Information Processing and Management 61 (2024) 103807

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Entity-centric multi-domain transformer for improving generalization in fake news detection

Parisa Bazmi [a,*], Masoud Asadpour [a], Azadeh Shakery [a,b], Abbas Maazallahi [a]

[a] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
[b] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

ARTICLE INFO

Keywords:
Cross-domain
Domain generalization
Entity abstraction
Fake news detection
Knowledge entities
Mixture-of-experts
Multi-domain

ABSTRACT

Fake news has become a significant concern in recent times, particularly during the COVID-19 pandemic, as spreading false information can pose significant public health risks. Although many models have been suggested to detect fake news, they are often limited in their ability to extend to new emerging domains since they are designed for a single domain. Previous studies on multidomain fake news detection have focused on developing models that can perform well on multiple domains, but they often lack the ability to generalize to new unseen domains, which limits their effectiveness. To overcome this limitation, in this paper, we propose the Entity-centric Multi-domain Transformer (EMT) model. EMT uses entities in the news as key components in learning domain-invariant and domain-specific news representations, which addresses the challenges of domain shift and incomplete domain labeling in multidomain fake news detection. It incorporates entity background information from external knowledge sources to enhance fine-grained news domain representation. EMT consists of a Domain-Invariant (DI) encoder, a Domain-Specific (DS) encoder, and a Cross-Domain Transformer (CT) that facilitates investigation of domain relationships and knowledge interaction with input news, enabling effective generalization. We evaluate the EMT's performance in multi-domain fake news detection across three settings: supervised multi-domain, zero-shot setting on new unseen domain, and limited samples from new domain. EMT demonstrates greater stability than state-of-the-art models when dealing with domain changes and varying training data. Specifically, in the zero-shot setting on new unseen domains, EMT achieves a good F1 score of approximately 72 %. The results highlight the effectiveness of EMT's entity-centric approach and its potential for real-world applications, as it demonstrates the ability to adapt to various training settings and outperform existing models in handling limited label data and adapting to previously unseen domains.

1. Introduction

In recent years, the spread of fake news has become a significant concern in society, prompting many researchers to investigate the problem of fake news detection. While many fake news detection models achieve impressive results within their specific domains, their ability to generalize to new or emerging domains is limited. This limitation arises because these models are often trained on data from a single domain. In practice, fake news spreads across various domains and topics. Word distribution, sentence structure and patterns do

* Corresponding author.
E-mail address: bazmi.parisa@ut.ac.ir (P. Bazmi).

https://doi.org/10.1016/j.ipm.2024.103807
Received 23 January 2024; Received in revised form 12 April 2024; Accepted 5 June 2024
Available online 14 June 2024

vary across these domains, a phenomenon known as domain shift (Pan & Yang, 2010), which renders previously trained models inefficient for detecting fake news in emerging news domains. Generalizing to unseen domains is crucial for early detection of fake news, especially considering the often limited volume of labeled data available for training within specific domains during the initial stages.

To address domain shift challenge, recent years have seen the development of multi-domain fake news detection models. Several multi-domain fake news detection models (Shu et al., 2022; Silva et al., 2021; Wang et al., 2018; Wu et al., 2024) aim to improve generalization across multiple domains by extracting Domain-Invariant (DI) features from these domains. These methods employ domain adversarial learning or contrastive learning (Wu et al., 2024) to extract DI features from news content and social context for fake news detection. However, previous studies (Tang et al., 2020; Zhu et al., 2019) have shown that it is hard to obtain shared information that can be generalized to multiple domains, particularly for unseen domains. In addition, common feature space for multiple domains may wash out informative characteristics of individual domains i.e., domain-specific (DS) features, which are crucial for accurate fake news identification.

Other studies on multi-domain fake news detection (Nan et al., 2021; Suprem & Pu, 2022; Zhu et al., 2022) mostly utilized the idea of Mixture-of-Expert (MoE) model (Jacobs et al., 1991; Ma et al., 2018) to extract DS features for each individual domain. The MoE model uses multiple expert networks, each designed to handle a specific subset (domain) or characteristics of the entire dataset, and a trainable gating network to select and weight a subset of these experts for each input, acting as a router that directs the data to the chosen expert. These models have several limitations. Firstly, these models primarily focus on achieving good performance on multiple known domains, requiring labeled data from each domain. They often assume that the training and testing domains are identical, neglecting the importance of generalizability to unseen domains. Furthermore, some of these methods (Nan et al., 2021; Suprem & Pu, 2022) rely solely on the explicit single news domain label provided by the fake news dataset (e.g., PolitiFact, COVID, GossipCop) for domain adaptation, while neglecting the potential diversity of topics within each news article. For instance, in Suprem and Pu (2022), the data of each domain is separately used for constructing each domain expert network. However, each news article may belong to different domains. Fig. 1 shows some news pieces from the COVID dataset that can be classified as both COVID-related and political news articles. (Zhu et al., 2022) referred to this issue as incomplete domain labeling in fake news datasets. These examples demonstrate that data from individual domains could enhance fake news detection in other domains, i.e., generalization using correlated and shared information between domains. However, relying on incomplete domain labels hinders the knowledge interaction between different news domains.

To address the mentioned challenges, in this paper, we focus on developing a multi-domain fake news detection model that leverages news pieces from multiple domains to enhance its generalizability. As a result, the proposed model aims to achieve accurate classification not only for seen domains but also for unseen domains. Our goal is to efficiently extract domain-specific features, while considering the diversity of topics in each news rather than relying on explicit domain label from dataset.

News articles often include diverse entities that play a crucial role in conveying the topic and domain of the news article (Bazmi et al., 2023). Notably, in the examples of Fig. 1, the named entities such as “US,” “Donald Trump,” and “Coronavirus” provide essential information about the domain of the news articles. Understanding the entities within the news and their relationships can help the model better grasp the meaning and domain relevance of the news.

For further investigation, we extracted the word clouds of the top 200 words from three distinct domains in fake news datasets, as depicted in Fig. 2: the PolitiFact, GossipCop (Shu et al., 2020), and COVID (Li et al., 2020) fake news datasets. PolitiFact contains political news, GossipCop includes fact-checked news about celebrities, and MM-Covid contains fake news about COVID-19. It is evident from this figure that the most frequently used words differ across domains i.e., domain shift, with entities being a prominent feature in each domain. Specifically, we notice that entities in each domain (such as Coronavirus, Trump, Meghan Markle, …) are core building blocks of domain-specific knowledge. Therefore, by reducing the model’s reliance on these named entities, we can obtain domain-invariant features for fake news detection.

According to above explanation, in this paper, we propose a novel **E**ntity-centric **M**ulti-domain **T**ransformer (EMT) for domain generalization in fake news detection. In EMT, entities serve as the key to learn both domain-invariant and domain-specific news representation. We leverage entity background information obtained from external knowledge source (Wikipedia in our experiments) for fine-grained news domain representation, thereby addressing the issue of incomplete news domain labeling. We inspired from MoE model with each expert capturing discriminative features based on news entities rather than domain labels. By construction of experts based on the entities, the model focuses on finer-grained topics, leading to a more comprehensive understanding of data characteristics

- US President Donald Trump did not announce a Coronavirus vaccine was ready.
- Joe Biden said, "People who have never died before are now dying from Coronavirus".
- For the third time in recent weeks, Speaker Nancy Pelosi is attempting to exploit the Coronavirus crisis to push her abortion agenda on the American people.

Fig. 1. Examples of news pieces from COVID dataset, with named entities highlighted in red. These examples can be classified as both COVID-related and political news articles, highlighting incomplete domain labeling in fake news datasets. The highlighted entities play a pivotal role in identifying the domain of the news pieces.

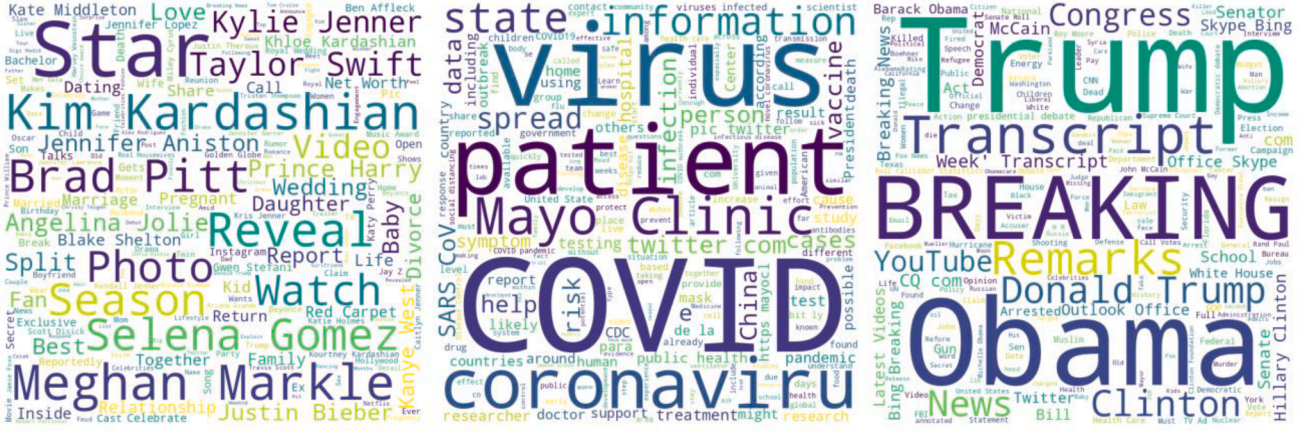

Fig. 2. The word clouds of the top 200 most frequent words in GossipCop, COVID, and PolitiFact datasets, revealing domain shift at the word level. Named entities emerge as the most repetitive words within each domain.

across domains. This finer-grained analysis improves generalizability and mitigates the effects of domain shift. additionally, we propose to adoptively model the relationship between domain-specific features (from each expert) and target news examples inspired by the works on adoptive MoE (Guo et al., 2018; Zhou et al., 2022) to improve generalization.

