HiPo: Detecting Fake News via Historical and Multi-Modal Analyses of Social Media Posts

Tianshu Xiao
School of Computer Science and Technology, Shandong University
Qingdao, China
201922180266@mail.sdu.edu.cn

Sichang Guo
School of Computer Science and Technology, Shandong University
Qingdao, China
201900140039@mail.sdu.edu.cn

Jingcheng Huang
School of Computer Science and Technology, Shandong University
Qingdao, China
hjc2019@mail.sdu.edu.cn

Riccardo Spolaor*
School of Computer Science and Technology, Shandong University
Qingdao, China
rspolaor@sdu.edu.cn

Xiuzhen Cheng
School of Computer Science and Technology, Shandong University
Qingdao, China
xzc@sdu.edu.cn

ABSTRACT

In recent years, fake news has been a primary concern since it plays a significant role in influencing the political, economic, and social spheres. The scientific community has proposed several solutions to detect such fraudulent information. However, such solutions are unsuitable for social media posts since they cannot extract sufficient information from one-line textual and graphical content or are highly dependent on prior knowledge, which may be unavailable in the case of unprecedented events (e.g., breaking news).

This paper tackles this issue by proposing HiPo, a novel multi-modal historical post-based fake news detection method. By combining the features extracted from the graphical and textual content, HiPo assesses the truthfulness of a social media post by building its historical context from prior off-label posts with high similarity, therefore achieving online detection without maintaining a context or knowledge database. We evaluate the performance of HiPo via an exhaustive set of experiments involving four real-world datasets. Our method achieves a detection accuracy higher than 84%, outperforming state-of-the-art methods in most experimental instances.

CCS CONCEPTS

• **Information systems** → *Social networks*; • **Security and privacy** → *Social aspects of security and privacy*; • **Computing methodologies** → Neural networks.

KEYWORDS

Multi-modal information retrieval; Fake news detection; Social media

ACM Reference Format:
Tianshu Xiao, Sichang Guo, Jingcheng Huang, Riccardo Spolaor, and Xiuzhen Cheng. 2023. HiPo: Detecting Fake News via Historical and Multi-Modal Analyses of Social Media Posts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3583780.3614914

1 INTRODUCTION

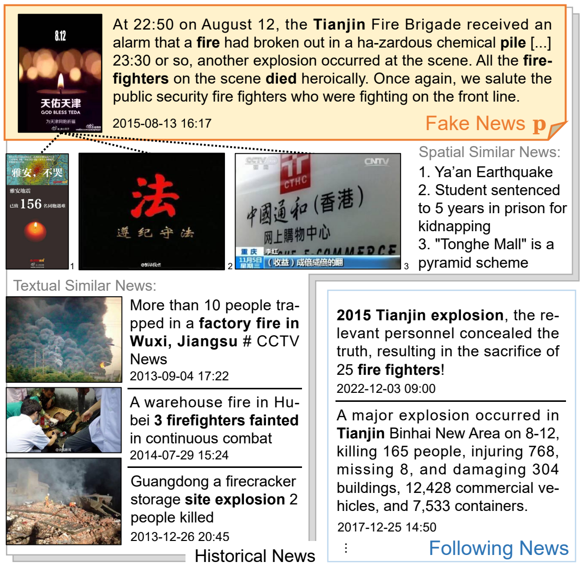
A comprehensive definition of fake news could be "news articles that are intentionally and verifiably false and could mislead readers" [1]. As fake news intentionally distorts facts [2], they have posed growing threats, especially with the emergence of Internet news agencies, the "echo chamber" effect, and multiple growing crises where people could become prey to disinformation, rumors, and polarized fake news. Attackers can also leverage fake news to lure unsuspecting users into becoming victims of attacks, such as phishing or malware, or unaware pawns in a distributed denial of service attack (DDoS) [3]. Moreover, online users' massive use of social networks has further exacerbated the spread of fake news in recent years. During the Covid-19 pandemic, fake news and misinformation through social media sites have spread so fast that the World Health Organization (WHO) calls it an "infodemic". The effects include confusion and support for behaviors that can harm health, lead to mistrust in science, and ultimately undermine the public health response to the pandemic [41].

Using techniques such as deep fusion for fake images and text-generating models, including GPT [4], the dangerous fake news can be planted in the truth automatically. However, fake news and misinformation have also affected AI-generated content; for example, a factual error made by the new Google Bard AI chatbot has led to an 8% fall in Alphabet shares in 2023 [13]. To identify fraudulent content, social media platforms have employed a considerable workforce from the community and third-parties checkers [23] with, unfortunately, limited success due to the astonishing number of posts to be verified. Therefore, researchers have recently proposed automatic fake news detection methods that aim to guarantee the correctness of information shared on social media posts. We can categorize such methods by the features extracted from

[*] Corresponding Author.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3614914



At 22:50 on August 12, the **Tianjin** Fire Brigade received an alarm that a **fire** had broken out in a ha-zardous chemical **pile** [...] 23:30 or so, another explosion occurred at the scene. All the **firefighters** on the scene **died** heroically. Once again, we salute the public security fire fighters who were fighting on the front line.

2015-08-13 16:17

Spatial Similar News:

1. Ya'an Earthquake
2. Student sentenced to 5 years in prison for kidnapping
3. "Tonghe Mall" is a pyramid scheme

Textual Similar News:

More than 10 people trapped in a **factory fire in Wuxi, Jiangsu** # CCTV News
2013-09-04 17:22

A warehouse fire in Hubei **3 firefighters fainted** in continuous combat
2014-07-29 15:24

Guangdong a firecracker storage **site explosion** 2 people killed
2013-12-26 20:45

2015 Tianjin explosion, the relevant personnel concealed the truth, resulting in the sacrifice of 25 **fire fighters**!
2022-12-03 09:00

A major explosion occurred in **Tianjin** Binhai New Area on 8-12, killing 165 people, injuring 768, missing 8, and damaging 304 buildings, 12,428 commercial vehicles, and 7,533 containers.
2017-12-25 14:50

Figure 1: An example of fake news p from Weibo and its related similar posts. We also report some news posted before (on the center and bottom left) and after (bottom right) the post time of p.

social media posts they leverage: news content (*content-driven methods*) [14, 20, 25, 50], user-to-user interactions *(social context-driven methods*) [8, 18, 32, 49], and knowlegde extraction from identified entities (*knowledge-driven methods*) [44, 46, 52].

On the one hand, expanding the perception domain for various news features allows models to cope with the *labeling* bias due to the expensive labeling process and possible mislabeled news [9, 11]. In particular, previous work has relied on multiple feature channels, where the interrelationship of news in the training set is discovered through the memory capacity of neural networks [52]. However, insufficient effort has been made outside of the training set to leverage the interdependence of labeled news with similar unlabeled information.

On the other hand, time bias occurs when assessing a model's performance with "temporally inconsistent evaluations that integrate future knowledge about the testing objects into the training phase or create unrealistic settings" [27]. In particular, such a bias may have led to the misevaluation of fake news detectors in the literature [6, 14, 17, 37] where the training-testing partitioning has not taken into account the temporal information of posts. In addition, limited prior information is a common scenario for virus-like propagation of fake news, which in turn spreads disbelief and unwillingness to follow the collection of restrictions among the public [16]. Figure 1 shows an example in which similar historical news is significantly less informative for the latest accident than the following news. Therefore, the real-world gap between historical and upcoming news requires temporally uncoupling training, validation, and testing sets to avoid time bias.

