Original Article

POMS

Production and Operations Management
2025, Vol. 34(8) 2448–2470

DOI: 10.1177/10591478251319686
journals.sagepub.com/home/pao


MT-GPD: A Multimodal Deep Transfer Learning Model Enhanced by Auxiliary Mechanisms for Cross-Domain Online Fake News Detection

Dongsong Zhang[1], Guohou Shan[2], Minwoo Lee[3], Lina Zhou[1] and Zhe Fu[4]

Abstract

The proliferation of fake news, more recently multimodal fake news, poses a significant threat to individuals, organizations, and society. While online social media platforms have employed automated methods to combat fake news, they face two notable challenges: the scarcity of labeled data and the diversity of news domains. To enhance the effectiveness and efficiency of online platforms in mitigating the spread of fake news, this study proposes MT-GPD (multimodal deep transfer learning with gating network, model patch, and domain classifier) for cross-domain fake news detection. MT-GPD integrates three novel design artifacts as auxiliary mechanisms for enhancing multimodal deep transfer learning, including a gating network that captures the relative importance of textual and visual components of individual news articles for dynamic fusion; a customized model patch that balances detection performance and computational efficiency; and a domain classifier that adapts multimodal representations to a target news domain. We evaluate the performance of MT-GPD using news datasets spanning four different domains. The results demonstrate the efficacy and robustness of MT-GPD, providing strong evidence for the impacts of the proposed auxiliary mechanisms on improving fake news detection performance.

Keywords

Fake News, Multimodal News, Social Media, Deep Transfer Learning, Auxiliary Mechanisms

Date received 9 April 2024; accepted 21 January 2025 after two revisions

Handling Editor: Vijay Mookerjee

1 Introduction

People increasingly rely on online platforms (e.g., social media) for news, which also serve as a fertile ground for fake news (Ng et al., 2023; Zhou et al., 2004). We define fake news as news that is deliberately fabricated and disseminated to mislead others. Fake news can lead to serious economic, political, social, mental, and societal issues. An economic study reports that fake news costs US$78 billion to the global economy annually (Brown, 2019). Despite increasing public awareness of online fake news, its detection remains challenging. A survey showed that humans' ability to detect fake news dropped from 39% in 2016 to 26% in 2019 (Watson, 2023). People tend not to question the credibility of information unless it violates their preconceptions, largely attributed to humans' cognitive biases (Zhou et al., 2004).

The challenge in fake news detection can be potentially addressed by AI technology. Some online platforms have deployed machine learning and deep learning models for this purpose (https://www.logically.ai/factchecks). For example, Meta uses AI algorithms to fact-check user-generated content (Meta, 2020). However, online platforms also face

[1]Belk College of Business & School of Data Science, The University of North Carolina at Charlotte, Charlotte, NC, USA
[2]D'Amore-McKim School of Business, Northeastern University, Boston, MA, USA
[3]Department of Computer Science & School of Data Science, The University of North Carolina at Charlotte, Charlotte, NC, USA
[4]Department of Software and Information Systems, The University of North Carolina at Charlotte, Charlotte, NC, USA

Corresponding author:
Guohou Shan, D'Amore-McKim School of Business, Northeastern University, Boston, MA 02115, USA.
Email: g.shan@northeastern.edu

notable challenges when implementing those machine learning or deep learning methods for fake news detection. First, fake news detection in a multi-domain setting remains rarely explored (Goel et al., 2021). The vast majority of existing research ignores significant differences in the language style, structure, and terminologies of news articles across different domains. As a result, the models trained with news samples in one domain may not be effective for detecting fake news in another domain. Nor is it effective to apply general models to detect fake news in specific domains. On the other hand, given that fake news creators may employ similar strategies in crafting fake news content, there may be intrinsic and shared characteristics among fake news articles across different domains. Thus, the knowledge gained in fake news detection from one domain can potentially be helpful and leveraged for detecting fake news in another domain. Second, fake news detection is typically viewed as a classification problem, distinguishing between fake and real news. The scarcity of labeled news data, however, presents a constraint on fake news detection research. Third, news articles are increasingly created in multiple modalities, such as text and images, to enrich their content and attract readers' interest. Images in news can stir emotions of readers and foster public outcry like no other means of expression (Zillman et al. 1999). According to Inforgraphics Statistics (www.demandsage.com/infographic-statistics), posts with images have a 650% higher engagement rate. Today, more and more news articles include images. Ruhl Ibarra et al. (2024) collected 1607 news articles on tragic incidents, where 60% of those articles included at least one image. Despite emerging studies on multimodal fake news detection (e.g., Hua et al., 2023; Singhal et al., 2020), effective fusion of characteristics of news content in different modalities beyond simple concatenation remains significantly under-explored.

Transfer learning (TL) (Bozinovski and Fulgosi, 1976), which focuses on applying knowledge learned from solving one problem/task to another different yet related one, is promising to address the first two challenges mentioned above. TL has the potential to tackle the heterogeneity of different news domains and mitigate the problem of the lack of labeled training data in a target domain in fake news detection by identifying their similarities. Several recent studies have explored TL models for fake news detection (Goel et al., 2021; Ng et al., 2023; Singhal et al., 2019), which, however, have several limitations. First, they mainly fine-tuned pre-trained models (PTMs) (e.g., BERT and XLNet) with fake news. The "transferred" knowledge is essentially the general text and image representations learned by those PTMs from large open-domain datasets rather than knowledge specific to fake news detection. As a result, the generality of those PTMs is confined to the similarity between general corpora and domain-specific news articles (Orhan, 2021). To the best of our knowledge, few studies have investigated multimodal TL for cross-domain fake news detection by adapting the fake news knowledge learned from one news domain to another, which is the primary objective of this research. Second, effective TL requires facilitation. However, existing research has rarely explored ways to improve TL via auxiliary mechanisms. Third, although some TL models have integrated multimodal pre-trained models (Goel et al., 2021; Singhal et al., 2019), none has considered the relative importance of multimodal news content, such as text and image, for fake news detection. Also, whether and how a TL model can benefit from the dynamic fusion of multimodal representations of news content and auxiliary mechanisms for cross-domain fake news detection remains severely under-studied.

To address these research limitations and gaps, we propose a novel multimodal deep transfer learning model augmented by three auxiliary mechanisms, including a gating network, a model patch, and a domain classifier (MT-GPD), to improve TL effectiveness. The design of MT-GPD follows the design science research paradigm (Hevner et al., 2004). MT-GPD first learns the knowledge about fake news detection from a source news domain, then adapts it to a target news domain facilitated by the three proposed auxiliary mechanisms. Among the three mechanisms, a gating network is designed to capture the relative importance of the textual and visual content of each news article. It dynamically assigns different importance weights to the latent representations of text and images in each news article to reflect their different roles in fake news detection. The customized model patch helps reduce the computational complexity of TL without compromising model performance. The domain classifier adapts the multimodal representations learned from a source domain to a different news domain without requiring specific domain knowledge.

This study makes several contributions to emerging research in operations management (OM) and information systems (IS) interfaces, including language models, AI, deep learning, content moderation, and online platforms. First, we design an effective TL-based model for fake news detection, which empowers online platform operations with automatic approaches to detecting fake news across different domains. Specifically, our proposed model could optimize the operation process of online platforms in mitigating the spread of fake news by providing the platforms with a well-performed system to automatically detect fake news. Second, our proposed auxiliary mechanisms for facilitating transfer learning can assist online platforms in strategically allocating their resources to improve operation efficiency in automatically detecting fake news on their platforms. Third, our study contributes to the increasing research that employs machine learning and deep learning techniques in solving OM-related problems (e.g., Choi et al., 2018; Lee et al., 2018; Ng et al., 2023; Zhang et al., 2022). The proposed approach is generalizable and can be applied to the detection of other online misinformation or fraud (e.g., financial disinformation detection (Zhang et al. 2022)). It can also advance big data analytic techniques in OM when addressing uncertainty with learning (e.g., caused by limited labelled data) (Ng et al. 2023), or when deciphering financial news for stock or market prediction or discovering

product defects from social media content (Abrahams et al. 2015), etc. To the best of our knowledge, this is the first study to design and comprehensively examine several auxiliary mechanisms for enhancing fake news detection that can support platforms in managing misinformation.

2 Related Work

2.1 Automatic Fake News Detection

Machine learning techniques have been widely used for automatic fake news detection to overcome human biases. Earlier studies mainly used traditional classification techniques, such as support vector machines (Zhang et al., 2022), which rely on manual and ad hoc feature selection. The common input features of those models are classified into four categories: (a) t*extual features* capturing the characteristics of news textual content; (b) *visual features* comprising count-based and content-based features of news images; (c) *network features* representing the propagation patterns of news and URLs; and (d) s*ender features* characterizing the behavior and demographics of news creators. Among these four types of features, textual features have been most commonly used, while sender and network features have been used much less because the latter are more difficult to acquire due to privacy concerns.

In recent years, deep learning techniques have been increasingly deployed to detect fake news. For example, Wang et al. (2018) designed an event-based adversarial neural network for fake news detection by deploying CNN models. Some more recent studies fine-tuned pre-trained transformer-based models, such as BERT (Essa et al., 2023) and XLNet (Athithan et al., 2023), with news samples. However, unlike this study, their focus is not on cross-domain fake news detection.

2.2 Multimodal Models for Fake News Detection

There are emerging efforts to develop multimodal models that integrate text and visual representations of news for fake news detection (e.g., Hua et al., 2023). For example, Wang et al. (2018) improved Jin et al. (2017)'s approach by deploying a CNN as the core module of a textual feature extractor and a pre-trained VGG-19 model for extracting visual features from a news article, then concatenating the generated representations to form the multimodal feature representation as the input of a fake news detector. Giachanou et al. (2020) first used BERT and VGG-16 to create representations of news's textual and visual content, respectively, then concatenated them into a similarity vector as the input to a softmax layer.

There are two major issues with existing multimodal models for fake news detection. First, they concatenate textual and image feature representations directly by overlooking the varying importance of text and image content of individual news articles to fake news detection. Second, despite the initial efforts in applying them to fake news detection and their success in many other applications, PTMs are prone to adopting shallow heuristics that succeed for the majority of training samples, instead of learning the underlying generalizations that they intend to capture (Wang et al., 2018). Therefore, they may suffer from potential "upstream overfitting" caused by intra-class semantic differences (Feng et al., 2022).

2.3 Deep Transfer Learning for Fake News Detection

A common assumption of machine learning models is that training and testing data are drawn from the same feature space and share the same distributions. When the distribution changes, models have to be rebuilt from scratch (Yu et al., 2020). This is problematic due to the varying feature distributions across news domains and the scarcity of labeled news data in specific domains.

Transferring knowledge learned from one domain to another extends a model beyond its original creation (Yang et al., 2020). By reusing the domain knowledge previously gained from a source task, TL can significantly reduce the required data, time, and computing resources for completing a similar task in another domain. Deep TL leverages deep neural networks to find invariants for the successful adaptation of knowledge gained from a source domain to a target domain.

A small number of studies (e.g., Goel et al., 2021) have explored TL for fake news detection. They all deployed PTMs, such as BERT and XLNet, to learn generic text and image representations from large-scale open datasets first, then fine-tuned those PTMs with fake news datasets (normally the last layer only). For example, Shu et al. (2022) proposed a BERT-based TL model that fused meta information associated with news (i.e., comments and user-news interactions) to enhance news representation. However, that model focused on news text only. Similarly, Ng et al. (2023) focused on linguistic features of news textual content only. A few recent studies (e.g., Singhal et al., 2020) explored multimodal transfer learning. For example, SpotFake+ used XLNet to derive a textual feature vector and VGG-19 to get a visual feature vector, then concatenated them into a 200-dimension multimodal feature vector (Singhal et al., 2020). However, the transferred knowledge learned by those PTMs does not pertain to fake news detection, which could pose challenges when PTMs are used in the target news domain.

It is crucial to address the fundamental challenges in fake news detection, including those arising from the scarcity of labeled data and domain diversity, particularly multimodal data. Despite increasing efforts in building TL models for fake news detection, we have identified several limitations. First, existing TL models mainly fine-tune general PTMs for fake news detection rather than adapt the knowledge about fake news detection learned from one news domain to another (e.g., Shu et al., 2022; Singhal et al., 2020; Singhal et al., 2021; Wu et al., 2021). The knowledge transferred in those models comprises general representations of text or images obtained from large-scale open-domain datasets rather than domain-specific knowledge related to fake news detection. The representations generalized by PTMs may fall short due to

potential upstream overfitting (Feng et al., 2022; Wang et al., 2018). Second, a learner can react selectively to the salient elements of a problem while disregarding irrelevant or unimportant elements (Rumbaugh et al., 2008). However, there is a dearth of attention to the design of facilitative mechanisms to improve the effectiveness of TL in the current state of research (e.g., Liu et al., 2023). For example, Liu et al. (2023) shared cross-domain knowledge by learning transferable knowledge features, but their model aligned inter-domain knowledge features without assessing the relative importance of text and image modalities of news articles to misinformation detection. Ignoring the importance of features in knowledge transfer in existing work can introduce significant noises that degrade model performance (Wu et al., 2020).

