Enhancing Multi-Domain Fake News Detection: Causal Intervention & Cross-Attention Mechanism

Anonymous ACL submission

Abstract

With the accelerating growth of information on social platforms, the looming threat of fake news to the real world intensifies. Traditional fake news detection methods, rooted in single-domain training, often prove suboptimal in the face of the multi-domain Fake News Detection (MFND) challenge. This underscores the imperative to craft specialized multi-domain detection models. Our research identifies two primary challenges whithin the current MFND problem. First, existing studies predominantly focus on extracting information from a single event, neglecting the potentially utility of information from other events in the domain. Second, we unveil non-causal confounding effects in the domain fusion feature extraction process. To tackle the former challenge, we introduce a multi-head cross-attention module with multi-domain storage. Addressing the latter, we propose a causal intervention module for deconfounding. Finally, our Causal and Cross-attention Multi-domain Fake News Detection model (C^2MFND) is formulated. Experiments validate the substantial performance improvement of our method in the realm of MFND. Our code is available at https://anonymous.4open.science/r/C2MFND-151D.

1 Introduction

With the popularity of the Internet and social platforms, people are becoming increasingly accustomed to obtaining information online, providing the perfect platform for the spread of fake news. In a real scenario, a platform releases tens of thousands of news articles in various domains daily, including science and technology, military, education, etc. The distribution of data in these domains differs, a phenomenon known as domain shift, which is reflected in terms of word frequency, style, and even the mode of transmission (Zhu et al., 2023b). To illustrate this difference, we created cloud maps for nine domains on Weibo21, as shown in Figure 6. It is evident that the most frequent word in the military domain is "military", while in the education domain, it is "student". Therefore, a model trained on a single domain cannot be effectively transferred to a Multi-domain Fake News Detection (MFND) problem (Nan et al., 2021). In other words, unlike natural language, the distribution consistency hypothesis is not valid in this context.

In our study, we have identified two primary challenges within the existing literature: (1) Previous studies on the MFND problem can generally be categorized into two types. The first type focuses on extracting domain-shared features (Wang et al., 2018) or extracting both domain-shared and domain-specific features (Silva et al., 2021). However, these features are exclusively extracted from a single event (article) and are then employed for its own detection. The second type aims to extract specialized features by various experts for each event (Ma et al., 2018; Nan et al., 2021; Zhu et al., 2023b), which are then combined with different weights in different domains for the detection task. In a similar vein, in this process, only the distinct representations of a single event itself are fused. In conclusion, previous work overlooks the necessity of feature fusion among events (articles) in domains to obtain more informative domain features. Therefore, in the quest to determine determine the authenticity of one event in a domain, they undoubtedly sacrifice potentially valid information carried by other events in the same domain.

(2) On the other hand, attempting to extract domain fusion features on a domain background composed of multiple events can lead to a false correlation between fusion features and prediction results. This false correlation arises from biases introduced by the domain distribution, impacting both the fusion features and the classification label. Consequently, a non-causal confounding effect emerges between the fusion features and the classification label. To elucidate the confounding effect,

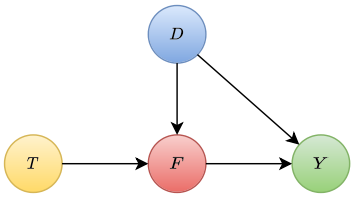


Figure 1: An illustration of the causal graph of MFND.

as depicted in Figure 1, we present the fake news detection process as a causal graph. Beyond the influence of the fusion feature F on the news label Y, which concerns the MFND method, two additional edges originate from the original text input feature T and the domain positive rate D, respectively, pointing to Y.

On one hand, for the branch $D \rightarrow Y$, we observe a significant difference in positive rates across different domains. This difference is often adopted by a model as a shortcut for classification, given that models built with imbalanced data tend to exhibit bias towards the majority class. On the other hand, for branches $T \rightarrow F$ and $D \rightarrow F$, we calculate the fusion feature F based on the similarity between input text and other events in the same domain, making F determined by both T and D. Thus, as shown in Figure 1, there exists a backdoor path $F \leftarrow D \rightarrow Y$, introducing a false correlation between the fusion features and news labels. To mitigate this non-causal confounding effect, we employ a backdoor adjustment using do-calculus $P(Y|do(F))$ (Pearl et al., 2016) to estimate the causal effect. Details are shown in appendix A.

To tackle the two challenges, we design the Causal and Cross-attention Multi-domain Fake News Detection model, denoted as C^{2}MFND. It comprises a sentence-level feature extractor, a multi-domain storage module, a multi-head cross-attention module, a causal intervention module, and a fake news classifier. With the Multi-head Cross-attention module, our model can adeptly extract domain-shared and domain-specific features on the domain background, and seamlessly fuse them with event-specific features. The incorporation of multi-domain storage mitigates the risk of noise introduced by direct fusion among events, enabling the model to not only prioritize recent events but also consider past events. Besides, to rectify the confounding effect stemming from the domain background, specifically the effect caused by uneven distributions in storage, we adopt the causal intervention module to refine the fusion features. Our contributions can be summarized as follows:

(1) We identified and analyzed a false correlation within MFND arising from uneven distributions in domains. **(2)** We introduced a novel MFND approach featuring Multi-domain storage based on the Cross-attention mechanism. Additionally, we applied a backdoor adjustment for causal intervention, effectively eliminating the confounding effect induced by uneven domain distributions. To the best of our knowledge, this is the first implementation of such an approach in the MFND problem. **(3)** We conducted extensive experiments on two datasets in Chinese and English, demonstrating superior performance compared to state-of-the-art baselines. The results substantiate the reliability and efficacy of our C^{2}MFND approach.