To address the biases mentioned above, this paper introduces time awareness as a fundamental architectural design element in fake news detection through the design of HiPo, a **Hi**storical **Po**sts-based multi-modal fake news detection method. Our model leverages a wide-range background of unlabeled historical news to ensure the quality of the information on social media platforms. Given a post to be verified, our method retrieves similar posts from a historical background dataset to cope with the insufficient information within the setting of time bias countering. To mitigate the labeling bias, HiPo's channels of news perception include textual and spatial modality, which encourages the discovery of in-depth inter-modal relationships among posts. We aggregate the features of a selection of similar posts in an intra-modal fashion to obtain the feature fusion results, followed by an inter-channel fusion layer.

The main contributions of this paper are:

- We present a historical posts-based multi-modal fake news detection model (HiPo) that integrates similar historical posts into the feature embedding process for the multi-channel historical-based perception.
- We evaluate the detection performance of HiPo with a thorough set of experiments on four extensive multi-lingual datasets of social media posts. Such an evaluation includes assessing the robustness of HiPo against time and labeling biases. Our results show that HiPo outperforms other content-based competitors in most experiment instances.
- To mitigate the *labeling* bias in state-of-the-art methods, we propose a historical perception-specified plugin based on HiPo's module that integrates information from unlabeled posts. We experimentally demonstrate that such a plugin improves their average accuracy by 1.8%.

2 RELATED WORK

In this section, we introduce previous research on content-driven fake news detection, where different approaches aim to integrate the limited information extracted from posts with additional sources. While they achieve significant detection performance by only relying on a post's content, they do not adequately address the time and labeling bias. In what follows, we survey the state of the art of content-driven fake news detection, organizing this work by their modality.

2.1 Single-Modality Fake News Detection

The evolution of neural networks contributes to the development of content-driven methods based on neural networks. Ma et al. in [20] are among the first researchers to utilize a neural network (i.e., RNN of multiple types) to improve the performance of single-modal detectors. Subsequently, the same research group [21, 51] uses a CNN in the multitask learning approach of model design. We can further categorize other consequential work into two main categories.

On positive effects of context granularity on model performances. [29] explores the contribution of visual information-based detection of fake news. [39] introduces graph neural networks for intersentence logic perceptions. [5] elaborates on the role of visual information from a broader perspective. [55] retrieves semantic information with a high dimension to provide targeted fake news detection, where multifaceted semantics based on a universal language model (BERT) and two domain-specific models cover related news contents of the given topic.

On the generalization ability for broader applications. [22] enhances models' robustness through an additional adversarial training process. [45] utilizes the reinforcement learning mechanism for better performance in the real-world setting of breaking events. [55] mitigates the impact of separate news entities through the identification of casual effects. [34] studies on the news environment from their context information. [25] improves the cross-domain generalization capability of the model through a multistage training process.

2.2 Multi-Modality Fake News Detection

First proposed in [14], the multi-modal fake news detection task focuses on further leveraging inter- and intra-modality news features. In that work [14], textual features generated by RNN and other multi-modal features are inputted into the attention layer for prediction. Subsequently, [42] proposes an adversarial neural network-based approach and [17] introduces the VAE method. It is worth noticing that [37] introduces the pre-trained models of BERT and VGG-19 into neural-based multi-modal feature extraction for textual and spatial news information, respectively. The work that extends such a method can be further divided into two different approaches:

On improving the multi-modal feature fusion. [36] proposes replacing BERT with a more advanced and text-based pre-trained model of AXNet. [54] calculates the similarity between textual and spatial features as a kind of auxiliary information. [48] assists fake news detection through multi-modal information consistency detection. [47] attempts to regenerate the human reading process through a multi-layer co-attention mechanism in the fusion process. [28] introduces multiple approaches to infer image-text relations for relation-specified feature fusion. [30] leverages the hierarchical semantic output of hidden layers in BERT to improve text-semantic perception. [50] similarly leverages the fusion of the outputs of the ResNet and BERT hidden layer. [6] relies on a cross-modal ambiguity learning module to adaptively aggregate unimodal features and cross-modal correlations.

On generalization abilities. [43] and [50] utilize the meta-learning mechanism and the topic memory module to improve models' cross-domain performances separately.

2.3 Inter-Modality and Intra-Modality Pre-training

While intra-modality pre-trained models are used to extract features from news context, inter-modality pre-training frameworks are used to embed model perception in the multi-modal post. Both approaches aim to comprehensively model news representations with specialization for the fake news detection task. On the one hand, fine-tuned intra-modality models, including BERT [7], VGG-16/19 [35] and ResNet [12], provide the intra-modality understanding for most of the recent work. On the other hand, the CLIP framework proposed in [31] provides a method for pre-training the perception of text-image relations for multi-modal modules.

The main limitation of the work mentioned above lies in not considering the abundant news' historical background as a viable source of information. To the best of our knowledge, our method is the first to detect fake news by also leveraging the historical background of the post under test. Moreover, our method is the first to consider and mitigate time and labeling biases by design.

3 THE HIPO SYSTEM

The ultimate goal of our system is to classify a given social media post as 'fake' or 'legitimate'. Therefore, we build a binary classifier that leverages information extracted from such a post p and the historical news background $\mathcal{H}$, i.e., the other posts available in the dataset (excluding the post p). Since a social media post can contain text and images, our method extracts information from both textual and spatial contents, respectively. Moreover, we extract information perceived from other similar posts in a dataset as historical news background. In summary, our multi-modal fake news detection method relies on information from three channels: textual, spatial, and perceptional.

In Figure 2, we depict the overall architecture of our system, which can be divided into four modules:

- **Multi-modal feature extraction** uses pre-training models to extract features from a given post and its contemporary posts.
- **Historical similar news retrieval** uses a multi-modal similarity metric to identify contemporary posts with high similarity to the post under test.
- **Historical fusion** involves previously selected posts that hold similar multi-modal features to improve the perception of three different modalities in the given post.
- **Inter-modal fusion** and **Classification** aggregate the multi-modal features and provide a classification for the given post.

In what follows, we provide a detailed description of each of the aforementioned modules.

3.1 Multi-Modal Feature Extraction Module

In this module, we extract the textual and spatial feature vectors from social media posts in the dataset.

Textual Feature Extraction. From the textual content of a post x, we extract a feature vector T_x by applying the BERT model [7] for natural language processing, which provides an effective representation of textual content. We fix the number of words w (i.e., the internal parameter of BERT) as twice the average number of words of the posts in the entire dataset. For a single word, BERT outputs a feature vector v of size 768 (i.e., the hidden size of BERT). Therefore, we concatenate the feature vectors for each word to obtain $T_x = v_1 \parallel \cdots \parallel v_w$ of total size $|T_x| = w \cdot 768$.

Spatial Feature Extraction. Social media posts often include images that contain information that may not be present in their textual contents. For this reason, we also extract from images of a post x a spatial feature vector S_x. For this task, we employ the VGG-19 model [35] and obtain a vector whose values are the ones of the last hidden layer (i.e., a total feature with a size of 1000).