3 Method

3.1 Design Rationales

To address the limitations and gaps of the literature outlined in the previous section, we design the proposed model with several considerations. First, we consider transfer learning as a process of transferring knowledge about fake news learned from one news domain to another. We predict that incorporating the knowledge learned from a source news domain should improve the initial state of a detection model trained for a target news domain, which in turn leads to better model performance. Second, text and images within news articles may offer complementary functions or properties of news. Thus, the proposed model should be a multimodal TL model that leverages both textual and visual content of news articles for fake news detection. Third, the textual and image content of individual news articles may carry different levels of importance to fake news detection. More important content in a specific modality of a news article should carry more weight during transfer learning by making it more salient and contributive to the detection task. Fourth, the model should promote resource efficiency to facilitate cross-domain transfer without sacrificing detection performance. Fifth, while news articles across different domains (e.g., entertainment, politics, and health) may share common elements, they also exhibit some distinct differences. Thus, the multimodal TL model should align latent representations of different domains to minimize the representation gap between those domains, as illustrated in Figure 1.

By following the design rationales, our proposed MT-GPD model integrates adaptive and dynamic fusion of multimodal representations, customized TL, and three novel auxiliary mechanisms, which include a gating network, a model patch, and a domain classifier. This strategic amalgamation addresses the challenges of fake news detection outlined in the introduction section. It offers a comprehensive TL-based solution that improves both model adaptability and efficiency while enhancing its effectiveness across various domains. Figure 2 shows the overall architecture of MT-GPD.

3.2 The Baseline Deep Transfer Learning Models

We first developed baseline deep TL models for fake news detection. For text-only TL, we trained CNN models with the textual content of news from a source news domain and stored the learned weights. Then, we built a text-only CNN model for a target domain by loading the previously trained source-domain model (i.e., transferring weights), adding a dense layer to the network, and fine-tuning the dense layer while preserving other learned weights. We followed the same procedure to construct an image-only CNN model. The baseline multimodal TL model concatenates the embeddings learned by the text-only and image-only CNN models before feeding them into the dense layer. Because MT-GPD is intended for cross-domain fake news detection that involves source and target news domains, it employs CNN and LSTM instead of PTMs. Based on the result, the CNN models consistently outperformed the LSTM models in detection precision, recall, F1 score, and accuracy ($p < 0.001$). Thus, we chose CNNs over LSTMs to build MT-GPD. Because existing studies on TL for fake news detection mainly use PT, we also implemented PTM-based models as another set of baseline models, including BERT for news text and VGG-19 for news images, and concatenated them by following previous studies (Goel et al., 2021; Ng et al., 2023).

3.3 Auxiliary Mechanism #1: A Gating Network

Different news domains may have different characteristics to attract readers' attention (Dai et al., 2018). The importance of text and images to fake news detection may vary among individual news articles and news domains. In addition, fabricators may attempt to build a sense of credibility by mixing fake and real content (Zhou et al., 2004) by not fabricating news text and images to the same extent. For example, a fake news creator may include a fake image, or keep the image intact but manipulate the textual content, of a news article. Furthermore, news fabrication strategies may depend on the cost and difficulty of fabrication. For example, it might be easier and more persuasive to manipulate images by simply flipping them than manipulating text in some news, or vice versa. Therefore, we predict that individual modalities of a news article may bear different levels of importance to fake news detection, and that those importance levels of individual modalities may vary from one news article to another. It would be desirable for a multimodal fake news detection model to take potentially varying roles of individual modalities in a news article into consideration by assigning different importance weights for different modalities dynamically.

By following the above rationale, we propose a gating network (GN) design that unveils the importance of multimodal news content. This design empowers MT-GPD with the ability to identify certain parts of high-dimensional inputs that carry important information for a target outcome, and then assigns larger (or smaller) weights to the more (or less) important parts through dynamic adaptation and exploitation of input-specific

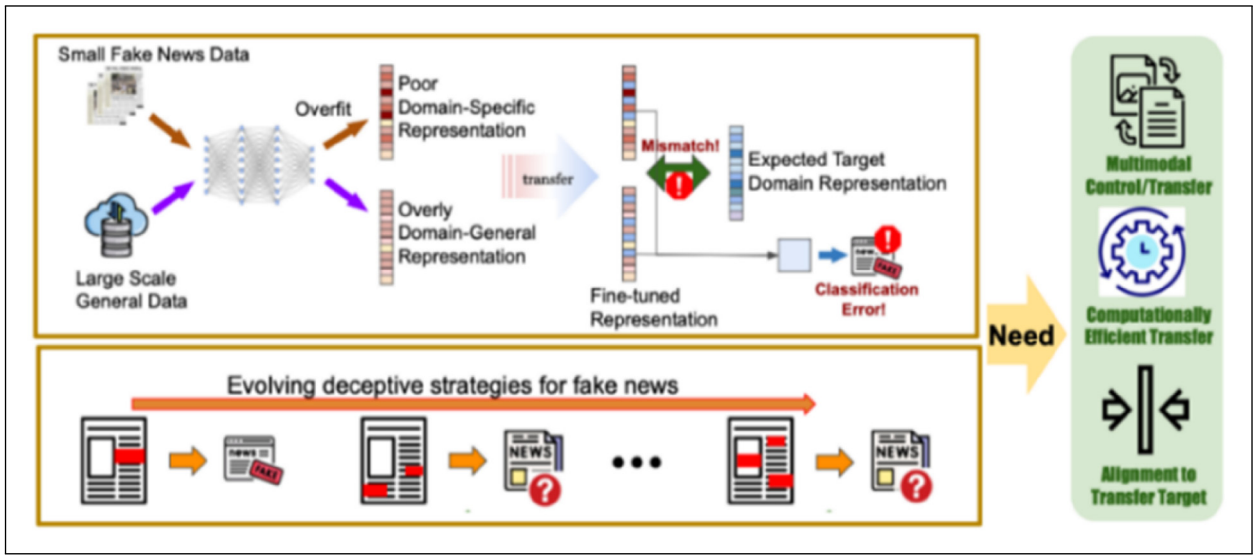

Figure 1. Three key needs identified from the research gaps of the existing PTM-based TL models.

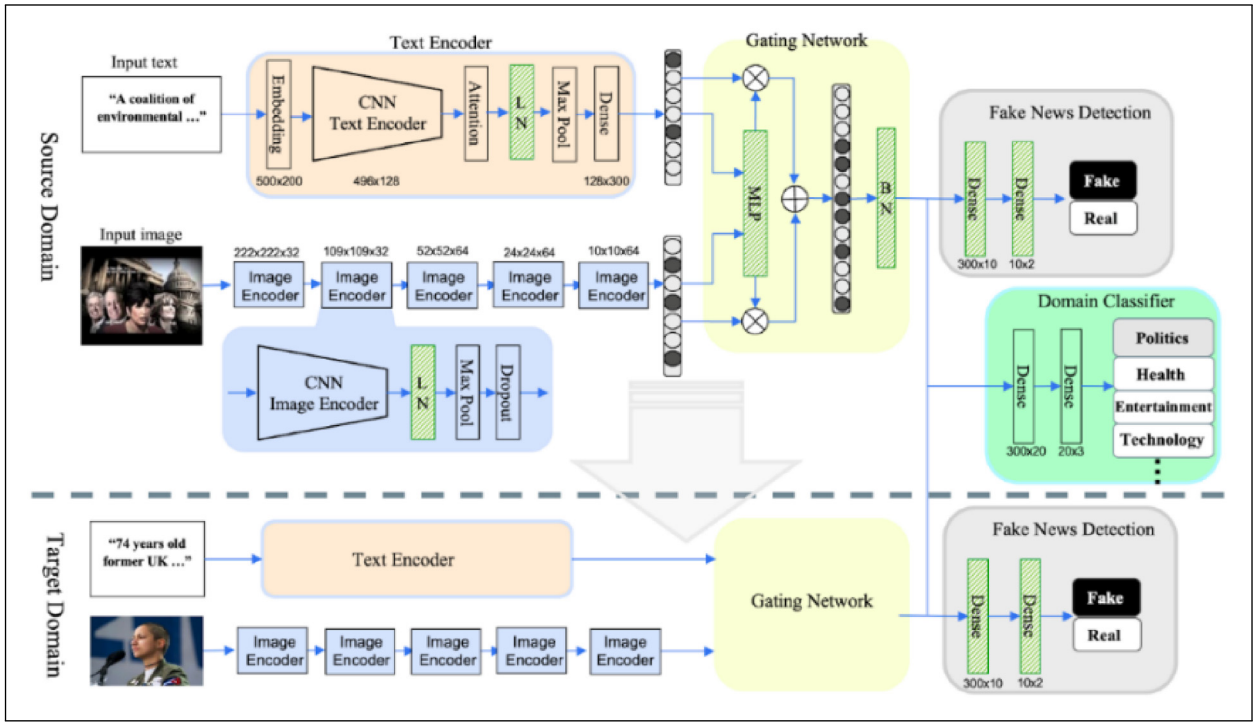

Figure 2. The overall architecture of MT-GPD.

characteristics. As a result, the less relevant or less important features of news content will have less impacts on the fake news detection decision. Gating enables MT-GPD to ensemble multiple neural networks or latent features and contributes to the improvement of model efficiency by developing a multimodal scoring rule for each news article.

Figure 3 illustrates how the GN discerns the varying significance of text and image modalities of each news article, obtaining both domain-specific and domain-independent scoring rules. The proposed GN captures the ratios of text-image cues (represented by blue bars for text and red bars for images in Figure 3) of each news article for fake news detection. The text and image cues are the information carried by news text and image used for detecting fake news, represented by their latent representations, and cue ratios represent the relative importance of news text and image for detecting fake news.

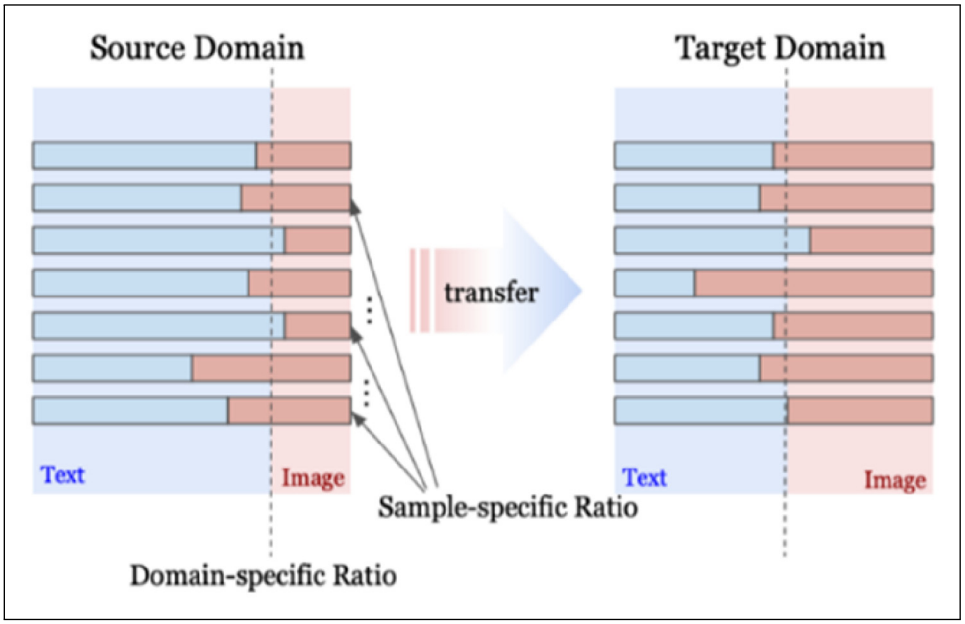

Figure 3. The gating network auxiliary mechanism.

Once the GN learns the cue ratios from the source domain, through TL, the GN imparts overarching domain-specific multimodality control rules, which is the process of multiplying the text and image latent representations with their relative dynamic gating scores (as shown in Equations (1) and (2)), to adapt the relative importance of text and image modalities of each news article while sharing the domain-independent rules. By transferring the GN, MT-GPD can prioritize the embeddings of text or images for effective multimodal representation based on the inherent characteristics of the target domain, which may lead to a more nuanced and accurate detection of fake news. As illustrated in Figure 2, the designed multimodal control for the GN plays a pivotal role in capturing the relative importance of text and image within individual articles.

