Expert Systems With Applications 261 (2025) 125487

Contents lists available at ScienceDirect

Expert Systems With Applications

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

ADA-UDA: A transferable transformer framework for rumor detection using Adversarial Domain Alignment within Unsupervised Domain Adaptation

Songlin Chen [a], Xiaoliang Chen [a,b,c,*], Duoqian Miao [b], Hongyun Zhang [b], Xiaolin Qin [d], Peng Lu [c]

[a] School of Computer and Software Engineering, Xihua University, Chengdu, 610039, PR China
[b] College of Electronic and Information Engineering, Tongji University, Shanghai, 201804, PR China
[c] Department of Computer Science and Operations Research, University of Montreal, Montreal, QC, H3C3J7, Canada
[d] Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610041, PR China

ARTICLE INFO

Dataset link: https://github.com/oulaxiaoge/ADA-UDA

Keywords:
Rumor detection
Online social networks
Unsupervised domain adaptation
Parameter transferable module
Adversarial domain alignment

ABSTRACT

The rapid spread of misinformation on social media has posed significant challenges, particularly in the early detection of rumors, which is critical to mitigating their negative impact. However, the limited data available during the initial stages of trending topics poses significant challenges, including analysis constraints and model overfitting. Therefore, the adoption of specific methods are needed to identify patterns from historical sample data and enable effective knowledge transfer, facilitating inductive reasoning with limited data on new topics. To address these challenges, we propose an Adversarial Domain Alignment Unsupervised Domain Adaptation (ADA-UDA) method based on the Transformer architecture. Our approach leverages labeled historical data from the source domain alongside a limited quantity of unlabeled data from new trending topics in the target domain. At the heart of our method is the Parameter Transferable Module (PTM), which guides the Transformer to focus on transferable and distinguishable features, thereby enhancing the model's ability to perform effective inductive reasoning with limited data. We conducted extensive experiments to evaluate our method, benchmarking it against current mainstream rumor detection techniques. The findings indicate that our ADA-UDA method outperforms existing approaches, underscoring its potential for early and accurate rumor detection in emerging topics.

1. Introduction

The rapid expansion of social media has transformed how information is disseminated, making these platforms the primary channel for news and updates. This shift has its downsides — most notably the accelerated spread of rumors, leading to considerable societal disruption during crises. Given the situation's urgency, the need for efficient and rapid rumor detection mechanisms becomes undeniably critical (Lotfi et al., 2020).

The rapid spread of misinformation on social media poses significant challenges, particularly in the early detection of rumors, which is crucial for mitigating their negative impacts. For instance, misinformation linking 5G technology to the coronavirus has caused real societal harm, including the vandalism of telecommunication infrastructure. This issue extends beyond a single case. In China, the widely circulated "salt rumor" triggered a nationwide panic-buying of salt, based on the false belief that consuming salt could prevent radiation poisoning. In another case, millions of people in Shanxi Province, China, took to the streets due to an unsubstantiated rumor about an impending earthquake. Additionally, the dairy industry in China suffered significant economic losses due to rumors of "leather milk powder", which raised widespread public concern. While many news agencies and social media platforms have developed rumor reporting systems, such as Sina's misinformation management center (https://service.account.weibo.com/?type=5&status=0), Snopes (http://www.snopes.com/), and Factcheck (http://www.factcheck.org/), these efforts remain hindered by manual verification processes, which cannot keep pace with the rapid flow of information on social media. As social media continues to expand, the need for automated systems capable of quickly and accurately identifying and curbing the spread of false information becomes increasingly urgent (Gereme & Zhu, 2019; Shu et al., 2017; Zhou et al., 2019).

Rumor detection encompasses a multidisciplinary approach combining Natural Language Processing (NLP), Machine Learning (ML), and Data Mining (Oshikawa et al., 2020). Its core objective is to automatically identify false or misleading content by analyzing and processing textual data. Early research primarily focused on content-based detection methods (Ferreira & Vlachos, 2016; Gupta et al., 2014),

* Corresponding author at: School of Computer and Software Engineering, Xihua University, Chengdu, 610039, PR China.
E-mail address: chenxl@mail.xhu.edu.cn (X. Chen).

https://doi.org/10.1016/j.eswa.2024.125487
Received 15 July 2024; Received in revised form 11 September 2024; Accepted 29 September 2024
Available online 18 October 2024

such as text writing style and text-image consistency, using predefined rules and features to identify rumors. Nevertheless, social media posts are often short and might not include images, limiting the effectiveness of content-based methods. To address these issues, researchers have proposed various rumor detection techniques, ranging from traditional models utilizing handcrafted features to those leveraging deep learning. Early studies utilized handcrafted features to capture key information in rumor propagation. For example, researchers used writing style, vocabulary choices, and source credibility to identify potential rumors, with these features often based on domain knowledge and experience. The advent of deep learning has led researchers to utilize neural networks for autonomously learning rumor representations (Kuter et al., 2018). Methods like Recurrent Neural Networks (RNN) (Bugueño et al., 2019; Lin et al., 2018; Wang, Guo et al., 2019) and Convolutional Neural Networks (CNN) (Bian et al., 2020; Ebrahimi Fard et al., 2019; Guo, Tang et al., 2021; Li et al., 2019) have been employed to extract semantic information from large text datasets, enabling a more comprehensive understanding of rumor content. This automatic learning approach not only improves model performance but also enhances generalization capabilities. Recently, researchers have focused on rumor propagation structures, proposing detection models based on propagation paths and influence (Guo et al., 2018). For instance, by constructing propagation networks (Xu et al., 2022), one can analyze information dissemination paths on social media to better understand rumor mechanisms. To improve detection accuracy, the idea of multi-source heterogeneous aggregation has been introduced, integrating multiple information sources such as text and images. By considering both text content and related images (Abdelnabi et al., 2022; Xuming et al., 2023), models can gain a more comprehensive understanding of rumor events, thereby enhancing detection performance. This approach has achieved significant progress in rumor detection. In addition, the advent of pre-trained language models like GPT and BERT (Devlin et al., 2019) has introduced new opportunities for rumor detection. These models, trained on extensive corpora, encapsulate deep semantic information and can be fine-tuned for particular rumor detection tasks, greatly enhancing detection effectiveness. Despite these advances, rumor detection still faces challenges. Rumors are diverse and flexible, making feature extraction and model training difficult. Additionally, real-time detection requires quick identification and processing of rumors in large-scale data streams. Cross-language and cross-cultural variations in rumor presentation also pose challenges.

In this paper, we aim to address several challenges that arise from data imbalance and scarcity, often accompanied by linguistic diversity and rapid evolution. Additionally, previous studies have overlooked the fact that not all features are transferable or distinguishable. To address these challenges, this paper introduces Unsupervised Domain Adaptation (UDA) methods aimed at improving rumor detection and addressing the problem of limited data for trending topics. New epidemic rumors often share similarities in expression and punctuation with historical rumors. Leveraging historical rumor data features to classify epidemic rumor data is effective due to the implicit relationship between them. In the early stages with limited rumor content, embedding the BERT model in our framework generates dynamic character-level vectors for rumor texts, alleviating vocabulary limitations. To handle complex semantic features of variable-length sequences, we integrate a BiLSTM model, which combines textual information and sentence order features to better extract semantic features of rumor texts. And we introduce a fine-grained local adversarial network for feature discrimination alignment, enhancing rumor detection accuracy.

Adversarial domain alignment is an adversarial learning framework comprising a feature extractor and a domain discriminator. The feature extractor extracts features from the data, while the domain discriminator distinguishes whether the features are from the source domain or the target domain. Yaroslav Ganin first introduced the concept of adversarial learning into transfer learning, utilizing a feature extractor to extract features from the data, and training the discriminator to become unable to differentiate between features from the source and target domains (Ganin et al., 2015). This alignment enables domain adaptation. However, when there is a significant distributional difference between the source and target domains, using a single feature extractor and adversarial network may result in poor model stability during the feature alignment process. To address this, Eric Tzeng et al. proposed a domain adaptation method. They initially trained a feature extractor and classifier on the source domain data using supervised learning (Tzeng et al., 2017). A separate feature extractor was then constructed to extract features from the target domain. Both feature extractors fed their outputs into the discriminator, and when the discriminator could no longer distinguish between the two domains, the target domain feature extractor, along with the pre-trained classifier, formed the final model to perform the target domain task. Previous works relied on a single feature extractor and adversarial learning to align features between domains, which may lead to domain-specific features being ignored or misaligned. To overcome this, Zhang, Tang et al. (2019) proposed a symmetric domain adaptation method, where features from the source and target domains are kept symmetric across multiple layers of the model. They also introduced the concept of a private-shared feature space, allowing both domains to maintain domain-specific features while learning a shared feature space where domain alignment is achieved.

In our work, we first achieve global feature alignment through global adversarial learning, which serves as a coarse-grained alignment. This approach enhances feature transferability to some extent, but its performance is limited, especially when there is a significant discrepancy between the target and source domains. The limitation arises because not all fine-grained feature representations that make up the global features are inherently transferable or distinguishable. Given that the sequential nature of Transformer models naturally provides finer-grained representations, we introduce local adversarial learning to focus the model on transferable and distinguishable fine-grained features, thereby improving the representation capacity of global features.

This study introduces a novel rumor detection framework utilizing the Transformer architecture's ADA-UDA method. By considering the diversity and complexity of social media content, our method overcomes existing limitations, providing an innovative and effective solution for addressing uncertainties and complexities in information dissemination. Our experimental results demonstrate that ADA-UDA outperforms existing methods across various unsupervised domain adaptation settings.

The main contributions of this paper are as follows:

- As far as we know, this is the first to incorporate adversarial networks within the Transformer architecture for UDA in rumor detection. This paper proposes a transferable Transformer framework for adversarial domain-aligned rumor detection based on unsupervised domain adaptation. The model aligns deeper features learned from the source domain with the target domain, addressing the issue overlooked by previous methods – that not all features are transferable or distinguishable – thereby enabling the model to learn higher-quality transferable features. We believe this work can provide a valuable reference for exploring the use of Transformer in other UDA tasks.
- We also introduced weight factors for global and local adversarial losses in the balanced loss function, allowing control over the weights of global and local adversarial losses simultaneously. This effectively mitigates the model's overfitting in domain feature alignment. Additionally, we fully leverage the inherent features of the Transformer by proposing the PTM module, which captures fine-grained transferable and discriminable local features, thereby enhancing the efficiency of UDA.
- Our method effectively utilizes historical rumor data features to detect the veracity of early trending topics. The proposed transferable Transformer model exhibits superior performance, surpassing benchmarks. Experiments on Chinese and English datasets

revealed that our model attained an accuracy of 81.67% for the Chinese dataset and 80.79% for the English dataset. Compared to existing models, our model's performance improved by 7.11% to 29.82%, demonstrating its effectiveness and providing a direction for future research.

This paper follows this structure: Section 2 reviews previous research on rumor detection. Section 3 outlines our proposed model architecture and its key components. Section 4 presents experimental results, comparing our method with benchmark models and analyzing the findings. Section 5 concludes the study and provides a summary of the work.

2. Related work

Rumors, as defined by Gist (1951), are unverified explanations or reasons targeting public concerns. In rumor detection, limitations of manual feature extraction, such as time consumption and feature bias, have shifted the focus towards deep learning methods. Notably, approaches based on RNN (Bugueño et al., 2019; Lin et al., 2018; Wang, Guo et al., 2019) and CNN (Bian et al., 2020; Ebrahimi Fard et al., 2019; Guo, Tang et al., 2021; Li et al., 2019) have shown significant promise in enhancing rumor detection capabilities.