2 The C^{2}MFND Model

In this section, we present our C^{2}MFND approach, illustrated in Figure 2, comprising five key components: the *sentence-level feature extractor*, *multi-domain storage*, *multi-head cross-attention module*, *causal intervention module*, and *fake news classifier*. The *sentence-level feature extractor* adopts pre-trained BERT (Liu et al., 2019) and TextCNN (Kim, 2014) to obtain sentence representations. The *multi-domain storage module* is designed to provide domain background features specific to different domains, while the *multi-head cross-attention module* adeptly extracts both domain-shared and domain specific features. Lastly, the *causal intervention module* is employed to mitigate the confounding effects resulting from uneven distributions. In the following sections, we delve into each component in details.

2.1 Sentence-level Feature Extractor

For raw text input, we generate word embeddings using a pre-trained BERT model RoBERTa(Liu et al., 2019) and transform them into sentence embeddings employing TextCNN (Kim, 2014) with convolution kernels of various sizes. The process unfolds as follows: Initially, prior to entering BERT, the input (textual news t) is encoded as a token vector $\mathbf{w} = \{w_1, w_2, \ldots, w_{n_w}\}$ ($n_w = |\mathbf{w}|$ indicating the maximum truncation length of a paragraph) based on the word bag. Subsequently, its word embeddings, denoted as $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n_w}\}$, are acquired through BERT, where $\mathbf{v} \in \mathbb{R}^{d_j}$ and d_j represents the dimension of word embeddings. Following that, TextCNN executes convolution and pooling opera-

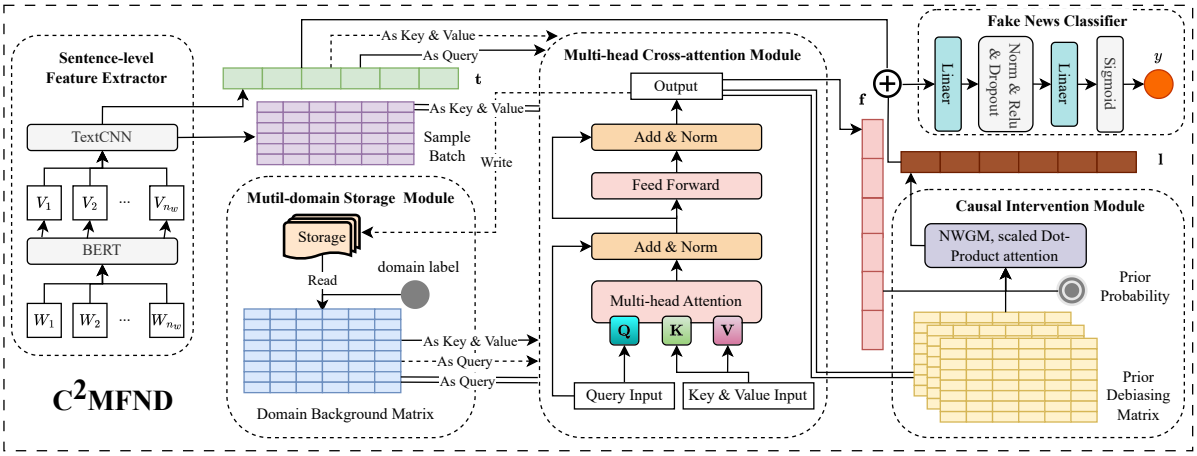


Figure 2: Overview of our C^{2}MFND Approach.

tions on the word embedding matrix to derive the sentence embedding $\mathbf{t}$ ($\mathbf{t} \in \mathbb{R}^{d_z}$), where d_z represents the dimension of sentence embeddings. It is noteworthy that the incorporation of BERT's multi-head attention and TextCNN's convolution kernels of different sizes aims to ensure that the model captures diverse perspectives, resulting in a broader field of view for obtaining enhanced sentence representations.

2.2 Multi-domain Storage Module

The objective of domain storage is to preserve the distinctive features of a domain and provide background features for a period of time to enhance news features. The limited-size external storage undergoes continuous updates during the training process, mitigating the risk of noise introduced by direct feature fusion among events. This simultaneous process broadens the model's field of view, allowing it to consider current and past domain information during feature fusion.

When a news article has a specific domain label, it signifies that the information it contains is biased toward that domain. Domains often include valuable information for detection tasks that may not explicitly present in the news article, thereby guiding the feature fusion direction. Inspired by this concept, we devised an external memory enhancement network (Santoro et al., 2016), featuring Cross-attention modules that control reading and writing, along with a Multi-domain storage. A single-domain storage comprises n_u memory units, each capable of storing a domain background embedding $\mathbf{b}$ ($\mathbf{b} \in \mathbb{R}^{d_z}$). Each memory unit's embedding characterizes a unique perspective of the domain. Thus, the domain background embedding within each single domain storage is represented as a matrix $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{n_u}\}$ ($\mathbf{B} \in \mathbb{R}^{n_u \times d_z}$). With n_d domains in MFND, the Multi-domain storage consists of n_d single-domain storages, and the Multi-domain background embedding can be represented as $\{\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_{n_d}\}$. The process of multi-domain storage is elaborated as follows:

Initialization: At the start, due to the use of ReLU activation function, we employ Kaiming initialization to initialize the domain background matrix $\mathbf{B}$ within each single domain storage.