Although both the GN and attention mechanism can be used to manage information flow and weighting in neural networks, they have distinct differences. Attention emphasizes the specific dimensions of a latent representation that may contribute more, while the GN focuses on effectively controlling different modalities of a news article to produce an overall optimal latent representation. Our exploratory investigation reveals that the GN demonstrates higher efficacy in facilitating transfer, whereas attention faces challenges in transferring shareable knowledge across different news domains. Given the dual modalities and their latent representations, we use a shallow network with the sigmoid function to calculate the gating scores (i.e., importance ratio) (See eq. (1) and (2)):

$$g_{\text{text}} = \frac{\sigma(w_{\text{text}} * h_{\text{text}} + b_{\text{text}})}{\sigma(w_{\text{text}} * h_{\text{text}} + b_{\text{text}}) + \sigma(w_{\text{image}} * h_{\text{image}} + b_{\text{image}})} \quad (1)$$

$$g_{\text{image}} = \frac{\sigma(w_{\text{image}} * h_{\text{image}} + b_{\text{image}})}{\sigma(w_{\text{text}} * h_{\text{text}} + b_{\text{text}}) + \sigma(w_{\text{image}} * h_{\text{image}} + b_{\text{image}})} \quad (2)$$

where h_{text} and h_{image} denote the latent text and image representations, respectively; w_{text}, w_{image}, b_{text}, and b_{image} denote the learned parameters (i.e., weights and biases) of the gating network; $\sigma(\cdot)$ is the Sigmoid function; and g_{text} and g_{image} are the gating scores of the text and image subnetworks, respectively. We chose sigmoid as the gating function, as it facilitates the calculation of gating scores with a bounded range, providing interpretability (see online Appendix A in the E-companion). The gating network calculates gating scores for multimodal content dynamically and multiplies them with the corresponding latent representations (e.g., text and image embeddings) of each news article, which enables MT-GPD to make dynamic adaptation and effective fusion of multimodal representations. In addition, we insert a batch normalization (BN) layer after the concatenation to improve model stability and efficiency. We save the model parameters (e.g., w_{text}, w_{image}, b_{text}, and b_{image}) after training the gating network in the source domain, then transfer the gating network structure with the saved model parameters to a target domain, where those parameters get updated with the target-domain training data.

3.4 Auxiliary Mechanism #2: Customized Model Patch

MT-GPD deploys a customized model patch for parameter-efficient fine-tuning of an overparameterized multimodal model for cross-domain transfer. It adapts (i.e., fine-tunes) a small set of pre-trained parameters dispersed throughout a

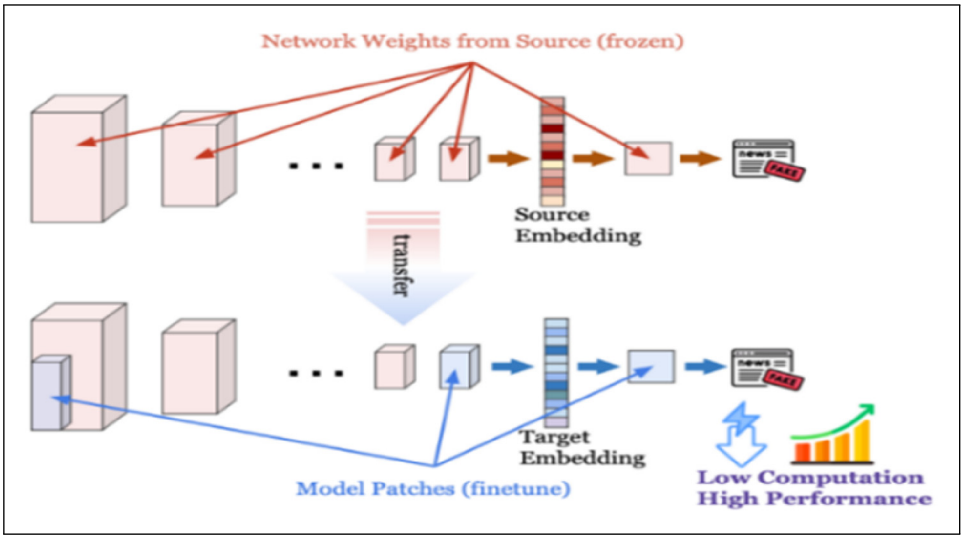

Figure 4. The model patch auxiliary mechanism.

network, rather than all the parameters in the network, aiming to not only reduce computational overhead (Mudrakarta et al., 2019), but also enhance asymptotic performance in the target news domain. There are three major motivations for our model patch design, including improving the computational efficiency of fine-tuning, avoiding the loss of the pre-learned knowledge, and reducing the redundancy in the network. First, the computational cost of neural network training or fine-tuning is proportional to the number of weights that need to be adjusted (Han et al., 2015). Therefore, fine-tuning an entire pre-trained model after knowledge transfer for the target task can be memory and computationally intensive. Second, full fine-tuning, which updates all parameters in a deep learning model, may cause catastrophic forgetting of the pre-learned/trained knowledge (Kemker et al., 2018). In comparison, fine-tuning a small subset of parameters can better preserve the valuable pre-trained parameters (i.e., knowledge). Third, overparameterized models often have shared knowledge among parameters, with as little as 5% of model parameters potentially enough to predict the majority of other parameters within a neural network while not sacrificing the model performance (Denil et al., 2013). Thus, it is important to strategically distill crucial information from the parameters that require tuning for the target domain.

Our model patch design shares a philosophy with emerging parameter-efficient fine-tuning (PEFT) techniques, which aim to minimize the parameters for fine-tuning while freezing the majority of a large network. For example, adapter-tuning (Pfeiffer et al., 2020) introduces a small number of additional trainable parameters for the target tuning while keeping the source network frozen. Cai et al. (2020) suggest an efficient parameter adjustment, focusing specifically on fine-tuning the bias terms at each layer. Hu et al. (2022) present LoRA, which enables PEFT by re-parameterizing some of the weight matrices in a large network to introduce low-rank updates. While these PEFT techniques hold promise for efficient transfer, there is no established de facto method, let alone for fake news detection. Importantly, the existing PEFT methods overlook the effective transfer of multimodal representations that focus on improving computation efficiency while achieving satisfactory performance. According to Han et al.'s (2024) taxonomy, PEFT can be categorized into additive fine-tuning for adapters and prefixes, re-parameterized fine-tuning for LoRA and its variants, and selective fine-tuning for pruning and FishMask. Model patch falls under selective fine-tuning. A critical motivation of our MP design in MT-GPD is to reduce computational complexity through the reduction of parameters that need to be fine-tuned without sacrificing or even potentially enhancing model performance. We believe that normalizing the representations of individual modalities and patching the fused representation can maximize the sharable knowledge and facilitate the process of multimodal cross-domain transfer. As a result, we design a customized MP, as shown in Figure 4.

Our customized MP fine-tunes only the normalization layers, gating network, and the two dense layers for several reasons. First, MT-GPD preserves the key knowledge for fake news detection learned from the source domain within the frozen layers. Second, those tunable MP layers, such as BN and dense layers, contribute to the generalizability of our model patch design for TL by transforming domain-specific knowledge for a target domain. Third, those tunable layers carry a small portion of the parameters of the entire model, thereby enhancing the computational efficiency of fine-tuning.

Our tailored MP design freezes the majority of perception layers of image- and text-only models and focuses on tuning post-transfer multimodal representations to improve the quality of domain-specific multimodal representations. To avoid *ad hoc* MP design, which often requires an exhaustive brute force search, we simplify the patch (i.e., LN + Gating + BN + Dense). It strategically incorporates

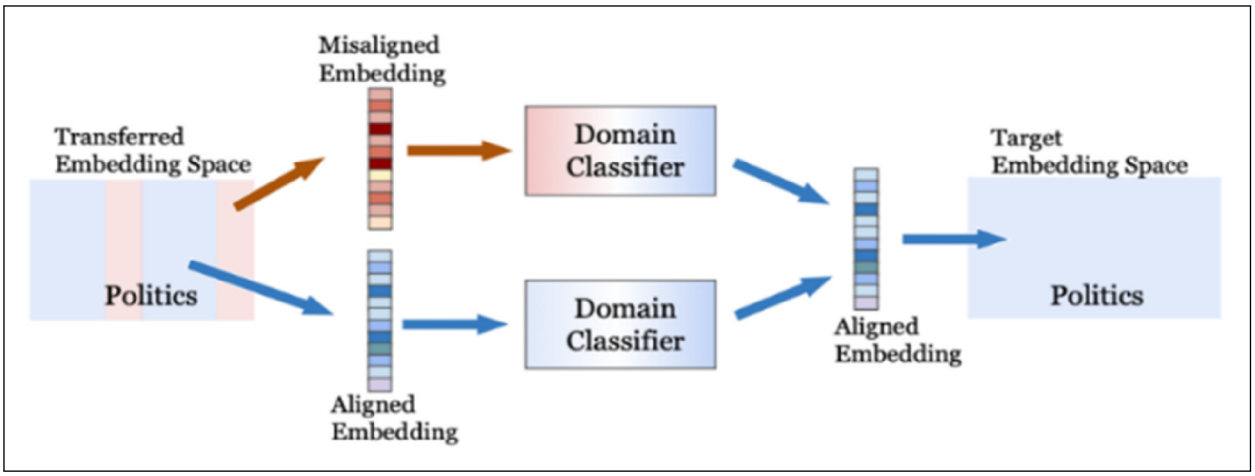

Figure 5. The domain classifier for aligning the latent representation with the target domain.

layer normalization (LN) (Ba et al., 2016) and BN (Ioffe and Szegedy, 2015) for effective and efficient model adaptation via scale-and-bias patch. To capture the basic statistics of the content in each modality of individual news articles, we introduce LN layers within text and image network layers. Accordingly, LN separates domain-specific and domain-independent knowledge for each modality of a news article. We also inject a BN layer for the gating features to learn the statistics of each dimension of the multimodal representation, which facilitates the sharing and fine-tuning of multimodal representation within the latent space. The condensed multimodal fusion rules embedded in the single-layer gating network are required to swiftly adapt to domain-specific sample distributions through TL and fine-tuning. The final layers that transform multimodal representations into a fake news detection decision are fine-tuned to enhance the accuracy of detection in a target news domain. In the end, MT-GPD fine-tunes only 48,033 parameters while freezing other 2,127,312 parameters.

3.5 Auxiliary Mechanism #3: A News Domain Classifier

News articles are in different domains, each exhibiting some unique characteristics and properties. Reinemann et al. (2012) categorize news into hard news, soft news, and general news. Hard news (e.g., political news) is characterized by immediate reporting and short lifespan; soft news (e.g., entertainment news) can be reported at any time with little or no intrinsic social importance; and general news (e.g., health news) is important but not necessarily urgent for reporting.

The content of news and the ways of fabricating news are expected to vary considerably across different domains. To cope with such heterogeneities, in MT-GPD, we propose a news domain classifier (DC) (Figure 5) as the third auxiliary mechanism to facilitate multimodal TL for fake news detection by automatically adjusting latent news representations, enabling the detection of diverse fabrication patterns.

The pre-trained DC enables MT-GPD to align multimodal representations with a target news domain. The proposed DC distinguishes itself as a multi-class classifier that predicts the specific target news domain to which a news article belongs. It is specifically designed to fortify the precise alignment of latent multimodal representations of a news article after the transfer of knowledge by facilitating the backpropagation process to effectively shift the multimodal representations from a source to a target domain without fine-tuning. When a transferred network generates a misaligned latent representation of a news article in the target domain, it risks misclassifying fake news due to differing fabrication patterns. Backpropagation through the DC enhances the alignment process, thereby improving fake news detection. Note that this DC mechanism does not require explicit domain information. Instead, it adapts the learned news characteristics of each news domain without requiring domain-specific knowledge. To minimize possible negative transfer, we define a combined loss function, as shown in Eq (3):

$$L = \frac{1}{N}\sum_{n=1}^{N}(\alpha \times l1(\theta; y1^{(n)}, o1^{(n)}) + (1-\alpha) \times l2(\theta; y2^{(n)}, o2^{(n)})) \quad (3)$$

where α is a convex weight that indicates the relative importance of latent feature representation of each modality to correctly detect a news article as fake or real; θ denotes all the parameters; $l1$ and $l2$ are loss functions; $y\,1^{(n)}$ and $o1^{(n)}$ are the actual label and predicted output for fake news detection in the target news domain, respectively; and $y2^{(n)}$ and $o2^{(n)}$ are the actual and predicted target news domain labels for news domain classification, respectively. The back-propagation process, aimed at minimizing the loss function (see Eq (3)), yields

gradients to update the latent representations generated by the upper hidden layers of the neural network. Such gradient updates, guided by the second term in Eq. (3) involving the domain classifier, ensure an improved representation alignment with the target domain. In instances where the generated latent representations are already well aligned, the loss from the second term becomes zero, resulting in no further updates to the latent representation. Consequently, the well-aligned latent representations minimize the need for fine-tuning the bottom classification layers, thereby facilitating transfer learning. We built a multi-class CNN-based classifier by using a diverse set of authentic news articles to detect the specific domain of individual news articles automatically by using the multimodal representations of news text and image. The architecture of the news DC mirrors that of our baseline text- and image-only CNNs. We used *Adam* as the optimizer and the sparse categorical cross entropy loss function and set the number of training epochs as 10 with a batch size of 32.