EMT comprises three modules: 1) Domain-Invariant encoder with entity abstraction strategy to improve the generalizability of news representation, 2) Domain-Specific encoder to learn the discriminative characteristics of each domain separately. The DS encoder uses multi-expert networks, with each expert capturing discriminative features based on news entities. 3) Cross-Domain Transformer (CT) that acts as a dynamic domain adapter for the input news. It contains a cross-domain encoder for investigation of domain relationships, and a cross-domain decoder that dynamically decodes domain experts' knowledge based on the input news piece. This approach enables the model to pay more attention to domains that share similarities with the input news, resulting in the dynamic decoding of domain-specific knowledge for classifying news from unseen domains. This can improve the proposed method's ability to generalize.

We evaluate the EMT's performance in multi-domain fake news detection through experiments in three different settings: supervised multi-domain, zero-shot setting on new unseen domain, and experiment with limited samples from new domain. Employing data from PolitiFact, GossipCop, and COVID domains, we train the model under various conditions to assess its adaptability and generalization capabilities. In the supervised multi-domain setting, the model is trained on all three domains and evaluated on their respective test sets. In the zero-shot setting, the model is trained on PolitiFact and GossipCop datasets, but tested on the unseen COVID dataset. Finally, in the experiment with limited samples from new domain, we train the model on a small portion of the COVID dataset, combined with the full PolitiFact and GossipCop datasets, simulating a real-world scenario with limited labeled data. Our goal is to understand the model's ability to adapt to different training settings and generalize to new domains with varying amounts of available data. The model exhibits competitive performance in detecting fake news across various domains, effectively generalizes to unseen domains, and performs well in situations with limited labeled data from new domains. The main contributions of this paper can be summarized as follows:

- We introduce entities in news as a key component in learning both domain-invariant and domain-specific news representations. We design entity abstraction to model domain-invariant news representation and use knowledge entity embeddings from Wikipedia to incorporate entity background information for fine-grained news domain representation.
- We propose a novel architecture for extracting domain-specific features without supervision using Transformer architecture and knowledge entity embeddings.
- We propose Transformer decoding scheme that enables interaction between the DI expert and the DS experts using a cross-attention mechanism. By leveraging this approach, domains that share similarities with the input news are given more weight in the attention output, resulting in the dynamic decoding of domain-specific knowledge for inferring news from unseen domains. This significantly improves the proposed method's ability to generalize.
- We conduct extensive experiments using different settings to accommodate varying amounts of available data for new domains. The results demonstrate the effectiveness of our method compared to state-of-the-art models on the multi-domain dataset for fake news detection in all settings.

2. Related work

In this section, we initially present an overview of prior research conducted on fake news detection, followed by an exploration of domain adaptation in fake news detection.

2.1. Fake news detection

The detection of fake news has been the subject of considerable research efforts. Methods for detecting fake news can be broadly classified into two categories: content-based and context-based approaches (Zhou et al., 2019). Content-based approaches rely on the analysis of the linguistic and visual features of news articles which include news style features (Castillo et al., 2011; Zhu et al., 2022), and semantic cues such as the use of emotionally charged language (Kolev et al., 2022; Luvembe et al., 2023; Zhu et al., 2022) or various deep language models for news text embedding (Dun et al., 2021; Koloski et al., 2022; Zhu et al., n.d.). In addition, other models use the combination of visual and textual features of news, which have been shown to be a crucial factor in identifying fake news (Hua et al., 2023; Jing et al., 2023).

Social-context-based methods for detecting fake news utilize different social contexts related to news, such as users' stances from their comments (Sengan et al., 2023; Zou et al., 2023), user features (Shu & Wang, 2019) and user interactions, which can be represented using various graph types. For instance, numerous investigations have modeled news dissemination patterns by incorporating distinct user representations (Dou et al., 2021; Guo et al., 2023; Kapadia et al., 2022). Other graph-based models have used the interaction between underlying users and news sources to obtain social context-based representations of news (Bazmi et al., 2023; Nguyen et al., 2020; Shu et al., 2019).

2.2. Domain adaptation in fake news detection

Although current fake news detection models have shown impressive performance within a single domain, they struggle to handle multiple domains or generalize to unseen domains due to the domain shift problem. This challenge has led to growing interest in

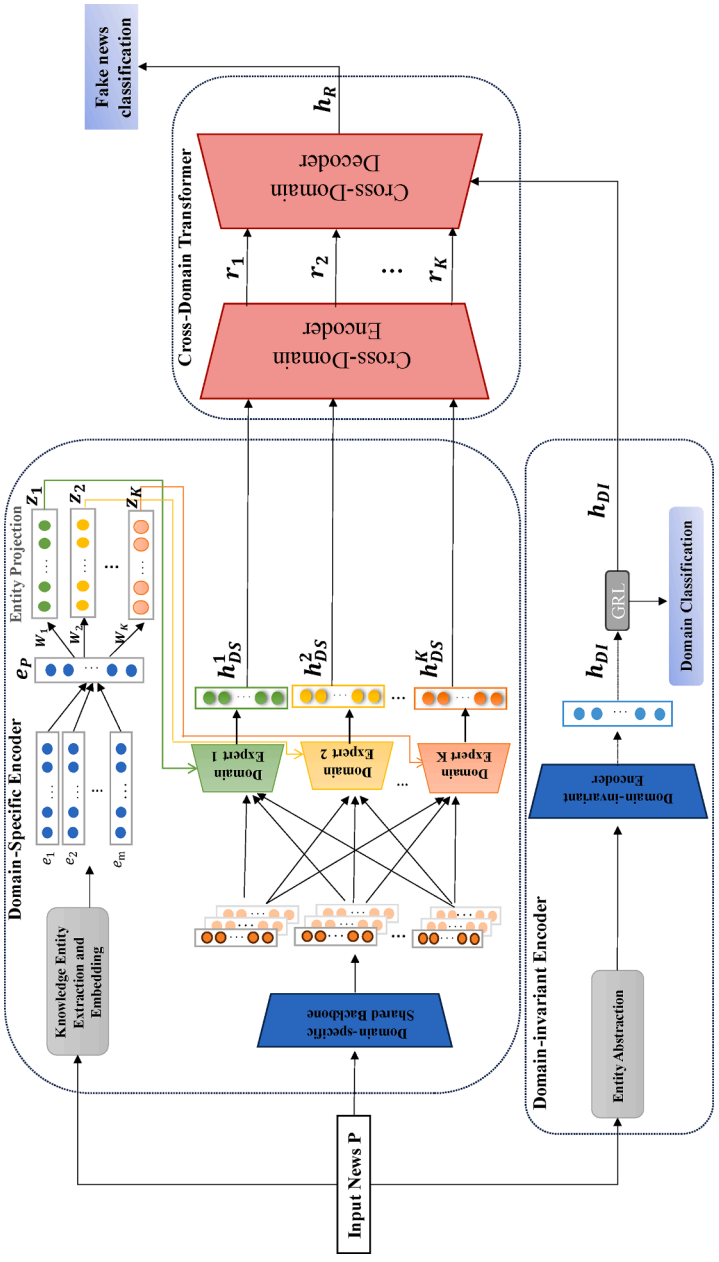

Fig. 3. The overall structure of the EMT model contains three components: 1- Domain-invariant encoder, 2-Domain-specific encoder, 3-Cross-domain Transformer.

domain adaptation techniques for fake news detection in recent years. Two lines of research dominate this field: multi-domain models and cross-domain models. Multi-domain models attempt to build a single, comprehensive model trained on data from multiple domains simultaneously. Conversely, cross-domain models focus on transferring generalizable knowledge from a source domain (where labeled data is abundant) to a target domain (where data might be scarce) to enhance performance in the target domain (Huang et al., 2021; Nan et al., 2022; Ng et al., 2023; Zeng et al., 2024). Unlike multi-domain models that build a single model for multiple domains at once, these techniques aim for knowledge transfer specifically to the target domain. Consequently, they require further training or fine-tuning on target domain samples. Additionally, these methods often train on just two domains (source and target) and might lack generalizability across multiple domains, potentially limiting their effectiveness even for handling both source and target domains simultaneously. Our model focuses on multi-domain fake news detection to enhance domain generalization.