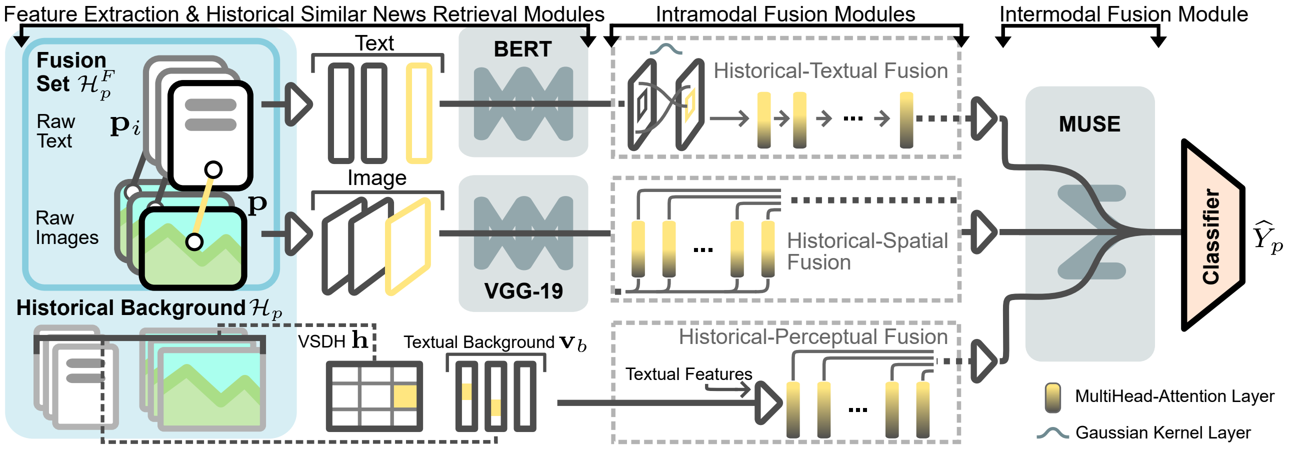

Feature Extraction & Historical Similar News Retrieval Modules
Intramodal Fusion Modules
Intermodal Fusion Module
Fusion Set $\mathcal{H}_p^F$
Raw Text
$\mathbf{p}_i$
Raw Images
$\mathbf{p}$
Text
BERT
Image
VGG-19
Historical-Textual Fusion
Historical-Spatial Fusion
Historical Background $\mathcal{H}_p$
VSDH $\mathbf{h}$
Textual Background $\mathbf{v}_b$
Historical-Perceptual Fusion
Textual Features
MUSE
Classifier
$\widehat{Y}_p$
MultiHead-Attention Layer
Gaussian Kernel Layer

Figure 2: The diagram of the main components of our Historical Posts Based Multi-modal Fake News Detection (HiPo) method. We mark the information from the post under test p in yellow.

3.2 Historical Similar News Retrieval Module

In this module, we adopt an approach based on micro environment [34] to identify a set $\mathcal{H}_p$ of high similarity to the post under test p. We first compress the textual feature vector T_x of a post x by applying a pooling layer to obtain the vector T'_x. As a similarity metric, we compute the L2-norm between the feature vectors of post p and another post $q \in \mathcal{H}$. Therefore, we select the k_{sim} most similar posts applying the k-nearest neighbors algorithm. Using the compressed T'_q and the spatial vector S_q for each post q in $\mathcal{H}$, we apply kNN to obtain the sets $\mathcal{H}_p^T$ and $\mathcal{H}_p^S$, respectively. Hence, we obtain a set of posts with high similarity $\mathcal{H}_p = \mathcal{H}_p^T \cup \mathcal{H}_p^S$ with size $k_{sim} \leq |\mathcal{H}_p| \leq 2 \cdot k_{sim}$ since some posts may appear in both $\mathcal{H}_p^T$ and $\mathcal{H}_p^S$.

Subsequently, we extract comprehensive historical background features (hereafter also referred to as *perceptional* channel) from the image and textual content of the retrieved posts in $\mathcal{H}_p$. For image content, we apply the VSDH (Visual Similarity Distribution Histogram) [15] to extract the similarity distribution with fine granularity. In particular, we calculate the cosine similarity $s(\cdot,\cdot)$ between the spatial feature vector of p and the one of each post in $\mathcal{H}_p$. Hence, we distribute such similarities (whose values range from $[0, 1]$) into 128 equal-interval bins and count them, obtaining the following vector:

$$\mathbf{h} = \left\{ \sum_{\mathrm{q} \in \mathcal{H}_p} \mathbb{1}\left[s(S_p, S_q) \in n\text{-th bin}\right] \right\}_{n=0}^{128} \tag{1}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. VSDH serves as an auxiliary inference to identify images with a similar context to the one of post $\mathbf{p}$ or image reuse. It is worth noting that we can align the output $\mathbf{h}$ (of size 128) with those from other perceptional channels (having also size 128) of the subsequent historical fusion module.

For the textual content, we aim to assess the similarity of p to the posts in $\mathcal{H}_p$ to grasp the related news trends despite their different text lengths. We integrate the average textual feature vector, the lower border, and upper border of textual features in the textual background vectors $\mathbf{v}_b$ with $3 \times |v|$ dimension, wherein weights in averaging the textual features T'_i are implemented as cosine similarities $\{s(T'_p, T'_i)\}_{i=0}^{k_{sim}}$ with the given post $\mathbf{p}$.

3.3 Historical Fusion Modules

In these modules, we generate a fusion of feature vectors from the target post p and a representative set of posts within $\mathcal{H}_p$. In particular, we identify such a set $\mathcal{H}_p^F$ containing k_f posts with $k_f < k_{sim}$ (i.e., $\mathcal{H}_p^F \subset \mathcal{H}_p$). The value of k_f is a hyper-parameter of HiPo. We design three distinct modules based on a multi-head attention mechanism [40] for the textual, spatial, and perceptional channels. For each module, we rely on multi-head attention to build bottom-up fusion integrated post-wise features of post x and $\mathcal{H}_p^F$ as building blocks. In what follows, we describe in detail each historical fusion module.

3.3.1 *Historical-Spatial Fusion Module.*

The spatial channel gives image-based indications, which may provide limited information. However, such information is crucial in a multi-modal channel scenario to enhance the overall detection performance. Given the spatial information of $\mathbf{p}$ and posts in $\mathcal{H}_p^F$, we use a paralleled multi-head attention layer to aggregate the historical spatial feature with a total size of $|\mathcal{H}_p^F| \cdot |S|$ for efficient spatial information processing, with the output of each head defined by:

$$\text{head}_x = \text{Attention}(\mathbf{Q}_k \mathbf{W}_i^Q, \mathbf{K}_k \mathbf{W}_i^K, \mathbf{V}_k \mathbf{W}_i^V), \tag{2}$$

where the k-th attention receives a tuple of the query, key, and value. Specifically, we configure the input to the applied multi-head as:

$$\mathbf{Q}_k = S_k \qquad \mathbf{K}_k = \mathbf{V}_k = S_p, \tag{3}$$

where we fuse the spatial features of $\mathbf{p}$ with historical posts separately. We then concatenate and align the resulting features with the other fusion to generate the historical-spatial fusion.