The proposed domain classifier introduces an innovative role in facilitating TL, ushering in the paradigm of self-supervised learning. By leveraging unlabeled data (i.e., without requiring fake news labels) for training, the self-supervised approach ensures effective learning of representations for downstream tasks. In our design, MT-GPD leverages the domain classifier to improve the domain-specific representations of the news text and image. Auxiliary learning delves into the exploration of diverse auxiliary tasks, aiming to leverage these tasks for learning latent representations that are not only useful but also highly effective. This enhanced effectiveness contributes to MT-GPD's enhanced generalization capabilities and facilitates its rapid convergence, which has never been explored in the literature on multimodal transfer learning for fake news detection. Although the term domain classifier has been used in a few other studies, it was designed for different purposes in those studies. For instance, Ng et al. (2023) deployed a binary classifier to determine which one of the two subsequent processing should be performed. Shu et al. (2022) used a classifier to force BERT to generate a domain-independent representation. Moreover, both studies proposed text-only based models instead of multimodal models. Although the proposed domain classifier primarily addresses the adaptation of latent embeddings to different news domains, it can be pre-trained to classify news into subcategories or other categorizations beyond general news domains, such as articles favoring different political parties, depending on specific requirements and availability of data or domain expertise.

3.6 A Summary of Design Novelties of MT-GPD

Unlike existing literature (See online Appendix B in the E-companion), MT-GPD incorporates several novel methodological designs. First, in contrast to previous multimodal fake news detection methods that directly concatenate text and image representations, MT-GPD incorporates a gating network that dynamically calculates the gating scores of news text and images reflecting their relative importance, then uses those scores to adjust the weights of text and image representations, leading to customized adaptation and fusion of multimodal features. Second, MT-GPD provides a solution and guidance on what to transfer and what to fine-tune when transferring a classification model for fake news detection. The model patch design is aimed at minimizing the number of parameters to be fine-tuned while striving for optimal model performance by fine-tuning only the normalization layers and maximizing the shared weights. The design artifact facilitates MT-GPD in establishing domain-independent knowledge while also developing domain-specific knowledge. Third, MT-GPD deploys self-supervised auxiliary learning via a domain classifier that leverages the inherent and characteristic differences among news articles across different domains to achieve aligned knowledge representation. Last, the novel multimodal TL approach used by MT-GPD addresses the limitations of current PTM-based TL, such as the transfer of domain-independent generic representations and *ad hoc* practice of fine-tuning for downstream tasks.

4 Evaluation

4.1 Datasets

In this study, we used news datasets collected from four diverse domains to evaluate MT-GPD. Among them, the political and entertainment news datasets were drawn from the FakeNewsNet (Shu et al., 2020); the health dataset from the CoAID and ReCovery datasets (Feng et al., 2022), and the technology-related news dataset from Desai et al. (2022). We performed data cleaning tasks, including removing logos from news articles and filtering news samples lacking images. After data cleaning, the political news dataset consists of 320 real and 164 fake news articles (1582 words/article on average); the entertainment news dataset consists of 2620 real and 2581 fake news articles (697 words/article on average); the health dataset consists of 927 real and 309 fake news articles (761 words/article on average); and the technology dataset consists of 305 real and 823 fake technology-related news articles (1131 words/article on average).

We select those datasets for several reasons. First, they come from diverse domains, which serve the objective of this study well. Second, they vary in sample sizes, average news lengths, and real-fake news ratios, allowing us to evaluate the robustness of MT-GPD. Third, they are multimodal news datasets, in which each news article includes both textual and image content. Fourth, previous studies have used those datasets, as mentioned above. The sizes of our datasets are comparable to, or larger than, those used in many prior studies, such as Giachanou et al. (2020) (2745 fake, 2714 real), Pham et al. (2021) (819 fake, 4018 real), Cruz et al. (2020) (1603 fake, 1603 real), Goel et al. (2021) (240 fake, 240 real), Silva et al., (2021) (1673 fake, 4264 real), Meel and Vishwakarma

(2021) (2121 fake, 1867 real), and Hua et al. (2023) (562 fake, 1297 real). Because the general notion of TL is to transfer knowledge learned from one domain/task with larger data to a target domain/task with less data (Yu et al., 2020), we select the entertainment dataset as the source-domain data, and other three datasets as separate target-domain datasets.

4.2 Modeling Tasks and Settings

We design and perform four incremental tasks to assess the effectiveness of MT-GPD and its auxiliary mechanism designs. Task 1 focuses on building a news DC (domain classifier). We randomly split each dataset into training (80%) and testing (20%) data. We also measure the cross-domain news similarity. Task 2 compares the performance of the baseline single-modal and multimodal TL models without any of the proposed auxiliary mechanisms with those of the baseline models without TL in cross-domain fake news detection. Task 3 investigates the impact of each auxiliary mechanism, namely the gating network (GN), customized model patch (MP), and DC, on MT-GPD performance via ablation experiments. Task 4 compares the performance of MT-GPD against fourteen baseline models, including EANN (Wang et al., 2018), IMD (Singhal et al., 2021), SpotFake+ (Singhal et al., 2020), EMFEND (Qi et al., 2021), MVAE (Khattar et al., 2019), MCAN (Wu et al., 2021), MMFND (Giachanou et al., 2020), both fine-tuned and not fine-tuned LLaVA (Liu et al., 2024) and LLaMA-3 (Touvron et al., 2023), attRNN (Jin et al., 2017), FT2, and FT2-GD (Appendix H in the E-companion). All but attRNN and non-fine-tuned LLaVA and LLaMA-3 are PTM-based multimodal TL models. FT2 and FT2-GD, the two MT-GPD variational models, are double fine-tuned PTM baseline models, while other PTM-based models are fine-tuned only once with target-domain news.

We downloaded the code of all the baseline models that are publicly available on GitHub, except FT2, FT2-GD, LLaVA, and LLaMA-3, and then fine-tuned and tested them using our datasets directly without making any changes. The baseline models were optimized by the Adam optimizer. For the sake of modeling complexity and paper length, we only employed the technology news dataset in tasks 1 and 4.

4.3 Evaluation Metrics

We use precision (P), recall (R), F1-measure (F1), and accuracy (A) as metrics to evaluate model performance. Precision measures the proportion of detected fake news articles actually being fake. Recall measures the proportion of actual fake news detected correctly. F1 is a harmonic mean of precision and recall (i.e., $2 * \text{Precision} * \text{Recall}/(\text{Precision} + \text{Recall})$). Accuracy is measured as the percentage of news articles being detected correctly. Due to the imbalanced distributions of real and fake news articles in our datasets, we use macro precision, macro recall, and macro-F1 (Mortaz, 2020) because they are more suitable for imbalanced data.

We deploy Monte Carlo cross-validation (Xu and Liang, 2001), also known as repeated random subsampling cross-validation, for the target domains. The training/testing process was repeated 30 times, generating 30 different training and testing data partitions for each target domain. The reported model performance is the average model performance of 30 runs.

5 Results

5.1 Task 1: Constructing the Domain Classifier and Measuring News Similarity Across Domains

The evaluation results of the proposed pre-trained auxiliary DC show an overall accuracy of 94.7%. The precision, recall, and F1 scores are 92.9%, 95.4%, and 94.1% for the politics domain, 91.7%, 93.9%, and 92.8% for the entertainment domain, 92.5%, 93.9%, and 93.2% for the health domain, and 93.6%, 92.7%, and 93.1% for the technology domain, respectively.

The similarity between the source (entertainment) and each target domain (i.e., politics, health, and technology) may influence TL performance. The lower the similarity, the more challenging the TL. We computed radial basis function-centered kernel alignment (CKA) similarity scores based on the text and image embeddings separately (see online Appendix C in the E-companion). We make several major observations from Table 1. First, there are varying degrees of similarity among those domains. Second, the levels of similarity among fake news articles across different domains are higher than those among real news articles. One possible reason is that the creators of fake news may deploy similar fabrication strategies across different news domains, which is less likely for real news. Third, the textual content exhibits much higher levels of similarity compared to the image content across various news domains.

5.2 Task 2: Examining the Impact of Transfer Learning on Detection Performance

5.2.1 Comparison of Baseline Models with TL vs. Without TL. We built three baseline deep TL models using news text only, news image only, and multimodal (i.e., both text and image) news representation, respectively. For all three TL models, the entertainment news articles were used as the source domain data, and the politics and health news were used as two separate target domain datasets. To explore the generalizability of fake news detection models across news domains and the benefit of TL, we also developed 1) three corresponding models without TL as baselines that used the data from the same target domains for both training and testing (referred to as *w/o TL* models); and 2) another three corresponding models without TL that were trained with the source-domain training data and then directly tested with the test data of target-domain data (referred to as *w/o_d* models).

Table 1. CKA similarity scores between the source (entertainment) and target domains.

Target domains	Modality	Fake news	Real news	All news	Fake vs. real news
Politics	Text	0.563	0.425	0.324	0.557
	Image	0.236	0.133	0.089	0.202
Health	Text	0.369	0.159	0.125	0.371
	Image	0.139	0.050	0.041	0.066
Technology	Text	0.281	0.178	0.113	0.403
	Image	0.019	0.050	0.011	0.035

The baseline models with and without TL were all constructed with CNN. The text-only baseline models without TL (i.e., w/o TL and w/o_d models) were composed of embedding (Emb), convolutional (Conv), LN, max pooling (MaxP), and dense (Den) layers (see details in the online Appendix D in the E-companion). The optimal parameters of different layers were empirically learned via a grid search. Specifically, the *Emb* layer took a sequence of n words from news text starting from the beginning as the input. We experimented with a number of different values of n in the range of 100 to 1500 words, with an increment of 50. The models achieved the optimal performance when the text length was 500 words, which we selected as the maximum length. If the length of a news article exceeds this maximum length, the news would be truncated - only the first 500 words would be considered. If the length of a news article is fewer than 500 words, then padding would be used. The optimal word embedding dimension was empirically set to 200 via a grid search in the range of [50, 500], with an increment of 50. The *Conv* layer iteratively takes a consecutive sequence of h words (window size) in the sentence S as the input and generates a feature vector as the output. The optimal filter and window size were 296 and 5, respectively; the *LN* layer follows the Conv layer, aiming to accelerate network training by reducing internal covariate shift and the dependence of gradients on the scale of parameters; the *MaxP* layer lowers the computational burden by reducing the number of connections between convolutional layers; and the *Den* layer deploys the ReLU activation function. The optimal number of nodes in the Den layer was empirically determined as 300. The final dense layer consisted of a single node that indicates whether a news article was fake or not. The models performed the best when using the batch size of 24 and running 30 training epochs.

The architecture of the image-only baseline CNN models without TL contained five *Conv* layers, consisting of 32, 32, 32, 64, and 64 nodes, respectively (see online Appendix E in the E-companion). Each *Conv* layer was followed by an *LN* layer, a dropout layer, and a *MaxP* layer. To avoid overfitting, we introduced a dropout layer that regularized the network by randomly deactivating a set of hidden units of a network layer and then training different subnetworks iteratively while sharing weights. The final two dense layers were identical to those used in the text-based CNN baseline model. The dimension size of image embedding was 1,600, and network parameters were optimized by minimizing the loss function.

Table 2 presents the performances of the baseline models. P, R, F1, and A in Table 2 represent macro precision, macro recall, macro F1-measure, and accuracy, respectively. We first compared *w/o TL* and *w/o_d* models without transfer learning across different modalities. The results show that *w/o TL* models outperform *w/o_d* models across all four performance measures on both politics and health datasets ($p < .001$). Therefore, we only use the *w/o TL* models for subsequent analyses. To examine the effects of TL and modality on fake news detection, we performed repeated measures ANOVA. The results reported in Table 3 yield significant main effects of TL ($p < .001$ or .05), modality ($p < .001$), and their interaction ($p < .001$ or .05) across all measures for both the politics and health datasets, except for the accuracy of political news ($p > .05$) and the precision of health news ($p > .05$).

Given the significant interaction of TL and modality, we performed post hoc contrast analyses on the effect of TL for each modality separately. The results show that TL improves all four performance measures of the text-only model ($p < .001$) for both target datasets. Similarly, TL improved all the measures of the multimodal models with both politics and health datasets (i.e., target domains). However, TL has a negative influence on all the performance measures of the image-only model ($p < .001$) except the accuracy of health datasets ($p > .05$). These results reveal different effects of TL for different modalities. First, it is important to examine the effect of multimodal models relative to unimodal counterparts to gain a deeper understanding of the effect of TL. Second, there is a need to consider the different roles of textual and image news content in fake news detection.

5.2.2 Effects of Modality. Given the significant main effect of modality, we performed post hoc multiple comparisons of modality with Bonferroni adjustments on the politics and health datasets. The results are reported in Table 4.

Results reveal that the baseline multimodal model consistently outperforms the unimodal counterparts (i.e., text-only and image-only models) in all four performance measures across both datasets ($p < .001$). In addition, the text-only model outperforms the image-only model ($p < .001$). Because the baseline multimodal TL model performs the best among all baseline models with and without TL, we only focus on the multimodal TL model for the remaining modeling and analyses.