Read function: The reading operation involves two steps. Firstly, the background embedding $\mathbf{B}_i$ corresponding to the domain label i of news t is extracted from the Multi-domain storage. Then, using the news sentence embedding $\mathbf{t}$ as the input for Query and $\mathbf{B}_i$ as the input for Key and Value, the fusion representation $\mathbf{f}$ is obtained through the Cross-attention module (Cross-attention function, denoted as $crossattention(.)$, is thoroughly introduced in Section 2.3):

$$\mathbf{f} = crossattention(\mathbf{t}, \mathbf{B}_i) \tag{1}$$

Write function: The write operation is the inverse of the read operation. By taking $\mathbf{B}_i$ as the input for Query and $\mathbf{t}$ as the input for Key and Value, the updated domain background embedding $\mathbf{B}'_i$ is obtained through the same cross-concern module. Finally, $\mathbf{B}'_i$ is used to replace $\mathbf{B}_i$:

$$\mathbf{B}'_i = crossattention(\mathbf{B}_i, \mathbf{t}) \tag{2}$$

2.3 Multi-head Cross-attention Module

Cross-attention mechanism (Gheini et al., 2021; Lin et al., 2022), an extension from Self-attention (Vaswani et al., 2017), integrates information from various sources by introducing additional input sequences for more accurate modeling. Simultaneously, the Multi-head ensures that the model can extract multi-view features. Therefore, we designed

a Multi-head Cross-attention module. In one head, the news sentence embedding $\mathbf{t}$ serves as the the input for Query, while the domain background matrix $\mathbf{B}$ serves as the input for Key and Value. The process is represented as in Equation 3:

$$\begin{aligned} \mathbf{f}' &= Softmax(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}})\mathbf{V} \\ \mathbf{Q} &= \mathbf{t}\mathbf{W}_q,\ \mathbf{K} = \mathbf{B}\mathbf{W}_k,\ \mathbf{V} = \mathbf{B}\mathbf{W}_v \end{aligned} \tag{3}$$

where $\mathbf{Q} \in \mathbb{R}^{1\times d_k}$, $\mathbf{K},\ \mathbf{V} \in \mathbb{R}^{n_u\times d_k}$ ($\mathbf{W}_q$, $\mathbf{W}_k$ and $\mathbf{W}_v \in \mathbb{R}^{d_z\times d_k}$ are learnable weight parameters), and d_k indicates the dimension of the latent layer.

Next, the output of all the heads is concatenated to obtain Multi-head output $\mathbf{f}_m$. After linear layer variation, residual connections, layer normalization ($Norm(.)$), and feedforward operations ($FFN(.)$), we obtain the fused feature $\mathbf{f}$, as represented in Equation 4:

$$\begin{aligned} \mathbf{f}_m &= [\mathbf{f}'_1, \mathbf{f}'_2, \ldots, \mathbf{f}'_{n_{head}}] \\ \mathbf{f} &= FFN\left(Norm\left(\mathbf{t} + Linear(\mathbf{f}_m)\right)\right) \end{aligned} \tag{4}$$

where $\mathbf{f}_m \in \mathbb{R}^{d_m}$, n_{head} indicates the number of heads, $d_m = n_{head} \times d_k$, and $\mathbf{f} \in \mathbb{R}^{d_z}$.

2.4 Causal Intervention Module

In Figure 1, we identify the presence of a confounder, denoted as D, in the branch $F \rightarrow Y$. This confounder can introduce misleading correlations between fusion features F and news labels Y through the likelihood $P(Y|F)$. To better understand the impact of confounding factors, we apply Bayes' theorem:

$$\begin{aligned} P(Y|F) &= \sum_d P(Y|F,d)p(d|F) \\ &\propto \sum_d P(Y|F,d)P(Y|d)P(d) \end{aligned} \tag{5}$$

Deconfounded training is then conducted on the $F \rightarrow Y$ branch, and backdoor adjustments (Pearl et al., 2016) with do-calculus are applied to F to compute the corresponding intervention distribution. Since $D \rightarrow F$ is truncated, we obtain:

$$Y_f = P(Y|do(F)) = \sum_d P(Y|F,d)P(d) \tag{6}$$

To estimate Y_f given the fusion features F's representation $\mathbf{f}$ and the confounder $\mathbf{d}$, Equation 6 is expressed as $\sum_d P(Y|\mathbf{f},\mathbf{d})P(\mathbf{d})$. Here, $P(Y|\mathbf{f},\mathbf{d})$ is the prediction of the news detection model $g(\cdot)$:

$$P(Y|\mathbf{f},\mathbf{d}) = \sigma(g(\mathbf{f},\mathbf{d})) \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function, mapping the output of $g(\cdot)$ to the range $(0,1)$. In general, Equation 6, when implemented, is defined as:

$$P(Y|do(F)) = \mathbb{E}_d[P(Y|F,d)] = \mathbb{E}_d[\sigma(g(\mathbf{f},\mathbf{d}))] \tag{8}$$

Note that $\mathbb{E}_d[\sigma(g(\mathbf{f},\mathbf{d}))]$ requires expensive sampling. Following recent works (Zhu et al., 2023a; Chen et al., 2023), we can use Normalized Weighted Geometric Mean (NWGM) (Xu et al., 2015) to approximate the expectation:

$$P(Y|do(F)) \overset{NWGM}{\approx} \sigma(\mathbb{E}_d[g(\mathbf{f},\mathbf{d})]) \tag{9}$$

Meanwhile, we can approximate the conditional probability with a linear model (Tian et al., 2022; Chen et al., 2023), i.e. $g(\mathbf{f},\mathbf{d}) = \mathbf{f}\mathbf{W}_f + h(\mathbf{d})\mathbf{W}_d$, where $h(\mathbf{d})$ is the feature transformation of $\mathbf{d}$, $\mathbf{W}_f$ and $\mathbf{W}_d \in \mathbb{R}^{d_z\times d_z}$ are learnable weight parameters. In this case, $\mathbb{E}_d[g(\mathbf{f},\mathbf{d})] = \mathbf{f}\mathbf{W}_f + \mathbb{E}_d[h(\mathbf{d})]\mathbf{W}_d$. Calculating $\mathbb{E}_d[h(\mathbf{d})]$ involves an approximation using scaled Dot-Product attention (Vaswani et al., 2017), leading to the derivation of deconfounded fusion features $\mathbf{l}$:

$$\begin{aligned} \mathbf{l} &= \mathbb{E}_d[h(\mathbf{d})] = \sum_d [Softmax(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}})\mathbf{D_d}]P(\mathbf{d}) \\ \mathbf{Q} &= \mathbf{f}\mathbf{W}_q, \mathbf{K} = \mathbf{D_d}\mathbf{W}_k \end{aligned} \tag{10}$$

where $\mathbf{Q} \in \mathbb{R}^{1\times d_k}$, $\mathbf{K} \in \mathbb{R}^{n_u\times d_k}$ ($\mathbf{W}_q$ and $\mathbf{W}_k$ $\in \mathbb{R}^{d_z\times d_k}$ are learnable weight parameters), and d_k indicates the scaling factor. Next, we elucidate the procedure for obtaining the prior probability and prior deconfounding matrix in Equation 10:

Prior probability $P(\mathbf{d})$: In a specified domain, considering the overall domain positive rate p_{domain}, a batch size n_{bat}, and the method of generating domain background features, we can model the sampling process without replacement as a hypergeometric distribution. When the overall sample size is sufficient large, and drawing a sample group minimally affects the domain positive rate in the remaining sample (typically true for large datasets), we approximate this distribution using a binomial distribution $P(\mathbf{d}) \sim Bin(n_{bat}, p_{domain})$. Consequent;y, we obtain a prior probability dictionary $P(\mathbf{d})$, where $\mathbf{d}$ is the domain positive rate, and $\mathbf{d} \in \{0/n_{bat}, 1/n_{bat}, \ldots, n_{bat}/n_{bat}\}$.

Prior deconfounding matrix $\mathbf{D_d}$: Initially, we trained an effective fake news detection models without the causal intervention module and saved its sentence embedding modules (BERT and TextCNN) and external memory enhancement network (Cross-attention modules and Multi-domain storage) for sampling. In a specific domain, given a sampling frequency n_{freq} and the

batch size n_{bat}, we sample according to the following rule: For all possible ratios (e.g. 0 positive n_{bat} negative, 1 positive $n_{bat}-1$ negative, ..., n_{bat} positive 0 negative), we sample n_{freq} times each time on the training set to obtain text sample groups $\{\mathcal{S}_{0/n_{bat}}, \mathcal{S}_{1/n_{bat}}, \ldots, \mathcal{S}_{n_{bat}/n_{bat}}\}$. Here, each $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{n_{freq}}\}$ and $\mathbf{S}$ is a sample batch. Subsequently, a sample batch $\mathbf{S}$ undergoes processing through BERT and TextCNN to get the sentence representation $\{\mathcal{T}_{0/n_{bat}}, \mathcal{T}_{1/n_{bat}}, \ldots, \mathcal{T}_{n_{bat}/n_{bat}}\}$. Each $\mathcal{T} = \{\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_{n_{freq}}\}$ exhibits a one-to-one correspondence between $\mathbf{T} \in \mathbb{R}^{n_{bat} \times d_z}$ and $\mathbf{S} \in \mathbb{R}^{n_{bat} \times n_w}$. The prior deconfounding matrix is then acquired through the external memory enhancement network, as shown in Equation 11:

$$\mathbf{U} = crossattention(\mathbf{B}, \mathbf{T}) \tag{11}$$

Thus, we obtain a prior deconfounding matrix dictionary $\mathbf{D_d} = \{\mathcal{U}_{0/n_{bat}}, \mathcal{U}_{1/n_{bat}}, \ldots, \mathcal{U}_{n_{bat}/n_{bat}}\}$, where each $\mathcal{U} = \{\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_{n_{freq}}\}$, and $\mathbf{U} \in \mathbb{R}^{n_u \times d_z}$. Here, $\mathbf{d}$ denotes the domain positive rate, and $\mathbf{d} \in \{0/n_{bat}, 1/n_{bat}, \ldots, n_{bat}/n_{bat}\}$.

2.5 Fake News Classifier

Finally, the features $\mathbf{l}$ and $\mathbf{t}$ are incorporated as inputs to the fake news detection module. The module comprises a multi-layer Perceptron (MLP) network and a sigmoid output layer. The MLP consists of normalization, dropout operation, and ReLU activation function. The output of this module is the predicted label of the news t, $\hat{y}$.

$$\hat{y} = Sigmoid(MLP(\mathbf{l} + \mathbf{t})) \tag{12}$$

Given that fake news detection essentially invovles binary classification, as expressed in Equation 13, we employ Binary Cross-Entropy Loss (BCELoss) as the Loss during training. y_t and $\hat{y}_t$ represent the true label and predicted label of news t, respectively.

$$L = -\sum_{i}^{N}(y^i \log \hat{y}^i + (1 - y^i)\log(1 - \hat{y}^i)) \tag{13}$$

3 Experiments

In this section, we conduct experiments on two widely used multi-domain fake news detection datasets in two languages to evaluate the effectiveness of our approach.

3.1 Datasets

Following previous work (Zhu et al., 2023b), we conducted experiments on two separate datasets: **Weibo21:**[1] this was publicly released for the multi-domain fake news detection task in Chinese. It was collected from Sina Weibo and manually tagged with nine topics based on content; **En-3**[2]: An English Multi-domain fake news detection dataset consisting of FakeNewsNet and COVID. It includes three domains: GossipCop, PolitiFact, and COVID. The statistics for the two dataset are present in appendix B.1.

Moreover, in our work, the data preprocessing for both datasets and the division of training, validation and test sets (in a ration of $6:2:2$) with consistent domain distribution in each set follow the procedures outlined in previous works (Nan et al., 2021; Zhu et al., 2023b).

3.2 Evaluation Metrics

Similar to prior research (Zhu et al., 2023b), we utilize the macro F1 score (F1) as an indicate for each domain. Additionally, we employ accuracy (Acc), macro F1 score (F1), and Area Under ROC (AUC) as indicators for the entire dataset. These metrics are more appropriate for evaluating model performance in scenarios with imbalanced sample distributions.