3.3.2 *Historical-Textual Fusion Module.*

To improve the effectiveness of textual feature extraction, we design the historical-based textual fusion module, which is separated into word-level and text-level fusion, with a kernel attention layer from [49] and a multi-head

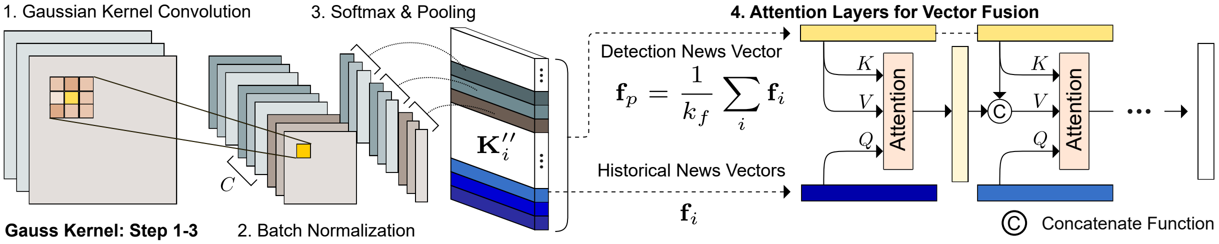


Figure 3: The detailed illustration of the historical-textual fusion module of HiPo.

attention layer implemented at the corresponding level. The word-level fusion aims to generate the post-wise full-text feature for all posts in $\{\mathbf{p}\} \cup \mathcal{H}_p^F$ using the input of the word vectors v contained in T_i for each $\mathbf{p}_i$ in the above compound set, and the purpose of the next text-level fusion is to fuse the full-text features for a single text-level feature output of $\mathbf{p}$.

In the first fusion, the transformation of word-level similarities based on word pairs from $T_p \times T_i$, where $\times$ denotes the Cartesian product, is generated from Gaussian kernels in the first attention layer, where we denote the similarity of each word pair as $s_i^{k,l} := s(v_k, v_l)$ with v_k and v_l from $\mathbf{p}$ and $\mathbf{p}_i$, separately. The similarities after Gaussian kernel transformation become the kernel attention on the word-level fusion for a single news post. We depict the diagram of bi-level fusion in Figure 3, where we use cross-attention to further leverage kernel attention from the word-level fusion. To enhance learning, the C kernels independently learn the transform matrix, where the combination of matrices represents the comprehensive attention on all word pairs, reflecting the predicted importance thereof in the word-level fusion:

$$\mathbf{K}_{i,j,k,l} := \text{Kernel}_j(s_i^{k,l}) = \exp\left(-\frac{(s_i^{k,l} - \mu_j)^2}{2\sigma_j^2}\right), \tag{4}$$

where the aggregated kernel attention matrix is denoted by $\mathbf{K}_{i,j,k,l}$ for the transformed similarities of the word pair (v_k, v_l) from different $\mathbf{p}$ and $\mathbf{p}_i$, and μ_j and σ_j denote the kernel mean values and widths, which are fixed as $\mu_j = 0$ and $\sigma_j = 5$ for the stable mutation of Gaussian kernels. The attention of word pairs within different kernels is integrated into single-valued word-level attention by summing over i and j, k of $\mathbf{K}_{i,j,k,l}$. We first use the average function to aggregate different words v_l of the historical post $\mathbf{p}_i$:

$$\mathbf{K}'_{i,j,k} := \log \sum_{l=1}^{w_i} \mathbf{K}_{i,j,k,l}, \tag{5}$$

where w_i is the word length of post $\mathbf{p}_i$ and w_p for $\mathbf{p}$. Then, with neural-based functions, the resulting matrix $\mathbf{K}_{i,j,k}$ is aggregated over different kernels and historical posts:

$$\mathbf{K}''_i := \frac{1}{w_p} \sum_{j,k} \text{softmax}_j(W \cdot \phi(\mathbf{K}'_{i,j,k}) + b), \tag{6}$$

where we use a Softmax layer to normalize the kernel weights generated by a CNN layer ϕ. Finally, we obtain the word-level fusion regarding different historical posts $\mathbf{p}_i$ as historical-textual fusion using post-wise textual features with kernel attention, giving the word-level fusion for $\mathbf{p}_i$ as $\mathbf{f}_i = \mathbf{K}''_i \cdot T_i$. Additionally, the average feature fusion in $\mathcal{H}_p^F$ is calculated as $\mathbf{f}_p = 1/k_f \sum_i \mathbf{f}_i$ for the $\mathbf{p}$'s word-level fusion and all word-level fusion are used as the corresponding kernel-based full-text features in the text-level fusion. We use $\mathbf{f}_p$ as the word-level feature in this module, which is used in the next stage of text-level fusion and the next multi-modal fusion module. For the text-level fusion, different from the parallel attention layer described above in Eq. 2, we connect the input and output of adjoining layers in a streamlined fashion by setting:

$$\mathbf{Q}_k = \mathbf{f}_i \qquad \mathbf{K}_k = \mathbf{f}_p \qquad \mathbf{V}_k = \mathbf{H}_k \oplus \mathbf{f}_p. \tag{7}$$

The full-text feature regarding $\mathbf{p}_i$, the memory of test post $\mathbf{p}$ and the previously generated text-level feature are fused, where the historical posts are fed to the inputs in decreasing similarity regarding $\mathbf{p}$. Finally, we use the last output of the multi-head attention layer $\mathbf{H}_{k_f}$ as the text-level feature of $\mathbf{p}$, which represents the comprehensive relationship among $\mathbf{p} \cup \mathcal{H}_p^F$ with the attention of general full-text features, apart from the word-level kernel attention fusion $\mathbf{f}_p$.

3.3.3 *Historical-Perceptual Fusion Module.*

In this module, we leverage the informative vectors of textual background $\mathbf{v}_b$ to fuse with the kernel perception $\mathbf{f}_p$ independently by following the same approach in the historical-spatial fusion above, where we implement three multi-head attention in parallel with the concatenated outputs after alignment. The histogram $\mathbf{h}$ for background images is used directly as the background spatial feature, which we fuse with the textual feature in the next module.

3.4 Inter-modal Fusion Module

We apply the MUSE attention paradigm [53] to fuse features from the textual, spatial, and perceptional channels. In particular, MUSE deploys multiple parallel CNNs with shared parameters to enhance the feature hierarchy within the self-attention mechanism.

In our implementation, we use three parallel convolutions (each consisting of two layers) and a self-attention part. In the first layer, we assign each convolution a distinct kernel size (i.e., 1, 3, and 5) and provide as input the concatenation of feature vectors from the three above channels. In the second layer, we use a point-wise kernel (i.e., with size = 1) for all three parallel convolutions.

The self-attention part applies Eq. 2 on the concatenation of five feature vectors (i.e., spatial, text- and word-level textual, and image- and text-perceptual) for the query, key, and value. The value is passed accordingly to the input of convolutions. Finally, we consider the multi-modal fusion as the weighted sum of self-attention and CNNs outputs.

Table 1: The dataset partitioning for experiments.

Dataset	Training	Validation	Testing	Background	Pre-train	Total
Fakeddit	8,191	2,584	2,527	563,999	29,184	577,301
IFND	1,536	512	512	54,154	13,568	56,714
Twitter	1,856	640	640	110,464	22,016	113,600
Weibo	6,967	2,560	2,560	175,000	28,928	187,087

3.5 Classification of Fake Posts

The classification of the given post $\mathbf{p}$ based on its multi-channel feature fusion $\mathbf{F}_p$ is generated by:

$$\widehat{Y}_p = \mathrm{Softmax}\left(\tanh\left(W(\mathbf{F}_p) + b\right)\right). \tag{8}$$

With the predicted label Y_i of the i-th training post, the training loss we use in learning is defined as:

$$\mathcal{L}(\Theta) = \sum_i -\left(Y_i \log \widehat{Y}_i + (1 - Y_i)\log(1 - \widehat{Y}_i)\right). \tag{9}$$

4 EXPERIMENTAL EVALUATION

In this section, we evaluate the performance of HiPo compared to other methods of the state of the art and varying the time and mislabeling biases.

4.1 Experimental Method

We describe the implementation and parameters of our method, the datasets, and the competitor methods considered in the experimental evaluation.