Table 2. Descriptive statistics for models without and with baseline transfer learning.

		Politics				Health			
Models	Modality	P	R	F1	A	P	R	F1	A
w/o_d	Text-only	0.530	0.506	0.516	0.459	0.656	0.505	0.549	0.595
	Image-only	0.491	0.490	0.475	0.582	0.479	0.478	0.475	0.587
	Multimodal	0.498	0.468	0.478	0.457	0.574	0.499	0.527	0.586
w/o_TL	Text-only	0.816	0.705	0.757	0.774	0.840	0.718	0.774	0.785
	Image-only	0.603	0.585	0.584	0.654	0.622	0.623	0.621	0.711
	Multimodal	0.836	0.735	0.782	0.797	0.853	0.735	0.790	0.806
with_TL	Text-only	0.835	0.723	0.775	0.802	0.861	0.737	0.794	0.814
	Image-only	0.479	0.488	0.450	0.608	0.579	0.549	0.548	0.716
	Multimodal	**0.859**	**0.755**	**0.804**	**0.826**	**0.878**	**0.755**	**0.812**	**0.834**

Note: The bold indicates the best performance.

Table 3. Effects of transfer learning and modality.

	P		R		F1		A	
Variables	*F*-Value	*p*-Value	*F*-Value	*p*-Value	*F*-Value	*p*-Value	*F*-Value	*p*-Value
(a) Politics								
TL	29.7	<0.001	1622.1	<0.001	81.0	<0.001	1.321	>0.05
Modality	1383.5	<0.001	43.6	<0.001	2407.1	<0.001	956.7	<0.001
TL* modality	105.6	<0.001	217.3	<0.001	265.2	<0.001	70.9	<0.001
(b) Health								
TL	0.027	>0.05	36.1	<0.001	25.0	<0.001	86.5	<0.001
Modality	2675.5	<0.001	1801.6	<0.001	3230.0	<0.001	828.6	<0.001
TL* modality	64.0	<0.001	278.7	<0.001	216.9	<0.001	15.7	<0.001

Table 4. Results of mean comparisons between unimodal and multimodal TL models.

	Modalities		Politics			Health		
Measures	(I)	(J)	(I-J)	SE	*p*-Value	(I-J)	SE	*p*-Value
Precision	Multi-modal	Text	0.028	0.001	<0.001	0.015	0.002	<0.001
		Image	0.276	0.005	<0.001	0.265	0.005	<0.001
	Text	Image	0.248	0.005	<0.001	0.250	0.005	<0.001
Recall	Multi-modal	Text	0.031	0.001	<0.001	0.018	0.001	<0.001
		Image	0.209	0.005	<0.001	0.159	0.004	<0.001
	Text	Image	0.178	0.005	<0.001	0.141	0.003	<0.001
F1-measure	Multi-modal	Text	0.028	0.001	<0.001	0.016	0.001	<0.001
		Image	0.276	0.005	<0.001	0.217	0.004	<0.001
	Text	Image	0.248	0.005	<0.001	0.200	0.004	<0.001
Accuracy	Multi-modal	Text	0.023	0.002	<0.001	0.020	0.001	<0.001
		Image	0.180	0.005	<0.001	0.107	0.003	<0.001
	Text	Image	0.157	0.005	<0.001	0.086	0.003	<0.001

5.3 Task 3: Evaluation of Individual Auxiliary Mechanisms

5.3.1 Effects of Individual Auxiliary Mechanisms on Detection Performance. We conduct an ablation experiment to investigate the effect of the proposed individual auxiliary mechanisms on the performance of the multimodal TL model in cross-domain fake news detection. The descriptive statistics of the model performances with various combinations of the proposed auxiliary mechanisms are reported in Table 5, where '+' (or '−') denotes adding (or not adding) the corresponding auxiliary mechanism to the baseline multimodal *TL* model.

We ran a repeated ANOVA by using DC, GN, and MP as within-subjects variables, all taking binary values. The results reported in Table 6 show that each of the three proposed

Table 5. Performances of MT-GPD with different auxiliary mechanisms.

Auxiliary mechanisms			Politics				Health			
GN	MP	DC	P	R	F1	A	P	R	F1	A
−	−	−	0.859	0.755	0.804	0.826	0.878	0.755	0.812	0.834
−	−	+	0.871	0.781	0.824	0.845	0.896	0.824	0.858	0.854
−	+	−	0.875	0.803	0.837	0.851	0.896	0.800	0.845	0.854
−	+	+	0.915	0.846	0.879	0.874	0.894	0.835	0.864	0.864
+	−	−	0.877	0.784	0.827	0.850	0.894	0.775	0.830	0.854
+	−	+	0.889	0.815	0.850	0.874	0.898	0.855	0.876	0.869
+	+	−	0.912	0.828	0.867	0.875	0.919	0.818	0.866	0.880
+	+	+	**0.916**	**0.860**	**0.885**	**0.890**	**0.923**	**0.898**	**0.911**	**0.902**

Note: P, R, F1, and A denote precision, recall, F1-measure, and accuracy. The bold indicates the best performance.

Table 6. Effects of auxiliary mechanisms on model performance.

	Precision				Recall			
	Politics		Health		Politics		Health	
Auxiliary mechanisms	*F*(1,29)	*P*-Value	*F*(1,29)	*P*-Value	*F*(1,29)	*P*-Value	*F*(1,29)	*P*-Value
(a) Macro precision and macro recall								
GN	244.5	<0.001	570.2	<0.001	83.2	<0.001	612.7	<0.001
MP	419.0	<0.001	467.9	<0.001	470.0	<0.001	715.0	<0.001
DC	94.8	<0.001	60.7	<0.001	187.1	<0.001	3122	<0.001
GN*MP	0.213	>0.05	108.7	<0.001	5.6	<0.05	28.6	<0.001
GN*DC	49.1	<0.001	4.5	<0.05	0.217	>0.05	75.5	<0.001
MP*DC	11.9	<0.01	28.9	<0.001	2.5	>0.05	41.3	<0.001
GN*MP*DC	53.2	<0.001	33.5	<0.001	2.0	>0.05	36.7	<0.001
	F1				Accuracy			
	Politics		Health		Politics		Health	
Auxiliary mechanisms	*F*(1,29)	*P*-Value	*F*(1,29)	*P*-Value	*F*(1,29)	*P*-Value	*F*(1,29)	*P*-Value
(b) Macro F1 and accuracy								
GN	127.5	<0.001	1104	<0.001	319.5	<0.001	487.3	<0.001
MP	926.4	<0.001	1116	<0.001	393.4	<0.001	407.9	<0.001
DC	250.8	<0.001	2241	<0.001	194.0	<0.001	215.6	<0.001
GN*MP	4.5	<0.05	75.4	<0.001	6.5	<0.05	44.5	<0.001
GN*DC	9.0	<0.01	52.6	<0.001	.317	>0.05	2.0	>0.05
MP*DC	5.6	<0.05	80.2	<0.001	.982	>0.05	1.5	>0.05
GN*MP*DC	13.2	<0.01	71.9	<0.001	7.1	<0.05	11.7	<0.01

mechanisms has a positive effect on all the performance measures for both datasets ($p < .001$). Given the positive roles of the three auxiliary mechanisms, the proposed MT-GPD integrates all of them. To understand whether those facilitative mechanisms have additive effects, we conducted another ablation experiment. To this end, we performed repeated measures ANOVA by using an auxiliary mechanism as the independent variable, followed by post hoc multiple comparisons between MT-GPD and other multimodal TL models incorporating fewer mechanisms with Bonferroni adjustments.

The results reported in Table 7 show that with the health target domain, MT-GPD outperforms all other models in all the performance measures ($p < .001$) except the precision of the TL + GN + MP (i.e., removing DC from MT-GPD) model ($p > .05$). For the politics dataset, first, MT-GPD outperforms all the other models incorporating one or no auxiliary mechanism across all the performance measures ($p < .001$). It also consistently outperforms TL + GN + DC (i.e., removing MP) across all the performance measures ($p < .001$). Second, MT-GPD outperforms TL+MP+DC in terms of recall and accuracy ($p < .001$), but the improvements in precision and F1 are insignificant ($p > .05$). Similarly, MT-GPD outperforms TL + GN + MP in recall ($p < .001$), F1 ($p < .001$), and accuracy ($p < .05$), but not precision ($p > .05$). The results show that MT-GPD can better facilitate TL for fake news detection

Table 7. Multiple comparison results of auxiliary mechanisms.

	Politics			Health		
(J)	(MT-GPD-J)	SE	*p*-Value	(MT-GPD-J)	SE	*p*-Value
(a) Precision						
TL	0.057	0.002	<0.001	0.046	0.002	<0.001
TL+DC	0.046	0.003	<0.001	0.028	0.001	<0.001
TL+GN	0.039	0.002	<0.001	0.029	0.001	<0.001
TL+MP	0.041	0.002	<0.001	0.028	0.001	<0.001
TL+MP+DC	0.001	0.002	>0.05	0.029	0.001	<0.001
TL + GN + DC	0.028	0.002	<0.001	0.025	0.001	<0.001
TL + GN + MP	0.005	0.003	>0.05	0.005	0.002	>0.05
(b) Recall						
TL	0.105	0.003	<0.001	0.143	0.003	<0.001
TL+DC	0.079	0.003	<0.001	0.074	0.003	<0.001
TL+GN	0.077	0.009	<0.001	0.124	0.003	<0.001
TL+MP	0.058	0.003	<0.001	0.098	0.003	<0.001
TL+MP+DC	0.014	0.004	<0.05	0.063	0.005	<0.001
TL + GN + DC	0.045	0.003	<0.001	0.043	0.003	<0.001
TL + GN + MP	0.033	0.005	<0.001	0.08	0.003	<0.001
(c) F1						
TL	0.081	0.003	<0.001	0.099	0.003	<0.001
TL+DC	0.062	0.003	<0.001	0.052	0.003	<0.001
TL+GN	0.058	0.003	<0.001	0.081	0.003	<0.001
TL+MP	0.048	0.003	<0.001	0.065	0.003	<0.001
TL+MP+DC	0.006	0.003	>0.05	0.047	0.003	<0.001
TL + GN + DC	0.035	0.003	<0.001	0.034	0.003	<0.001
TL + GN + MP	0.018	0.003	<0.001	0.045	0.003	<0.001
(d) Accuracy						
TL	0.065	0.003	<0.001	0.068	0.001	<0.001
TL+DC	0.046	0.003	<0.001	0.048	0.002	<0.001
TL+GN	0.041	0.005	<0.001	0.048	0.002	<0.001
TL+MP	0.040	0.003	<0.001	0.048	0.002	<0.001
TL+MP+DC	0.016	0.003	<0.001	0.038	0.001	<0.001
TL + GN + DC	0.017	0.002	<0.001	0.033	0.001	<0.001
TL + GN + MP	0.015	0.004	<0.05	0.022	0.002	<0.001

by incorporating the proposed auxiliary mechanisms. Incorporating more auxiliary mechanisms leads to better performance.

5.3.2 The Benefits of Model Patch. We conducted a sensitivity test on our current model patch design (i.e., the baseline) by gradually adding (or removing) network layer(s) from the proposed model patch, and then tested the updated MT-GPD model with test news samples in the technology dataset. The sensitivity analysis results presented in Table G1 in Online Appendix G in the E-companion show that our current model patch design leads to the optimal model performance in fake news detection than other alternative designs of model patch (MP). We also assessed the impact of MP on computational cost measured by (a) FLOPS (Floating Point Operations Per Second) and (b) memory usage required for fine-tuning of MT-GPD, (c) inference time (i.e., the time that MT-GPD uses to predict), and d) model performance. We fine-tuned and assessed the model using the technology dataset on a server, including four NVIDIA RTX A5000 GPU cards with 128 CPUs. As shown in Table 8, compared to MT-GPD with full-parameter fine-tuning, MT-GPD with the proposed MP required significantly less FLOPS and memory for fine-tuning, took similar inference time (because both models shared the identical network architecture), and achieved better detection performance. These results clearly demonstrate the benefits of the proposed model patch for fine-tuning MT-GPD.

Lastly, considering that there are other PEFT methods, such as LoRA (Hu et al., 2022), we implemented and tested another variant of MT-GPD by replacing the MP with LoRA. The results, as reported in Table 9, show that replacing MP with LoRA in MT-GPD significantly worsens the performance of MT-GPD in fake news detection across different target domains.

Table 8. Comparison of partial (with MP) vs. full-parameter fine-tuning (w/o MP) of MT-GPD.

Scale of fine-tuning	Memory usage	FLOPS	Inference time (seconds)	P	R	F1	A
Partial fine-tuning via MP	2,761 MB	96.2	0.729	0.975	0.934	0.952	0.964
Complete fine-tuning	8,580 MB	12,900	0.733	0.918	0.915	0.931	0.936

Table 9. The performance of MT-GPD with different PEFT methods and target datasets.

