., 2018), Politifact[1], and Gossipcop[2] datasets for cross-domain fake news detection. FakeNewsAMT is a multi-domain dataset containing samples from technology, education, business, sports, politics, and en-

[1] https://www.politifact.com [2] https://www.gossipcop.com

tertainment domains. The Celeb dataset has been derived from the web, and contains news about celebrities. Politifact is a web-scrapped dataset containing political news, while Gossipcop contains news extracted from the web, manually annotated via crowd-sourcing and by experts.

We use the Unison model (Colnerič and Demšar, 2020) to annotate all datasets with the core emotions from Ekman's (Ekman, 1992) (6 emotions: *Joy*, *Surprise*, *Anger*, *Sadness*, *Disgust*, *Fear*) and Plutchik's (Plutchik, 1982) (8 emotions: *Joy*, *Surprise*, *Trust*, *Anger*, *Anticipation*, *Sadness*, *Disgust*, *Fear*) emotion theories. During preprocessing, we convert text to lowercase, remove punctuation, and de-contract verb forms (eg. *"I'd"* to *"I would"*).

3.2 Multi-task Learning

We use multi-task learning (MTL) to incorporate emotion classification as an auxiliary task to our fake news detection branch. Multi-task learning enables a model to learn the shared features between two or more correlated tasks for improved feature extraction and performance. We use Ekman's or Plutchik's emotions labels for emotion classification branch in our MTL models to see which performs better, and compare the performance with the corresponding single-task (STL) models in domain-adaptive and non-adaptive settings.

3.3 Emotion-guided Domain-adaptive Framework

We propose the cumulative use of domain adaptation and emotion-guided feature extraction for cross-domain fake news detection. Our approach aims to improve the feature alignment between different domains using adversarial domain adaptation, by leveraging the correlation between the emotion and the veracity of a text (as shown in Figure 1(4)). Figure 2 shows our proposed framework. We use an LSTM-based (Hochreiter and Schmidhuber, 1997) feature extractor, which is trained using the accumulated loss from fake news classifier, emotion classifier and the discriminator (aids in learning domain-invariant features). LSTM can be replaced with better feature extractors. We used it specifically for easier comparison to non-adapted emotion-guided and non-adapted single-task models. The domain classifier acts as the discriminator. In our proposed framework, we do not use the truth labels for the target dataset for domain adaptation. However, we utilize the target domain emotion labels in our approach to better align the two domains using the emotion labels for supervised learning. The fake news classification loss, emotion classification loss, adversarial loss, and total loss are defined as in Equations 1, 2, 3, and 4:

$$L_{FND} = \min_{\theta_l,\theta_f} \sum_{i=1}^{n_s} L_f^i \tag{1}$$

$$L_{emo} = \min_{\theta_l,\theta_e} \sum_{i=1}^{n_s} L_{es}^i + \sum_{j=1}^{n_t} L_{et}^j)) \tag{2}$$

$$L_{adv} = \min_{\theta_d}(\max_{\theta_l}(\sum_{i=1}^{n_s} L_{ds}^i + \sum_{j=1}^{n_t} L_{dt}^j)) \tag{3}$$

$$L_{Total} = (1-\alpha-\beta)*L_{FND} + \alpha * (L_{adv}) + \beta * (L_{emo}) \tag{4}$$

where n_s and n_t are number of samples in source and target sets; θ_d, θ_f, θ_e and θ_l are parameters for discriminator, fake news classifier, emotion classifier and LSTM feature extractor; L_{d_s} and L_{d_t} are binary crossentropy loss for source and target classification; L_{es} and L_{et} are crossentropy loss for emotion classification; L_f is binary crossentropy loss for Fake News Classifier; α and β are weight parameters in L_{Total}. We optimised α and β for each setting for optimal performance.

We used 300 dimension GloVe (Pennington et al., 2014) embeddings. All models were trained for up to 50 epochs, stopped when the peak validation accuracy for the in-domain validation set was achieved. We used a batch size of 25 for every experiment. Each model used the Adam optimizer with learning rate 0.0025. We used an LSTM layer with 256 units for feature extraction, while both fake news detection and emotion classification branches consisted of two dense layers each.

4 Experimental Analysis & Results

We evaluated our proposed approach on various combinations of source and target datasets. Each model was optimized on an in-domain validation set from the source set. Table 1 illustrates our results proving the efficacy of using emotion-guided models in non-adapted and domain-adapted cross-domain settings. It compares non-adaptive models, domain-adaptive models, and our emotion-guided domain-adaptive models in various settings. MTL(E) and MTL(P) refer to emotion-guided multi-task frameworks using Ekman's and Plutchik's emotions respectively. STL refers to the single-task framework. DA refers to the use of the domain-adaptive framework, containing a discriminator. Some findings observed are:

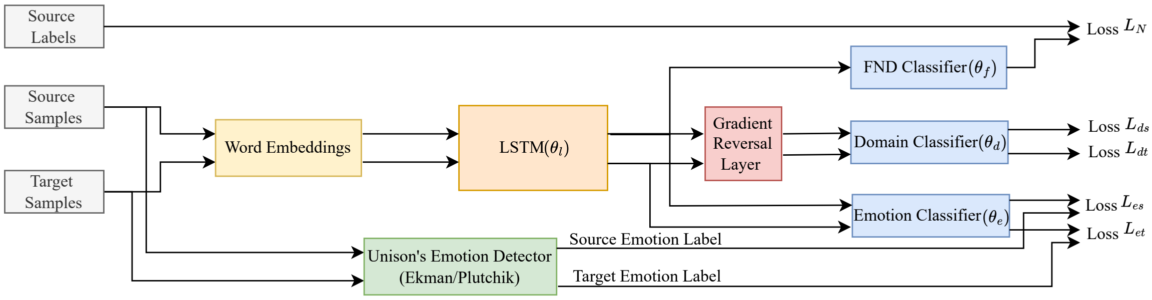

Figure 2: Pictorial depiction of our emotion-guided domain-adaptive approach for cross-domain fake news detection.