Hardware and Software. We implement our model in Python 3.8.15 and the development is based on Pytorch 1.10.0 with the support of CUDA 11.3 as a verification tool. We also use Pandas and NumPy libraries for the dataset with table processing and CPU-only tasks. The training process is based on the Adam optimizer and runs on a server with two Intel XEON W2295 (18 cores) and four NVIDIA Geforce RTX 3090. The version of our model with `256batch_size` reaches a maximum memory usage of 30GB RAM on the CPU and 14GB video RAM on the GPU.

Datasets. For our experimental evaluation, we use the following dataset of social media posts: *Fakeddit* [24], *IFND* [33], *Weibo* [34, 42] and *Twitter* [19, 26, 38]. The posts on IFND and Fakeddit datasets are in English, while the ones on Weibo are in Chinese. The Twitter dataset is multi-lingual, with the majority of posts in English. Due to computational limitations, we crop the length of posts' textual content at twice the average length of 36 and 120 words or characters for English and Chinese, respectively. Since our implementation of BERT is based on the majority of languages in a single dataset, the HiPo relies on parameterized uncased BERT for textual processing (i.e., `bert-base-chinese` for Chinese dataset Weibo and `bert-base-uncased` for the other datasets). To avoid time bias, we sort the posts in each dataset by their publication date. Hence, we split a dataset into temporally continuous sets for training, validation, testing, and background. For the pre-training process, we use a subset of the background set. While the Fakeddit dataset is imbalanced (i.e., 66% of the posts are fake news), the other three datasets are well balanced (i.e., 50% legit and 50% fake news posts). We report in Table 1 the number of examples in each of the considered datasets.

Table 2: Fine-tuned model parameters on different datasets.

Dataset	L. Rate	Grad. Decay	MultiHead Size	Dropout	Boundary
Fakeddit	$6 \cdot 10^{-4}$	10^{-3}	[2, 4, 8]	10^{-1}	[1.8, 80]
IFND	$5 \cdot 10^{-6}$	10^{-2}	[4, 1, 12]	10^{-1}	[1.6, 33]
Twitter	$9 \cdot 10^{-6}$	10^{-3}	[8, 8, 8]	10^{-1}	[1.5, 70]
Weibo	$4 \cdot 10^{-6}$	10^{-4}	[8, 2, 4]	$3 \cdot 10^{-1}$	[3.5, 100]

Model parameters. We select the following parameters of HiPo independently from the dataset: learning rate $\{10^{-i}, 2 \cdot 10^{-i}, 3 \cdot 10^{-i}, \cdots ,9 \cdot 10^{-i}\}_{i=3}^{6}$; gradient decay $\{0, 10^{-1}, \cdots, 10^{-6}\}$; multi-head size in attention layers $\{1, 2, 4, 8, 12\}$ (i.e., triple for text, image and perceptional channels); model dropout $\{0, 0.1, 0.3, 0.5, 0.01\}$; minimum similarity boundary of historical background $\mathcal{H}_p$ (i.e., k_{sim}) is within the interval $[0, 100]$ (i.e., couple for text and image channels); and numbers of similar historical posts set to $k_{sim} = 10$ and $k_f = 5$. The selection of parameters for the next experiment based on different datasets is listed in Table 2.

Competitors. For the performance comparison, we consider seven state-of-the-art fake news detection methods. We can divide such methods into two categories:

- *Single-modal competitors*: The bi-directional LSTM-based **Bi-LSTM** [10], BERT-base-uncased pre-trained model-based **BERT** [7] and image feature extractor based on **VGG-19** [35]. For these three methods, we apply a full connection and a softmax layer to obtain their final predictions.
- *Multi-modal competitors*: **Att-RNN** [14] uses LSTM to process news text and VGG-19 to process spatial information by including an attention mechanism for multi-modal fusion. **MVAE** [17] uses a bi-modal variational autoencoder coupled with a binary classifier. **SpotFake** [37] utilizes the concatenation of textual and spatial features as input for a linear and a Softmax layer. **CAFE** [6] first uses an encoding-decoding style ambiguity-aware network to update uni- and multi-modal features, and then employs a fusion layer for the final prediction.

For the above methods, we use the parameter values that achieve the best performance as reported in the respective work. Since we consider four distinct datasets, we incorporate adaptive parameter optimizations for each dataset.

Evaluation Metrics. To evaluate the models, we use the following metrics: accuracy $Ac = (TP + TN)/(TP + TN + FP + FN)$, precision $Pr = TP/(TP + FP)$, and recall $Re = TP/(TP + FN)$, where True Negative (TN) and False Negative (FN) are the number of posts correctly and incorrectly classified as "legitimate", respectively; and True Positive (TP) and False Positive (FP) are the number of posts correctly and wrongly classified as "Fake", respectively. In addition, we calculate the Area Under the receiver operator characteristic Curve (AUC).

4.2 Performance Evaluation of the Basic Model

In this evaluation, we assess the performance of HiPo in the datasets considered and compare it with the other competitors in the state of the art. In the same settings, we perform an ablation study to assess the contribution of each information channel to HiPo performance.

Table 3: The performance evaluation of the model and competitors.

Model	Fakeddit				IFND				Twitter				Weibo			
	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC
HiPo (Ours)	**86.7**	71.6	83.9	**90.6**	**93.4**	**90.8**	**95.2**	**97.6**	**85.2**	**86.5**	83.1	**88.7**	**88.2**	**91.7**	85.8	**95.0**
VGG-19 [35]	76.9	48.8	68.7	76.3	72.7	61.1	77.4	77.8	82.4	80.7	81.4	85.4	63.4	67.5	62.4	67.5
BERT [7]	85.5	73.9	78.5	90.3	89.8	91.5	87.6	96.7	81.3	80.8	78.6	86.4	86.7	83.5	89.2	94.0
Bi-LSTM [10]	84.8	74.5	75.9	89.9	90.2	90.3	89.6	96.3	80.7	73.3	81.5	87.7	85.6	82.2	88.3	93.1
AttRNN [14]	86.1	72.5	81.5	85.8	81.6	77.3	83.4	89.5	81.3	76.3	81.2	84.9	83.3	80.2	85.4	90.5
SpotFake [37]	86.2	70.1	83.5	90.7	81.3	77.7	82.4	88.6	83.2	80.0	82.1	87.8	84.3	83.1	85.2	91.6
MVAE [17]	85.3	65.4	85.2	90.1	78.9	79.4	77.5	87.3	82.8	79.3	82.1	87	84.5	80.2	87.7	91.6
CAFE [6]	85.4	**87.4**	**91.9**	83.9	92.0	90.6	94.4	92.4	82.7	82.2	**87.2**	84.2	**88.2**	85.7	**91.9**	88.4

4.2.1 Comparison Study. In this study, we compare the performance of HiPo with competitors on the four datasets. In Table 3, we report the empirical results of such a comparison according to the metrics considered. We can see that HiPo outperforms other state-of-the-art methods in the majority of the instances. Overall, HiPo achieves the best accuracy, and AUC on the four datasets. However, the precision of HiPo is lower than CAFE on Fakeddit and Twitter datasets (i.e., 8.0% and 4.1% lower, respectively) and lower than BERT, Bi-LSTM, MVAE and CAFE on the Weibo dataset (i.e., 2.6%, 1.7%, 1.1% and 6.1% lower, respectively). This means that HiPo tends to identify challenging posts as "fake", leading to false positives. Although HiPo has a lower recall than CAFE (i.e., 15.8% lower) on Fakeddit, it achieves a significantly higher AUC (i.e., 6.7% higher). This suggests that HiPo's performance is more favorable on balanced datasets, and underperforms on the skewed Fakeddit dataset. However, Hipo achieves the highest accuracy and AUC over Fakeddit, showing stable performance under imbalanced datasets. In particular, the posts in the Twitter dataset are the most challenging to learn from due to their multi-lingual nature. Hence, the feature extraction and similar news retrieval methods may be inaccurate, since they would extract information using non-corresponding BERT cases. On this note, with insufficient textual information of each language for training, the multi-modal HiPo can achieve relatively stable performance in the historical-based spatial feature fusion, which aids the degraded textual channel. This result is also confirmed by our ablation study in Section 4.2.2.