3.3 Model Settings

The baseline methods for comparison are shown in appendix B.2, and hyperparameters for these baseline methods are set consistently with (Zhu et al., 2023b), while ours are shown in appendix B.3.

To ensure the repeatability of the experiment and the fairness of the comparison, we selected 12 consecutive random number seeds (2013-2024), trained 12 models for each method, and calculated the average scores of the 12 models as the final display performance in Tables 1 and 2.

3.4 Overall Comparative Results

The comprehensive comparative results with the baselines on the two datasets can be found in Tables 1 and 2. Here are our key findings:

(1) Our C^{2}MFND model consistently outperforms the optimal baseline model across various scenarios. This suggests that our model not only

[1] https://github.com/kennqiang/MDFEND-Weibo21.
[2] https://github.com/ICTMCG/M3FEND.

Method	Science	Military	Education	Disasters	Politics	Health	Finance	Ent.	Society	Total		
										F1	AUC	Acc
BiGRU	0.8076	0.8971	0.8254	0.8385	0.8698	0.8780	0.8581	0.8561	0.8647	0.8737	0.9459	0.8738
TextCNN	0.8090	0.9248	0.8781	0.8557	0.8737	0.8982	0.8731	0.8284	0.8692	0.8809	0.9480	0.8810
RoBERTa	0.7566	0.9418	0.8558	0.8626	0.8815	0.9179	0.8633	0.8800	0.8825	0.8934	0.9584	0.8935
StyleLSTM	0.8200	0.9237	0.8733	0.8710	0.8911	0.9091	0.8775	0.8920	0.8982	0.9023	0.9646	0.9024
DualEmo	**0.8788**	0.9296	0.8850	0.8895	0.8988	0.9154	0.8969	0.8981	0.9074	0.9126	0.9702	0.9127
EANN	0.7917	0.9383	0.8808	0.8678	0.8902	0.9137	0.8811	0.8788	0.8993	0.9017	0.9629	0.9018
MoSE	0.8104	0.8848	0.8594	0.8802	0.8743	0.9033	0.8488	0.8766	0.8777	0.8877	0.9533	0.8878
EDDFN	0.8291	0.9006	0.8843	0.8763	0.8822	0.9086	0.8449	0.8586	0.8743	0.8887	0.9528	0.8888
MDFEND	0.8207	**0.9430**	0.8844	0.8930	0.8900	0.9500	0.8913	0.8983	0.9126	0.9172	0.9717	0.9173
M^3FEND	0.7909	0.9382	0.8843	0.8969	0.8858	0.9496	0.8887	0.9114	**0.9199**	0.9197	0.9755	0.9197
C^2MFND	0.8432	0.9419	**0.8926**	**0.8991**	**0.9041**	**0.9525**	**0.9109***	**0.9158**	0.9186	**0.9254****	**0.9769***	**0.9255****

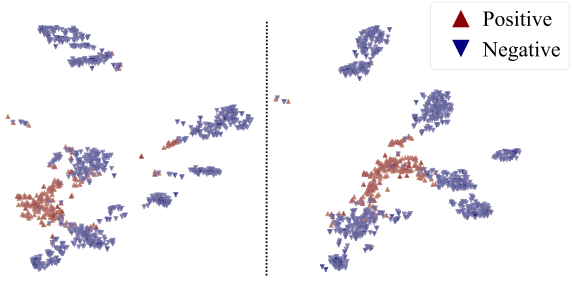
Table 1: The comparative results on Weibo21. Best result is highlighted in boldface and the runner-up is underlined. Statistical significance of pairwise differences between C^2MFND (as the best) vs. the runner-up is determined by a paired t-test (***, ** and * for p-value $< 0.001, 0.01, 0.1$, respectively).

Method	Politifact	Gossipcop	COVID	Total		
				F1	AUC	Acc
BiGRU	0.7272	0.7807	0.8850	0.8014	0.8996	0.8627
TextCNN	0.6891	0.7827	0.8803	0.8006	0.8921	0.8676
RoBERTa	0.8001	0.7934	0.9060	0.8189	0.9143	0.8783
StyleLSTM	0.7925	0.7991	0.9171	0.8258	0.9228	0.8806
DualEmo	0.7922	0.8013	0.9002	0.8236	0.9198	0.8803
EANN	0.7325	0.7852	0.8863	0.8050	0.9010	0.8673
MoSE	**0.8600**	0.8105	**0.9395**	0.8419	0.9298	0.8924
EDDFN	0.8527	0.8019	0.9308	0.8336	0.9250	0.8890
MDFEND	0.8482	0.8014	0.9164	0.8303	0.9197	0.8898
M^3FEND	0.8398	0.8112	0.9335	0.8408	0.9310	0.8919
C^2MFND	0.8192	**0.8315*****	0.9338	**0.8547*****	**0.9364****	**0.9012*****

Table 2: The comparative results on En-3. Best result is highlighted in boldface and the runner-up is underlined. Statistical significance of pairwise differences between C^2MFND (as the best) vs. the runner-up is determined by a paired t-test (***, ** and * for p-value$< 0.001, 0.01, 0.1$, respectively.)

Method	Weibo21	En-3
C^2MFND	0.9254	0.8547
Single-domain Storage	0.9193	0.8421
w/o Storage	0.9158	0.8395
w/o Multi-head Cross-attention Module	0.9149	0.8512
w/o Causal Intervention Module	0.9229	0.8502

Table 3: Results of ablation study on Weibo21 and En-3.