Source	Target	Setting	Accuracy
FakeNewsAMT	Celeb	STL	0.420
		MTL(E)	0.520
		MTL(P)	0.530
		DA STL	0.560
		DA MTL(E)	0.540
		DA MTL(P)	**0.600**
Celeb	FakeNewsAMT	STL	0.432
		MTL(E)	0.471
		MTL(P)	0.476
		DA STL	0.395
		DA MTL(E)	0.501
		DA MTL(P)	**0.551**
Politifact	Gossipcop	STL	0.527
		MTL(E)	0.555
		MTL(P)	0.603
		DA STL	0.585
		DA MTL(E)	**0.698**
		DA MTL(P)	0.671
Celeb	Gossipcop	STL	0.488
		MTL(E)	0.501
		MTL(P)	0.490
		DA STL	0.525
		DA MTL(E)	0.555
		DA MTL(P)	**0.587**
FakeNewsAMT	Gossipcop	STL	0.451
		MTL(E)	0.652
		MTL(P)	0.620
		DA STL	0.790
		DA MTL(E)	**0.805**
		DA MTL(P)	0.795
FakeNewsAMT	Politifact	STL	0.363
		MTL(E)	0.450
		MTL(P)	0.530
		DA STL	0.621
		DA MTL(E)	**0.704**
		DA MTL(P)	0.621

Table 1: Cross-domain evaluation of non-adaptive and adaptive models on FakeNewsAMT, Celeb, Politifact and Gossipcop datasets. Emotion-guided domain-adaptive models (DA MTL(E) and DA MTL(P)) outperform their corresponding STL models in cross-domain settings. Domain-adaptive MTL models consistently outperform baseline STL, non-adaptive MTL and domain-adaptive STL models.

MTL(E) and MTL(P) models outperform their STL counterparts in cross-domain settings, as seen in Table 1. This indicates improved extraction of generalizable features by the emotion-guided models, which aids in improved fake news detection across datasets from different domains.

DA STL models generally outperform STL models in cross-domain settings across multiple combinations of datasets. However, we see the STL model outperformed the DA STL model for Celeb dataset as the source dataset and FakeNewsAMT dataset as target, confirming that unguided adaptation can sometimes lead to negative alignment, reducing the performance of the model.

DA MTL(E) and DA MTL(P) models improve performance in cross-domain settings. Table 1 shows improved results obtained using the emotion-guided adversarial DA models over their non-adaptive counterparts. This shows the scope for improved feature extraction even after using DA, and emotion-guided models act as a solution aiding in correct alignment of the samples and features extracted by the adaptive framework from different domains. Emotion-guided DA models mitigated the issue of negative alignment when Celeb dataset was the source and FakeNewsAMT dataset the target, where STL model outperformed the DA STL model. The emotion-guided DA models helped correctly align the two domains, leading to significantly improved performance.

5 Conclusion

In this work, we showed the efficacy of emotion-guided models for improved cross-domain fake news detection, and presented an emotion-guided domain-adaptive fake news detection approach, evaluating it against baseline STL, emotion-guided MTL, DA STL and emotion-guided DA MTL models for various source and target combinations from four datasets. Our proposed approach led to improved cross-domain fake news detection accuracy, indicating that emotions are generalizable across domains and aid in better alignment of different domains during domain adaptation.

References

Arjun Choudhry, Inder Khatri, and Minni Jain. 2022. An emotion-based multi-task approach to fake news detection (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12929–12930.

Niko Colnerič and Janez Demšar. 2020. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE Transactions on Affective Computing*, 11(3):433–446.

P. Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2015. Domain-adversarial training of neural networks.

Chuan Guo, Juan Cao, Xueyao Zhang, Kai Shu, and Miao Yu. 2019. Exploiting emotions for fake news detection on social media. *ArXiv*, abs/1903.01728.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Yinqiu Huang, Min Gao, Jia Wang, and Kai Shu. 2021. DAFD: domain adaptation framework for fake news detection. In *Neural Information Processing - 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8-12, 2021, Proceedings, Part I*, volume 13108 of *Lecture Notes in Computer Science*, pages 305–316. Springer.

Anoop K, Deepak P, and Lajish V L. 2020. Emotion cognizance improves health fake news identification. In *Proceedings of the 24th Symposium on International Database Engineering & Applications*, IDEAS '20. Association for Computing Machinery.

Yichuan Li, Kyumin Lee, Nima Kordzadeh, Brenton Faber, Cameron Fiddes, Elaine Chen, and Kai Shu. 2021. Multi-source domain adaptation with weak supervision for early fake news detection. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 668–676.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401. Association for Computational Linguistics.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5).

Tanik Saikh, Arkadipta De, Asif Ekbal, and Pushpak Bhattacharyya. 2019. A deep learning approach for automatic detection of fake news. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 230–238, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 395–405, New York, NY, USA. Association for Computing Machinery.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Tong Zhang, Di Wang, Huanhuan Chen, Zhiwei Zeng, Wei Guo, Chunyan Miao, and Lizhen Cui. 2020. Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection. In *IJCNN*.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, WWW '21, page 3465–3476, New York, NY, USA. Association for Computing Machinery.