Several multi-domain fake news detection models utilized the idea of MoE model. Multi-expert approach can help in capturing diverse characteristics of the data, making the model more robust to domain shifts. MoE models have proven to be a valuable tool for domain adaptation in various fields, including natural language processing (Guo et al., 2018), and computer vision (Zhou et al., 2021; Zhou et al., 2022).

For fake news detection, MDFND (Nan et al., 2021) proposed a MoE model that uses a set of expert networks with a weighted linear gating function. M^{3}FEND (Zhu et al., 2022) is also inspired by MoE and defines three expert networks, each responsible for extracting different views of news articles and their comments, including semantic, emotional, and stylistic perspectives. The model utilizes a domain memory to capture domain characteristics based on previously seen news domain. This memory helps select appropriate features from each expert network for a specific domain. Notably, this model extracted complicated style and emotion features from not only the news content but also its comments. Consequently, the model's performance may suffer when there are limited comments available for each domain. Both of MDFEND and M^{3}FEND need data from all domains during training for domain representation. This hinders their ability to generalize to unseen domains. MiDAS (Suprem & Pu, 2022) utilizes a domain-invariant encoder with multiple domain-specific decoders, each trained on individual domain data. An adaptive model selector estimates decoder relevance for new samples. While this model considers adaptive domain expert selection for new samples, aiding generalization, it does not take into account the diversity of topics within each domain, as it treats each domain with a single decoder as the domain expert. Consequently, this model does not exploit the correlation between different news domains.

Several other domain adaptation methods focus on learning DI features from news content or social context to detect fake news across multiple domains and improve generalization. EANN (Wang et al., 2018) achieves DI features through joint supervision of a fake news detector and a domain discriminator. However, it neglects the discriminative characteristics specific to individual source domains (DS features). Both EDDFN (Silva et al., 2021) and LIMFA (Wu et al., 2024) address this by using adversarial feature disentanglement to separate DI and DS features for fake news detection. EDDFN employs domain adversarial feature alignment learning to obtain DI features. LIMFA, inspired by contrastive learning, proposes a center-based feature alignment across multiple domains to obtain DI features from a larger number of domains. However, a limitation of these two models is their use of a single vector to represent all DS features. This restricts the model's ability to consider the wide range of DS features within each domain.

RDGT-GAN (Varshini et al., 2023) is proposed to improve generalization by generating different adversarial representations of real data. However, these models struggle with domain shift. They cannot generate adversarial examples relevant to unseen data distributions, nor can they effectively handle entirely different domains present even during training. Knowledge graph-based multi-domain model KG-MFEND (Fu Et Al., 2023) integrates knowledge graph triples into BERT embeddings to enrich news representations using domain background knowledge. However, the model is not well-structured to fully address the domain shift problem.

3. Problem definition

The problem addressed in this paper is the development of a multi-domain fake news detection model that can accurately classify news articles as real or fake, even when presented with new, previously unseen domains. Consider a news piece denoted as P, which is classified into an explicit domain with a domain label D, where D is an element of the domain set $DS = \{D_1, D_2, \ldots, D_N\}$. Here, N is the total number of domains obtained from fake news dataset. Each news article is assigned a ground-truth label y, where $y \in \{0, 1\}$, with 1 indicating that the news article is fake, while 0 denotes that it is real. Given a news article P, the objective of multi-domain fake news detection is to identify the veracity of the news, i.e., whether it is fake or real. The model is trained on news pieces from various source domains, and subsequently, it is deployed on news pieces from different target domains, which may not necessarily align with the source domains. This approach aims to ensure that the model can effectively adapt and perform accurately on diverse domains.

4. Method

In this section, we provide a detailed explanation of the Entity-centric Multi-domain Transformer (EMT) model for domain generalization in fake news detection. Fig. 3 illustrates the entire structure of the model. First, we present an overview of the model. Then, we detail the components of EMT, which include a Domain-Invariant (DI) encoder, a Domain-Specific (DS) encoder, a Cross-domain Transformer (CT) for investigating domain relationships, and dynamic extraction of domain-specific knowledge based on the input news piece. The useful representations extracted from CT are then fed into the fake news predictor.

4.1. Model overview

Our proposed classification model, illustrated in Fig. 3, takes a news piece P as input. P is a sequence of words represented as $P =$

$\{w_1, w_2, \ldots, w_n\}$. The model employs two distinct encoders, the DI and DS encoders, to process this sequence and extract multiple representations of the news P. The DI encoder, F_{DI}, focuses on capturing the general domain-invariant knowledge present in the news piece P for fake news detection. $F_{DI} : P \rightarrow h_{DI}$, transforms the sequence P into a domain-invariant news representation vector $h_{DI} \in R^o$, where o is the dimension of the DI embedding subspace. The DS encoder, F_{DS}, aims to capture domain-specific knowledge relevant to fake news detection within the news piece P. It leverages K expert networks, each specializing in identifying discriminative features associated with a particular implicit domain. These expert networks map the input sequence P to a set of vectors: $F_{DS} : P \rightarrow \{h^1_{DS}, h^2_{DS}, \ldots, h^K_{DS}\}$, where $h^i_{DS} \in R^{d_i}$ and d_i represents the dimensionality of the embedding subspace for the ith implicit domain obtained from expert network i. F_{DS} enables the representation of domain-specific knowledge for each domain, allowing P to be effectively adapted to various contexts. Finally, as shown in Fig. 3, the outputs from both encoders – the domain-invariant representation h_{DI} and the set of domain-specific representations $\{h^1_{DS}, h^2_{DS}, \ldots, h^K_{DS}\}$ – are fed into the Cross-domain Transformer (F_{CT}). By analyzing the relationships between these representations, F_{CT} can identify the distinguishing characteristics across all relevant domains that are most informative for classifying the input news piece P as real or fake.

4.2. Domain-invariant encoder

The encoder F_{DI} is responsible for learning domain-invariant features that are shared across all domains through a two-step process: entity abstraction and adversarial feature learning.

4.2.1. Entity abstraction

The entity abstraction aims to generalize news representation by substituting specific entity tokens with their corresponding entity types, resulting in a domain-invariant news representation. The entity types are defined based on the entity's semantic meaning, such as a person, organization, location, and so on. For example, the news piece "*US President Donald Trump did not announce a coronavirus vaccine was ready*" is converted to "*GPE President PERSON did not announce a PRODUCT was ready*".

By abstracting entities to their types, the model focuses on broader semantic categories rather than specific entities, reducing its reliance on domain-specific information. This is particularly beneficial for robustness against domain shifts, as entities often differ significantly across domains as shown in Fig. 2. In addition, entity abstraction can help address the issue of entity bias in model generalization, as determined by (Zhu et al., n.d.), for future data by reducing the model's reliance on specific entities for prediction.

For each news piece P, represented by a sequence of words $\{w_1, w_2, \ldots, w_n\}$, we can define the entity abstraction process as follows:
For each w_i in P:

- If w_i is an entity token, replace w_i with its corresponding entity type $E(w_i)$.
- Otherwise, retain w_i, so $E(w_i)$ is equal to w_i.

The resulting transformed news piece P' can be represented as $P' = \{(E(w_1), E(w_2), \ldots, E(w_n)\}$. We use the SpaCy library in Python (Honnibal et al., 2020) to extract entities from the news content and map them to their corresponding entity types in SpaCy. In this process, entities of type Cardinal, Ordinal, and Date in SpaCy are not taken into consideration.

Given the transformed news piece P', we can encode its sequence of tokens as $H_{DI} = \left\{h^{DI}_1, h^{DI}_2, \ldots, h^{DI}_N\right\}$, where $h^{DI}_i \in R^o$, and o denotes the embedding dimension. We use RoBERTa (Liu et al., 2019), an enhanced BERT-based pre-trained model, for encoding the news content P'. Two Transformer layers are added on top of RoBERTa (without fine-tuning) to obtain the domain-invariant news sequence representation. In line with other classification tasks using BERT-based models, we just use the hidden feature vector from [CLS] token, denoted as h_{DI}, for sequence representation. The CLS token embedding vector is considered to represent the overall sequence and is often used for classification tasks. Therefore, we utilize h_{DI} as the domain-invariant news representation. h_{DI} is the output of F_{DI} as shown in Eq. (1):

$$F_{DI}(P) = h_{DI} \tag{1}$$

4.2.2. Adversarial feature learning

To ensure that the sequence representations obtained from the prior entity abstraction process are domain-invariant, we incorporate adversarial domain classification in this component. To achieve this, we utilize a Gradient Reversal Layer (GRL) (Ganin et al., 2016) to process the feature outputs, h_{DI}, for domain classification. Specifically, the GRL is applied to the feature outputs to obtain a domain-invariant representation, which is then fed into a Multilayer Perceptron (MLP) for classification. The GRL is used to encourage the learning of domain-invariant representations by reversing the gradient during backpropagation. This reversal of gradients creates a domain confusion mechanism, making it difficult for the model to differentiate between different domains. We use cross-entropy loss as the objective function for domain classification. The process is described in Eqs. (2)-(4):

$$h_{grl} = \mathrm{GRL}(h_{DI}), \tag{2}$$

$$P_d = \mathrm{Softmax}\big(\mathrm{MLP}\big(h_{grl}\big)\big), \tag{3}$$

$$L_d = \mathrm{CrossEntropy}(P_d, y_d), \tag{4}$$

where y_d is a news domain label obtained from the news category in our dataset i.e., COVID, GossipCop, or PolitiFact. Ultimately, h_{DI} from this component serves as a domain-invariant news representation and is then used for extracting relevant features for news piece P from various domains in a cross-domain transformer (Section 4.4).