	Politics				Health				Technology			
PEFT methods	P	R	F1	A	P	R	F1	A	P	R	F1	A
LoRA	0.896	0.783	0.803	0.870	0.393	0.502	0.437	0.730	0.745	0.678	0.662	0.785
MP	0.916	0.860	0.885	0.890	0.923	0.898	0.911	0.902	0.975	0.934	0.952	0.964

Table 10. Performances of the baseline models vs. MT-GDP.

	Politics				Health				Technology			
Models	P	R	F1	A	P	R	F1	A	P	R	F1	A
SpotFake+	0.759	0.709	0.709	0.770	0.740	0.638	0.601	0.744	0.746	0.700	0.657	0.772
EANN	0.765	0.732	0.735	0.772	0.739	0.647	0.654	0.782	0.837	0.799	0.805	0.854
IMD	0.768	0.705	0.709	0.766	0.834	0.550	0.511	0.772	0.752	0.614	0.777	0.835
EM-FEND	0.740	0.686	0.676	0.763	0.741	0.632	0.636	0.786	0.835	0.807	0.812	0.856
MVAE	0.786	0.754	0.756	0.791	0.736	0.697	0.701	0.794	0.872	0.864	0.864	0.893
MCAN	0.790	0.757	0.761	0.797	0.767	0.652	0.661	0.797	0.830	0.798	0.802	0.851
MMFND	0.790	0.679	0.684	0.765	0.706	0.568	0.549	0.770	0.809	0.773	0.784	0.846
LLaVA	0.577	0.596	0.501	0.508	0.515	0.519	0.468	0.489	0.549	0.558	0.481	0.487
LLaMA-3	0.513	0.510	0.507	0.670	0.511	0.512	0.511	0.625	0.500	0.500	0.500	0.500
FT LLaVA	0.834	**0.917**	0.855	0.875	0.859	0.715	0.751	0.847	0.932	**0.965**	0.946	0.956
FT LLaMA-3	0.**924**	0.786	0.828	**0.890**	0.750	0.663	685	0.801	0.898	0.886	0.892	0.916
attRNN	0.778	0.686	0.690	0.761	0.744	0.614	0.607	0.772	0.812	0.787	0.791	0.840
FT2	0.813	0.715	0.761	0.771	0.786	0.677	0.728	0.740	0.933	0.796	0.836	0.890
FT2-GD	0.913	0.853	0.882	0.889	0.918	0.895	0.906	0.901	0.964	0.896	0.922	0.944
MT-GPD	0.916	0.860	**0.885**	**0.890**	**0.923**	**0.898**	**0.911**	**0.902**	**0.975**	0.934	**0.952**	**0.964**

Note: The bold indicates the best performance.

5.4 Task 4: Comparison of MT-GPD vs. the PTM-Based TL Baselines

5.4.1 Performance Comparison. We compared the effectiveness of MT-GPD in cross-domain fake news detection against that of 14 baseline models introduced in Section 4.2 using the politics, health, and technology datasets as the target domain. Those datasets differ in not only news domains but also sample sizes and real-fake news ratios, which can provide insights into the robustness of MT-GPD. The results are reported in Table 10.

We compared the performances between MT-GPD and individual baseline models. The results reported in Table 11 reveal that MT-GPD consistently and significantly outperforms all non-LLM baseline models across all measures for all three target domains ($p < .001$), with the exception of insignificant improvements in F1 and accuracy for political news and in recall and accuracy for health news when compared to FT2-GD ($p > 0.05$). In addition, both double fine-tuned FT2 and FT2-GD models generally outperform the other three models with single fine-tuning, and FT2-GD outperforms FT2. These results demonstrate the superior performance and robustness of our proposed individual auxiliary mechanisms, as well as MT-GPD as a whole. Furthermore, the results of FT2 and FT2-GD suggest that PTM-based TL models should undergo double fine-tuning with both source and target domain data, instead of single fine-tuning with target domain data only, which has been commonly deployed in existing studies. MT-GPD also significantly outperforms both LLaMA-3 and LLaVA without fine-tuning (i.e., direct testing).

To gain more insights, we also tested the performance of the fine-tuned LLaMA-3 and LLaVA using LoRA. Fine-tuning those large models is very computationally intensive and time-consuming. For example, fine-tuning LLaMA and LLaVA on the server with four NVIDIA A100 GPUs using the technology dataset alone and Monte Carlo cross-validation took approximately 30 days, in contrast to approximately 504 min required for fine-tuning MT-GPD on less powerful GPUs (e.g., four NVIDIA RTX A5000) with the same dataset. Thus, we only fine-tuned and tested LLaMA and LLaVA with each of

Table 11. Multiple comparisons between MT-GPD and individual baselines.

		Politics			Health			Technology		
Baselines (J)	M	(MT-GPD-J)	SE	*p*-Value	(MT-GPD-J)	SE	*p*-Value	(MT-GPD-J)	SE	*p*-Value
SpotFake+	P	0.158	0.017	<0.001	0.183	0.023	<0.001	0.229	0.045	<0.001
	R	0.152	0.014	<0.001	0.260	0.017	<0.001	0.235	0.029	<0.001
	F1	0.176	0.016	<0.001	0.310	0.020	<0.001	0.294	0.039	<0.001
	A	0.121	0.009	<0.001	0.158	0.016	<0.001	0.192	0.030	<0.001
EANN	P	0.151	0.010	<0.001	0.185	0.009	<0.001	0.138	0.007	<0.001
	R	0.128	0.010	<0.001	0.252	0.009	<0.001	0.136	0.012	<0.001
	F1	0.151	0.009	<0.001	0.256	0.009	<0.001	0.147	0.129	<0.001
	A	0.118	0.008	<0.001	0.120	0.005	<0.001	0.110	0.098	<0.001
IMD	P	0.148	0.011	<0.001	0.089	0.012	<0.001	0.223	0.018	<0.001
	R	0.156	0.013	<0.001	0.348	0.017	<0.001	0.320	0.023	<0.001
	F1	0.176	0.015	<0.001	0.400	0.019	<0.001	0.175	0.009	<0.001
	A	0.124	0.010	<0.001	0.130	0.008	<0.001	0.128	0.006	<0.001
EM-FEND	P	0.177	0.148	<0.001	0.182	0.049	<0.001	0.140	0.007	<0.001
	R	0.174	0.109	<0.001	0.266	0.070	<0.001	0.128	0.009	<0.001
	F1	0.209	0.139	<0.001	0.274	0.078	<0.001	0.139	0.007	<0.001
	A	0.127	0.069	<0.001	0.116	0.022	<0.001	0.108	0.005	<0.001
MVAE	P	0.130	0.050	<0.001	0.188	0.036	<0.001	0.103	0.005	<0.001
	R	0.107	0.061	<0.001	0.201	0.056	<0.001	0.070	0.008	<0.001
	F1	0.130	0.061	<0.001	0.209	0.056	<0.001	0.088	0.006	<0.001
	A	0.099	0.048	<0.001	0.108	0.025	<0.001	0.071	0.004	<0.001
MCAN	P	0.127	0.042	<0.001	0.157	0.039	<0.001	0.145	0.007	<0.001
	R	0.104	0.062	<0.001	0.247	0.062	<0.001	0.137	0.011	<0.001
	F1	0.125	0.060	<0.001	0.250	0.072	<0.001	0.149	0.009	<0.001
	A	0.093	0.045	<0.001	0.105	0.017	<0.001	0.113	0.005	<0.001
MMFND	P	0.151	0.006	<0.001	0.153	0.004	<0.001	0.128	0.007	<0.001
	R	0.070	0.007	<0.001	0.193	0.026	<0.001	0.125	0.017	<0.001
	F1	0.206	0.012	<0.001	0.343	0.011	<0.001	0.178	0.014	<0.001
	A	0.125	0.007	<0.001	0.132	0.004	<0.001	0.117	0.007	<0.001
attRNN	P	0.155	0.007	<0.001	0.151	0.003	<0.001	0.134	0.006	<0.001
	R	0.082	0.009	<0.001	0.154	0.012	<0.001	0.122	0.010	<0.001
	F1	0.199	0.012	<0.001	0.297	0.012	<0.001	0.165	0.010	<0.001
	A	0.129	0.007	<0.001	0.130	0.003	<0.001	0.123	0.007	<0.001
FT2	P	0.104	0.001	<0.001	0.137	0.001	<0.001	0.042	0.003	<0.001
	R	0.145	0.003	<0.001	0.221	0.003	<0.001	0.139	0.007	<0.001
	F1	0.125	0.003	<0.001	0.183	0.002	<0.001	0.116	0.006	<0.001
	A	0.119	0.003	<0.001	0.162	0.001	<0.001	0.074	0.004	<0.001
FT2-GD	P	0.004	0.001	<0.05	0.006	0.001	<0.001	0.011	0.003	<0.001
	R	0.008	0.003	<0.05	0.004	0.003	>0.05	0.038	0.007	<0.001
	F1	0.004	0.003	>0.05	0.005	0.002	<0.05	0.030	0.006	<0.001
	A	0.002	0.003	>0.05	0.001	0.003	>0.05	0.020	0.004	<0.001

the three target-domain datasets once, instead of repeating the test for 30 runs like other baseline models. As a result, we were not able to perform a statistical test in performance differences between MT-GPD and LLaMA-3/LLaVA. The performances of the fine-tuned LLaMA-3 and LLaVA shown in Table 10 suggest that overall, MT-GPD outperforms both in fake news detection. This is not totally surprising in that those large language/multimodal models are not pre-trained for this specific down-streaming task.

5.4.2 The Impact of Source-Domain and Target-Domain Data Sizes and Class Balance on the Performance of MT-GPD in Fake News Detection. To investigate the impact of the sample size of the source-domain dataset on the performance of MT-GPD, we collected approximately 5000 additional entertainment news. Half of them were fake news collected from the Onion website (theonion.com), and the other half were real news collected from FakeNewsNet that did not overlap with our current entertainment news dataset. Then, we ran a sensitivity analysis on MT-GPD's performance by increasing the size of the source-domain dataset from 1000 to 10,000 randomly selected entertainment news articles, with an increment of 1000 news articles.

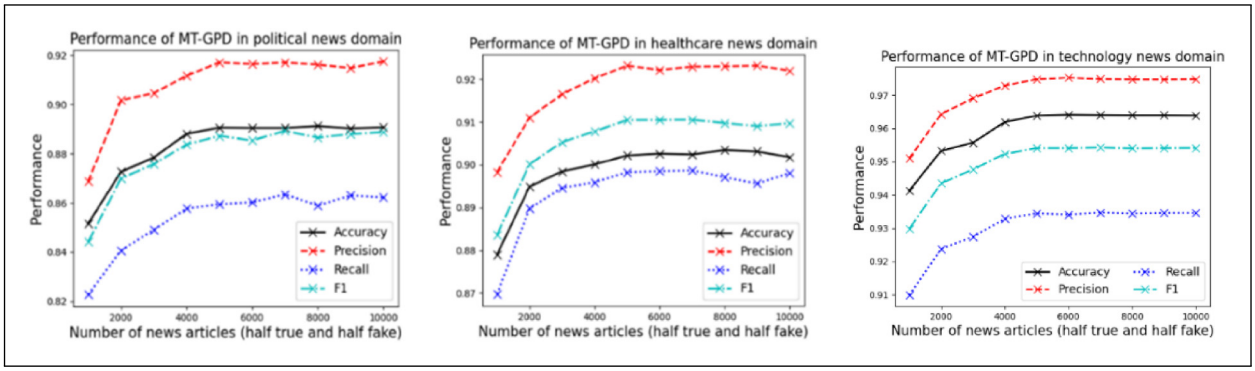

Figure 6. Results of sensitivity analysis on the impact of source-domain data size on MT-GPD performance.

Table 12. Performances of MT-GDP on various target-domain data sizes.

	Politics				Health				Technology			
Data	P	R	F1	A	P	R	F1	A	P	R	F1	A
1/3 of data	0.827	0.786	0.805	0.814	0.838	0.806	0.821	0.819	0.873	0.842	0.855	0.874
2/3 of data	0.883	0.836	0.858	0.862	0.871	0.878	0.874	0.868	0.934	0.896	0.915	0.924
Full data	0.916	0.860	0.885	0.890	0.923	0.898	0.911	0.902	0.975	0.934	0.952	0.964

Note. P = Precision; R = Recall; F1 = F1-score; A = Accuracy.

Table 13. Performances of MT-GDP with balanced and imbalanced target-domain datasets.

	Politics				Health				Technology			
Data	P	R	F1	A	P	R	F1	A	P	R	F1	A
Balanced	0.914	0.859	0.885	0.888	0.921	0.895	0.908	0.890	0.970	0.930	0.950	0.960
Imbalanced	0.916	0.860	0.885	0.890	0.923	0.898	0.911	0.902	0.975	0.934	0.952	0.964

Note. P = Precision; R = Recall; F1 = F1-score; A = Accuracy; imbalanced class is the full data.