(a) w/o CI (b) w/ CI

Figure 3: News visualizations: (a) w/ and (b) w/o causal intervention module, on the COVID domain using t-SNE on En-3. The F1-scores are 0.9114 and 0.9392, respectively.

enhances overall detection performance but also demonstrates specific improvements within individual domains. The incorporation of domain background extraction and feature fusion through specialized cross-attention modules enriches the input features for the classifier. **(2)** Single-domain models (BiGRU, TextCNN and RoBERTa) consistently underperforms multi-domain approaches, validating the necessity of designing specialized models to address the challenges posed by the MFND problem. **(3)** EANN's failure on En-3 highlights the difficulty of extracting interdomain-shared features from highly imbalanced domains. While MoSE, EDDFN, and MDFEND address this challenge differently, C^2MFND stands out by utilizing multi-head attention to form a soft sharing network, effectively capturing both domain-shared and domain-specific features. **(4)** Both C^2MFND and M^3FEND employ external memory enhancement networks, but they serve distinct purposes. C^2MFND utilizes it as the background, while M^3FEND employs it as a knowledge bank. The Cross-attention mechanism in C^2MFND enables the integration of features from other news in the domain background, providing additional information beyond a single news article. **(5)** In domains with uneven positive rates, such as Weibo21's finance and En-3's Gossipcop, our model outperforms the optimal baseline significantly. This indicates that the causal intervention module plays a crucial role in improving detection effectiveness by addressing confounding issues.

3.5 Ablation Study

In this section, we analyze the impacts of different modules in our C^2MFND model by considering four variants: *Single-domain Storage* exclusively

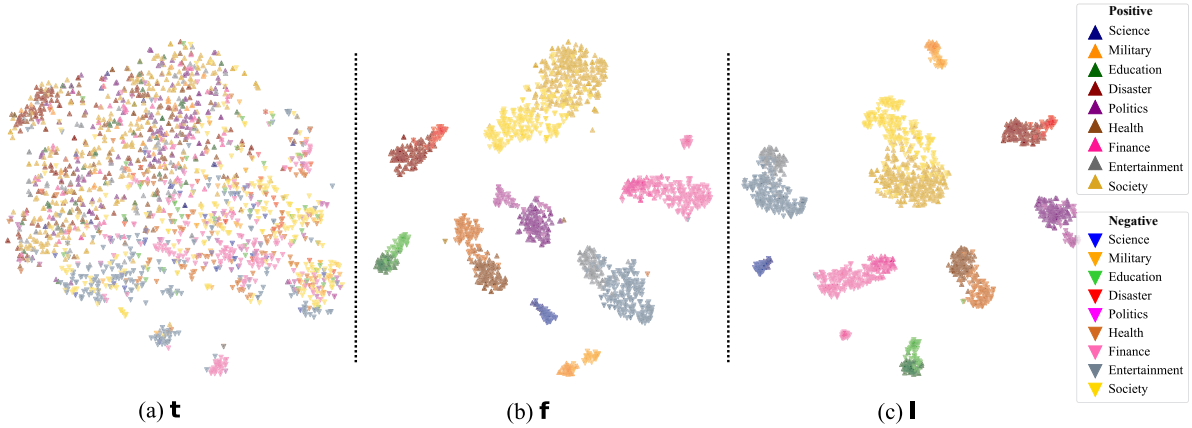


Figure 4: News visualizations in terms of different representations: (a) $\mathbf{t}$, (b) $\mathbf{f}$, and (c) $\mathbf{l}$, using t-SNE on Weibo21.

performs feature fusion on the shared domain background with shared storage among domains; *w/o Storage* eliminates the multi-domain storage module, aggregating event embeddings with the same domain label to create a domain background matrix for the corresponding domain; *w/o Multi-head Cross-attention Module* uses a linear layer instead of Cross-attention to get fusion weight matrix; and *w/o Causal Intervention Module* omits the causal intervention module.

Experimental results in terms of overall F1-scores on two datasets are present in Table 3, yielding the following observations: **(1)** Poor results after using single storage indicate that setting up storage separately for each domain is necessary, possibly because a distinct background matrix is insufficient in describing the differences among domains. **(2)** Model performance decreases notably after invalidating memory, especially on the larger En-3 dataset (longer sentences, more events), suggesting that external storage introduced as domain background knowledge is beneficial, particularly in complex tasks. **(3)** Replacing the Multi-head Cross-attention module with a common linear layer results in decreased model performance, especially on Weibo21 with a larger number of domains. This indicates that the Cross-attention module can extract more comprehensive information from diverse domain backgrounds, leading to better fusion features. **(4)** The causal intervention module significantly enhances the model's validity regarding F1-score, especially in domains with imbalanced positive rates. In the most imbalanced COVID domain in two datasets which is shown in table 5, deconfounding notably improves model performance, increasing the F1-score from 0.91 to 0.93 on average. Figure 3, illustrates that after adding the module, the overlapping area of positive and negative samples is reduced, facilitating the upper layer classifier in learning the classification patterns.

3.6 Further Model Analysis

Here, we provide a detailed analysis and explanations regarding the effectiveness of feature fusion in the domain background and causal intervention module, along with the discussion on the sensitivity of hyperparameters. For the first task, we visualize the news on Weibo21 using different representations (sentence representation $\mathbf{t}$, fused features $\mathbf{f}$ after the cross-attention module, and deconfounded fusion features $\mathbf{l}$ with the causal intervention module) through the dimension reduction method t-SNE. Refer to Figure 4 for the detailed results.

3.6.1 Feature fusion in domain background

We introduce Multi-domain storage and Cross-attention for effective feature fusion. A comparison between Figures 4(a) and 4(b) reveals that we have achieved feature fusion for different news based on the background matrix of their domains, resulting in a natural clustering of these news. The news within each domain exhibit closer relationships, while those from different domains are distantly placed. This facilitates learned representation having domain characteristics, making it convenient to learn specific detection models according to the unique characteristics of each domain.