4.3. Domain-specific encoder

The DS encoder, F_{DS}, is designed to model the discriminative characteristics of individual domains in fake news detection, drawing inspiration from MoE models. Typically, MoE models consist of a shared backbone network used across all expert networks and multiple expert networks, each responsible for a specific domain. However, training a separate model for each source domain leads to a significant increase in the number of model parameters. Furthermore, due to incomplete domain labeling of news, it is not possible to assign each input news piece to a single expert network. Therefore, our goal is to efficiently extract domain-specific features without supervision (without explicit single domain labels from dataset).

As previously mentioned, entities provide crucial information about the domain of a news article. Hence, we utilize them to create an implicit representation of the news domain, rather than relying on explicit single domain labels from the dataset. The core idea of the DS encoder is to divide the embedding space, which represents the news content, into multiple subspaces using expert networks. Each expert network focuses on capturing discriminative features based on entities within the news content. By constructing experts based on entities, the model focuses on finer-grained topics, leading to a more comprehensive understanding of data characteristics across domains and capturing more diverse domain-specific features. This finer-grained analysis improves generalizability and mitigates the effects of domain shift.

4.3.1. Knowledge entity extraction and embedding

We use the SpaCy library to extract entities for each news piece. In this paper, we utilize Luke (Yamada et al., 2020) for knowledge entity embedding. LUKE is a pre-trained language model based on the Transformer (Vaswani et al., 2017) architecture, specifically designed for the purpose of entity embedding. Luke leverages a large entity-annotated corpus from Wikipedia as its primary training source to generate contextualized representations of entities. Since LUKE utilizes the knowledge of entities from the Wikipedia knowledge-base to obtain entity representations, we refer to them as knowledge entities.

Consider a news piece P containing m entities with representation vectors $E_P = \{e_1, e_2, \ldots, e_m\}$ obtained from LUKE model. To obtain a fixed-length entity vector for each news piece that represents the implicit domain of the news, we average the entity vectors of P to form $e_P \in R^{d_e}$, which serves as the entity vector for news piece P. d_e is the dimension of the entity embedding space.

4.3.2. Entity projection

The DS encoder aims to divide the embedding space, representing the news content, into several latent subspaces using expert networks, with each expert capturing discriminative features based on news entities. To achieve this, we define multiple projection heads in the entity embedding space, with each head implicitly covering a subset of entities (e.g., political entities, COVID-related entities) within the latent space. These projected heads act as guides for the individual experts. In this model, each expert is designed as a lightweight, knowledge-guided attention head (inspired from multi-head attention in Transformers) to efficiently model domain-specific features.

Our goal is to project the entity vector, e_P, into multiple subspaces, as illustrated in Fig. 3. Each subspace corresponds to a set of related entities. To achieve this, we use a set of K (number of expert networks) projection matrices $W = \{W_1, W_2, \ldots, W_K\}$ to generate K entity projection heads to guide the K expert network, where $W_i \in R^{(d_e \times E_i)}$, d_e is the dimension of the entity embedding, and E_i is the dimension of the ith entity head's subspace. The projection of the news entity vector, e_p, into the ith entity head subspace is computed as Eq. (5):

$$z_i = e_P W_i, \tag{5}$$

where $z_i \in R^{E_i}$ is the ith entity head (top part of domain-specific encoder in Fig. 3).

Nesxt, we calculate the domain-specific expert features for P using the attention mechanism in Transformer (Vaswani et al., 2017). Before describing each expert network in DS architecture in detail, we first describe the multi-head attention in Transformer architecture.

4.3.3. Multi-head attention in transformer

The multi-head attention module in Transformer (Vaswani et al., 2017) enables different attention heads to focus on distinct representation subspaces, allowing the model to learn the input from multiple perspectives. In a multi-head attention mechanism with n heads, the Q, K, and V matrices are linearly projected n times using distinct learned linear projections for each subspace. Attention scores are computed for each projected Q and K pair using Eq. (6). The resulting values from the different heads are then aggregated using Eq. (8).

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \tag{6}$$

$$head_i = \text{Attention}\left(QW_i^q, KW_i^k, VW_i^V\right), \tag{7}$$

$$\text{MultiHeadAttention}(Q,\ K,\ V) = \text{Concat}(head_1, head_2, \ldots, head_n)W^o, \tag{8}$$

where W_i^q, W_i^k, and W_i^V are the learned linear projections for the ith head, and W^o is the learned linear projection for the output. d_k is the dimension of the key vectors.

4.3.4. Multi-head attention with domain-aware head aggregation in EMT

As shown by previous works, each head in multi-head attention in BERT-like language models is responsible for attending to a different aspect of the input, such as semantic, positional, or syntactic dependency modeling (Clark et al., 2019; Htut et al., 2019; Voita et al., 2019; Zhang et al., 2022). The discriminability of these aspects varies across different news domains when it comes to fake news detection. For instance, the stylistic and syntactic viewpoint serves as a discriminative feature for identifying the veracity of news in the Science Domain, but it does not hold the same discriminative power in the Entertainment Domain, as demonstrated in (Zhu et al., 2022). Hence, we exploit the heads in multi-head attention of BERT-based embedding for modeling discriminative characteristics of each domain to tackle the problem of domain shift. In the EMT model, the aggregation of heads for each token is performed based on the knowledge of news domain (instead of Eq. (8)). In fact, EMT exploits the multi-head attention mechanism with domain-aware aggregation to capture domain-specific features. This allows the model to effectively capture the specific features unique to each domain, mitigating the impact of domain shift on veracity detection. The process is described in detail in the subsequent sections.

4.3.5. Domain-specific shared encoder

The DS encoder comprises a shared backbone Transformer model with multiple attention heads per token. The shared backbone, denoted by F_{shared}, includes two Transformer layers on top of RoBERTa (without fine-tuning) to obtain the sequence representation of tokens for the news piece P, as shown in Eq. (9),

$$\text{F}_{\text{shared}}(P) = \left\{h_i^j\right\}_{i=1,2,\cdots,t}^{j=1,2,\cdots,H} \tag{9}$$

where $h_i^j \in R^C$, and C denotes the head embedding dimension from multi-head attention in the Transformer layer, H is the number of heads, and t is the number of tokens. Therefore, for a news piece P, the output of the shared backbone is a set of H head embedding vectors for each token, before aggregating the heads according to Eq. (8).

4.3.6. Knowledge-guided attention experts

For a news piece P, we utilize each entity head z_k ($k \in 1, 2, \ldots, \text{K}$) as a query and all H heads of each token as keys and values, and compute the attention score using Eq. (6) to weigh the contribution of the entity head z_k to each head for each token. This process creates K separate expert networks, denoted by $\text{F}_{\text{DS}}^1,\ \text{F}_{\text{DS}}^2, \ldots, \text{F}_{\text{DS}}^{\text{K}}$. Each of these networks processes each token $h_i^{j=1,2,\ldots,H}$ (considering all H heads for each token) as shown in Eq. (10):

$$\text{F}_{\text{DS}}^k\left(h_i^{j=1,2,\ldots,H}\right) = \text{Attention}\left(z_k, h_i^{j=1,2,\ldots,H}, h_i^{j=1,2,\ldots,H}\right), \quad k \in 1, 2, \cdots, \text{K} \tag{10}$$

where Attention is computed according to Eq. (6).

For all tokens in the news piece P, the function F_{DS}^k is applied, resulting in the following output:

$$F_{\text{DS}}^k(P) = \left\{\text{F}_{\text{DS}}^k\left(h_1^{j=1,2,\ldots,H}\right), \text{F}_{\text{DS}}^k\left(h_2^{j=1,2,\ldots,H}\right), \ldots, \text{F}_{\text{DS}}^k\left(h_t^{j=1,2,\ldots,H}\right)\right\}, \tag{11}$$

The output of each expert network F_{DS}^k, which consists of the knowledge-modulated token representation vectors, is averaged to obtain a single vector as the domain-specific news features of domain k, denoted by h_{DS}^k. Thus, the output of the DS encoder, containing the outputs of all K domain experts, can be represented as:

$$F_{DS}(P) = \{h_{DS}^1, h_{DS}^2, \ldots, h_{DS}^K\} \tag{12}$$

By incorporating knowledge entities and expert networks in the DS Encoder, we can effectively model the discriminative characteristics of individual source domains in fake news detection, overcoming the limitations of traditional MoE models.