Several studies have utilized sequential models to extract features from multimedia content in social media posts. Many existing models focus only on text features. For instance, Ma et al. introduced an RNN (Ma et al., 2016) to capture continuous representations of Weibo events, effectively capturing time-series information from original posts, retweets, and comments to automatically discern hidden features. Building on this, Chen et al. embedded an attention mechanism within the RNN to emphasize temporal hidden features from sequential posts, assisting in early rumor detection (Chen et al., 2018). Chen et al. proposed a novel hybrid model XGA (namely XLNet-based Bidirectional Gated Recurrent Unit (BiGRU) network with Attention mechanism) for Cantonese rumor detection on Twitter (Chen et al., 2020). First, XLNet produces text-based and sentiment-based embeddings at the character level. Then, perform joint learning of character and word embeddings to obtain the words' external contexts and internal structures. Leverage BiGRU and the attention mechanism to obtain important semantic features and use the Cantonese rumor dataset to train the model.

In addition to text-based methods, recent research has explored multimodal approaches to better understand social media practices. For instance, multimodal methods (Jin et al., 2017; Khattar et al., 2019; Wang et al., 2018) propose models that integrate image and text features from posts to detect rumors. This fusion approach enables models to gain a more comprehensive understanding of rumor events, significantly enhancing detection performance.

Additionally, to validate the authenticity of posts, some studies employ external resources. For instance, Fang et al. combined multimodal data and knowledge graphs to enhance reasoning capability and improve accuracy in rumor detection by supplementing background knowledge and semantic connections (Fang et al., 2019). Hierarchical attention mechanisms have also been employed to enhance detection accuracy. Lan et al. used a hierarchical attention-based RNN model to detect rumors on social media, effectively distinguishing rumors from non-rumors by automatically extracting key semantic and temporal features (Lan et al., 2018). Another innovative approach, introduced by Khan et al. employed bidirectional graph convolutional networks within a deep learning framework to detect rumors on social media (Khan et al., 2024).

Methods focusing on propagation structures have also garnered considerable attention. Rao et al. proposed the LGAM-BERT model, which uses hierarchical attention masks on BERT to detect rumors, leveraging comments as auxiliary features and reducing language noise by limiting interaction between posts and comments in the lower layers (Rao et al., 2021). On the other hand, researchers have also focused on methods based on propagation structures. Ma et al. utilized tree structures for modeling information propagation and introduced a propagation tree kernel method to distinguish rumors from other information (Ma et al., 2017). They subsequently enhanced this approach by incorporating a tree-structured Recursive Neural Network (RvNN) to represent and process rumor propagation structures (Ma et al., 2018). Additionally, Lu et al. introduced Graph-aware Co-Attention Networks (GCAN), a neural network model that combines user features and tweet encoding features to predict information authenticity (Lu & Li, 2020). Fang et al. proposed the Kernel Graph Attention Network (KGAT), which combines edge and point kernels with attention mechanisms based on BERT, integrating Graph Convolutional Networks (GCN) and hierarchical attention mechanisms (Fang et al., 2015). This method significantly improves fact-checking accuracy and model interpretability. Another notable framework for rumor detection (Tu et al., 2021), introduced by Tu et al. merges text representation learning with propagation structure learning. This framework constructs a large propagation graph, integrating the propagation structures of all tweets, then uses network embedding methods to learn node vector representations and employs a convolutional neural network to simultaneously learn features of textual content and propagation structures, thereby enhancing rumor detection performance.

Despite these advances, several challenges remain. One major challenge is data imbalance, where labeled data for actual rumors and non-rumors on social media is often unequal, affecting model generalization. Additionally, the diversity and rapid evolution of social media language pose difficulties for model training and generalization. Lu et al. proposed a model called Subjective Information Enhanced Reinforcement Learning (SIFTER) using multi-task learning to assimilate external knowledge explicitly detailing rumors (Lu et al., 2022). This model employs reinforcement learning combined with existing rumor detection models and implements a sequential training mode to address propagation inconsistency issues, enhancing the trained model's robustness to noisy comments. The SIFTER framework improves detection accuracy and real-time performance and excels in cross-domain and continuous prediction scenarios. Liu et al. proposed the Dual-Attention GCN (DAGCN) method, combining dual-attention mechanisms with GCN for rumor detection on social networks (Liu et al., 2023). This method constructs event propagation graphs and uses GCN to retrieve structural information from every event-related tweet. It then integrates these data with features from the original posts to create interactive semantic text features. Additionally, attention mechanisms are utilized to minimize false and irrelevant interactions. But if the data is scarce or of poor quality, their performance may decline because the effectiveness of both models relies heavily on the training data's quality and richness.

Further advancements in rumor detection have explored semi-supervised and unsupervised methods to overcome the lack of labeled data. For instance, Alzanin et al. employed a semi-supervised and unsupervised expectation–maximization algorithm to identify rumors in Arabic tweets, using limited labeled data along with a large amount of unlabeled data to improve detection accuracy (Alzanin & Azmi, 2019). In contrast, our research focuses on unsupervised rumor detection. Ran et al. employed contrastive learning and cross-attention techniques to bring feature representations of the same class samples from different domains closer while pushing different class samples apart, achieving unsupervised cross-domain rumor detection (Ran & Jia, 2023). However, their method presumes that the source and target domains have an identical class space and necessitates clustering techniques to create pseudo-labels for the target domain, potentially introducing noise. Guo et al. developed a rumor detection system using explainable adaptive learning. This system builds dynamic classifiers with Graph-based Adversarial Learning (GAL) concepts and detailed feature spaces employing graph-level encoding, and incorporates ongoing adversarial training between the generator and unsupervised decoders to tackle scenarios with insufficient labeled training data (Guo, Yu et al., 2021).

Fang et al. proposed an unsupervised rumor detection technique utilizing the Propagation Tree Variational AutoEncoder (PTVAE) to tackle the issue of scarce reliable pre-labeled datasets (Fang et al., 2023). This method captures high-order propagation patterns and reconstructs trees through message-passing strategies, encoding and decoding trees, aligning multiple modalities such as tree structures and propagation features to output final predictions. Zhang et al. introduced a novel approach called the Multimodal Disentangled Domain Adaptation Method (MDDA), which incorporates multimodal disentangled representation learning and UDA (Zhang et al., 2021). This method breaks down multimedia posts into distinct content and rumor-related features. It removes elements unique to the content and filters out event-specific attributes, while preserving common rumor characteristics across different events. In addition to these methods, Xiao et al. proposed a graph-based contrastive learning self-supervised LSTM model (Xiao et al., 2023). This model constructs positive and negative sample pairs and utilizes graph neural networks to capture complex data relationships, extracting useful features from large amounts of unlabeled data to improve detection accuracy. Although this study focuses on Medicare fraud detection, its self-supervised learning framework effectively captures data structures and behavior patterns, providing valuable insights and references for rumor detection.

Generative models have been utilized in rumor detection. Yang et al. addressed rumor detection on social media using generative models by considering news veracity and user credibility as latent random variables and utilizing user interactions to assess news authenticity (Yang et al., 2019). Initially, a Bayesian network model captures the relationships among news veracity, user opinions, and user credibility. Then, an interaction graph is constructed between users and news, using comments and retweets to model their opinions and credibility. Lastly, The approach employs collapsed Gibbs sampling to deduce news veracity and user credibility, eliminating the requirement for labeled data. Zhang et al. built upon Yang et al.'s approach by incorporating tweet authorship, a critical factor in tweet propagation (Zhang, Wang et al., 2019). A common drawback of these methods is their failure to recognize that not all features are transferable and discriminable. Thus, it is crucial to focus on both capabilities during the transfer process.

This section provides an overview of existing approaches to rumor detection, discussing both text-based and multimodal methods. It also addresses the limitations of previous studies, such as their inability to capture transferable and discriminable features effectively, which leads to the justification for the proposed model.

3. Problem definition

3.1. Rumor detection

In this study, a 'rumor' is articulated as information propagated through Online Social Networks (OSN) that is not substantiated by corroborating evidence and typically exerts a considerable adverse impact on the collective societal fabric. A rumor encompasses elements that are officially confirmed as false information and parts that remain unverified or undisclosed, as illustrated in Fig. 1. The process of rumor detection can be defined as the methodological delineation of protocols designed to identify and categorize information on OSNs as either spurious or credible. This involves parsing the messages into two distinct categories: rumors and non-rumors. The current methodologies for detecting rumors can be categorized into two primary types: identifying rumors in individual posts or in clusters of posts (Wang et al., 2020). The model proposed herein is specifically designed to detect rumors in individual posts.

Formally, the set $S = (s_1, s_2, \ldots, s_l, \ldots, s_r)$ delineates the sample spectrum of the entire post collection, where $s_l = (w_1, w_2, \ldots, w_i, \ldots, w_n)$ typifies the content of an individual post s_l encompassing n characters, and each w_i encapsulates the ith word, numeral, or symbol. Our model aims to train a classification function F that transmutes s_l into a feature vector m^f in f-dimensional space, simplified hereafter as m. The study employs f-dimensional matrix vectors universally across the realm of rumor detection, ultimately distilling the evaluation to a binary classification of the data into non-rumors (0) and rumors (1), as shown in Eq. (1).

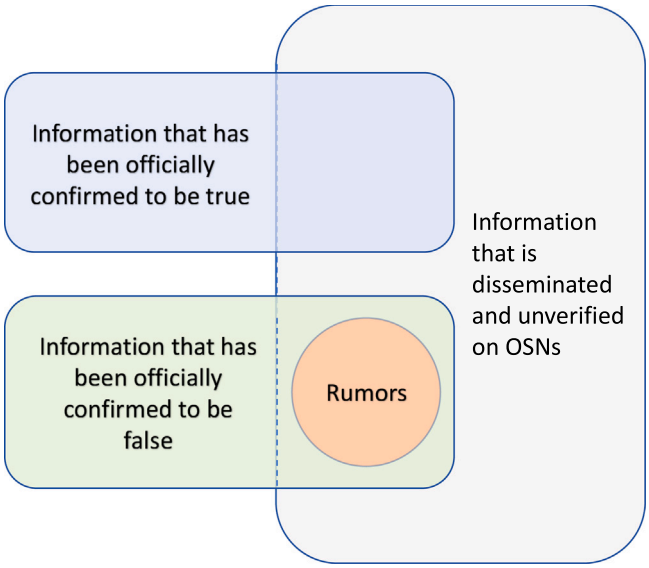

Fig. 1. Structure of rumors in OSN.

$$F_{(s_l)} = \begin{cases} 0, & \text{if } s_l \text{ is not a rumor} \\ 1, & \text{others} \end{cases} \tag{1}$$

3.2. Dataset division

The data division in this study differs from traditional methods in certain respects. The UDA methodology primarily divides the data into three distinct categories: domain adaptation network training, domain adaptation network validation, and testing. The dataset under consideration comprises B types of data, each with $C + E$ instances. Of these, C instances are labeled, while E instances are unlabeled. In this context, the support set comprises $B \cdot C$ instances, while the query set contains $B \cdot E$ instances. Training in the source domain takes place on the support set and is validated there, whereas testing in the target domain occurs on the query set. The objective is to migrate features acquired from multiple events in the source domain to the target domain, thereby achieving domain adaptation.