Moreover, in the experiment, we observed that, under the same conditions of selecting optimal model parameters and early-stop criteria based on the validation set, our model achieved slightly better results on the validation set compared to other benchmark methods. However, our indicators on the test set exhibited a much smaller drop. This suggests that our model demonstrates stronger ro-

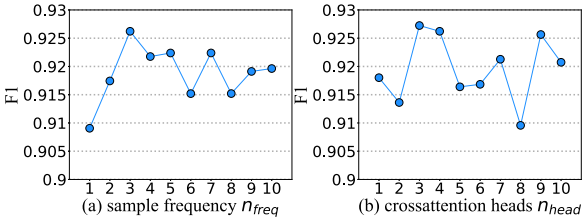


Figure 5: Performance (F1) of C^{2}MFND by varying hyperparameters (a) n_{freq} and (b) n_{head} on Weibo21.

bustness and generalization ability, attributed to the Multi-domain memory module aggregating more generalized domain background features during training, and the Cross-attention module extracting superior fusion features with a broader field of view.

3.6.2 Causal intervention

With the causal intervention module, the updated fusion features $\mathbf{f}$ effectively mitigate the confounding effect caused by uneven distributions. A comparison between Figures 4(b) and 4(c) reveals a reduction in the overlap of positive and negative samples in each domain. The detection performance of our C^{2}MFND is also enhanced with causal intervention module as shown in Figure 3

Additionally, it is noteworthy that the domains in Figure 4(c) appear to be more tightly integrated, and the number of outliers is further reduced. This is attributed to the causal intervention module, which acts as a form of clustering and provides an auxiliary benefit for Multi-domain detection.

3.6.3 Hyperparameter Sensitivity

On Weibo21, we tested the sensitivity of two important hyperparameters, including the number of heads n_{head} in the Cross-attention module, and the sampling frequency n_{freq} in the causal intervention module, as depicted in Figure 5.

It is noticeable that, with an increase in the number of heads n_{head} or the sampling frequency n_{freq}, the model's performance exhibits oscillations. n_{head} determines the size of the Cross-attention module. A larger n_{head} results in a module with more parameters and strong fitting ability, potentially leading to overfitting. Conversely, a smaller n_{head} may result in underfitting. n_{freq} represents the deconfounding "strength" of the causal inference module to some extent. A larger n_{freq} involves more scaled dot-product attention operations between the fusion feature $\mathbf{f}$ and the prior deconfounding matrix $\mathbf{D_d}$. Therefore, it is crucial to select appropriate values for n_{head} and n_{freq} to ensure optimal model performance.

4 Related Work

Our work is related to two areas of research: multi-domain fake news detection and causal inference.

Multi-domain fake news detection: to enhance model performance across multiple domains, Nan et al. (2021) employed a domain gate to aggregate representations extracted by a mixture of experts. Liang et al. (2022) constructed fuzzy domain labels for news items to discover potential labels. Additionally, Zhu et al. (2023b) proposed a Domain Memory Bank to enrich domain information for identifying potential domain labels. Focusing on news without domain labels, Wang et al. (2023) generated soft labels to improve detection performance. And, Yu et al. (2022) built a history and mainstream news environment and a perception module for mixing perceptual representations via domain gates.

Causal inference: As a powerful statistical modeling tool, it identifies true causal relationships between variables by mitigating confounding biases. Applied in various fields like recommender systems and natural language processing, causal inference includes causal intervention and counterfactual inference. For example, Zhu et al. (2023a) used causal inference to eliminate name bias in machine reading comprehension. In fake news detection, Hu et al. (2023) addressed image-text matching bias, Chen et al. (2023) introduced causal intervention to eliminate the false correlations from psychological language confounding factors, and employed counterfactual reasoning to mitigate image bias.

5 Conclusions

In this paper, to more effectively exploit domain background information and mitigate the confounding factor caused by uneven distributions in multi-domain fake news detection problem (MFND), we proposed our C^{2}MFND approach. It comprises Cross-attention and Multi-domain storage modules for better domain background information learning, and a causal intervention module for deconfounding. Extensive experiments on two datasets, En-3 and Weibo21, validate the effectiveness of our model designs and demonstrate the superior performance of our approach over the state-of-the-art baselines on MFND problem.

Limitations

In our method, the causal intervention module eliminates the non-causal correlation between the fusion feature F and news label Y, but does not eliminate the bias on classification results caused by uneven domain positive rate D itself, and this effect is affected by the quality of $\mathbf{D_d}$. In the future we will look for better methods to get better $\mathbf{D_d}$.

References

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Mozhdeh Gheini, Xiang Ren, and Jonathan May. 2021. Cross-attention is all you need: Adapting pretrained transformers for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1765.

Linmei Hu, Ziwei Chen, Ziwang Zhao, Jianhua Yin, and Liqiang Nie. 2023. Causal inference for leveraging image-text matching bias in multi-modal fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11141–11152.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Chaoqi Liang, Yu Zhang, Xinyuan Li, Jinyu Zhang, and Yongqi Yu. 2022. Fudfend: fuzzy-domain for multi-domain fake news detection. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 45–57. Springer.

Hezheng Lin, Xing Cheng, Xiangyu Wu, and Dong Shen. 2022. Cat: Cross attention in vision transformer. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1930–1939.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*, pages 3818–3824.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Piotr Przybyla. 2020. Capturing the style of fake news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 490–497.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning*, pages 1842–1850. PMLR.

Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 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*, volume 35, pages 557–565.

Bing Tian, Yixin Cao, Yong Zhang, and Chunxiao Xing. 2022. Debiasing nlu models via causal intervention and counterfactual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11376–11384.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Daokang Wang, Wubo Zhang, Wenhuan Wu, and Xiaolei Guo. 2023. Soft-label for multi-domain fake news detection. *IEEE Access*.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 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*, pages 849–857.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057. PMLR.

Wencheng Yu, Jike Ge, Zhaoxu Yang, Yan Dong, Yujie Zheng, and Haojun Dai. 2022. Multi-domain fake news detection for history news environment perception. In *2022 IEEE 17th Conference on Industrial Electronics and Applications (ICIEA)*, pages 428–433. IEEE.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, pages 3465–3476.