4.2.2 Ablation Study. In this study, we evaluate the accuracy of HiPo across different channels, namely textual, spatial, and perceptional channels. We report the results of the ablation study in Figure 4. These results offer the following insights:

- Using a single channel, HiPo achieves similar accuracy to its single-modal competitors. However, a single channel is not sufficient to provide comprehensive historical-based information.
- The contribution to the performance of different channels varies significantly. Compared to textual and historical channels, the spatial channel provides a limited contribution.
- Not considering the spatial channel (i.e., T+P) results in a notable decrease in accuracy compared to the three-channel HiPo (i.e., All). This suggests that the perceptional and textual channels provide different information, whose diversity could benefit from the spatial channel.

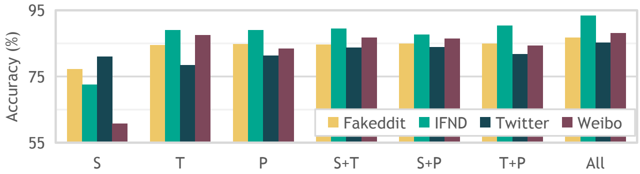


Figure 4: Accuracy for the ablation study. We consider the features from the textual (T), spatial (S), and perceptional (P) channels.

4.3 Effects of Labeling Bias

In its basic version, HiPo extracts information from posts in the historical background without considering whether they are fake or legit. As the label for such posts may not be available, we can obtain their labels as classifications from one of our competing methods or via semi-supervised learning. By relying on such additional information, we can incur labeling bias, since the obtained labels may be wrong (i.e., false positive or false negative), thus uncertain.

In this analysis, we modify HiPo to integrate learning with posts in historical background labeled by a competitor (i.e., *labeler*) and assess the effect of labeling bias. To integrate HiPo with the labeled posts, we add a similarity matrix $\mathbf{W}_{i,j} = s(T_i, T_j)$ to the perceptional channel, where we assign the labeler's classification $\widetilde{Y}_p$ to diagonal values $\mathbf{W}_{i,i}$.

Table 4 reports the results of this analysis. Comparing the results with Basic HiPo (without labelers), we notice that the information from labeled background sets provides comparable or minor improvements on the considered datasets, with only a major precision improvement on the Twitter dataset. In summary, HiPo is robust against labeling bias as it achieves stable performance, while minor improvements are related to the quality of the individual labeler. As an example of mislabeling, most labelers on the IFND dataset assign identical labels to more than half of the background news, which results in constant performance (see Table 4).

4.4 Robustness under Time Bias

In this section, we study the effect of time bias on detection performance by considering the partitioning of time-mixed datasets and different numbers of similar historical posts.

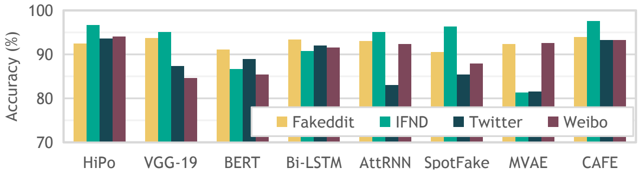


Figure 5: Accuracy with time-mixed training and testing set.

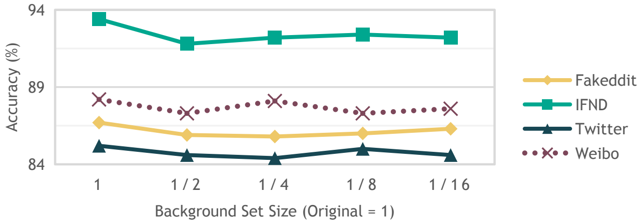


Figure 6: Accuracy varying the size of similar historical posts.

4.4.1 Time-mixed Dataset Partitioning. In a real-world setting, a dataset may have issues with the time of posts (e.g., the timestamp is missing (or wrong), only includes posts from discontinuous time intervals, or is temporally highly concentrated. Therefore, our method needs to be robust even on datasets with the above issues, where the time bias cannot be avoided. In this evaluation, we compare the detection performance of HiPo and other competitors in time-mixed datasets (i.e., posts are shuffled before partitioning into training and testing sets) partitioned using the same proportions in Table 1. From the results reported in Figure 5, we can observe that HiPo outperforms competitors, except for VGG-19 and Bi-LSTM in the Fakeddit dataset. HiPo is robust against time-mixed datasets, as it achieves an accuracy higher than 92.5% across all datasets. Under time-mixed partitioning settings, CAFE achieves the highest accuracy on Fakeddit and IFND, but HiPo outperforms it on Twitter and Weibo (i.e., 0.3% and 0.8% higher, respectively). In comparison, other methods achieve either an overall lower accuracy or a significantly lower accuracy on one or more datasets.

4.4.2 Downsizing the Historical Background and Training Sets. The amount of posts available as historical background (i.e., Background set) and for training is another aspect influenced by time bias.
Historical Background. In Figure 6, we illustrate how exponentially reducing the size of the background set (i.e., $1/2^i$ of its original size with $i = 0, \ldots, 4$) influences the accuracy of HiPo on the considered datasets. We can see that the accuracy of HiPo remains unchanged in the Fakeddit dataset. Additionally, after an initial slight accuracy drop at half the size of the original background set (i.e., 1/2), HiPo maintains a stable accuracy on the IFND, Weibo, and Twitter datasets despite further reductions.
Training Set. Considering the Fakeddit dataset, we reduce the size of the training set using the same exponential reduction as above. The results in Figure 7 show that HiPo maintains a higher and quite stable accuracy compared to the other methods (i.e., only a 3.3% decrease between the original size and 1/16). Such results suggest that the information from the historical background compensates for the reduced number of posts in the training set.

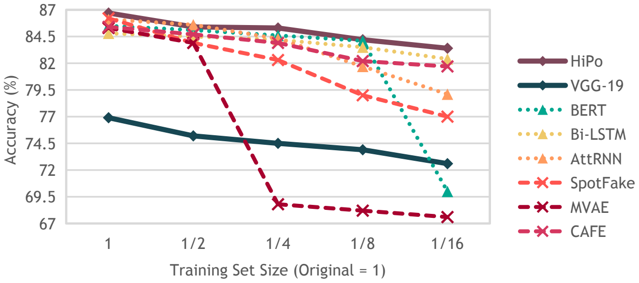


Figure 7: Accuracy varying the size of Fakeddit training set.

5 EXTENDED APPLICATIONS OF HIPO

In this section, we discuss two possible extensions for the HiPo method related to the use of pre-trained historical fusion by other methods and transfer learning between different domains.

