An Emotion-Guided Approach to Domain Adaptive Fake News Detection Using Adversarial Learning (Student Abstract)

Arkajyoti Chakraborty[*1], **Inder Khatri**[*1], **Arjun Choudhry**[*1], **Pankaj Gupta**[1], **Dinesh Kumar Vishwakarma**[1], **Mukesh Prasad**[2]

[1] Biometric Research Laboratory, Delhi Technological University, New Delhi, India
[2] School of Computer Science, FEIT, University of Technology Sydney, Sydney, Australia
{arkajyotichakraborty_2k19ep022, dinesh}@dtu.ac.in, {inderkhatri999, choudhry.arjun}@gmail.com,
mukesh.prasad@uts.edu.au

Abstract

Recent works on fake news detection have shown the efficacy of using emotions as a feature for improved performance. However, the cross-domain impact of emotion-guided features for fake news detection still remains an open problem. In this work, we propose an emotion-guided, domain-adaptive, multi-task approach for cross-domain fake news detection, proving the efficacy of emotion-guided models in cross-domain settings for various datasets.

Introduction

Over the years, our reliance on social media as an information source has increased, leading to an exponential increase in the spread of *fake news*. To counter this, researchers have proposed various approaches for fake news detection (FND). Models trained on one domain often perform poorly on datasets from other domains due to the domain shift (Figure 1(1)). Some works show the efficacy of domain adaptation for cross-domain FND by extracting domain-invariant features (Figure 1(2)) for classification (Zhang et al. 2020). However, adapting domains does not ensure that features in different classes align correctly across domains, which sometimes has a negative impact on performance. Some works have shown a correlation between fake news and their intrinsic emotions (Guo et al. 2019; Choudhry, Khatri, and Jain 2022) (Figure 1(3)), having successfully used it for fake news detection. However, these works are restricted to in-domain settings and don't consider cross-domain evaluation. We propose the use of emotion-guided multi-task models for improved cross-domain fake news detection, experimentally proving its efficacy, and present an emotion-guided domain adaptive approach for improved cross-domain fake news detection by leveraging better feature alignment across domains due to the use of emotion labels (Figure 1(4)).

Proposed Methodology

Datasets, Emotion Annotation & Preprocessing

We use the FakeNewsAMT & Celeb (Pérez-Rosas et al. 2018), Politifact[1], and Gossipcop[2] datasets. We annotate
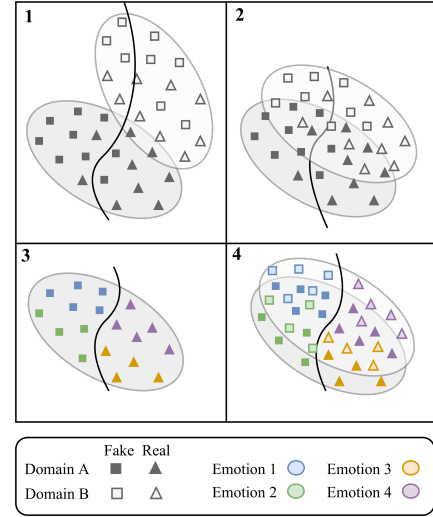
Figure 1: (1) Cross-domain texts not aligned. (2) Domain adaptation for improved alignment. (3) Emotion-guided classification. (4) Emotion-guided domain adaptation.

them 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. We use the Unison model (Colneric and Demsar 2018) for annotating the datasets with emotion tags. During preprocessing, we convert text to lower case, remove punctuation, and decontract verb forms (eg. *"I'd"* to *"I would"*).

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)).