4.4. Cross-domain transformer

As mentioned previously, each news piece may be associated with multiple domains. For this reason, it is critical to identify a subset of relevant features for news veracity detection from each domain expert (created in the DS encoder) and combine them. For this goal, a Cross-domain Transformer (CT) is designed, which contains an encoder and decoder modules. The CT encoder aims to model the relationships and connections among various domains. The encoder contains a multi-head self-attention (the queries, keys, and values are identical) block to model interaction among different domains and learns their dependency relationships. It takes the K domain-specific features $\{h_{DS}^1, h_{DS}^2, \ldots, h_{DS}^{\text{K}}\}$ generated by the DS encoder, as input and outputs a set of domain relationship features $R=\{r_1, r_2, \ldots, r_{\text{K}}\}$ using Eqs. (6)–(8), where $r_{\text{i}} \in R^{d_r}$ and d_r is the dimension of the feature space.

The CT decoder is designed to identify the relevant knowledge available in each domain expert and dynamically decode the domain experts' knowledge for inference of a news piece's veracity. Indeed, the CT decoder acts as a dynamic domain adapter for input news. The decoder utilizes multi-head cross-attention (queries are different from keys, and values). Specifically, the decoder utilizes the domain relationship feature vectors, R, as both keys and values and the domain-invariant news representation, h_{DI}, as the query. This approach allows the decoder to establish links among domain experts and the target news domains. According to the description, the cross-domain decoder computes the dynamic expert aggregation as follows:

$$\beta_i = \text{Softmax}\left(\frac{h_{DI} r_i^T}{\sqrt{d_r}}\right) \tag{13}$$

$$h_R = \sum_{i=1}^{K} (\beta_i r_i) \tag{14}$$

β_i is attention score for r_i, which reflects the importance of the ith domain expert's knowledge based on its relevance to the domain-invariant news representation (h_{DI}). The resulting representation vector h_R, captures the most relevant features across all pertinent domains for the specific news piece. Finally, h_R is passed through an MLP layer for news classification. This representation can be expected to encompass the distinguishing characteristics across all relevant domains for the classification of the input news piece P.

4.5. Fake news classification

With the resulting representation vector h_R, we employ the cross-entropy loss to predict whether the news piece is fake or real, as follows:

$$P_f = \text{sigmoid}(\text{MLP}(h_R)), \tag{15}$$

$$L_f = -y\log\left(P_f\right) - (1-y)\log\left(1-P_f\right), \tag{16}$$

where $y \in \{0, 1\}$ is the ground-truth news veracity label i.e., real or fake.

Final Loss: The overall loss of the model is the combination of domain-classification loss L_d from domain-invariant module and fake news classification loss L_f as follows:

$$L(\theta) = \lambda_f L_f + \lambda_d L_d \tag{17}$$

Here λ_f and λ_d are coefficients that control the weight given to the fake news classification loss and domain classification loss, receptively. θ determines all the model parameters.

5. Experiment

In this section, we will initially introduce the datasets employed in our experiments. Subsequently, we will outline the experiments conducted to address the following research questions:

RQ1: What is the performance of EMT in comparison to other multi and single-domain models in detecting fake news on a multi-domain dataset?
RQ2: How does the EMT model compare to other models in terms of generalization to detect fake news in new domains?
RQ3: What is the performance of EMT in comparison to other state-of-the-art multi-domain models when a low amount of data from one domain exists at an early stage for early fake news detection?
RQ4: What impacts do various elements of EMT have on fake news detection?

To answer the above questions, we explore the performance of EMT when trained on varying amounts of data from one domain. We conduct three types of experiments to evaluate the model's performance under different settings: supervised multi-domain, zero-shot setting on new domain, and limited samples on new domain to answer RQ1, RQ2, and RQ3, receptively.

5.1. Dataset

We evaluate EMT with baselines using the multi-domain fake news detection dataset which is used in (Zhu et al., 2022). The dataset is a combination of FakeNewsNet (Shu et al., 2020) and MM-COVID (Li et al., 2020) datasets. FakeNewsNet contains news gathered from two fact-checking platforms: PolitiFact,[1] which contains political news; and GossipCop,[2] which verifies news related to celebrities. MM-COVID contains fake news about COVID-19. The dataset statistics are presented in Table 1.

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

5.2. Baselines

In the experiments, we have employed a variety of baseline models for comparison. Following (Zhu et al., 2022), these baselines fall into three main categories:

A. **Single-domain methods**

These methods train separate models for each domain. The baselines employed in this category include:

BiGRU (Ma et al., 2016): A common text encoding baseline for fake news detection based on Recurrent Neural Networks (RNN). One BiGRU layer with a hidden size of 300 is used in this paper.

TextCNN (Kim, 2014): A common Convolutional Neural Networks (CNNs) based text encoder. TextCNN with 5 kernels of varying strides (1, 2, 3, 5, and 10), each with 64 channels is used in this paper.

RoBERTa (Liu et al., 2019): An enhanced BERT pretraining model that encodes news content tokens and uses average embeddings for final predictions.

B. **Mixed-domain methods**

These methods combine the data of all domains and treat them as a single domain. Following (Zhu et al., 2022) the single-domain models BiGRU, TextCNN, and RoBERTa are also used in this category, along with:

StyleLSTM (Przybyła, 2020): This model uses a BiLSTM to extract news representations from content and combines them with news style features (such as news readability, formality, credibility, interactivity, integrity, etc.) for final predictions.

DualEmo (Zhang et al., 2021): This model employs a BiGRU for news representation and combines it with emotion features (such as emotional category, sentiment score, emotional intensity, emotion score, number of pronouns, etc.) from news content and its comments to predict the veracity of news article.

C. **Multi-domain methods**

These methods are designed especially for multi-domain fake news detection:

EANN (Wang et al., 2018): This model uses adversarial training to learn domain-invariant news representations for fake news detection.

MMoE (Ma et al., 2018): A widely used multi-domain model that utilizes multi-gate MoE, where each domain is associated with a gate. In this model, both the experts and the gates are implemented as MLPs.

MoSE (Qin et al., 2020): This model utilizes multi-gate MoE like MMoE, however in this model experts are LSTM rather than MLP.

EDDFN (Silva et al., 2021): This model uses two distinct domain-specific and domain-invariant encoders for news representation through adversarial training. All encoders are MLPs.

MDFEND (Nan et al., 2021): A multi-domain fake news detection model incorporating MoE with a domain gate to assign weights to experts for individual news items. Each expert in the model is a CNN layer.

M^{3}FEND (Zhu et al., 2022): The latest state-of-the-art multi-domain fake news detection model that incorporates multi-view feature extraction from news text and its comments, including semantic, style, and emotion views. The model utilizes adaptive cross-view interactions between these different views by using memory to store and retrieve relevant information from various domains.

The dataset and baseline implementation and settings are obtained from the paper by Zhu et al. (2022) and the corresponding GitHub[3] repository. RoBERTa is utilized for news content embedding in all baselines except for BiGRU and TextCNN, which use word2vec (Mikolov et al., 2013) embeddings.

5.3. Experimental settings

Following (Zhu et al., 2022), We use accuracy, F1 scores, and Area Under the Curve (AUC) as performance evaluation metrics. As the class distribution is imbalanced, we report macro F1. We conduct 5-fold cross-validation and present the averaged outcomes. All models are trained for a maximum of 50 epochs, utilizing Adam optimization with early stopping. An Adam learning rate is set by grid search, ranging from 1e−2 to 1e−6. Apart from employing early stopping based on validation sets, no additional dataset-specific adjustments are made. The number of domain expert networks for EMT is set to 10, which is obtained by grid search in {5, 10, 15}. The RoBERTa base model is used for EMT news content embedding. The maximum length for sequences in all models is defined as 300. After performing a grid search, λ_f and λ_d are set to 1 and 5, receptively.

5.4. Supervised multi-domain setting

In this setting, we train the EMT using a supervised learning approach for all domains. The dataset for each domain is split into training, validation, and testing subsets separately and then combined to preserve the distribution of domains within each subset. The model is trained on the combination of all three domains and evaluated on their respective test sets to measure its multi-domain

[3] https://github.com/ICTMCG/M3FEND

Table 1
The statistics of the datasets.

	PolitiFact	GossipCop	MM-COVID	All
#Real	377	16,804	4750	21,931
#Fake	363	5067	1317	6747
Total	740	21,871	6067	28,678

performance. In this setting for EMT, similar to multi-domain baselines, three different MLP heads are defined on h_R for the classification of data from each of the three domains. Table 2 presents the results, including F1 scores for each domain as well as the F1, accuracy, and AUC for overall performance. The best and second-best results are indicated by bold and underlined formatting, respectively. Based on these findings, we can draw several conclusions:

- Comparing EMT to single-domain baselines, EMT consistently demonstrates superior performance across all metrics. This indicates that EMT effectively exploits correlations and shared information across domains.
- Furthermore, since each domain contains multiple subdomains, the proposed EMT algorithm can capture the distinct characteristics of each subdomain, leading to improved model performance for each domain.
- EMT also demonstrates better performance compared to mixed-domain baselines. Notably, it surpasses the StyleLSTM and DualEmo models, both of which incorporate additional features such as news style and emotion. However, the mixed-domain methods are still limited by their treatment of all domains as a single domain, which may not fully capture the domain-specific characteristics.
- When compared to multi-domain baselines, the results of a paired *t*-test on EMT versus the best baselines exhibit the best performance among all models for the PolitiFact dataset in terms of F1 score and overall accuracy across all datasets. However, for the GossipCop dataset, M^{3}FEND performs better in terms of F1 score. This indicates that style and emotion features from user feedback in their comments could be more important parameters than entities in detecting fake news for the GossipCop domain, which can be utilized for multi-domain news representation. For the COVID dataset, EMT performs comparably to the best baselines.

