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 Gupta1, Vishwakarma1, Mukesh Prasad2
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 improved alignment. (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), Politifact1, and Gossipcop2 datasets. We annotate 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)).

---

*These authors contributed equally.
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
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 (aims 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:

\[ L_{\text{FND}} = \min_{\theta_l, \theta_f} \sum_{i=1}^{n_d} L_{d_i} \]  

\[ L_{\text{emo}} = \min_{\theta_l, \theta_e} \sum_{i=1}^{n_d} L_{d_i} + \sum_{j=1}^{n_t} L_{e_j} \]  

\[ L_{\text{adv}} = \min_{\theta_f, \theta_l} (\max(\sum_{i=1}^{n_d} L_{d_i}^1 + \sum_{j=1}^{n_t} L_{e_j}^1) \]  

\[ L_{\text{Total}} = (1-\alpha-\beta) \cdot L_{\text{FND}} + \alpha \cdot (L_{\text{adv}}) + \beta \cdot (L_{\text{emo}}) \]

where \( n_d \) and \( n_t \) are number of samples in source and target sets; \( \theta_l, \theta_f, \theta_e, \) and \( \theta_t \) are parameters for discriminator, fake news classifier, emotion classifier, and LSTM feature extractor; \( L_{d_i} \) and \( L_{e_j} \) are binary cross entropy for source and target classification; \( L_{d_i} \) and \( L_{e_j} \) are cross entropy loss for emotion classification; \( L_f \) is binary cross entropy loss for Fake News Classifier; \( \alpha \) and \( \beta \) are weight parameters in \( L_{\text{Total}} \). We optimized \( \alpha \) and \( \beta \) 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 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**