The analysis results, as shown in Figure 6, reveal that in general, MT-GPD performs the best when using 5000 source-domain entertainment news articles for knowledge learning. Further increasing this sample size does not lead to significant improvements in model performance. This finding confirms that our initial 5201 entertainment news samples in the original source-domain dataset are adequate and appropriate for effective knowledge learning and transfer in MT-GPD. We also examined the impact of target-domain data size on the performance of MT-GPD. Considering the already small sample size in our target domains, we randomly split the data in each target domain into three folds and compared the performances of MT-GPD when using one-third, two-thirds, and full set of news articles in each target domain. The results shown in Table 12. reveal that the model performance increases as the target domain data size grows. In addition, the performance improvement from ⅔ data to full data is smaller than the performance improvement from ⅓ data to the ⅔ data). The results also demonstrate the high effectiveness of MT-GPD even when the target training dataset is very small (e.g., a couple of hundred news articles when only ⅓ of samples were used).

We also examined the impact of balanced target-domain data on the model performance by re-training the MT-GPD model with balanced data (i.e., equal numbers of fake and real news) in each target domain after removing the extra samples from the majority class. Table 13 suggest that the performance of MT-GPD trained on a balanced target-domain dataset is similar to that trained on an imbalanced dataset. This finding may be attributable to the relatively small size of each target-domain full dataset.

5.4.3 Robustness Check of MT-GPD in Detection of Fake News Created with Different Strategies. We also conducted a preliminary robustness check of MT-GPD by testing it using ten political and health-related multimodal fake news articles created by different fabrication strategies, such as the negation of real news text; real text with a fake image created by Generative AI; combining real text and a real image from two different news articles; and real text with a fabricated image for another

piece of news). The online Appendix F in the E-companion reports the descriptions of those ten fake news articles, the manipulation strategies used, the detection results from MT-GPD, and the probability of news being fake. The results show that the predicted probabilities of those sample articles being fake range from 0.705 to 0.905, which provides preliminary evidence for the high robustness of the MT-GPD model against different fake news fabrication strategies.

6 Discussion

6.1 Major Findings and Alternative Explanations

The main objective of this research is to design and evaluate a novel multimodal TL model enhanced by auxiliary mechanisms for cross-domain fake news detection. First, basic TL leads to improved fake news detection performance of the text-only and multimodal models, but worse performance of the image-only model. There are several possible explanations for the latter (negative) effect: 1) the CKA scores reported in the task 1 results indicate that the degrees of image similarity among news articles across different domains are lower than those of text similarity, which makes image-based TL more challenging. The basic TL approach may be inadequate in capturing and transferring pertinent knowledge from intricate image data for the detection of fake news; 2) manipulating images presents a greater challenge compared to fabricating textual content in news. Consequently, fake news is more likely to feature manipulated textual content than manipulated images. This suggests that the knowledge acquired from images in news articles from the source domain may be insufficient for fake news detection in a target domain; and 3) in our datasets, the class labels of news articles were annotated at the level of entire news articles, rather than at the level of individual news text or image components. There was no ground truth available regarding the presence of either manipulated text, image content, or both, within a fake news article. Thus, transferring fake news labels to image-only data may not work well.

Second, the multimodal TL baseline model without any auxiliary mechanisms consistently outperforms its unimodal counterparts. The observed performance improvement becomes even more salient after incorporating the three proposed auxiliary mechanisms.

Third, the conducted ablation experiment demonstrates that each of the three proposed auxiliary mechanisms contributes to improving the detection performance of the multimodal TL model. Generally, more auxiliary mechanisms lead to better detection performance. The findings of task 4 further indicate that integrating the proposed auxiliary mechanisms into existing PTM-based TL models with double fine-tuning can also improve detection performance. These results underline the efficacy, robustness, and generalizability of the proposed auxiliary mechanisms in support of TL.

Fourth, MT-GPD outperforms the PTM-based TL models. Our findings highlight the necessity and benefits of customized TL. Additionally, double fine-tuning of PTM (i.e., FT2 and FT2-GD) seem to be more effective than a single fine-tuning counterpart. MT-GPD also shows much better performance in fake news detection than LLaVA and LLaMA-3 when the latter two models are directly applied to fake news detection without fine-tuning. Even after fine-tuning those two large language/multimodal models, MT-GPD still generally outperforms both. Moreover, fine-tuning LLaVA and LLaMA-3 is much more computationally expensive and time consuming than fine-tuning MT-GPD.

Finally, the results of evaluation suggest that the performance in fake news detection depends on the similarity between fake and real news within the same domain. Based on the similarity values of fake vs. real news in the same domains, as reported in the last column of Table 1, the similarity between fake and real news within the politics domain is much higher than that within the health domain for both text and image, despite that the politics domain has a higher similarity to the source domain. The specific characteristics of fake news in each domain can also influence the detection performance. The health domain news dataset was in the specific context of Covid-19, whereas the political news covers a wide range of politics topics. Thus, the amount of available data, similarity between fake and real news within the same target domain, and the specificity of the target domain can all be potential determinants of the success of the fine-tuned fake news detection models.

Many factors may contribute to the differences in the performance of different models across different studies. For example, datasets used are definitely one of the primary reasons; variations in the architectural design of the models, hyperparameter tuning, training process, and the ways that models learn to extract features from data, which may result in different patterns being emphasized during learning, etc. In fact, one may even get different results when training the same deep learning model with the same datasets multiple times due to the stochastic nature of the training process. That is why we trained and tested our model 30 times and reported the average performance across those 30 runs as the model performance.

6.2 Research Contributions

From a design science research aspect, this study advances fake news detection and transfer learning techniques, contributing several novel research insights. First, this study introduces a novel TL model for cross-domain fake news detection. Current TL approaches primarily employ PTMs to generate domain-independent representations from very large yet open-domain datasets. The fine-tuning of PTMs using news data can be compromised by superfluous features inherent in word or image embeddings pre-learned by those models, thus resulting in sub-optimal model performance (Kim and Kang, 2022). In contrast, MT-GPD learns and transfers fake news patterns learned from a source domain, demonstrating greater

effectiveness in capturing both important domain-specific and domain-independent patterns for fake news detection.

Second, we propose and evaluate three auxiliary mechanisms as novel design artifacts to facilitate transfer learning, including gating network, model patch, and domain classifier. Ablation experiment results show that each of these artifacts contributes to the performance of MT-GPD positively:

1. Effective and dynamic multimodal fusion: the proposed GN innovatively and dynamically assesses the relative importance of text and image representations of individual news articles to fake news detection. By deriving gating scores for multimodal news content representations, the GN facilitates TL more effectively than the simple concatenation of multimodal representations.
2. Parameter-efficient TL: our MP mechanism sets the BN and LN layers to separate domain-specific and domain-independent knowledge. By restricting fine-tunable layers to the domain-specific layers only, MT-GPD achieves enhanced performance and computational efficiency. Fine-tuning the entire PTMs is often neither efficient nor generalizable. While our MP design requires refining the multimodal representation fusion module for a target news domain, this requirement could potentially be relaxed in the future as research progresses toward more robust and generalizable multimodal representations, which can further improve the efficiency of transfer.
3. Transformation of multimodal representation across different domains: the proposed DC learns domain-specific characteristics through self-supervised learning and transforms them directly to align the multimodal representation with a target news domain.

Third, this is the first study that empirically examines the impacts of a dynamic fusion of weighted multimodal representations of news content and individual auxiliary mechanisms on a multimodal transfer learning-based model for fake news detection. The evaluation results demonstrate the positive impacts of those design artifacts on detection performance. In addition, the superior performance of FT2-GD to FT2 indicates the effectiveness and generalizability of the auxiliary mechanisms in other deep network architectures employing diverse TL practices.

Last but not least, MT-GPD makes a noteworthy contribution to TL beyond the task of fake news detection. While the designs of the proposed auxiliary mechanisms are customized to fake news detection in this study, their underlying design principles are generic and extendable to other tasks involving multimodal TL. For example, this study introduces a novel application of a gating network to discern the relative importance of multimodal content in an article for a classification task, and a domain classifier that aids in alignment of the multimodal representations with a target domain without domain knowledge. The online Appendix B in the E-companion summarizes major differences between MT-GPD and some state-of-the-art multimodal and/or TL models for fake news detection, demonstrating the uniqueness and novelty of the former.

6.3 Practical Implications

Our research findings provide multi-fold practical implications. First, TL emerges as a promising solution to address the challenge posed by the lack of labeled news data and the poor generalizability issues of multimodal deep neural networks in cross-domain fake news detection due to the heterogeneous news domains. As this research verifies, news articles in different domains vary significantly in their content, which makes a "universal" or generic fake news detection model ineffective for news in some domains. Leveraging the knowledge learned from one news domain to build a model for fake news detection in another domain could significantly improve detection performance. Therefore, online platform managers, seeking to adopt automatic approaches to combat the spread of fake news on their platforms, are encouraged to develop and/or deploy TL-based models customized to the nuances of different news domains.

Second, our investigation of the gating scores of news text and images reveals that text gating scores generally surpass their image counterparts, implying greater importance of textual content for detecting fake news in our datasets. However, it is essential to underline that the inclusion of news image content is valuable to the fake news detection model, contributing to the overall model performance when integrated with textual content. This insight suggests that stakeholders, such as social media platforms, news agencies, and general news consumers, could enhance their fake news detection capabilities by incorporating image content alongside text analysis. This multi-modal approach presents a more comprehensive strategy for platforms to optimize their operations through automatic fake news detection methods.

Our research has broader practical implications beyond the detection of fake news on online platforms. Our research sheds light on big data analytics-related research focusing on processing multimodal unstructured data from various sources (Cezar et al., 2020; Choi et al., 2018; Ng et al., 2023; Zhang et al., 2022). Big data analytics is important to many OM-related problems, which often involve a large amount of data from various sources (Choi et al., 2018; Ng et al., 2023). Our proposed TL model, augmented by auxiliary mechanisms, can enhance the big data analytics capability in addressing many OM-related problems. In addition, MT-GPD enables early detection and intervention of fake news, minimizing its negative impact. Online platforms and websites grappling with managing the formidable challenge of combating online misinformation can leverage automated techniques like MT-GPD to improve the efficiency and effectiveness of screening and analyzing user-generated content.

The proposed MT-GPD model and its three individual auxiliary mechanisms are designed to be generic and independent of specific fake news characteristics or specific detection knowledge, making them applicable to other problems, particularly those with limited labeled data and scenarios that can benefit from TL. For example, the widespread use of social media platforms offers a promising venue for early detection of users' mental health problems for proactive and timely intervention (Chau et al., 2020). Previous work has built models for detecting mental health problems, such as post-traumatic stress disorder (PTSD) (Coppersmith et al., 2014), depression (Chau et al., 2020), schizophrenia (Bae et al., 2021), and suicidal ideation (Zhang et al., 2024) from social media posts. However, the vast majority of them focus on the textual content of those posts only, despite the fact that multimodal social media content has become increasingly popular nowadays. Research has shown that visual features manifested in images can indicate self-disclosure needs: expressions of emotional distress, calls for help, and displays of vulnerability (Manikonda and Choudhury, 2017). More importantly, despite some differences, various mental health problems share some common warning signs, such as stress, social isolation, excessive worrying or fear, and sleep problems (www.psychiatry.org/patients-families/warning-signs-of-mental-illness). Therefore, MT-GPD holds great potential for detecting various mental health problems from multimodal social media content. Moreover, it can transfer and adapt knowledge learned from detecting one type of mental health disorder (e.g., depression) to another (e.g., suicidal ideation).

While fine-tuning LLMs may tailor those models for specific domains and tasks, it remains a significant practical challenge due to their inherent complexity and the vast number of potential domains. The computational resources required for fine-tuning those models are substantial, making it economically prohibitive for many organizations, especially those with limited resources or niche domains. However, as computational costs continue to decrease and hardware becomes more powerful, accessible, and affordable, the effectiveness of fine-tuning LLMs for fake news detection may gradually increase. This is particularly true for scenarios where fine-tuning is a one-time process and the fine-tuned model is deployed in a large variety of settings and/or applicable for multiple tasks to compensate the considerable fine-tuning cost.

6.4 Limitations and Future Research

This research has limitations that offer future research opportunities. First, the current model only analyzes a single image within each news article because the vast majority of the news articles in our datasets only contain one image. Analyzing one image is also common in existing studies (e.g., Khattar et al., 2019; Wang et al., 2018). Future research should aim to develop more advanced methods for processing images within news articles for a multimodal fake news detection model. These methods should first identify the types of images and then extract relevant features while assessing their relative importance. This approach will enable the model to capture domain-specific image features of news articles with greater precision, enhancing its effectiveness in fake news detection.

Second, beyond text and images, news encompasses other features, such as user and social context features. Future research could benefit from the integration of those supplementary features, further enhancing the depth and breadth of knowledge in TL models.