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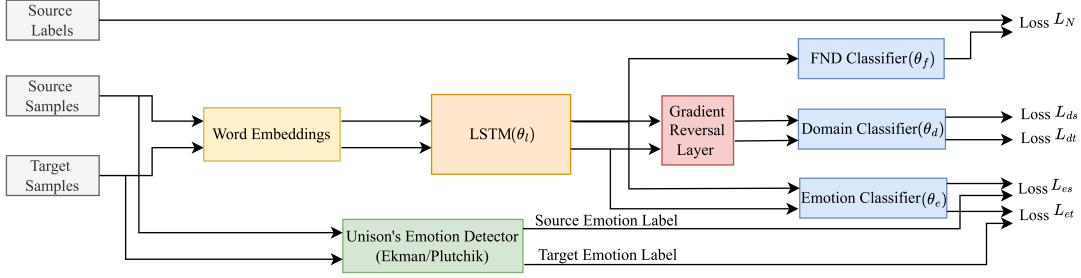[2]https://www.gossipcop.com/

Figure 2: Graphical representation of our emotion-guided domain-adaptive framework for cross-domain fake news detection.

Figure 2 shows our proposed framework. We use an LSTM-based multi-task learning (MTL) feature extractor which is trained by the cumulative losses from fake news classifier, emotion classifier, and discriminator (aids in learning domain-invariant features). LSTM can be replaced with better feature extractors. We use it specifically for easier comparison to non-adapted emotion-guided and non-adapted single-task models. The domain classifier acts as the discriminator. Fake news classification loss, emotion classification loss, adversarial loss, and total loss are defined as:

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

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

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

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

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 optimized α and β for each setting.

Experimental Results & Discussion

Each model used for evaluation was optimized on an in-domain validation set. Table 1 illustrates our results proving the efficacy of using emotion-guided models in non-adapted cross-domain settings. Table 2 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 single-task framework. DA refers to domain-adaptive framework with a discriminator. Non-DA refers to a non-adapted model. Some findings observed are:

Emotions-guided non-adaptive multi-task models outperform their single-task counterparts in cross-domain settings, as seen in Table 1, indicating improved extraction of features that are applicable across different datasets.

Emotion-guided domain-adaptive models improve performance in cross-domain settings. Table 2 shows the

Source	Target	Accuracy Non-DA STL	Accuracy Non-DA MTL(E)	Accuracy Non-DA MTL(P)
FAMT	Celeb	0.420	0.520	**0.530**
Celeb	FAMT	0.432	0.471	**0.476**

Table 1: Cross-domain evaluation of non-adaptive models on FakeNewsAMT (FAMT) & Celeb datasets. Emotion-guided models (MTL (E) and MTL (P)) outperform their corresponding STL models in cross-domain settings.

Source	Target	Accuracy Non-DA STL	Accuracy DA STL	Accuracy DA MTL(E)	Accuracy DA MTL(P)
FAMT	Celeb	0.420	0.560	0.540	**0.600**
Celeb	FAMT	0.432	0.395	0.501	**0.551**
Politi	Gossip	0.527	0.585	**0.698**	0.671
Celeb	Gossip	0.488	0.525	0.555	**0.587**
FAMT	Gossip	0.451	0.790	**0.805**	0.795
FAMT	Politi	0.363	0.621	**0.704**	0.621

Table 2: Cross-domain evaluation of non-adaptive, adaptive and emotion-guided adaptive models on various datasets.

advantage of emotion-guided adversarial domain-adaptive models over their non-adaptive counterparts. This shows the scope for improved feature extraction even after adversarial adaptation, and emotion-guided models act as a solution.

References

Choudhry, A.; Khatri, I.; and Jain, M. 2022. An Emotion-Based Multi-Task Approach to Fake News Detection (Student Abstract). *AAAI*, 36(11).

Colneric, N.; and Demsar, J. 2018. Emotion Recognition on Twitter: Comparative Study and Training a Unison Model. *IEEE Transactions on Affective Computing*, 11(3).

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

Guo, C.; Cao, J.; Zhang, X.; Shu, K.; and Yu, M. 2019. Exploiting Emotions for Fake News Detection on Social Media. *ArXiv*, abs/1903.01728.

Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; and Mihalcea, R. 2018. Automatic Detection of Fake News. In *COLING*. ACL.

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

Zhang, T.; Wang, D.; Chen, H.; Zeng, Z.; Guo, W.; Miao, C.; and Cui, L. 2020. BDANN: BERT-Based Domain Adaptation Neural Network for Multi-Modal Fake News Detection. In *IJCNN*.