Despite the state-of-the-art multi-domain model, M^{3}FEND, including additional features such as news style and the emotion of news and social comments, EMT, which only utilizes news content, demonstrates effective performance in handling multi-domain data. Specifically, for the PolitiFact dataset where entities play an important role, the performance of EMT is particularly impressive. However, since we use a pre-trained Luke model that was trained on the December 2018 version of Wikipedia, the context of some entities, particularly for the COVID datasets, may not be fully available in that version. Consequently, the performance of EMT could potentially be enhanced by incorporating more up-to-date entity knowledge. This is an area that we plan to explore in future research.

5.5. Zero-shot setting on new domain

The second experiment investigates the model's performance on an unseen domain. The results of the zero-shot setting demonstrate each model's efficacy in addressing the domain shift problem and their ability to generalize to unseen domains. In this case, we exclude the COVID dataset during the training phase. The model is trained only on the PolitiFact and GossipCop datasets and then evaluated on the COVID dataset, which is left out entirely until the testing phase. This experiment tests the model's ability to generalize to an unseen domain, commonly known as the zero-shot setting. Since the COVID dataset is newer than the others, we consider the model to be trained on older datasets for predicting the new domain which is COVID. As single-domain and mixed-domain models do not perform

Table 2
Results of supervised setting on the multi-domain dataset. * ($p \leq 0.05$) indicates paired *t*-test of EMT vs. the best baseline.

Methods		GossipCop	PolitiFact	COVID	Overall		
					F1	ACC	AUC
Single-domain	BiGRU	0.7690	0.7782	0.8845	0.7936	0.8668	0.8834
	TextCNN	0.7884	0.8002	0.9040	0.8131	0.8790	0.8982
	RoBERTa	0.7810	0.8860	0.9288	0.8154	0.8802	0.9108
Mixed-domain	BiGRU	0.7485	0.7380	0.7481	0.7574	0.8381	0.8506
	TextCNN	0.7539	0.7041	0.8328	0.7639	0.8342	0.8678
	RoBERTa	0.8022	0.7664	0.9159	0.8239	0.8823	0.9179
	StyleLSTM	0.8011	0.7626	0.9012	0.8193	0.8775	0.9120
	DualEmo	0.8049	0.7577	0.8854	0.8202	0.8786	0.9143
Multi-domain	EANN	0.7877	0.7125	0.8735	0.8025	0.8708	0.9023
	MMoE	0.8058	0.8573	**0.9473**	0.8373	0.8918	0.9270
	MoSE	0.7999	0.8385	0.9352	0.8292	0.8859	0.9185
	EDDFN	0.8002	0.8887	0.9431	0.8337	0.8920	0.9240
	MDFEND	0.7964	0.8784	0.9440	0.8323	0.8916	0.9264
	M^{3}FEND	***0.8223**	0.8813	0.9469	**0.8501**	0.8986	**0.9318**
	EMT	0.8145	***0.8910**	0.9404	0.8432	***0.9087**	0.9309

well on unseen data, we only compare the model's performance with multi-domain baselines. For MMoE and MoSE models, which consider different gates for each domain, the mean output of the PolitiFact and GossipCop gates is used for the COVID dataset. For M^{3}FEND and MDFEND, the mean of domain embeddings for the PolitiFact and GossipCop datasets is used for the COVID domain embedding.

The results of the zero-shot setting for the COVID dataset are shown in Table 3. The experiments were performed three times, and a paired *t*-test was computed to check the significance of the difference. The mean of experiments is reported in Table 3. From the results, it is observed that:

- With the results of the *t*-test, we find that EMT outperforms the best baseline i.e., M^{3}FEND significantly in terms of F1 (EMT: 0.7164, M^{3}FEND:0.6939). Both EMT and M^{3}FEND consider subdomains and subtopics and dynamically weight experts based on the input data, allowing them to generalize better than other baselines for unseen domains. M^{3}FEND obtains news domain representation by considering the similarity of news pieces with others in memory, acting as an implicit domain representation alongside an explicit one. However, EMT benefits from background knowledge of entities for domain representation. Entities can provide context to news content. Understanding the types of entities and their relationships can help the model better grasp the meaning and domain relevance of the news. Instead of relying solely on word-level features, entity knowledge can provide a richer representation by including the relationships and characteristics associated with those entities. These can be more generalizable across domains compared to individual word features. Therefore, EMT can potentially create a more robust and nuanced domain representation, enabling it to capture a wider range of domain-specific features and perform better across different unseen domains (less susceptible to domain shift).

  The other reason for EMT's superiority lies in its novel architecture that utilizes entity-based expert networks. These networks enable the model to focus on finer-grained domain-specific features by analyzing entities within the news content. This leads to a more comprehensive understanding of the data across domains and mitigates the effects of domain shift in EMT, as demonstrated by the results of the zero-shot setting.
- Other MoE models, such as MMoE, MoSE, and MDFEND, perform well in supervised settings. However, they struggle in zero-shot settings for unseen domains. This limitation can be attributed to their reliance on weighted linear gating functions to control expert contributions. These functions don't consider domain-specific features from each expert based on the input news article. The results demonstrate the effectiveness of using cross-attention in the CT decoder instead of the linear layer for domain generalization to unseen domains in fake news detection. The main reason is that the CT decoder acts as a dynamic domain adapter for input news. It aggregates useful features from different domain experts based on the specific news article, enabling better adaptation to unseen domains.
- In conclusion, the EMT model significantly outperformed baseline models in the zero-shot setting on the COVID dataset. This demonstrates its superior ability to generalize to unseen domains and effectively tackle the domain shift problem, a crucial challenge in real-world fake news detection.

5.6. *Experiment with limited samples on new domain*

The third experiment evaluates the model's performance on the COVID dataset when only a small amount of labeled data is available. We use approximately 20 % of the COVID dataset for training, combined with the full PolitiFact and GossipCop training sets. The remaining 80 % of the COVID dataset is reserved for testing. This experiment simulates a real-world scenario where limited labeled data is available for a specific domain and assesses the model's ability to learn effectively from limited examples. The results of this setting are shown in Table 4.

In the limited samples on new domain setting, the EMT model outperforms all other models in terms of AUC score, and it achieves comparable F1 scores with the best-performing models (EDDFN and MMoE). It shows the EMT model's effectiveness in learning from limited labeled data and its ability to generalize well to new domains with few examples. This highlights the model's potential for real-world applications where acquiring labeled data is limited or expensive. In addition, the results show that the cross-domain Transformer in EMT is effective in capturing the underlying relationships between different domains and using them to improve the model's generalization performance.

Both EDDFN and MMoE are effective in capturing domain-specific features and improving the model's generalization performance. EDDFN uses a combination of domain-invariant and domain-specific encoders, while MMoE employs a gating mechanism to control the contribution of each domain expert. However, their performance is not good when dealing with unseen domains in the zero-shot

Table 3
Results of zero-shot setting on the COVID dataset. * ($p \leq 0.05$) indicates paired *t*-test of EMT vs. the best baseline.

Models	F1	Accuracy	AUC
EANN	0.5665	0.8095	0.6802
MMoE	0.6660	0.8164	<u>0.8299</u>
MoSE	0.6522	0.7890	0.7236
EDDFN	0.5725	0.8209	0.7125
MDFEND	0.6210	0.7729	0.6939
M^{3}FEND	<u>0.6939</u>	**0.8367**	0.8042
EMT	***0.7164**	<u>0.8317</u>	**0.8355**

Table 4
Results of experiments with limited samples from the COVID dataset. * ($p \leq 0.05$) indicates paired *t*-test of EMT vs. the best baseline.

Models	F1	Accuracy	AUC
EANN	0.6913	0.8531	0.9313
MMoE	0.9166	0.9475	0.9772
MoSE	0.9096	0.9427	0.9703
EDDFN	**0.9207**	0.9444	0.9757
MDFEND	0.755	0.8748	0.9301
M^3FEND	0.8816	0.9233	0.9668
EMT	0.9163	**0.9479**	***0.9805**

setting.