3.3. Transferable transformer UDA

The proposed transformer-based ADA-UDA model is capable of accurately determining the veracity of emerging topics and performing classification tasks for the purpose of detecting rumors. The UDA task employs the labeled source domain $D_B \left\{ \left(s_c, y_c \right) \right\}_{c=1}^{N_s}$ derived from its historical process. In this context, s_c represents the source domain sample features, y_c denotes the classification targets, and N_s specifies the quantity of training samples. With regard to the unlabeled target domain $D_t \left\{ \left(s_j \right) \right\}_{N_s+1}^{r}$ of trending topics, s_j represents the target domain sample features, and $N_s + 1 \sim r$ denotes the limited number of samples within the target domain. The primary objective of UDA is to identify and learn distinguishing and invariant features between the two domains. These transferable features are then integrated into the multi-head attention module, ensuring accurate classification in the rumor detection task.

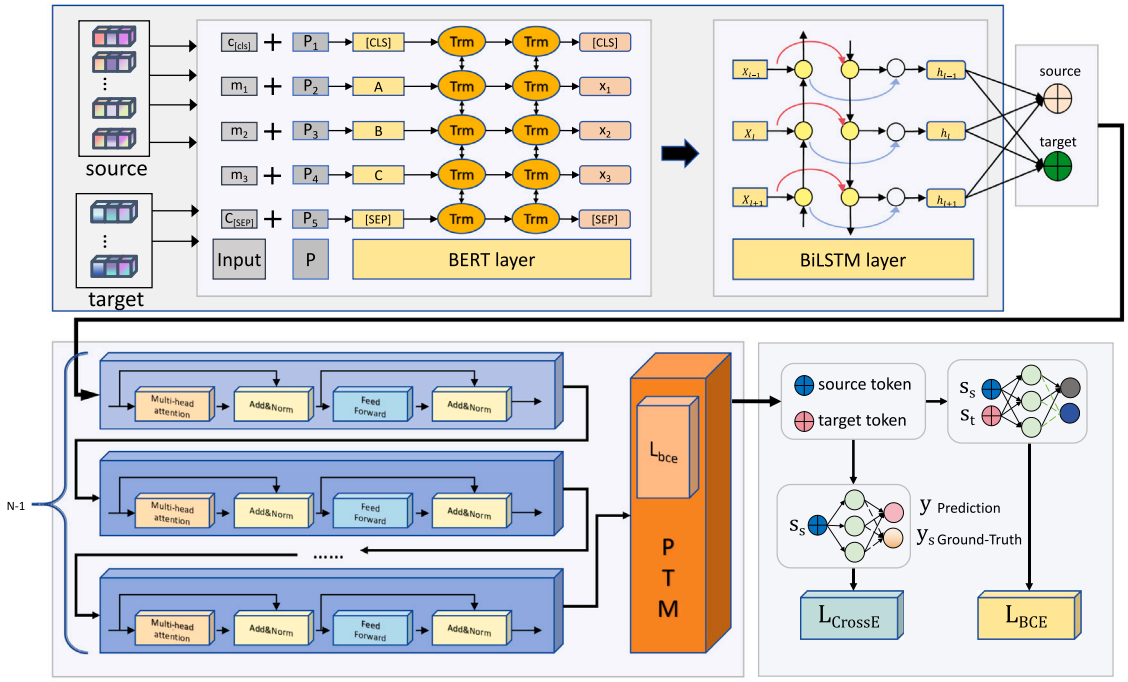

Fig. 2. Overview of the ADA-UDA model: Consists of three parts: the BERT + BiLSTM combination module, the parameter transferability module (PTM), and the adversarial alignment with global and local discriminators and the MLP classifier.

The problem definition clarifies the context in which rumor detection is performed, with a focus on Online Social Networks (OSNs). It outlines the goal of classifying messages into rumor and non-rumor categories and explains the structure of the dataset used in this study.

4. Methodology

This section aims to develop an ADA-UDA model framework based on the Transformer architecture, designed to identify whether a post is a rumor. This section provides a comprehensive description of the overall model, detailing the specific structure of each module. The specific explanations for the notations can be found in Table A.8 in the Appendix.

4.1. Overview

Fig. 2 depicts the overall architecture of the proposed ADA-UDA model, consisting of three main components. The initial module is a combination of BERT and BiLSTM, which converts input data from the source and target domains into vectors. This includes both labeled historical rumor data and unlabeled trending topic data (discussed in Section 4.2). The second part of the model involves global adversarial analysis and the final classification task, which is facilitated by a domain discriminator that encourages the formation of a domain-invariant feature space (discussed in Section 4.3). The third module is the PTM, which injects learned transferable features into the attention module, guiding the Transformer to focus on transferable and discriminable features (discussed in Section 4.4).

4.2. Embedding module based on BERT+BiLSTM

Research indicates that combined models generally outperform single-network models in feature extraction and text representation (Cheng et al., 2020; Guha et al., 2020). In our ADA-UDA model, the BERT model generates dynamic character-level vector representations of rumor text. This approach mitigates the issue of insufficient vocabulary for new epidemic rumors during the input stage. By integrating the BiLSTM model, we combine text information and sentence order features to derive semantic features of rumor texts. This combined model can handle variable-length sequences, representing more complex semantic features.

We utilize a pre-trained BERT model for character-level embeddings to convert textual features into vectors. Initially, the text is split into individual characters, each assigned a sequential positional encoding, commencing with 1. A [CLS] token is added at the beginning (position 0), and a [SEP] token is appended at the end. This positional encoding facilitates feature extraction in subsequent stages.

Given an input text of a single post $s_l = (w_1, w_2, \ldots, w_i, \ldots, w_n)$, the goal is to derive a vector representation $m = (x_1, x_2, \ldots, x_i, \ldots, x_n)$. The core framework of BERT, utilizing the bidirectional Transformer encoder structure, passes the input text through the initial sublayer, the multi-head self-attention layer, and then to the fully connected feedforward layer. Each sublayer incorporates normalization and residual connections, with the overall output of each sublayer computed according to Eq. (2):

$$Output_{sublayer} = LayerNorm(x + sublayer(x)) \tag{2}$$

The function $sublayer(x)$ represents the sublayer's specific function. In order to achieve effective residual connections, the outputs of all sublayers and embeddings in the model are set to $d = 512$ dimensions.

In the self-attention submodule, three vectors are used: query matrix (q), key matrix (k), and value matrix (v). Each submodule takes a set of (q), (k), and (v) as input, calculates the similarity between each word, and determines the corresponding weights based on the similarity. This enables each word to obtain related information from other words, thereby learning dependencies between different positions. The process involves the calculation of dot products between keys and queries, which are then subjected to the application of the $softmax$ function in order to obtain value weights. The attention layer is computed as described in Eq. (3):

$$Attention(q, k, v) = softmax\left(\frac{qk^T}{\sqrt{d_k}}\right)v \tag{3}$$

where d_k is the dimension of the (q) and (k). To comprehensively compute attention, a multi-head attention mechanism is applied, linearly mapping the input to generate query, key, and value matrices.

Each matrix calculates the scaled dot-product attention for every input sentence, with the results called heads.

The attention matrices are concatenated horizontally and multiplied by a weight matrix W^0 to compress them into a single matrix, allowing attention to different spatial representations at different positions. The calculations are presented in Eqs. (4) and (5):

$$head_i = Attention(qW_i^{Qm}, kW_i^{Km}, vW_i^{Vm}) \tag{4}$$

$$MultiHead(q,k,v) = Concat(head_1, head_2, \dots, head_i, \dots, head_m)W^0 \tag{5}$$

where $W_i^{Q_m}$, $W_i^{K_m}$, and $W_i^{V_m}$ are the weight matrices for the q, k, and v of the ith head, respectively. $Concat$ denotes the function concatenating multiple heads, and W^0 is the weight matrix used in concatenation. The multi-head attention layer's output is sent to a fully connected feedforward network, incorporating multiple activation functions to produce the final output. The calculations are presented in Eq. (6):

$$Output_{FN} = dropout(RELU(W_{MAL} * Attention(q,k,v)+b_{MAL}))W_{FF}+b_{FF} \tag{6}$$

where W_{MAL} represents the weights of the multi-head attention, and b_{MAL} represents its biases. The feedforward network layers have weights denoted as W_{FF} and biases denoted as b_{FF}. Finally, the output from BERT is fed into the BiLSTM model to extract features and mine information. The BiLSTM model comprises the following elements: character input x_t, the internal state C_t, the input state C_{t_z}, hidden state h_t, forget gate f_t, memory gate m_t, and the output gate o_t, all at time t. The internal calculation process of the BiLSTM model is detailed as follows, with the forget gate calculation given by Eq. (7):

$$f_t = \delta(W_f x_t + U_f h_{t-1} + b_f) \tag{7}$$

where δ is the activation function, and W_f, U_f, and b_f denote the parameters of the forget gate. The output of the hidden layer at time $t-1$ is h_{t-1}.

The memory gate calculation at time t is given by Eq. (8). C_{t_z} is mainly derived from the hidden layer output of the previous time step $t-1$ and the input at t, undergoing a linear transformation and tanh activation to produce a new state value, as shown in Eq. (9). The previous internal state at $t-1$ is then updated to C_t at t, as detailed in Eq. (10). The calculation process for the output gate o_t is provided in Eq. (11). The parameters for the input gate are W_i, U_i, and b_i; for the input state are W_c, U_c, and b_c; and for the output gate are W_o, U_o, and b_o.

$$m_t = \delta(W_i x_t + U_i h_{t-1} + b_i) \tag{8}$$

$$C_{t_z} = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{9}$$

$$C_t = C_{t-1} \cdot f_t + C_{t_z} \cdot m_t \tag{10}$$

$$o_t = \delta(W_o x_t + U_o h_{t-1} + b_o) \tag{11}$$

At time t, the hidden states of the forward and backward LSTM networks are $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$, respectively, as shown in Eqs. (12) and (13). The final hidden state h_t is computed as shown in Eq. (14). The hidden states $H = (h_1, h_2, \dots, h_T)$ of T serve as the input to the softmax classifier, resulting in a weight vector W calculated as shown in Eq. (15):

$$\overrightarrow{h_t} = o_{t-1} * \tanh(C_{t-1}) \tag{12}$$

$$\overleftarrow{h_t} = o_{t+1} * \tanh(C_{t+1}) \tag{13}$$

$$h_t = Concat(\overleftarrow{h_t}, \overrightarrow{h_t}) \tag{14}$$

$$W = softmax(W_{ss1} \tanh(W_{ss2} H^T) + b_s) \tag{15}$$

where $W_{ss1} \in \mathbb{R}^{d_s \times 2u}$ and $W_{ss2} \in \mathbb{R}^{d_s}$ are weight matrices, d_s is a hyper-parameter, u denotes the size of the hidden state of the unidirectional LSTM, and b_s is a bias. Thus, the final single text vector m is represented as shown in Eq. (16):

$$m = \sum_{t=1}^{T} W_t \cdot h_t \tag{16}$$

Subsequently, the text vectors from both the source and target domains are transmitted to the transferable Transformer module.

4.3. Global adversarial feature alignment

The Transformer architecture has achieved significant success in NLP tasks, showcasing exceptional performance across various language applications such as text classification and machine translation (Devlin et al., 2019; Zhou et al., 2020). This success is primarily attributed to the attention mechanism's feature extraction capabilities. However, the transferability of parameters and the incorporation of adversarial domain alignment in the Transformer architecture for UDA methods in rumor detection have not yet been explored. This paper focuses on analyzing knowledge transfer in the Transformer-based UDA method for rumor detection, leveraging the Transformer's multi-head self-attention mechanism to capture long-term dependencies. Vaswani et al. (2017). To accurately predict unlabeled target data, a common approach in domain adaptation tasks is to further optimize the loss function as shown in Eq. (17), aiming to enhance joint feature learning, domain adaptation, and classifier learning.

$$L_{c1}(s_s, y_s) + \beta L_{adv}(s_s, s_t) \tag{17}$$

where L_{c1} is a supervised classification loss, L_{adv} is a transfer loss, and β is a parameter used to control the importance of L_{adv}. In this paper, the L_{adv} adopted is an adversarial loss, which primarily forms an invariant feature space through the domain discriminator (Ganin et al., 2015).