Jiazheng Zhu, Shaojuan Wu, Xiaowang Zhang, Yuexian Hou, and Zhiyong Feng. 2023a. Causal intervention for mitigating name bias in machine reading comprehension. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12837–12852.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2023b. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge & Data Engineering*, 35(07):7178–7191.

Figure 6: Wordcloud for nine domains on Weibo21.

A Preliminaries

In this section, we begin by introducing the composition of the causal graph and the fundamental principles of intervention through backdoor adjustment on the causal graph (Pearl et al., 2016).

A.1 Causal Graph

A causal graph is a directed acyclic graph used to depict interactions among individual variables, expressed as a graph $G = \{N, E\}$, where nodes signify variables, edges denote causality (causal relationships), and N and E represent sets of nodes and edges, respectively. In Figure 1, the arrow $F \rightarrow Y$ signifies the direct causal effect of F on Y, while D denotes the confounding factor between F and Y due to the presence of $D \rightarrow F$ and $D \rightarrow Y$.

A.2 Causal Intervention

In the presence of confounders, to accurately measure the true causal effect among variables, generally a causal intervention is performed on the causal graph. This involves eliminating all the influences on a node, ensuring that its original parent has no impact on it. Using Figure 1 as an illustration, we eliminate the confounding effect of F and Y by D and employ backdoor adjustment with do-calculus to derive, where d is the possible values of D:

$$P(Y|do(F)) = \sum_{d} P(Y|F, d)P(d) \tag{14}$$

B Dataset Statistics & Experiment Details

B.1 Statistics

Figure 6 visualizes the word frequency statistics of each domain on Weibo21.

The statistics for the Weibo21 and En-3 are present in tables 4 and 5, respectively.

Domain	Science	Military	Education	Disasters	Politics	Health	Finance	Entertainment	Society	Total
Fake	93	222	248	591	546	515	362	440	1,471	4,488
Real	143	121	243	185	306	485	959	1,000	1,198	4,640
Total	236	343	491	776	852	1,000	1,321	1,440	2,669	9,128
Positive rate	0.39	0.65	0.51	0.76	0.64	0.52	0.27	0.31	0.55	0.49

Table 4: The statistics of Weibo21.

Domain	PolitiFact	GossipCop	COVID	Total
Fake	379	5,067	1,317	6763
Real	447	16,804	4,750	22001
Total	826	21,871	6,067	28,764
Positive rate	0.46	0.23	**0.22**	0.24

Table 5: The statistics of En-3.

B.2 Baselines

We compare our approaches with three categories of baselines: (1) Single-domain methods, including **BiGRU**, **TextCNN**, and **RoBERTa**). BiGRU (Ma et al.) applies GRU to encode text information, TextCNN (Kim, 2014) employs CNN for text processing, and RoBERTa (Liu et al., 2019) is a pre-trained language model. For fake news detection on Weibo21 and En-3, RoBERTa is pre-trained separately in Chinese (Cui et al., 2020, 2021) and English. (2) Mixed-domain methods, i.e., **StyleLSTM** and **DualEmo**, treat all domains as one single domain. StyleLSTM (Przybyla, 2020) uses a BiLSTM to extract text features, incorporating both text and style features for final predictions. On the other hand, DualEmo (Zhang et al., 2021) uses a BiGRU to extract text features while also considering emotional features; (3) Multi-domain methods includes **EANN**, **MoSE**, **EDDFN**, **MDFEND**, and **M^{3}FEND**. Specifically, EANN (Wang et al., 2018) is designed to learn event invariant representations. MoSE (Ma et al., 2018) employs a mixture-of-expert (MoE) shared by each domain to extract news representations. EDDFN (Silva et al., 2021) retains both domain-specific and domain-shared features using MLPs. MDFEND (Nan et al., 2021) uses domain gates to select useful experts in MoE. M^{3}FEND (Zhu et al., 2023b), as the most state-of-art approach and a competitive baseline, fuses multi-perspective features (semantic, emotional, and stylistic) of single news under the guidance of memory.

It is important to note that, for fair comparisons, the word embeddings, emotional features, and stylistic features used in our approach and all the baselines are identical.

B.3 Model Settings

The hyperparameters for our method are present in tables 6. In our method, we set the number of training rounds to 50, early stopping to 3, batch size n_{bat} to 64, maximum truncation length n_w to 170 and 512 for Weibo21 and En-3, respectively. The optimizer is set to Adam, the initial learning rate to 5e-4, and dropout to 0.2. The dimension of the word embedding of BERT d_j is 768. TextCNN utilizes five kernels with steps $1, 2, 3, 5, 10$, and the dimension of sentence embedding d_z is set to 320. The hidden layer dimension of the MLP is set to 384. The number of Multi-head Cross-attention heads n_{head} is set to 4, the dimension of each head d_k to 0.75*320=240, and the hidden layer dimension of the feedforward network to 1*320=320. The number of Multi-domain storage units n_u is set to 8. The sampling frequency n_{freq} for the deconfounding matrix $\mathbf{D_d}$ is set to 3. The text truncation length on En-3 and Weibo21 is set to 512 and 170, respectively.

Hyperparameters	Values
epochs	50
early stopping	3 (w/o CI), 10 (w/ CI)
n_{bat}	64
n_w	170 (Weibo21), 512 (En-3)
optimizer	Adam
initial learning rate	5e-4
dropout	0.2
d_j	768
kernels for TextCNN	$\{1, 2, 3, 5, 10\}$
n_u	8
d_z	320
n_{head}	4
d_k	240
hidden layer dimension for FFN	320
n_{freq}	3
hidden layer dimension for MLP	384

Table 6: hyperparameters settings for C^{2}MFND.