Third, how to bridge the gap between pre-training and task-specific fine-tuning is crucial. Efficient and effective task-specific fine-tuning is an important research direction for the future application of PTMs (Li et al., 2023). Additionally, it is worth exploring whether incorporating universal computation engines and adaptive fine-tuning could provide unified representations of news and further enhance the integration of multimodal features.

Fourth, deceivers are engaged in both strategic and non-strategic behaviors (Zhou et al., 2004). Despite the potential strategic adaptation of fake news strategies to various domains, non-strategic leakage can be generic and extensible across different domains. This observation reinforces the theoretical underpinning of TL for cross-domain fake news detection. Our comprehensive evaluation involving multiple news datasets from diverse domains demonstrates the high level of robustness of MT-GPD. To address the potential evolution of strategies for creating fake news, researchers need to continuously explore TL-based models capable of effectively adapting to future changes in the deceptive strategies employed by individuals creating fake news. Moreover, advancing incremental learning methods is essential, as they can progressively and continuously enhance the knowledge of existing models with newly acquired data.

Fifth, in this study, we used VGG-19 to represent images. Future research could focus on developing a domain-specific classifier for images to enhance the image classification algorithm (i.e., image-CNN). Additional features, such as the presence of human faces, gender, and emotions, may serve as indicators of news authenticity. Future studies should also explore theory-driven feature engineering for TL in a multimodal context.

Last but not least, the proposed MT-GPD model integrates deep learning and transfer learning, making the interpretation of its detection outcome challenging. While providing interpretation for the detection outcomes of MT-GPD is beyond the scope of this research, we recognize that the lack of explainability in fake news detection models presents a challenge to the field. Deep learning models are characterized as "black boxes" because they are often considered incapable of providing the rationale or explanations for their outcomes. This may negatively affect users' trust in and adoption of fake news detection models, as well as pose challenges for model refinement and enhancement. Thus, there is a strong need for advancing and leveraging explainable AI research to improve the interpretability of these models.

Acknowledgments

The authors thank the entire review team for their valuable comments and guidance throughout the review process. This work was partially supported by summer grants from the School of Data Science and Belk College of Business at UNC Charlotte and Trust Bank. Guohou Shan is the corresponding author.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Truist Business Research and Innovation Grant, School of Data Science, UNC Charlotte.

ORCID iDs

Guohou Shan https://orcid.org/0000-0002-3268-6348
Lina Zhou https://orcid.org/0000-0003-1864-0527
Zhe Fu https://orcid.org/0000-0002-3097-8451

Supplemental Material

Supplemental material for this article is available online (doi: 10.1177/10591478251319686).

References

Abrahams A, Fan W, Wang G, et al. (2015) An integrated text analytic framework for product defect discovery. *Production and Operations Management* 24(6): 975–990.

Athithan S, Sachi S, Singh AK, et al. (2023) Twitter fake news detection by using XLNET model. In: *Proceedings of 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pp.868–872.

Ba JL, Kiros JR and Hinton GE (2016) Layer normalization. arXiv preprint In: *Advances in NIPS 2016 Deep Learning Symposium*.

Bae YJ, Shim M and Lee WH (2021) Schizophrenia detection using machine learning approach from social media content. *Sensors* 21(17): 5924.

Bozinovski S and Fulgosi A (1976) The influence of pattern similarity and transfer learning upon training of a base perceptron b2. In: *Proceedings of Symposium Informatica*, pp.121–126.

Brown E (2019) *Online fake news is costing us $78 billion globally each year*. https://www.zdnet.com/article/online-fake-news-costing-us-78-billion-globally-each-year/.

Cai H, Gan C, Zhu L, et al. (2020) Tinytl: Reduce memory, not parameters for efficient on-device learning. In: *Proceedings of Advances in Neural Information Processing Systems*, pp.11285–11297.

Cezar A, Raghunathan S and Sarkar S (2020) Adversarial classification: Impact of agents' faking cost on firms and agents. *Production and Operations Management* 29(12): 2789–2807.

Chau M, Li TMH, Wong PWC, et al. (2020) Finding people with emotional distress in online social media: A design combining machine learning and rule-based classification. *MIS Quarterly* 44(2): 933–955.

Choi TM, Wallace SW and Wang Y (2018) Big data analytics in operations management. *Production and Operations Management* 27(10): 1868–1883.

Coppersmith G, Harman C and Dredze M (2014) Measuring post traumatic stress disorder in Twitter. In: *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, (1), pp.579–582.

Cruz JCB, Tan JA and Cheng C (2020) Localization of fake news detection via multitask transfer learning. In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp.2596–2604.

Dai Z, Taneja H and Huang R (2018) Fine-grained structure-based news genre categorization. In: *Proceedings of the Workshop Events and Stories in the News*, pp.61–67.

Denil M, Shakibi B, Dinh L, et al. (2013) Predicting parameters in deep learning. In: *Proceedings of Advances in Neural Information Processing Systems*, pp.2148–2156.

Desai V, Shah N, Jain J, et al. (2022) Analyzing and detecting fake news using convolutional neural networks considering news categories along with temporal interpreter. In: *Proceedings of Intelligent Systems and Sustainable Computing*, pp.387–403.

Essa E, Omar K and Alqahtani A (2023) Fake news detection based on a hybrid BERT and LightGBM models. *Complex & Intelligent Systems* 9(6): 6581–6592.

Feng Y, Jiang J, Tang M, et al. (2022) Rethinking supervised pretraining for better downstream transferring. In: *International Conference on Learning Representations (ICLR)*.

Giachanou A, Zhang G and Rosso P (2020) Multimodal multi-image fake news detection. In: *Proceedings of 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp.647–654.

Goel P, Singhal S, Aggarwal S, et al. (2021) Multi domain fake news analysis using transfer learning. In: *Proceedings of 2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pp.1230–1237.

Han S, Pool J, Tran J, et al. (2015) Learning both weights and connections for efficient neural network. In: *Proceedings of Advances in Neural Information Processing Systems*, pp.28.

Hevner AR, March ST, Park J, et al. (2004) Design science in information systems research. *MIS Quarterly* 28(1): 75–105.

Hu EJ, Shen Y, Wallis P, et al. (2022) LoRA: Low-rank adaptation of large language models. In: *International Conference on Learning Representations (ICLR)*.

Hua J, Cui X, Li X, et al. (2023) Multimodal fake news detection through data augmentation-based contrastive learning. *Applied Soft Computing* 136: 110125.

Ioffe S and Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167.

Jin Z, Cao J, Guo H, et al. (2017) Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: *Proceedings of the 25th ACM International Conference on Multimedia*, pp.795–816.

Kemker R, McClure M, Abitino A, et al. (2018) Measuring catastrophic forgetting in neural networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, (1), pp.3390–3398.

Khattar D, Goud JS, Gupta M, et al. (2019) MVAE: Multimodal variational autoencoder for fake news detection. In: *Proceedings of the World Wide Web Conference*, pp.2915–2921.

Kim G and Kang S (2022) Effective transfer learning with label-based discriminative feature learning. *Sensors* 22(5): 2025.

Lee SY, Qiu L and Whinston A (2018) Sentiment manipulation in online platforms: An analysis of movie tweets. *Production and Operations Management* 27(3): 393–416.

Li H, Zhu C, Zhang Y, et al. (2023) Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7454–7463.

Liu H, Wang W, Sun H, et al. (2023) Robust domain misinformation detection via multi-modal feature alignment. *IEEE Transactions on Information Forensics and Security* 19: 793–806.

Liu H, Li C, Wu Q, et al. (2024) Visual instruction tuning. In: *Proceedings of Advances in Neural Information Processing Systems*, vol. 36, pp.34892–34916.

Manikonda L and Choudhury M (2017) Modeling and understanding visual attributes of mental health disclosures in social media. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp.170–181.

Meel P and Vishwakarma DK (2021) Han, image captioning, and forensics ensemble multimodal fake news detection. *Information Sciences* 567: 23–41.

Meta M (2020) Using AI to detect COVID-19 misinformation and exploitative content. https://ai.facebook.com/blog/using-ai-to-detect-covid-19-misinformation-and-exploitative-content/.

Mortaz E (2020) Imbalance accuracy metric for model selection in multi-class imbalance classification problems. *Knowledge-Based Systems* 210(27): 106490.

Mudrakarta PK, Sandler M, Zhmoginov A, et al. (2019) K for the price of 1: Parameter-efficient multi-task and transfer learning. In: *Proceedings of International Conference on Learning Representations (ICLR)*.

Ng KC, Ke PF, So MK, et al. (2023) Augmenting fake content detection in online platforms: A domain adaptive transfer learning via adversarial training approach. *Production and Operations Management* 32(7): 2101–2122.

Orhan AE (2021) Compositional generalization in semantic parsing with pretrained transformers. arXiv preprint arXiv*:2109.15101*.

Pfeiffer J, Rücklé A, Poth C, et al. (2020) Adapterhub: A framework for adapting transformers. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp.46–54.

Pham ND, Le TH, Do TD, et al. (2021) Vietnamese Fake news detection based on hybrid transfer learning model and Tf-Idf. In: *Proceedings of 2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, pp.1–6.

Qi P, Cao J, Li X, et al. (2021) Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In: *Proceedings of the 29th ACM International Conference on Multimedia*, pp.1212–1220.

Reinemann C, Stanyer J, Scherr S, et al. (2012) Hard and soft news: A review of concepts, operationalizations and key findings. *Journalism* 13(2): 221–239.

Ruhl Ibarra G, van't Riet J, Kleemans M, et al. (2024) Picturing tragedy: A content analysis of the publication of graphic images in newspapers. *Mass Communication and Society*: 1–29.

Rumbaugh DM, Washburn DA, King JE, et al. (2008) Why some apes imitate and/or emulate observed behavior and others do not: Fact, theory, and implications for our kind. *Journal of Cognitive Education and Psychology* 7(1): 101–110.

Shu K, Mahudeswaran D, Wang S, et al. (2020) Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* 8(3): 171–188.

Shu K, Mosallanezhad A and Liu H (2022) Cross-domain fake news detection on social media: A context-aware adversarial approach. In: *Frontiers in Fake Media Generation and Detection*. Cham, Switzerland: Springer, 215–232.

Silva A, Luo L, Karunasekera S, et al. (2021) Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp.557–565.

Singhal S, Shah RR, Chakraborty T, et al. (2019) Spotfake: A multi-modal framework for fake news detection. In: *Proceedings of 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pp.39–47.

Singhal S, Kabra A, Sharma M, et al. (2020) Spotfake+: A multi-modal framework for fake news detection via transfer learning (student abstract). In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp.13915–13916.

Singhal S, Dhawan M, Shah RR, et al. (2021) Inter-modality discordance for multimodal fake news detection. In: *Proceedings of ACM Multimedia Asia*, pp.1–7.

Touvron H, Lavril T, Izacard G, et al. (2023) Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Wang Y, Ma F, Jin Z, et al. (2018) EANN: Event adversarial neural networks for multi-modal fake news detection. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp.849–857.

Watson A (2023) Level of confidence in distinguishing between real news and fake news among adults in the United States as of December 2020. https://www.statista.com/statistics/657090/fake-news-recogition-confidence/.

Wu L, Rao Y, Nazir A, et al. (2020) Discovering differential features: Adversarial learning for information credibility evaluation. *Information Sciences* 516: 453–473.

Wu Y, Zhan P, Zhang Y, et al. (2021) Multimodal fusion with co-attention networks for fake news detection. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pp. 2560–2569. Stroudsburg, PA: Association for Computational Linguistics.

Xu QS and Liang YZ (2001) Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems* 56(1): 1–11.

Yang W, Wang X, Lu J, et al. (2020) Interactive steering of hierarchical clustering. *IEEE Transactions on Visualization and Computer Graphics* 27(10): 3953–3967.

Yu N, Pan S, Yang CC, et al. (2020) Exploring the role of medias sources on COVID-19–related discrimination experiences and concerns among Asian people in the United States: Cross-sectional survey study. *Journal of Medical Internet Research* 22(11): e21684–e21684.

Zhang X, Du Q and Zhang Z (2022) A theory-driven machine learning system for financial disinformation detection. *Production and Operations Management* 31(8): 3160–3179.

Zhang D, Zhou L, Tao J, et al. (2024) KETCH: A knowledge-enhanced transformer-based approach to suicidal Ideation detection from social media content. *Information Systems Research*: 1–28.

Zhou L, Burgoon JK, Nunamaker JF, et al. (2004) Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation* 13: 81–106.

Zillman D, Gibson R and Sargent SL (1999) Effects of photographs in news-magazine reports on issue perception. *Media Psychology* 1(3): 207–228.

How to cite this article

Zhang D, Shan G, Lee M, Zhou L and Fu Z (2025) MT-GPD: A Multimodal Deep Transfer Learning Model Enhanced by Auxiliary Mechanisms for Cross-Domain Online Fake News Detection. *Production and Operations Management* 34(8): 2448–2470.