The encoder H_f is employed initially for feature learning, predominantly within the parameter-transferable Transformer, while H_c serves as the classifier for classification. The domain discriminator D_d is utilized for global feature alignment, focusing on the output states of class tokens in both the source and target domain text content. To enhance domain knowledge adaptation, this study implements a max–min game strategy between H_f and D_d, where H_f learns domain-invariant features to deceive D_d, and D_d aims to distinguish features between the source and target domains. The supervised classification loss is calculated as shown in Eq. (18), and the global adversarial loss for the parameter transferability module is calculated as shown in Eq. (19).

$$L_{c1}(s_s, y_s) = \frac{1}{N_s} \sum_{s_m \in D_B} L_{CrossE}(H_c(H_f(s_m)), y_m) \tag{18}$$

$$L_{adv}(s_s, s_t) = -\frac{1}{r} \sum_{s_\Delta \in D} L_{BCE}(D_d(H_f(s_\Delta)), y^*) \tag{19}$$

where $D = D_B \cup D_t$, s_s and s_t represent data from the source domain data and target domains, respectively, and y_s represents the source domain labels. The subscript of s_Δ can indicate either domain. y^* represents the domain label ($y^* = 1$ for the source domain and $y^* = 0$ for the target domain). L_{CrossE} and L_{BCE} denote the loss functions for cross-entropy and binary cross-entropy, respectively. Ablation experiments have shown that while global adversarial alignment improves transferability to some extent, it does not fully exploit the parameter transferability capabilities within the Transformer. To address more complex rumor detection scenarios, this study introduces fine-grained local adversarial alignment. During feature transfer, not all encoded positions have transferable and discriminable features. The Transformer's sequential feature transmission naturally provides finer-grained features. Thus, incorporating fine-grained local adversarial alignment allows a focus on both transferable and discriminable features. To solve the above issues, a PTM is further proposed, detailed in Section 4.4.

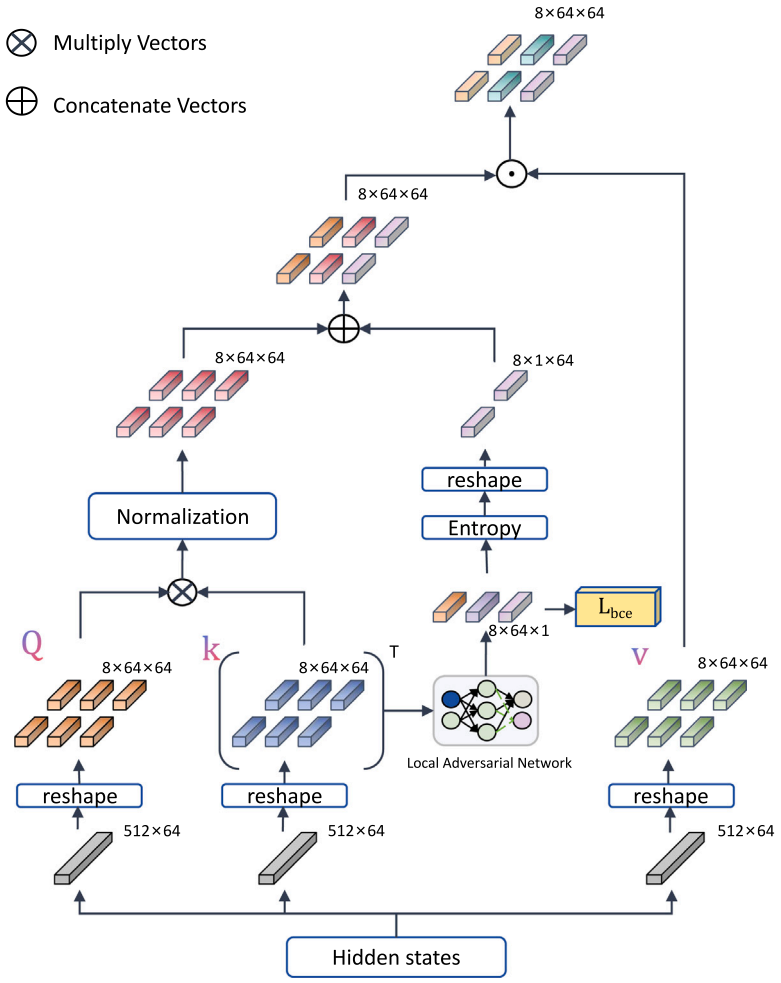

Fig. 3. Framework of the PTM.

4.4. Transferability feature module

This section analyzes the implementation of parameter-transferable features within the attention module, driving the Transformer to focus on transferable and discriminable features. To study the domain adaptation capabilities of the Transformer, we train our backbone network on labeled data from source domain and partially unlabeled data from the target domain. We validate the transferability performance using results on target data obtained from the detection of rumors. The detailed framework of the PTM is shown in Fig. 3.

The module primarily focuses on capturing transferable and semantically significant features. PTM leverages the intrinsic attention mechanism and sequential feature transmission advantages of the Transformer. Since tokens in the input text are treated as local features with fine-grained representations, the module can capture various aspects of features. Tokens in the input text have different semantic importance and transferability. The PTM module assigns varying weights to these tokens to guide the learned text representation, namely the output state of the class token, towards transferable and discriminable tokens. Although self-attention weights in the Transformer can serve as discriminative weights, a challenge arises as the transferability of each token is not inherently available. To address this issue, inspired by the ideas in Pei et al. (2018) and Wang, Li et al. (2019), we introduce a token-level domain discriminator D_{tl} to align cross-domain local features within the PTM module. The local adversarial loss calculation for the parameter transferability module is shown in Eq. (20).

$$L_{tok}(s_s, s_t) = -\frac{1}{rn}\sum_{s_\Delta \in D}\sum_{N=1}^{n} L_{bce}(D_{tl}(H_f(s_{\Delta N})), y_N^*) \tag{20}$$

where n represents the number of tokens, and $D_{tl}(\bullet)$ denotes the probability that the local region belongs to the source domain. During adversarial learning, D_{tl} assigns a label of 1 to source domain tokens and a label of 0 to target domain tokens. Essentially, after H_f learns domain-invariant features to deceive D_{tl} (for example, if D_{tl} is around 0.5, the token is easily deceived, indicating stronger transferability between domains and thus requiring better transferability assignment). Therefore, this study uses the standard entropy function $G(\bullet)$ to measure transferability, replacing $G(\bullet)$ with *Entropy* in Fig. 3, and using $G(D_{tl}(H_f(s_{\Delta_N}))) \in [0, 1]$ to evaluate the transferability of the Nth token of the Δth text.

By transforming the traditional multi-head attention mechanism into a transferable multi-head attention mechanism (T-MultiHead), we can inject learned transferable features into the class token attention weights. The T-MultiHead is built on the transferable attention mechanism (T-Attention), with the specific calculation shown in Eq. (21).

$$T - Attention = softmax\left(\frac{Qk^T}{\sqrt{d_k}}\right) \odot [1; G(k_{token})]v \tag{21}$$

Here, Q is the class token query, In the earlier sections of this paper, q refers to the query matrix, which differs from Q here, where it represents the class token query. k_{token} is the token key, $\odot$ is the Hadamard product, and $[;]$ indicates concatenation. Therefore, softmax $\left(\frac{Qk^T}{\sqrt{d_k}}\right)$ can represent the semantic importance of each token, while $[1; G(k_{\text{token}})]$ signifies each token's transferability. To jointly

capture transferability at various positions, the T-MultiHead calculation process is redefined as shown in Eq. (22), with $head_i$ redefined in Eq. (23).

$$T-MultiHead(Q,k,v)=Concat(head_1,head_2,\ldots,head_i,\ldots,head_m)W^o \tag{22}$$

$$head_i=T-Attention(QW_i^{Qm},kW_i^{Km},vW_i^{Vm}) \tag{23}$$

Applying the PTM in the last layer of the Transformer enables it to focus on fine-grained, transferable features essential for our classification task. The final PTM calculation result is presented in Eq. (24), where LN stands for layer normalization, and N indicates the total number of Transformer layers.

$$\begin{aligned} z^N &= MLP(LN(T-MultiHead(LN(z^{N-1}))+z^{N-1})) \\ &\quad +T-MultiHead(LN(z^{N-1}))+z^{N-1} \end{aligned} \tag{24}$$

After the above analysis, the target loss function of the proposed model is shown in Eq. (25),

$$Loss_{global}=L_{c1}(s_s,y_s)+\beta L_{adv}(s_s,s_t)+\lambda L_{tok}(s_s,s_t) \tag{25}$$

where the hyperparameters β and λ control the importance of the corresponding losses. The overall process of the model is provided in pseudocode 1.

Algorithm 1 Overall precedure of ADA-UDA

Input: Source domain, $D_B\left\{\left(s_c,y_c\right)\right\}_{c=1}^{N_s}$;
Target domain, $D_t\left\{\left(s_j\right)\right\}_{N_s+1}^{r}$;
Dimension of a query and key vector, d_k;
Sublayer and embedding layer dimension, d;
Output: classifier

$s_c,y_c,s_j=preprocess(S)$
for each post s_l in S **do**
  m^f = BERT + BiLSTM encoder s_l;
  obtain m^f of s_c, m^f of s_j;
end for
for $c\in 1,2,3,\ldots,N_s, j\in N_s+1,\ldots,r$ **do**
  assign m^f of s_c to D_B,assign m^f of s_j to D_t;
end for
input m^f from D_B and D_t into the $N-1$ layer of the transformer;
learn and extract features;
input to the PTM module;
use the H_f encoder to learn the features;
perform local adversarial alignment using the local domain discriminator D_{tl} ;
D_{tl} assigns the label of 1 to tokens in D_B and the label of 0 to tokens in D_t;
evaluate each token's transferability using the standard information entropy function $G(\cdot)$;
PTM ←Equation (24);
perform global adversarial domain alignment using the D_d global domain discriminator;
return classifier;

5. Experimental evaluation

This section details the datasets used in our experiments, the baseline techniques applied, and the parameter settings for the experiments. Following this, we analyze the results of the rumor detection experiments and assess the effectiveness of the modules through ablation experiments.

Table 1
The statistic of real-world datasets.

	DatasetCN			DatasetEN		
	HD	ED	All	HD	ED	All
RumorNum	4321	321	4642	2574	222	2796
NonRumorNum	4628	420	5048	3079	374	3453
Min.Length	41	30	30	35	16	16
Max.Length	275	167	275	143	129	143
Ave.Length	177	156	175	137	119	135

5.1. Data description

We assessed the ADA-UDA model's performance in rumor detection using datasets in both Chinese and English. The specific data statistics are presented in Table 1. During data preprocessing, we applied several quality control procedures to guarantee the integrity and reliability of the data. These included the removal of replies with fewer than three words and user replies containing more than 70% emoji. Additionally, we deleted a significant number of irrelevant posts.

For testing rumor detection on Chinese datasets, we employed the rumor dataset provided by Song et al. (2019) as our historical rumor dataset. In addition, we collected historical rumor data from the Sina Weibo False Information Reporting Platform to expand the dataset. The dataset CHECKED, provided by Yang et al. (2021), was used to construct the data set of rumors pertaining to the 2019 novel coronavirus (COVID-19), also known as the 2019-nCOV or simply the novel coronavirus. Additionally, rumors officially recognized by the Sina Community Management Center were collected to build the corpus of COVID-19 rumors.