5.1 Enhancing the Competing Methods via Pre-trained Historical Fusion Module

In the literature, researchers have used pre-trained models to improve the performance of other natural language processing methods [7, 31]. In this study, we first employ pre-training on the historical post-based attention mechanism provided by the historical fusion modules. We then enhance the competitors by replacing their feature extraction modules with the historical posts-based HiPo mechanism, where the multi-modality features are rebuilt from the output of historical-textual and historical-spatial fusion modules. The competitor's enhancement consists of two steps.
Pre-train the historical fusion module. We separately pre-train the textual and spatial modules, use them to extract a target post's features, and use such features to identify a set $\mathcal{H}_p$ of similar posts in the historical background. In particular, we enable the intra-modal features with pre-training by masking feature input [7] as follows: (i) the first quarter of features includes a random word in the post $\mathbf{p}$ to mark it as masked; (ii) the second quarter suffers 1/4 of its continuous words masked by a substring of a post $\mathbf{q} \in \mathcal{H}_p$ of equal length; (iii) the third quarter contains only 3/4 of random words from $\mathbf{p}$; and (iv) the fourth quarter is left unchanged.
Applying Pre-trained HiPo modules to competitors. We provide the output of our historical fusion modules as input for the feature processing modules of other competing models (i.e., we replace the output of their original feature extraction).

We apply this process to enhance the competing methods and report their results in Table 5. We can observe an overall performance improvement compared to their original versions in Table 3. In particular, the average accuracy of enhanced competitors increases by 0.6%, 3.1%, 1.0%, and 2.4% on Fakeddit, IFND, Twitter, and Weibo, respectively. In summary, the introduction of historical-based attention mechanisms enriches the information available to content-based methods, improving their detection performance. Considering that the cross-modal ambiguity learner integrated into CAFE is partly trained from the feature fusion module, our implementation replaces the input of the fusion module while conducting zero-based training for each dataset. This restricts the quality of training samples for the ambiguity module (especially datasets with

Table 4: The performance of HiPo under the labeling bias using classifications from labelers (i.e., competing methods).

Labeler	Fakeddit				IFND				Twitter				Weibo			
	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC
VGG-19 [35]	85.4	**79.2**	75.3	90.7	92.8	91.5	93.4	97.4	84.8	85.1	82.3	88.7	86.6	84.7	**88.1**	93.9
BERT [7]	86.0	75.7	79.0	90.5	92.8	**91.9**	93.0	**97.6**	83.4	85.0	80.6	88.6	86.5	84.8	87.7	94.3
Bi-LSTM [10]	86.4	76.4	79.8	91.0	92.8	**91.9**	93.0	**97.6**	84.2	84.6	81.5	**90.3**	86.6	89.8	87.5	94.4
AttRNN [14]	85.9	74.6	79.9	90.5	92.8	**91.9**	93.0	**97.6**	84.0	85.7	80.9	89.1	87.0	89.4	85.3	94.5
SpotFake [37]	85.7	74.5	79.0	**91.4**	92.8	**91.9**	93.0	97.5	83.6	82.1	82.1	89.7	86.5	86.8	86.3	94.1
MVAE [17]	85.4	73.3	78.5	90.6	92.8	**91.9**	93.0	**97.6**	84.4	**86.9**	80.9	89.4	86.6	86.0	87.0	94.1
CAFE [6]	86.3	76.8	79.0	91.1	93.2	**91.9**	93.8	**97.6**	83.6	85.9	79.8	89.4	86.2	**92.8**	82.0	94.4
HiPo (no bias)	**86.7**	71.6	**83.9**	90.6	**93.4**	90.8	**95.2**	**97.6**	**85.2**	86.5	**83.1**	88.7	**88.2**	91.7	85.8	**95.0**

Table 5: The performance of competitors enhanced with historical posts-based attention.

Enhanced Competitor	Fakeddit				IFND				Twitter				Weibo			
	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC	Accuracy	Recall	Precision	AUC
VGG-19 [35]	77.9	51.1	70.7	79.2	73.8	58.3	82.3	79.3	83.2	82.3	81.6	86.4	63.3	66.3	62.5	64.3
BERT [7]	85.5	74.1	78.1	90.4	90.2	87.5	91.9	**96.5**	81.4	81.4	78.5	84.7	88.2	89.1	87.5	88.3
Bi-LSTM [10]	85.4	73.1	76.9	89.2	90.4	**89.1**	90.9	**96.5**	82.2	80.4	80	86.7	**88.9**	86.7	**90.8**	**88.7**
AttRNN [14]	86.4	73.1	81.4	88.7	84.8	87.9	81.9	91.5	**83.8**	80.9	**83.4**	87.6	87.0	85.6	88.1	86.9
SpotFake [37]	86.6	69.4	85.0	**91.3**	82.2	72.1	89.0	90.9	83.4	83.5	80.3	**87.8**	86.5	84.7	87.8	86.2
MVAE [17]	86.6	71.0	84.1	88.8	**91.8**	88.7	**94.0**	96.4	83.6	80.8	82.2	88.4	88.2	**90.3**	86.7	88.5
CAFE [6]	**86.7**	**88.8**	**92.2**	85.2	83.2	83.8	83.8	83.2	82.2	**84.1**	83.1	82.0	88.2	87.5	88.9	88.1

Table 6: The performance of models under shifted training sets settings.

Training set	IFND, Twitter				Fakeddit, Twitter				Fakeddit, IFND			
Testing set	**Fakeddit**				**IFND**				**Twitter**			
Model	Ac	Re	Pr	AUC	Ac	Re	Pr	AUC	Ac	Re	Pr	AUC
HiPo (Ours)	**65.1**	15.7	**36.6**	**52.4**	**68.0**	87.5	61.9	72.9	**67.6**	69.1	63.1	**73.8**
VGG-19 [35]	46.3	50.8	29.9	46.1	46.5	21.1	39.7	42.4	52.2	75.9	48.2	50.8
BERT [7]	57.9	16.9	25.2	41.5	55.1	**99.6**	51.8	65.3	66.7	**83.9**	55.0	70.4
Bi-LSTM [10]	38.9	53.3	26.7	41.7	59.6	**99.6**	54.4	**73.5**	53.9	07.3	46.0	66.4
AttRNN [14]	35.2	**79.3**	30.3	46.0	52.9	11.7	55.8	57.0	52.7	00.4	05.9	49.7
FakeSpot [37]	46.4	58.9	31.2	49.0	58.2	51.4	57.5	62.9	66.8	63.8	63.3	69.5
MVAE [17]	54.1	21.0	23.8	40.0	63.5	85.8	58.2	67.6	54.7	30.2	50.0	58.3
CAFE [6]	59.3	41.7	8.6	51.3	54.7	54.9	**72.5**	54.9	44.4	49.4	**79.4**	29.1

a limited size, such as IFND and Twitter), thus it may impact its performance due to overfitting.

5.2 Transfer Learning in Shifted Labeled News Domain

We explore the learning transfer capabilities of HiPo by applying a model to a different dataset previously unseen during training. Considering the English-language datasets (i.e., Fakeddit, IFND, and Twitter), we run this experiment following three steps: (i) we reformat the fields of posts (dates, character encoding, etc.) in a uniform format across the three datasets; (ii) we train HiPo using two datasets as training sets (e.g., IFND and Twitter); and (iii) we measure the performance with the remaining dataset as testing set (e.g., Fakeddit). We repeat steps (ii) and (iii) for each dataset.

We report the results of transfer learning in Table 6. We can see that HiPo outperforms other methods in transfer learning on larger datasets (Fakeddit and Twitter). This result indicates that the use of our historical posts-based method leads to better transferability. While the learned perception of historical background is distributed in a wider range for the shifted training set, our model can leverage the unlabeled historical posts provided by the testing dataset in an online manner, therefore leading to better resilience to trans-domain adaptation and dataset shifts.