5.7. Ablation study

In this section, we address RQ4 by conducting ablation studies to examine the role and design of each component in unseen domain classification. Additionally, we explore the impact of various expert aggregation strategies on classifying data from an unseen domain. All ablation studies are performed in the zero-shot setting. We define seven variants of EMT by removing certain components from the original EMT model:

- EMT\DI: This variant excludes the DI encoder. The DS encoder's output is passed to the Transformer encoder, and the resulting output is averaged for classification.
- EMT\EA: In this variant, we remove the entity abstraction part from the DI encoder.
- EMT\ADV: The adversarial loss is removed from EMT in this variant.
- EMT\DS-CT: This version excludes the DS encoder and, since the CT input is the DS encoder output, also removes the CT. The output of the DI encoder is used for classification.
- EMT\CT: This variant excludes the CT component. The outputs of the DI encoder and DS encoder are concatenated for classification.
- EMT\CTE: In this version, we remove the encoder part of CT and use only the cross-attention between the DI encoder output and the DS encoder output (decoder part) for classification.
- EMT\CTD: This variant removes the decoder part of CT, using the average of the encoder output for classification.

Additionally, we introduce two other variants:

- ELL: In this variant, the Experts are Linear Layers (ELL), similar to expert networks in the MMoE and MDFEND models. The average embedding of news tokens encoded by RoBERTa is used as input to the experts.
- MGLA: To investigate the impact of expert aggregation strategies on unseen domain classification, we use Multi-Gate Linear Aggregation (MGLA) from Ma et al. (2018) instead of the CT component. MGLA includes different gates for various domains, each with its own classification head. Each gate consists of two linear layers with SoftMax to select and weight the experts' output for each domain. In this variant, the DI news representation h_{DI} is used as input to different gates to weight each expert's contribution for classifying data from each domain.

Table 5 shows the results of the ablation study in zero-shot setting on the COVID dataset. Based on the ablation study of DI, DS encoders, and CT, we find that all components are crucial for classifying data from an unseen domain.

Table 5
Results of ablation study in zero-shot setting on the COVID dataset.

Models	F1	Accuracy	AUC
EMT	**0.7164**	**0.8317**	**0.8355**
EMT\CTE	0.686	0.827	0.789
EMT\ADV	0.642	0.821	0.764
EMT\CTD	0.656	0.811	0.789
EMT\DI	0.625	0.808	0.690
EMT\CT	0.6233	0.798	0.7498
EMT\EA	0.588	0.808	0.776
EMT\DS-CT	0.496	0.742	0.486
ELL	0.617	0.779	0.784
MGLA	0.592	0.804	0.810

5.7.1. Effectiveness of domain-invariant encoder and its design

The EMT\DI results highlight the importance of the DI encoder in news classification. However, when the DI encoder is used alone in EMT\DS-CT for classification, the model's performance decreases significantly, indicating that the DI encoder must work in conjunction with the DS encoder. These results imply the difficulty of obtaining domain-independent information that generalizes well to unseen domains. This reinforces the significance of incorporating domain-specific features and modeling domain discrepancy for effective fake news classification.

To investigate DI encoder's design, we use the average embedding of news tokens encoded by Roberta with domain classification adversarial loss for DI encoding of news in EMT\EA. This change dramatically reduces the model's performance across all metrics, emphasizing the significance of entity abstraction in this architecture. Removing adversarial loss also decreases the model's performance, showing that the combination of entity abstraction and adversarial loss is well-designed for domain-invariant news representation.

5.7.2. Effectiveness of domain-specific encoder and its design

The EMT\DS-CT results reveal that removing the domain experts makes the model inefficient. To investigate the effectiveness of the expert network design, we implement the ELL model using linear expert networks rather than entity-based experts. Comparing ELL's performance with EMT demonstrates the importance of entity knowledge and the effectiveness of domain-aware head aggregation in DS encoder for domain generalization.

5.7.3. Impact of CT and expert aggregation strategies

The significance of CT in classifying news is demonstrated by using the concatenation of both DI and DS news representation without CT in EMT\CT. The CT encoder, which models the correlation between different domains, is effective based on the EMT\CTE results. In the CT decoder, the experts are dynamically decoded according to DI, which is important considering the EMT\CTD results. The results indicate that the CT decoder is more critical than the CT encoder. We investigate the effectiveness of CT by using a simpler aggregation method in MGLA, where a linear layer is employed on DI news representation instead of cross-attention to control each domain expert's contribution. The MGLA results emphasize the effectiveness of the dynamic mechanism using cross-attention for domain generalization, as the model's performance declines significantly.

6. Discussion and conclusion

In this paper, we have presented the Entity-centric Multi-domain Transformer (EMT), a novel approach for domain generalization in fake news detection. EMT leverages entities for learning domain-invariant and domain-specific news representations, addressing challenges such as domain shift and incomplete domain labeling. Our experiments under supervised multi-domain setting, zero-shot and limited samples on new domain settings demonstrate the effectiveness of EMT compared to state-of-the-art models on a multi-domain dataset for fake news detection. The performance of EMT demonstrates greater stability when faced with changes in domain and varying amounts of available training data, whereas other models exhibit fluctuating performance across different training settings.

The supervised multi-domain results show that the EMT model exhibits competitive performance in fake news detection across all three categories of baseline models, i.e., single-domain, mixed-domain, and multi-domain baselines. This can be attributed to its novel architecture, which takes into account the individual characteristics of domains and sub-domains, explores domain relationships, and adaptively decodes domain experts' knowledge based on the input news article. As a result, EMT effectively addresses the challenges associated with domain shift and incomplete domain labeling, leading to a robust and generalizable fake news detection model. The EMT model outperforms the baseline models in the zero-shot setting on the COVID dataset, demonstrating its ability to generalize effectively to an unseen domain. In the experiment with limited samples on new domain, the model is required to predict the veracity of news from a new domain, having only seen a few relevant samples during training. The results highlight the EMT model's effectiveness in learning from limited labeled data and its ability to generalize well to new domains with limited examples.

In conclusion, the experimental results show that the EMT model adeptly exploits correlations and shared information across domains, which with its entity abstraction strategy enhances the generalizability of news representation. The model's capacity to adapt to different training settings and generalize to new domains with varying amounts of available data reveals its potential for real-world applications.

For future work, we aim to enhance our model by focusing on explainable multi-domain fake news detection using auxiliary social knowledge sources and exploring reinforcement learning techniques to provide evidence for fake news detection.

CRediT authorship contribution statement

Parisa Bazmi: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Masoud Asadpour:** Writing – review & editing, Supervision, Conceptualization. **Azadeh Shakery:** Writing – review & editing, Validation, Supervision, Conceptualization. **Abbas Maazallahi:** Writing – review & editing, Resources.

Data availability

The data is publicly available at https://github.com/ICTMCG/M3FEND.

Acknowledgements

This research was in part supported by a grant from the School of Computer Science, Institute for Research in Fundamental Sciences, IPM, Iran (No. CS1403-4-05).

References

Bazmi, P., Asadpour, M., & Shakery, A. (2023). Multi-view co-attention network for fake news detection by modeling topic-specific user and news source credibility. *Information Processing and Management, 60*(1), Article 103146. https://doi.org/10.1016/j.ipm.2022.103146

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In *Proceedings of the 20th international conference companion on world wide web, WWW 2011* (pp. 675–684). https://doi.org/10.1145/1963405.1963500

Clark, K., Khandelwal, U., Levy, O., & Manning, C.D. (2019). *What does BERT look at? An analysis of BERT's attention*. 276–286. https://doi.org/10.18653/V1/W19-4828.

Dou, Y., Shu, K., Xia, C., Yu, P., & Sun, L. (2021). User preference-aware fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 2051–2055).

Dun, Y., Tu, K., Chen, C., Hou, C., & Yuan, X. (2021). KAN: Knowledge-aware attention network for fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(1), 81–89. https://ojs.aaai.org/index.php/AAAI/article/view/16080.

Fu, L., Peng, H., & Liu, S. (2023). KG-MFEND: An efficient knowledge graph-based model for multi-domain fake news detection. *Journal of Supercomputing*, 1–28. https://doi.org/10.1007/S11227-023-05381-2/TABLES/11

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., Dogan, U., Kloft, M., Orabona, F., & Tommasi, T. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research, 17*, 1–35.

Guo, J., Shah, D. J., & Barzilay, R. (2018). Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 conference on empirical methods in natural language processing, EMNLP 2018* (pp. 4694–4703). https://doi.org/10.18653/V1/D18-1498

Guo, Y., Ji, S., Cao, N., Chiu, D. K. W., Su, N., & Zhang, C. (2023). MDG: Fusion learning of the maximal diffusion, deep propagation and global structure features of fake news. *Expert Systems with Applications, 213*. https://doi.org/10.1016/J.ESWA.2022.119291

Honnibal, M., and Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python | BibSonomy*. https://www.bibsonomy.org/bibtex/2616669ca18ac051794c0459373696942/rerry.

Htut, P.M., Phang, J., Bordia, S., & Bowman, S.R. (2019). Do Attention Heads in BERT Track Syntactic Dependencies? *ArXiv Preprint ArXiv:1911.12246*.