For testing rumor detection on English datasets, we used the publicly available Twitter dataset and parts of the FakeNewsNet dataset. The Twitter dataset primarily comprises samples from Twitter15 and Twitter16, supplemented with rumor events from the Snopes platform to build the historical rumor dataset. Additionally, we constructed the COVID-19 rumor dataset using data from Cheng et al. (2021), which includes news web pages and Twitter data, with all rumors verified by fact-checking websites.

5.2. Experimental setup

5.2.1. Baseline methods

To more effectively assess the performance of our UDA method, we compared it against the following baseline methods. Each baseline experiment involved training on the source domain and evaluation on the target domain:

(1) **FastText** (Joulin et al., 2017): A text classification model that uses the bag-of-words method to vectorize text while considering word order between sentences. N-grams are used to vectorize adjacent words as auxiliary features to improve classification accuracy.
(2) **TextCNN** (Kim, 2014): A convolutional neural network model for text classification that effectively extracts local features and contextual information through convolution and pooling operations, enhancing classification performance.
(3) **TextRNN** (Liu et al., 2016): A text classification method using a recurrent neural network (RNN). After embedding the text samples to obtain word vector representations, the RNN model is used for modeling, and the output is mapped to the category space through a fully connected layer.
(4) **Att_TextRNN** (Zhou et al., 2016): A text classification method combining BiLSTM and attention mechanisms. It models text sequences through BiLSTM to capture long-distance dependencies and contextual information, and uses the attention mechanism to learn the importance weights of different parts of the text, focusing on information that contributes to the classification task, thus enhancing classification performance.

(5) **Transformer** (Vaswani et al., 2017): The Transformer model uses self-attention mechanisms to directly focus on different positions in the input sequence, capturing long-distance dependencies more effectively. It significantly improves parallel computing capabilities, training, and inference speed, and has achieved outstanding results in various NLP tasks.
(6) **KDCN** (Sun et al., 2023): A knowledge-guided dual consistency network for multimodal rumor detection, combining visual and textual information with entity information from external knowledge bases to achieve efficient performance in rumor detection tasks.
(7) **EBGCN** (Wei et al., 2021): A reinforced Bayesian graph convolutional network model for rumor detection incorporates an edge enhancement mechanism and Bayesian inference to model propagation uncertainty, thus boosting the performance and reliability of rumor detection.

Furthermore, an examination was conducted on three variants of ADA-UDA, with the objective of evaluating the effectiveness of the various model components.

(1) **ADA-UDA-a**: In the text representation part, ADA-UDA-a ignores the BERT embedding module and uses only BiLSTM to model variable-length sequences for rumor detection.
(2) **ADA-UDA-b**: In the feature transfer part, ADA-UDA-b ignores the PTM module and uses only the multi-head attention within the Transformer for feature transfer to complete rumor detection.
(3) **ADA-UDA-c**: In the global adversarial analysis part, ADA-UDA-c ignores global adversarial alignment to complete rumor detection.

5.2.2. Evaluation metrics

In this study, four metrics are used to evaluate the comparative methods: Accuracy, Precision, Recall, and Macro F1. Macro F1 is a balanced measure that addresses the issue of label imbalance, which is a common challenge in text classification.

5.2.3. Parameter settings

This section outlines the parameter settings involved in the ADA-UDA model. We implemented the ADA-UDA model using Python 3.8 and the Pytorch 1.11.0 deep learning framework. The parameters are as follows: $Dropout = 0.1$, $Train\_batch\_size = 500$, $Eval\_batch\_size = 64$, $Test\_batch\_size = 64$, $Num\_train\_epochs = 80$, $Learning\_rate = 3e{-}6$, and $Max\_seq\_length = 140$.

The experimental setup outlines the datasets, baseline methods, evaluation metrics, and parameter settings used to validate the proposed model. It provides the necessary details to replicate the experiments and understand the comparison with other state-of-the-art methods.

5.3. Experimental results

5.3.1. Overall performance

Tables 2 and 3 show the Accuracy, Precision, Recall, and Macro F1 scores of the proposed model compared to seven deep learning models on both the Chinese dataset (DatasetCN) and the English dataset (DatasetEN). Overall, our proposed ADA-UDA method shows superior performance compared to baseline methods on both datasets. The results demonstrate that our model, integrating UDA and adversarial domain alignment, enhances rumor detection performance. Bold text highlights the best results for each evaluation metric.

Experiments showed that the ADA-UDA method achieved an accuracy of 81.67% on the Chinese dataset and 80.79% on the English dataset. Compared to the seven existing models, ADA-UDA significantly improved rumor detection performance by 7.11% to 29.82%. This improvement stems from the ADA-UDA model's effective extraction of shared semantic features by combining a transferable attention module with extensive historical rumor data and limited trending rumor data. Among the other models, EBGCN performed best due to its edge-enhanced mechanism, which captures local relationships between different words more effectively. In contrast, Transformer had the worst performance, with accuracies of 51.85% on DatasetCN and 53.50% on DatasetEN. This result indicates that Transformer have difficulty capturing process dependencies on small-scale datasets. TextCNN outperformed TextRNN, achieving classification accuracies of 59.54% and 55.02% on the two datasets, respectively. TextCNN's ability to capture long-range dependencies through multiple convolutional layers gives it an advantage over RNN, which suffers from gradient vanishing issues and lacks the capability to capture complex features. Att_TextRNN improved performance by effectively capturing long-range dependencies with LSTM and using an attention mechanism to identify the most relevant parts of the input sequence for the task, achieving accuracies of 71.51% and 67.52% on DatasetCN and DatasetEN, respectively. KDCN, employing dual consistency learning and external knowledge guidance to extract complex features, achieved classification accuracies of 72.64% on DatasetCN and 71.32% on DatasetEN.

5.3.2. Ablation experiments

Fig. 4 shows the results of ADA-UDA and its variants in terms of Accuracy, Precision, Recall, and Macro F1.

We first evaluated the performance of ADA-UDA and ADA-UDA-a on both datasets. The variant ADA-UDA-a achieved accuracies of 79.29% on DatasetCN and 78.52% on DatasetEN, which are 2.38% and 2.27% lower than those of ADA-UDA, respectively. The experimental results indicate that BERT's bidirectional structure captures the left and right context at each time step, providing a more comprehensive semantic representation and improving rumor detection accuracy.

Subsequently, we assessed the performance of ADA-UDA and ADA-UDA-b. ADA-UDA outperformed ADA-UDA-b by 5.15% on DatasetCN and 5.47% on DatasetEN, demonstrating the superior performance of the PTM module. The PTM module effectively gathers both local transferable features and discriminable features, enabling finer-grained domain alignment and enhancing rumor detection capability.

Finally, we conducted a comparative analysis of the performance of ADA-UDA and ADA-UDA-c. The results demonstrated that ADA-UDA outperformed ADA-UDA-c by 3.75% on DatasetCN and 4.11% on DatasetEN. The incorporation of the global adversarial component into the ADA-UDA model resulted in the creation of an invariant feature space, leading to enhanced performance.

5.3.3. The impact of hyperparameters

We tested the sensitivity of ADA-UDA to the hyperparameters β and λ on two sets of cross-domain datasets. As shown in Fig. 5, β and λ range between (0,1) and are grouped in our experiments. In Fig. 5(a), the horizontal axis represents a group of β values from top to bottom, with λ fixed at its optimal value. In Fig. 5(b), the horizontal axis represents a group of λ values from top to bottom, with β fixed at its optimal value. ADA-UDA shows different sensitivities on different datasets. For instance, when λ is fixed at its optimal value, the model achieves the best performance on the Chinese dataset at $\beta = 0.6$ and on the English dataset at $\beta = 0.4$. When β is fixed at its optimal value, the model reaches the best performance on both Chinese and English datasets at $\lambda = 0.4$.

5.3.4. Impact of epoch changes on training and validation accuracy

We analyzed the impact of epoch variations on training and validation accuracy across two cross-domain datasets. As shown in Fig. 6, with epoch $\in [1, 100)$, the Chinese dataset Fig. 6(a) demonstrates that accuracy increases with the number of epochs during training, stabilizing around epoch 65. During validation, accuracy also improves with more epochs, peaking at epoch 69. Similarly, for the English dataset Fig. 6(b), training accuracy rises with more epochs, stabilizing

Table 2
Comprision with baseline in DatasetCN.

Method	Accuracy	Precision	Recall	Macro F1
Transformer (Vaswani et al., 2017)	0.5185	0.5508	0.5736	0.4935
TextRNN (Liu et al., 2016)	0.5786	0.5187	0.6202	0.4218
TextCNN (Kim, 2014)	0.5954	0.5934	0.6284	0.5467
FastText (Joulin et al., 2017)	0.6211	0.6339	0.6358	0.6071
Att_TextRNN (Zhou et al., 2016)	0.7151	0.6979	0.7395	0.7006
KDCN (Sun et al., 2023)	0.7264	0.7336	0.7378	0.7285
EBGCN (Wei et al., 2021)	0.7456	0.7579	0.7581	0.7331
ADA-UDA	**0.8167**	**0.8092**	**0.8132**	**0.8146**

Table 3
Comprision with baseline in DatasetEN.

Method	Accuracy	Precision	Recall	Macro F1
Transformer (Vaswani et al., 2017)	0.5350	0.6195	0.5850	0.5219
TextRNN (Liu et al., 2016)	0.5500	0.5917	0.6323	0.4451
TextCNN (Kim, 2014)	0.5502	0.6230	0.6331	0.5293
FastText (Joulin et al., 2017)	0.6250	0.6216	0.6583	0.4823
Att_TextRNN (Zhou et al., 2016)	0.6752	0.6614	0.7086	0.6693
KDCN (Sun et al., 2023)	0.7132	0.7328	0.7196	0.7035
EBGCN (Wei et al., 2021)	0.7352	0.7438	0.7449	0.7326
ADA-UDA	**0.8079**	**0.7983**	**0.8017**	**0.8031**

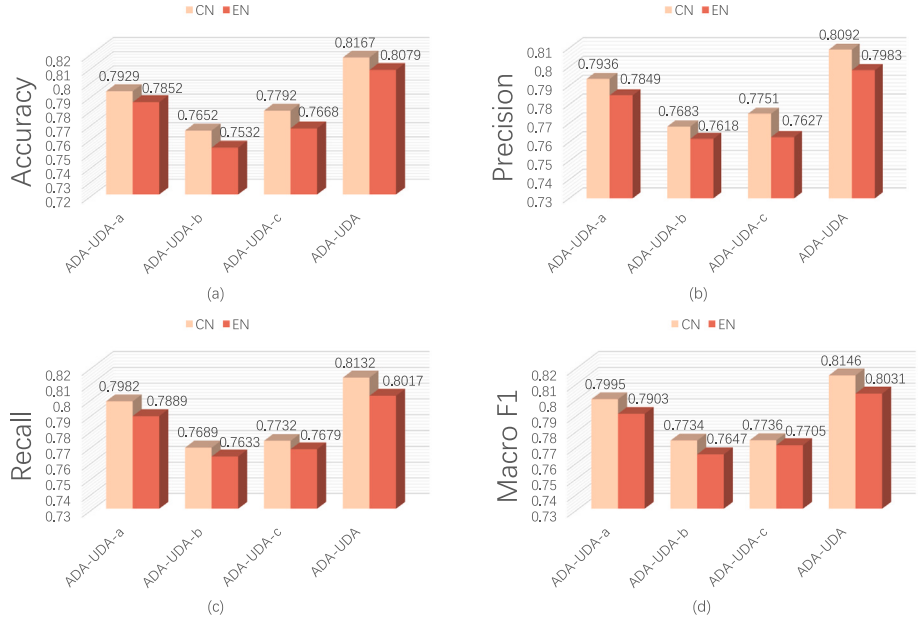

Fig. 4. Performance analysis of ADA-UDA and its variants on two datasets.