6 CONCLUSION AND FUTURE WORK

In this paper, we proposed HiPo, a historical post-based method to detect fake news in social media posts. We designed and fully implemented all its modules, including its multi-modal and historical background feature fusion modules. Under time-aware experimental settings, we performed an in-depth evaluation of HiPo performance and comparison with other state-of-the-art methods.

Empirical evidence shows that the multi-attention mechanism used in historical-based fusion modules effectively combines the information from a target post with its historical background, improving the model's performance in fake news detection. Moreover, we can use HiPo's pre-trained historical-based fusion modules to enhance another context- or knowledge-based multi-modality fake news detector.

As a future work, we intend to investigate the possible integration of HiPo with other learning approaches (e.g., reinforcement and semi-supervised learning) and other perception channels (e.g., social context, knowledge entities, and textual inferences). We may also improve the efficiency of embedding the context in the background news for detection by combining various data mining approaches and optimizing the model structure.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (May 2017), 211–236.

[2] Nicolas Belloir, Wassila Ouerdane, Oscar Pastor, Émilien Frugier, and Louis-Antoine de Barmon. 2022. A Conceptual Characterization of Fake News: A Positioning Paper. In *Research Challenges in Information Science*, Renata Guizzardi, Jolita Ralyté, and Xavier Franch (Eds.). Springer International Publishing, Cham, 662–669.

[3] Bogdan Botezatu. 2017. Beware of Fake News - From a Cybersecurity Standpoint. https://businessinsights.bitdefender.com/fake-news-cybersecurity accessed 2023.2.1.

[4] Johnny Botha and Heloise Pieterse. 2020. Fake News and Deepfakes: A Dangerous Threat for 21st Century Information Security. In *International Conference on Cyber Warfare and Security*. Academic Conferences International Limited, 57–XII.

[5] Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. 2020. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities* (2020), 141–161.

[6] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference*. 2897–2905.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[8] Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2051–2055.

[9] Karen N. DSouza and Aaron M. French. 2022. Social Media and Fake News Detection using Adversarial Collaboration. In *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS)*. 115–123.

[10] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks* 18, 5-6 (2005), 602–610.

[11] Chen Hajaj, Sixie Yu, Zlatko Joveski, Yifan Guo, and Yevgeniy Vorobeychik. 2019. Adversarial Coordination on Social Networks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 1515–1523.

[12] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[13] Aaron Hurst. 2023. Bard AI chatbot's wrong answer leads to drop in Alphabet shares. https://www.information-age.com/bard-ai-chatbots-wrong-answer-leads-to-drop-in-alphabet-shares-123501533/ accessed 2023.2.12.

[14] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*. 795–816.

[15] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017. Novel Visual and Statistical Image Features for Microblogs News Verification. *IEEE Transactions on Multimedia* 19 (2017), 598–608.

[16] Jean-Noel Kapferer. 2013. *Rumors: Uses, interpretations, and images.* Transaction Publishers.

[17] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*. 2915–2921.

[18] Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th annual meeting of the Association for computational linguistics*. 1173–1179.

[19] Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation. *arXiv* 2011.04088 (2020).

[20] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 3818–3824.

[21] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion proceedings of the the web conference 2018*. 585–593.

[22] Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The world wide web conference*. 3049–3055.

[23] Meta. 2020. Tips to Spot False News. https://www.facebook.com/formedia/blog/third-party-fact-checking-tips-to-spot-false-news accessed 2023.2.13.

[24] Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. (2020), 6149–6157.

[25] Qiong Nan, Danding Wang, Yongchun Zhu, Qiang Sheng, Yuhui Shi, Juan Cao, and Jintao Li. 2022. Improving Fake News Detection of Influential Domain via Domain-and Instance-Level Transfer. In *Proceedings of the 29th International Conference on Computational Linguistics*. 2834–2848.

[26] Dan Saattrup Nielsen and Ryan McConville. 2022. MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3141–3153.

[27] Feargus Pendlebury, Fabio Pierazzi, Roberto Jordaney, Johannes Kinder, and Lorenzo Cavallaro. 2019. TESSERACT: Eliminating experimental bias in malware classification across space and time. In *Proceedings of the 28th USENIX Security Symposium*. 729–746.

[28] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1212–1220.

[29] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting Multi-domain Visual Information for Fake News Detection. *2019 IEEE International Conference on Data Mining (ICDM)* (2019), 518–527.

[30] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 153–162.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[32] Yuxiang Ren, Bo Wang, Jiawei Zhang, and Yi Chang. 2020. Adversarial Active Learning based Heterogeneous Graph Neural Network for Fake News Detection. In *Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM)*. 452–461.

[33] Dilip Kumar Sharma and Sonal Garg. 2023. IFND: a benchmark dataset for fake news detection. *Complex & intelligent systems* 9, 3 (2023), 2843–2863.

[34] Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom Out and Observe: News Environment Perception for Fake News Detection. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (03 2022), 4543–4556.

[35] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. (2015).

[36] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13915–13916.

[37] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. SpotFake: A Multi-modal Framework for Fake News Detection. *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)* (2019), 39–47.

[38] Maciej Szpakowski and Renato Cordeiro. 2020. Fake News Corpus. https://github.com/several27/FakeNewsCorpus accessed 2023.1.21.

[39] Vaibhav Vaibhav, Raghuram Mandyam, and Eduard Hovy. 2019. Do Sentence Interactions Matter? Leveraging Sentence Level Representations for Fake News Classification. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*. 134–139.

[40] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *ArXiv* abs/1706.03762 (2017).

[41] Antonia M. Villarruel and Richard James. 2022. Preventing the spread of misinformation. https://www.myamericannurse.com/preventing-the-spread-of-misinformation/ accessed 2023.2.12.

[42] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 849–857.

[43] Yaqing Wang, Fenglong Ma, Haoyu Wang, Kishlay Jha, and Jing Gao. 2021. Multimodal emergent fake news detection via meta neural process networks. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 3708–3716.

[44] Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 international conference on multimedia retrieval*. 540–547.

[45] Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. 2020. Weak supervision for fake news detection via reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 516–523.

[46] Chenxi Whitehouse, Tillman Weyde, Pranava Madhyastha, and Nikos Komninos. 2022. Evaluation of fake news detection with knowledge-enhanced language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1425–1429.

[47] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*. 2560–2569.

[48] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management* 58 (2021), 102610 – 102610.

[49] Ruichao Yang, Xiting Wang, Yiqiao Jin, Chaozhuo Li, Jianxun Lian, and Xing Xie. 2022. Reinforcement subgraph reasoning for fake news detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2253–2262.

[50] Long Ying, Hui Yu, Jinguang Wang, Yongze Ji, and Shengsheng Qian. 2021. Fake News Detection via Multi-Modal Topic Memory Network. *IEEE Access* 9 (2021), 132818–132829.

[51] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2017. A convolutional approach for misinformation identification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 3901–3907.

[52] Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2019. Multimodal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM international conference on multimedia*. 1942–1951.

[53] Guangxiang Zhao, Xu Sun, Jingjing Xu, Zhiyuan Zhang, and Liangchen Luo. 2019. MUSE: Parallel Multi-Scale Attention for Sequence to Sequence Learning. *ArXiv* abs/1911.09483 (2019).

[54] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multimodal Fake News Detection. (2020), 354–367.

[55] Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2120–2125.