Hua, J., Cui, X., Li, X., Tang, K., & Zhu, P. (2023). Multimodal fake news detection through data augmentation-based contrastive learning. *Applied Soft Computing, 136*, Article 110125. https://doi.org/10.1016/J.ASOC.2023.110125

Huang, Y., Gao, M., Wang, J., & Shu, K. (2021). DAFD: Domain adaptation framework for fake news detection. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 13108 LNCS* (pp. 305–316). https://doi.org/10.1007/978-3-030-92185-9_25/COVER

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3*(1), 79–87. https://doi.org/10.1162/NECO.1991.3.1.79

Jing, J., Wu, H., Sun, J., Fang, X., & Zhang, H. (2023). Multimodal fake news detection via progressive fusion networks. *Information Processing and Management, 60*(1), Article 103120. https://doi.org/10.1016/J.IPM.2022.103120

Kapadia, P., Saxena, A., Das, B., Pei, Y., & Pechenizkiy, M. (2022). Co-attention based multi-contextual fake news detection. In *Springer proceedings in complexity* (pp. 83–95). https://doi.org/10.1007/978-3-031-17658-6_7/COVER

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *EMNLP 2014 - 2014 conference on empirical methods in natural language processing, proceedings of the conference* (pp. 1746–1751). https://doi.org/10.3115/V1/D14-1181

Kolev, V., Weiss, G., & Spanakis, G. (2022). FOREAL: RoBERTa model for fake news detection based on emotions. In *Proceedings of the 14th international conference on agents and artificial intelligence* (pp. 429–440). https://doi.org/10.5220/0010873900003116

Koloski, B., Stepišnik Perdih, T., Robnik-Šikonja, M., Pollak, S., & Škrlj, B. (2022). Knowledge graph informed fake news classification via heterogeneous representation ensembles. *Neurocomputing, 496*, 208–226. https://doi.org/10.1016/j.neucom.2022.01.096

Li, Y., Jiang, B., Shu, K., & Liu, H. (2020). *MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized bert pretraining approach*. http://arxiv.org/abs/1907.11692.

Luvembe, A. M., Li, W., Li, S., Liu, F., & Xu, G. (2023). Dual emotion based fake news detection: A deep attention-weight update approach. *Information Processing and Management, 60*(4), Article 103354. https://doi.org/10.1016/J.IPM.2023.103354

Ma, J., Gao, W., Mitra, P., Kwon, S., J.Jansen, B., Wong, K.-F., & Cha, M (2016). Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence (IJCAI-16) detecting* (pp. 3818–3824).

Ma, J., Zhao, Z., Yi, X., Chen, J., Hong, L., & Chi, E. H. (2018). Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In , *18*. *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1930–1939). https://doi.org/10.1145/3219819.3220007

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st international conference on learning representations, ICLR 2013 - workshop track proceedings*.

Nan, Q., Cao, J., Zhu, Y., Wang, Y., & Li, J. (2021). MDFEND: Multi-domain fake news detection. In *Proceedings of the 30th ACM international conference on information & knowledge management* (pp. 3343–3347). https://doi.org/10.1145/3459637.3482139

Nan, Q., Wang, D., Zhu, Y., Sheng, Q., Shi, Y., Cao, J., & Li, J. (2022). Improving fake news detection of influential domain via domain-and instance-level transfer. In *Proceedings of the 29th international conference on computational linguistics* (pp. 2834–2848).

Ng, K. C., Ke, P. F., So, M. K. P., & Tam, K. Y. (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. https://doi.org/10.1111/poms.13959

Nguyen, V.-H., Sugiyama, K., Nakov, P., & Kan, M.-Y. (2020). FANG: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 1165–1174).

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Przybyła, P. (2020). Capturing the style of fake news. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(01), 490–497. https://doi.org/10.1609/AAAI.V34I01.5386

Qin, Z., Cheng, Y., Zhao, Z., Chen, Z., Metzler, D., & Qin, J. (2020). Multitask mixture of sequential experts for user activity streams. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 3083–3091). https://doi.org/10.1145/3394486.3403359

Sengan, S., Vairavasundaram, S., Ravi, L., AlHamad, A. Q. M., Alkhazaleh, H. A., & Alharbi, M. (2023). Fake news detection using stance extracted multimodal fusion-based hybrid neural network. *IEEE Transactions on Computational Social Systems*, 1–12. https://doi.org/10.1109/TCSS.2023.3269087

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (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., & Liu, H. (2022). *Cross-domain fake news detection on social media: A context-aware adversarial approach* (pp. 215–232). https://doi.org/10.1007/978-981-19-1524-6_9.

Shu, K., & Wang, S. (2019). The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 436–439). https://doi.org/10.1109/MIPR.2018.00092

Shu, K., Wang, S., & Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. In *WSDM '19: proceedings of the twelfth ACM international conference on web search and data mining* (pp. 312–320).

Silva, A., Luo, L., Karunasekera, S., & Leckie, C. (2021). Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *35th AAAI conference on artificial intelligence, AAAI 2021, 1* (pp. 557–565). https://doi.org/10.1609/aaai.v35i1.16134

Suprem, A., & Pu, C. (2022). *MiDAS: Multi-integrated domain adaptive supervision for fake news detection*.

Tang, H., Liu, J., Zhao, M., & Gong, X. (2020). Progressive layered extraction (PLE): A novel multi-task learning (MTL) model for personalized recommendations. In *RecSys 2020 - 14th ACM conference on recommender systems* (pp. 269–278). https://doi.org/10.1145/3383313.3412236

Varshini, U. S. S., Sree, R. P., Srinivas, M., & Subramanyam, R. B. V. (2023). RDGT-GAN: Robust distribution generalization of transformers for COVID-19 fake news detection. *IEEE Transactions on Computational Social Systems*. https://doi.org/10.1109/TCSS.2023.3269595

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Kaiser, Ł. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *ACL 2019 - 57th annual meeting of the association for computational linguistics, proceedings of the conference* (pp. 5797–5808). https://doi.org/10.18653/v1/p19-1580

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (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).

Wu, D., Tan, Z., Zhao, H., Jiang, T., & Qi, M. (2024). LIMFA: Label-irrelevant multi-domain feature alignment-based fake news detection for unseen domain. *Neural Computing and Applications, 36*(10), 5197–5215. https://doi.org/10.1007/s00521-023-09340-z

Yamada, I., Asai, A., Shindo, H., Takeda, H., & Matsumoto, Y. (2020). LUKE: Deep contextualized entity representations with entity-aware self-attention. In *EMNLP 2020–2020 conference on empirical methods in natural language processing, proceedings of the conference* (pp. 6442–6454). https://doi.org/10.18653/v1/2020.emnlp-main.523

Zeng, H., Yue, Z., Shang, L., Zhang, Y., & Wang, D. (2024). Unsupervised domain adaptation via contrastive adversarial domain mixup: A case study on COVID-19. *IEEE Transactions on Emerging Topics in Computing*. https://doi.org/10.1109/TETC.2024.3354419

Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., & Shu, K. (2021). Mining dual emotion for fake news detection. In *The Web conference 2021 - proceedings of the world wide web conference, WWW 2021*. https://doi.org/10.1145/3442381.3450004

Zhang, X., Shen, Y., Huang, Z., Zhou, J., Rong, W., & Xiong, Z. (2022). Mixture of attention heads: Selecting attention heads per token. In *Proceedings of the 2022 conference on empirical methods in natural language processing, EMNLP 2022* (pp. 4150–4162). https://arxiv.org/abs/2210.05144v1.

Zhou, K., Yang, Y., Qiao, Y., & Xiang, T. (2021). Domain adaptive ensemble learning. *IEEE Transactions on Image Processing, 30*, 8008–8018. https://doi.org/10.1109/TIP.2021.3112012

Zhou, Q., Zhang, K. Y., Yao, T., Yi, R., Ding, S., & Ma, L. (2022). Adaptive mixture of experts learning for generalizable face anti-spoofing. In *MM 2022 - proceedings of the 30th ACM international conference on multimedia* (pp. 6009–6018). https://doi.org/10.1145/3503161.3547769

Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2019). Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the twelfth ACM international conference on web search and data mining, December 2018* (pp. 836–837). https://doi.org/10.1145/3289600.3291382

Zhu, Y., Sheng, Q., Cao, J., Li, S., Wang, D., & Zhuang, F. (n.d.). *Generalizing to the future: Mitigating entity bias in fake news detection; generalizing to the future: Mitigating entity bias in fake news detection*. https://doi.org/10.1145/3477495.3531816.

Zhu, Y., Sheng, Q., Cao, J., Nan, Q., Shu, K., Wu, M., Wang, J., & Zhuang, F. (2022). Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering, 14*(8), 1–14. https://doi.org/10.1109/TKDE.2022.3185151

Zhu, Y., Zhuang, F., & Wang, D. (2019). Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 5989–5996. https://doi.org/10.1609/AAAI.V33I01.33015989

Zou, T., Qian, Z., Li, P., & Zhu, Q. (2023). Cross-modal adversarial contrastive learning for multi-modal rumor detection. In *ICASSP 2023–2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1–5). https://doi.org/10.1109/ICASSP49357.2023.10096883