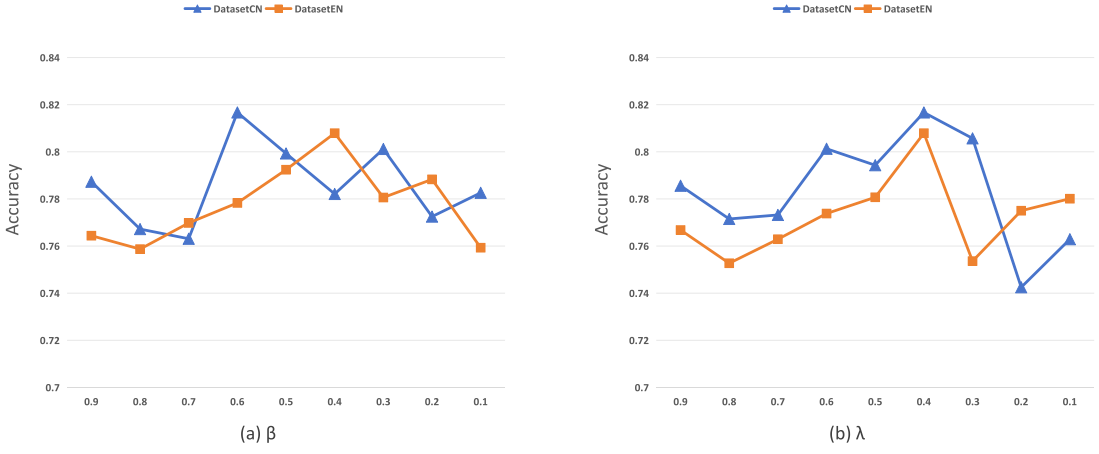

Fig. 5. The Impact of hyperparameters β and λ on ADA-UDA performance.

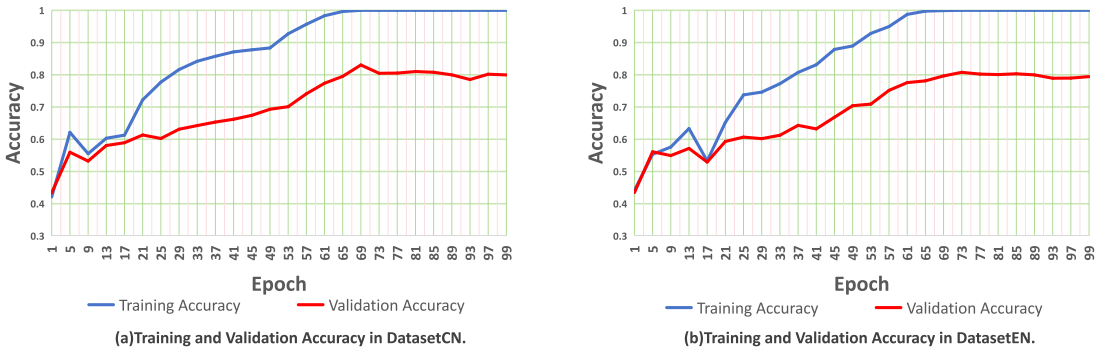

Fig. 6. Impact of epoch changes on training and validation accuracy.

around epoch 69. Validation accuracy reaches its optimal level at epoch 75.

This section presents the results of the experiments, comparing the performance of ADA-UDA with existing models. It includes detailed discussions of the ablation studies, the impact of hyperparameters, and the influence of the number of epochs on model performance.

6. Conclusion and future work

This study addresses the critical issue of rumor detection during epidemic outbreaks, where the rapid spread of misinformation can significantly disrupt social order. The development of high-precision detection methods is crucial for effectively preventing and controlling the dissemination of rumors. Given the limited availability of early-stage trending topic data and the uncertainty of their veracity, it is imperative to leverage traditional historical data to identify the truthfulness of such topics. In order to achieve this, we propose the ADA-UDA model for the detection of rumors, which employs both local and global adversarial networks in order to identify early-stage epidemic rumor posts.

The experimental results demonstrate the following key findings:

(1) The proposed ADA-UDA model significantly enhances rumor detection performance, particularly in scenarios with limited trending data lacking relevant labels. The model outperforms several state-of-the-art models in terms of Accuracy, Precision, Recall, and Macro F1. Specifically, the model achieved the best performance with an Accuracy of 81.67%, Precision of 80.92%, Recall of 81.32%, and Macro F1 of 81.46% on the Chinese dataset (DatasetCN). This represents an improvement of approximately 9.59% in Accuracy, 7.11% in Precision, 7.36% in Recall, and 8.15% in Macro F1 compared to the best-performing baseline models. On the English dataset (DatasetEN), the model reached Accuracy of 80.79%, Precision of 79.83%, Recall of 80.17%, and Macro F1 of 80.31%, showing improvements of 9.44% in Accuracy, 7.40% in Precision, 7.31% in Recall, and 8.04% in Macro F1. These substantial improvements across all metrics validate the model's effectiveness and robustness.
(2) The local adversarial alignment component of the model markedly improves rumor detection performance by effectively identifying and extracting local features, which are crucial for distinguishing between truthful and deceptive information.
(3) The global adversarial component employs domain discriminators to form invariant feature spaces, thereby achieving better stability and generalization across different domains.

During epidemic outbreaks, the proliferation of rumor posts among vast amounts of information can cause significant public anxiety. The proposed ADA-UDA model offers a robust solution for detecting rumors on OSNs such as Sina Weibo. This study not only provides a practical tool for rumor detection but also sets a new research direction for future work in this field. Future research could explore the integration of additional data sources and the refinement of adversarial components to further enhance detection accuracy and adaptability.

CRediT authorship contribution statement

Songlin Chen: Conceptualization, Data curation,Formal analysis,Software,Supervision,Visualization, Writing – original draft, Writing – review & editing. **Xiaoliang Chen:** Funding acquisition, Investigation, Methodology, Project administration, Writing – review & editing. **Duoqian Miao:** Funding acquisition. **Hongyun Zhang:** Validation. **Xiaolin Qin:** Funding acquisition, Resources. **Peng Lu:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the Science and Technology Program of Sichuan Province, PR China (Grant no. 2023YFS0424), the National Key R and D Plan "Key Special Project of Cyberspace Security Governance", PR China (No. 2022YFB3104700), the National Natural Science Foundation, PR China (Grant nos. 62076182, 61976158, 62376198), the Science and Technology Service Network Initiative, PR China (No. KFJ-STS-QYZD-2021-21-001), and the Talents by Sichuan provincial Party Committee Organization Department, PR China, and Chengdu - Chinese Academy of Sciences Science and Technology Cooperation Fund Project, PR China (Major Scientific and Technological Innovation Projects).

Appendix

A.1. Summary of related works

The method of related works is summarized in Table A.4

A.2. Comparison tables

To maintain the flow and clarity of the main text, we have placed three key tables in the appendix for further reference. These tables provide additional detailed experimental results. Table A.5 presents the results of the ablation experiments. Table A.6 shows the impact of key hyperparameters(α and β) on the model's performance. Table A.7 displays the evolution of training and validation accuracy on DatasetCN and DatasetEN across different epochs.

Table A.4
Summary of related work methods.

Methods	Strengths	Weaknesses
RNN-RD (Ma et al., 2016)	Accurately capture temporal sequences and detect dynamic features	Demands extensive data for training
Call attention to rumors (Chen et al., 2018)	Employs a soft attention mechanism to highlight the temporal hidden representations across consecutive posts	Depends on the integrity of the temporal data
CHRD-T (Khattar et al., 2019)	Generates text-based and emotion-based embeddings at the character level	Unable to fully extract the emotional features of some rumors
Mvae (Jin et al., 2017)	Discovers cross-modal associations by learning shared representations between text and images	Fails to incorporate social context or event-specific information
att-RNN (Wang et al., 2018)	Proposes an end-to-end attention-based RNN leveraging multimodal content	The feature alignment between text and images is not clear
Eann (Fang et al., 2019)	Integrated an event discriminator and a fake news detector	The fusion of text and image features is merely a simple concatenation
MKEMN (Lan et al., 2018)	Utilizing an event memory network to measure the differences between various events has improved performance and generalization in detecting new events	Relies on external knowledge graphs to supplement the background knowledge of the text
MSV-RNN (Khan et al., 2024)	Utilizes a hierarchical attention mechanism to detect content that is rich and textually distinct	High computational overhead
Rpf-gcns (Rao et al., 2021)	Bidirectional graph convolutional networks provide a more comprehensive understanding of rumor propagation structures	High computational complexity
STANKER (Ma et al., 2017)	The integration of BERT helps mitigate the interference from noisy information	Complex preprocessing
PSKL (Ma et al., 2018)	Using propagation trees to represent rumor spread paths captures the dynamic characteristics of rumor propagation more effectively	Depends on the quality of the propagation tree
TRNN (Lu & Li, 2020)	Leverages a tree-structured RNN to effectively capture the hierarchical relationships within rumor propagation	Depends on the quality of the data
GCAN (Fang et al., 2015)	Integrates graph convolutional networks with co-attention mechanisms to model latent user interactions, exhibiting robust performance in processing short texts	Depends on the accuracy of the graph structure
Word-of-mouth understanding (Tu et al., 2021)	Provides a visualization of content and opinions associated with event entities	The emotional interpretation of some images may be biased
Rumor2vec (Lu et al., 2022)	Employs a joint graph approach to mitigate the problem of feature sparsity	Depends on the quality of the preprocessing
Sifter (Liu et al., 2023)	Introducing external subjective information effectively addresses the issue of domain shift	Both training and inference demand substantial computational resources
DAM-GCN (Alzanin & Azmi, 2019)	By extracting noise-resistant features, the model is better equipped to manage complex relationships within propagation graphs	Depends on the quality of the propagation graph
SSUM (Ran & Jia, 2023)	Combines semi-supervised and unsupervised expectation–maximization algorithms	Depends on the choice of initial parameters
UCD-CLCA (Guo, Yu et al., 2021)	Incorporates cross-domain contrastive learning and cross-attention mechanisms	Depends on the accuracy and quality of the generated pseudo-labels
EAL-FDS (Fang et al., 2023)	Integrates generative adversarial networks with graph embeddings	The training process lacks stability
PTVAE (Zhang et al., 2021)	Incorporates sentiment analysis and propagation features	Demands substantial computational resources
MDDA (Yang et al., 2019)	Separating the content and rumor-specific features of multimedia posts solves the issue of highly entangled event content	The discriminative power of the visual style space is lower than that of the textual style space
GCLF (Zhang, Wang et al., 2019)	Combines Medicine Graph, GCN, and contrastive learning	Limited by the Medicine Graph
UFND (Morone et al., 2016)	Not rely on pre-labeled datasets	Performs poorly in handling complex textual content
INCC (Wang et al., 2020)	Employs the improved network constraint coefficient to quantify local node advantages while maintaining global connectivity through tenacity	The parameter selection process is complex

Table A.5
Ablation experiments: performance comparison for different methods.

Methods	CN				EN			
	Accuracy	Precision	Recall	Macro F1	Accuracy	Precision	Recall	Macro F1
ADA-UDA-a	0.7929	0.7936	0.7982	0.7995	0.7852	0.7849	0.7889	0.7903
ADA-UDA-b	0.7652	0.7683	0.7689	0.7734	0.7532	0.7618	0.7633	0.7647
ADA-UDA-c	0.7792	0.7751	0.7732	0.7736	0.7668	0.7627	0.7679	0.7705
ADA-UDA	0.8167	0.8092	0.8132	0.8146	0.8079	0.7983	0.8017	0.8031

Table A.6
Accuracy on DatasetCN and DatasetEN for different values of β and α.

Value	Accuracy for β		Accuracy for α	
	DatasetCN	DatasetEN	DatasetCN	DatasetEN
0.9	0.7873	0.7644	0.7856	0.7668
0.8	0.7672	0.7587	0.7715	0.7527
0.7	0.7631	0.7698	0.7732	0.7629
0.6	0.8167	0.7783	0.8013	0.7738
0.5	0.7993	0.7924	0.7943	0.7807
0.4	0.7821	0.8079	0.8167	0.8079
0.3	0.8012	0.7806	0.8057	0.7536
0.2	0.7725	0.7883	0.7425	0.7775
0.1	0.7826	0.7593	0.7629	0.7801

Table A.7
Training and validation accuracy across epochs for DatasetCN and DatasetEN.

Epoch	DatasetCN		DatasetEN	
	Training accuracy	Validation accuracy	Training accuracy	Validation accuracy
1	0.4213	0.4345	0.4415	0.4347
5	0.6214	0.5598	0.5534	0.5611
9	0.5553	0.5324	0.5759	0.5494
13	0.6032	0.5806	0.6332	0.5716
17	0.6127	0.5893	0.5327	0.5293
21	0.7223	0.6133	0.6524	0.5932
25	0.7772	0.6024	0.7377	0.6067
29	0.8163	0.6311	0.7463	0.6019
33	0.8421	0.6425	0.7724	0.6125
37	0.8575	0.6534	0.8074	0.6431
41	0.8711	0.6621	0.8312	0.6322
45	0.8776	0.6743	0.8788	0.6683
49	0.8833	0.6931	0.8895	0.7041
53	0.9272	0.7011	0.9324	0.7091
57	0.9567	0.7411	0.9497	0.7517
61	0.9832	0.7738	0.9874	0.7758
65	0.9965	0.7952	0.9975	0.7812
69	1.0000	0.8305	0.9992	0.7965
73	1.0000	0.8047	1.0000	0.8077
77	1.0000	0.8053	1.0000	0.8022
81	1.0000	0.8101	1.0000	0.8007
85	1.0000	0.8079	1.0000	0.8031
89	1.0000	0.7998	1.0000	0.7997
93	1.0000	0.7854	1.0000	0.7898
97	1.0000	0.8018	1.0000	0.7944
99	1.0000	0.7997	1.0000	0.7944

A.3. N-gram analysis

We provide detailed N-gram plots to further analyze the performance of different models on both the DatasetCN and DatasetEN, as shown in Fig. A.7. These plots illustrate the classification accuracy across various N-gram settings, ranging from 1-gram to 5-gram. The visualizations allow for a clear comparison of how different models—such as Transformer, TextRNN, TextCNN, FastText, AttTextRNN, KDCN, EBGCN, and ADA-UDA—perform when classifying verified and unverified false information. The results demonstrate that the ADA-UDA model consistently achieves superior accuracy across different N-gram settings, highlighting its effectiveness in rumor detection tasks.

A.4. Main notations used in this paper

Main notations as shown in Table A.8

A.5. Accuracy Comparison at Different Percentages of Training Data

Accuracy Comparison at Different Percentages of Training Data as shown in Table A.9.

A.6. Performance of Large Language Models

Performance of Large Language Models as shown in Table A.10.

A.7. Performance comparison with different models

The newly added comparison models from 2023 and 2024, which address data imbalance and involve attention mechanisms, are shown in Table A.11.

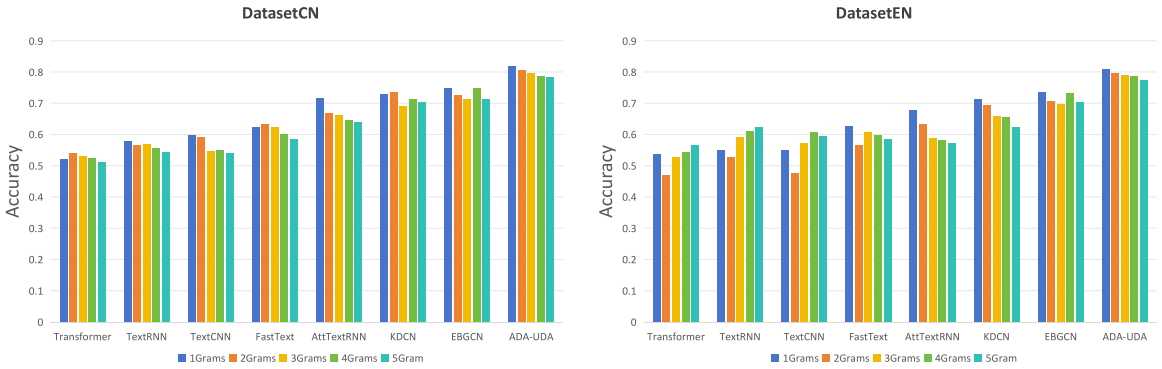

Fig. A.7. N-gram analysis of classification performance across models on DatasetCN and DatasetEN.

Table A.8
Main notations used in this paper.

Notation	Description
S	Information of the entire post sample
s_l	Information of the lth post sample
w_i	ith character in a post
m^f	f-dimensional matrix–vector
F	Prediction function
B	Number of classes contained in an episode
C	Number of labeled instances in each class
E	Number of unlabeled instances in each class
D_B	Source domain space
D_t	Target domain space
s_c	Input source domain sample
s_j	Input target domain sample
N_s	Number of posts in the source domain
r	Total number of posts
y_c	Output classification target
x_i	Vectorized representation of the ith character in a post
q	Query matrix–vector
k	Key matrix–vector
v	Value matrix–vector
d_k	Dimension of a query and key vector
W_i	Weight matrices corresponding to the ith header
W	Weight
b	Bias value
C_t	Storage cell state of BiLSTM
C_{t_z}	Temporary cell state of BiLSTM
h_t	Hidden state
f_t	Forget gate
m_t	Memory gate
O_t	Output gate
δ	Activation function
H^T	T hidden states of the whole BiLSTM
y^*	Domain label
HD	Historical rumor dataset
ED	Epidemic rumor data
s_s	Data from the source domain
s_t	Data from the target domain
H_c	Representation classifier
H_f	Feature encoder
D_d	Domain discriminator
D_{tl}	Local feature domain discriminator
Q	Query matrix vector of class token
$G(\cdot)$	Standard information entropy function
n	Number of tokens

Table A.9
Accuracy comparison at different percentages of training data.

Our model	Accuracy (50%)	Accuracy (60%)	Accuracy (70%)	Accuracy (80%)	Accuracy (90%)	Accuracy (100%)
ADA-UDA (DatasetCN)	0.7386	0.7458	0.7623	0.7841	0.8014	0.8167
ADA-UDA (DatasetEN)	0.7249	0.7391	0.7604	0.7739	0.7903	0.8079

Table A.10
Performance of large language models on DatasetCN and DatasetEN.

Method	DatasetCN				DatasetEN			
	Accuracy	Precision	Recall	MacroF1	Accuracy	Precision	Recall	MacroF1
LLaMA-13B	0.5339	0.5017	0.5142	0.4936	0.5548	0.5219	0.5425	0.5023
Claude	0.5631	0.5509	0.5240	0.5326	0.5513	0.5460	0.5311	0.5460
ChatGPT	0.6204	0.6392	0.6317	0.6228	0.6437	0.6345	0.6338	0.6182

Table A.11
Performance comparison for different methods on DatasetCN and DatasetEN.

Method	DatasetCN				DatasetEN			
	Accuracy	Precision	Recall	MacroF1	Accuracy	Precision	Recall	MacroF1
PN+CNN (Wang et al., 2023)	0.6822	0.6617	0.6724	0.6542	0.6634	0.6734	0.6441	0.6215
GNN+PHEME (Bilal et al., 2024)	0.7318	0.7409	0.7416	0.7211	0.7201	0.7351	0.7226	0.7130
ADA-UDA	0.8167	0.8092	0.8132	0.8146	0.8079	0.7983	0.8017	0.8031

Data availability

Data are available for download at the following web links. https://github.com/oulaxiaoge/ADA-UDA.

References

Abdelnabi, S., Hasan, R., & Fritz, M. (2022). Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *2022 IEEE/CVF conference on computer vision and pattern recognition* (pp. 14920–14929). http://dx.doi.org/10.1109/CVPR52688.2022.01452.

Alzanin, S. M., & Azmi, A. M. (2019). Rumor detection in arabic tweets using semi-supervised and unsupervised expectation-maximization. *Knowledge-Based Systems*, *185*, http://dx.doi.org/10.1016/J.KNOSYS.2019.104945.

Bian, T., Xiao, X., Xu, T., Zhao, P., Huang, W., Rong, Y., & Huang, J. (2020). Rumor detection on social media with bi-directional graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, 549–556. http://dx.doi.org/10.1609/aaai.v34i01.5393.

Bilal, I. M., Nakov, P., Procter, R., & Liakata, M. (2024). Generating zero-shot abstractive explanations for rumour verification. arXiv preprint arXiv:2401.12713.

Bugueño, M., Sepulveda, G., & Mendoza, M. (2019). *An empirical analysis of rumor detection on microblogs with recurrent neural networks* (pp. 293–310). http://dx.doi.org/10.1007/978-3-030-21902-4_21.

Chen, X., Ke, L., Lu, Z., Su, H., & Wang, H. (2020). A novel hybrid model for cantonese rumor detection on Twitter. *Applied Sciences*, *10*(20), http://dx.doi.org/10.3390/app10207093.

Chen, T., Li, X., Yin, H., & Zhang, J. (2018). Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In M. Ganji, L. Rashidi, B. C. M. Fung, & C. Wang (Eds.), *Trends and applications in knowledge discovery and data mining* (pp. 40–52). Cham: Springer International Publishing.

Cheng, M.-Y., Kusoemo, D., & Gosno, R. A. (2020). Text mining-based construction site accident classification using hybrid supervised machine learning. *Automation in Construction*, *118*, Article 103265, URL https://api.semanticscholar.org/CorpusID:219758051.

Cheng, M., Wang, S., Yan, X., Yang, T., Wang, W., Huang, Z., Xiao, X., Nazarian, S., & Bogdan, P. (2021). A COVID-19 rumor dataset. *Frontiers in Psychology*, *12*, Article 644801.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1423, URL https://aclanthology.org/N19-1423.

Ebrahimi Fard, A., Mohammadi, M., Chen, Y., & Walle, B. (2019). Computational rumor detection without non-rumor: A one-class classification approach. *IEEE Transactions on Computational Social Systems*, *PP*, 1–17. http://dx.doi.org/10.1109/TCSS.2019.2931186.

Fang, L., Feng, K., Zhao, K., Hu, A., & Li, T. (2023). Unsupervised rumor detection based on propagation tree VAE. *IEEE Transactions on Knowledge and Data Engineering*, *35*(10), 10309–10323. http://dx.doi.org/10.1109/TKDE.2023.3267821.

Fang, Q., Qian, S., & Xu, C. (2019). *Multi-modal knowledge-aware event memory network for social media rumor detection* (pp. 1942–1951). http://dx.doi.org/10.1145/3343031.3350850.

Fang, Q., Xu, C., Sang, J., Hossain, M. S., & Muhammad, G. (2015). Word-of-mouth understanding: Entity-centric multimodal aspect-opinion mining in social media. *IEEE Transactions on Multimedia*, *17*, 1. http://dx.doi.org/10.1109/TMM.2015.2491019.

Ferreira, W., & Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 1163–1168). San Diego, California: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N16-1138, URL https://aclanthology.org/N16-1138.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. S. (2015). Domain-adversarial training of neural networks. In *Journal of machine learning research*. URL https://api.semanticscholar.org/CorpusID:2871880.

Gereme, F. B., & Zhu, W. (2019). Early detection of fake news "before it flies high". In *Proceedings of the 2nd international conference on big data technologies* (pp. 142–148). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3358528.3358567.

Gist, P. N. P. (1951). Rumor and public opinion. *American Journal of Sociology*, *57*(2), 159–167.

Guha, R., Ghosh, M., Singh, P. K., Sarkar, R., & Nasipuri, M. (2020). A hybrid swarm and gravitation-based feature selection algorithm for handwritten indic script classification problem. *Complex & Intelligent Systems*, *7*, 823–839, URL https://api.semanticscholar.org/CorpusID:218581215.

Guo, H., Cao, J., Zhang, Y., Guo, J., & Li, J. (2018). Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 943–951). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3269206.3271709.

Guo, Z., Tang, L., Guo, T., Yu, K., Alazab, M., & Shalaginov, A. (2021). Deep graph neural network-based spammer detection under the perspective of heterogeneous cyberspace. *Future Generation Computer Systems*, *117*, 205–218. http://dx.doi.org/10.1016/j.future.2020.11.028.

Guo, Z., Yu, K., Jolfaei, A., Bashir, A. K., Almagrabi, A. O., & Kumar, N. (2021). Fuzzy detection system for rumors through explainable adaptive learning. *IEEE Transactions on Fuzzy Systems*, *29*(12), 3650–3664. http://dx.doi.org/10.1109/TFUZZ.2021.3052109.

Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P. (2014). TweetCred: Real-time credibility assessment of content on Twitter. *Springer International Publishing*.

Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017). *Multimodal fusion with recurrent neural networks for rumor detection on microblogs* (pp. 795–816). http://dx.doi.org/10.1145/3123266.3123454.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. In M. Lapata, P. Blunsom, & A. Koller (Eds.), *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers* (pp. 427–431). Valencia, Spain: Association for Computational Linguistics, URL https://aclanthology.org/E17-2068.

Khan, Z., Gwak, J., Iltaf, N., Pedrycz, W., & Jeon, M. (2024). RPf-GCNs: reciprocal perspective driven fused GCNs for rumor detection on social media. *Journal of Big Data*, *11*(1), 12. http://dx.doi.org/10.1186/S40537-023-00866-6.

Khattar, D., Singh, J., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. (pp. 2915–2921). http://dx.doi.org/10.1145/3308558.3313552,

Kim, Y. (2014). Convolutional neural networks for sentence classification. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing* EMNLP, (pp. 1746–1751). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1181, URL https://aclanthology.org/D14-1181.

Kuter, S., Akyürek, Z., & Weber, G.-W. (2018). Retrieval of fractional snow covered area from MODIS data by multivariate adaptive regression splines. *Remote Sensing of Environment*, *205*, 236–252. http://dx.doi.org/10.1016/j.rse.2017.11.021.

Lan, T., Li, C., & Li, J. (2018). Mining semantic variation in time series for rumor detection via recurrent neural networks. In *20th IEEE international conference on high performance computing and communications; 16th IEEE international conference on smart city; 4th IEEE international conference on data science and systems,*

HPCC/smartCity/DSS 2018, exeter, United kingdom, June 28-30, 2018 (pp. 282–289). IEEE, http://dx.doi.org/10.1109/HPCC/SMARTCITY/DSS.2018.00068.

Li, Q., Zhang, Q., & Si, L. (2019). Rumor detection by exploiting user credibility information, attention and multi-task learning. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1173–1179). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1113, URL https://aclanthology.org/P19-1113.

Lin, D., Ma, B., Cao, D., & Li, S. (2018). Chinese microblog rumor detection based on deep sequence context: Rumor detection sequence context. *Concurrency Computations: Practice and Experience*, *31*, Article e4508. http://dx.doi.org/10.1002/cpe.4508.

Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence* (pp. 2873–2879). AAAI Press.

Liu, X., Zhao, Z., Zhang, Y., Liu, C., & Yang, F. (2023). Social network rumor detection method combining dual-attention mechanism with graph convolutional network. *IEEE Transactions on Computational Social Systems*, *10*(5), 2350–2361. http://dx.doi.org/10.1109/TCSS.2022.3184745.

Lotfi, R., Zare Mehrjerdi, Y., Pishvaee, M., Sadegheih, A., & Weber, G.-W. (2020). A robust optimization model for sustainable and resilient closed-loop supply chain network design considering conditional value at risk. *Numerical Algebra*, http://dx.doi.org/10.3934/naco.2020023.

Lu, M., Huang, Z., Li, B., Zhao, Y., Qin, Z., & Li, D. (2022). SIFTER: A framework for robust rumor detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 429–442. http://dx.doi.org/10.1109/TASLP.2022.3140474.

Lu, Y.-J., & Li, C.-T. (2020). GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 505–514). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.48, URL https://aclanthology.org/2020.acl-main.48.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, J., Wong, K.-F., & Cha, M. (2016). *Detecting rumors from microblogs with recurrent neural networks*.

Ma, J., Gao, W., & Wong, K.-F. (2017). *Detect rumors in microblog posts using propagation structure via kernel learning*. http://dx.doi.org/10.18653/v1/P17-1066.

Ma, J., Gao, W., & Wong, K.-F. (2018). *Rumor detection on Twitter with tree-structured recursive neural networks*. http://dx.doi.org/10.18653/v1/P18-1184.

Morone, F., Min, B., Bo, L., Mari, R., & Makse, H. (2016). Collective influence algorithm to find influencers via optimal percolation in massively large social media. *Scientific Reports*, *6*, http://dx.doi.org/10.1038/srep30062.

Oshikawa, R., Qian, J., & Wang, W. Y. (2020). A survey on natural language processing for fake news detection. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 6086–6093). Marseille, France: European Language Resources Association, URL https://aclanthology.org/2020.lrec-1.747.

Pei, Z., Cao, Z., Long, M., & Wang, J. (2018). Multi-adversarial domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), http://dx.doi.org/10.1609/aaai.v32i1.11767, URL https://ojs.aaai.org/index.php/AAAI/article/view/11767.

Ran, H., & Jia, C. (2023). Unsupervised cross-domain rumor detection with contrastive learning and cross-attention. In B. Williams, Y. Chen, & J. Neville (Eds.), *Thirty-seventh AAAI conference on artificial intelligence, AAAI 2023, thirty-fifth conference on innovative applications of artificial intelligence, IAAI 2023, thirteenth symposium on educational advances in artificial intelligence, EAAI 2023, washington, DC, USA, February 7-14, 2023* (pp. 13510–13518). AAAI Press, http://dx.doi.org/10.1609/AAAI.V37I11.26584.

Rao, D., Miao, X., Jiang, Z., & Li, R. (2021). STANKER: stacking network based on level-grained attention-masked BERT for rumor detection on social media. In M. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing, EMNLP 2021, virtual event / punta cana, dominican Republic, 7-11 November, 2021* (pp. 3347–3363). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2021.EMNLP-MAIN.269.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*, http://dx.doi.org/10.1145/3137597.3137600.

Song, C., Yang, C., Chen, H., Tu, C., Liu, Z., & Sun, M. (2019). CED: Credible early detection of social media rumors. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3035–3047.

Sun, M., Zhang, X., Ma, J., Xie, S., Liu, Y., & Yu, P. S. (2023). Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 12736–12749. http://dx.doi.org/10.1109/TKDE.2023.3275586.

Tu, K., Chen, C., Hou, C., Yuan, J., Li, J., & Yuan, X. (2021). Rumor2vec: A rumor detection framework with joint text and propagation structure representation learning. *Information Sciences*, *560*, 137–151. http://dx.doi.org/10.1016/J.INS.2020.12.080.

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *2017 IEEE conference on computer vision and pattern recognition* CVPR, (pp. 2962–2971). URL https://api.semanticscholar.org/CorpusID:4357800.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Neural information processing systems*. URL https://api.semanticscholar.org/CorpusID:13756489.

Wang, X., Du, Y., Chen, D., Li, X., Chen, X., li Lee, Y., & Liu, J. (2023). Constructing better prototype generators with 3D CNNs for few-shot text classification. *Expert Systems with Applications*, *225*, Article 120124. http://dx.doi.org/10.1016/j.eswa.2023.120124, URL https://www.sciencedirect.com/science/article/pii/S0957417423006267.

Wang, Z., Guo, Y., Li, Z., Tang, M., Qi, T., & Wang, J. (2019). *Research on microblog rumor events detection via dynamic time series based GRU model* (pp. 1–6). http://dx.doi.org/10.1109/ICC.2019.8761457.

Wang, X., Li, L., Ye, W., Long, M., & Wang, J. (2019). Transferable attention for domain adaptation. *Vol. 33*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 5345–5352).

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* (pp. 849–857). http://dx.doi.org/10.1145/3219819.3219903.

Wang, Y., Qian, S., Hu, J., Fang, Q., & Xu, C. (2020). *Fake news detection via knowledge-driven multimodal graph convolutional networks* (pp. 540–547). http://dx.doi.org/10.1145/3372278.3390713.

Wei, L., Hu, D., Zhou, W., Yue, Z., & Hu, S. (2021). Towards propagation uncertainty: Edge-enhanced Bayesian graph convolutional networks for rumor detection. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 3845–3854). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-long.297, URL https://aclanthology.org/2021.acl-long.297.

Xiao, S., Bai, T., Cui, X., Wu, B., Meng, X., & Wang, B. (2023). A graph-based contrastive learning framework for medicare insurance fraud detection. *Frontiers of Computer Science*, *17*(2), Article 172341.

Xu, W., Wu, J., Liu, Q., Wu, S., & Wang, L. (2022). Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM web conference 2022* (pp. 2501–2510). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3485447.3512122.

Xuming, H., Guo, Z., Chen, J., Wen, L., & Yu, P. (2023). *MR2: A benchmark for multimodal retrieval-augmented rumor detection in social media* (pp. 2901–2912). http://dx.doi.org/10.1145/3539618.3591896.

Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., & Liu, H. (2019). Unsupervised fake news detection on social media: A generative approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 5644–5651. http://dx.doi.org/10.1609/aaai.v33i01.33015644.

Yang, C., Zhou, X., & Zafarani, R. (2021). CHECKED: Chinese COVID-19 fake news dataset. *Social Network Analysis and Mining*, *11*(1), 58.

Zhang, H., Qian, S., Fang, Q., & Xu, C. (2021). Multimodal disentangled domain adaption for social media event rumor detection. *IEEE Transactions on Multimedia*, *23*, 4441–4454. http://dx.doi.org/10.1109/TMM.2020.3042055.

Zhang, Y., Tang, H., Jia, K., & Tan, M. (2019). Domain-symmetric networks for adversarial domain adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition* CVPR, (pp. 5026–5035). URL https://api.semanticscholar.org/CorpusID:104291876.

Zhang, D., Wang, Y., & Zhang, Z. (2019). Identifying and quantifying potential superspreaders in social networks. *Scientific Reports*, *9*, 1–11. http://dx.doi.org/10.1038/s41598-019-51153-5.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In K. Erk, & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 207–212). Berlin, Germany: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P16-2034, URL https://aclanthology.org/P16-2034.

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* (pp. 836–837). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3289600.3291382.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2020). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI conference on artificial intelligence*. URL https://api.semanticscholar.org/CorpusID:229156802.