Unveiling Implicit Deceptive Patterns in Multi-modal Fake News via Neuro-Symbolic Reasoning

Yiqi Dong[1], Dongxiao He[1,2], Xiaobao Wang[2*], Youzhu Jin[3], Meng Ge[4], Carl Yang[5], Di Jin[1,2]

[1]School of New Media and Communication, Tianjin University, Tianjin, China,
[2]Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China,
[3]Beijing-Dublin International College, Beijing University of Technology, Beijing, China,
[4]Saw Swee Hock School of Public Health, National University of Singapore, Singapore,
[5]Department of Computer Science, Emory University, Georgia, USA.
{dongyiqi, hedongxiao,wangxiaobao}@tju.edu.cn, youzhu.jin@ucdconnect.ie, gemeng@nus.edu.sg, j.carlyang@emory.edu, jindi@tju.edu.cn.

Abstract

In the current Internet landscape, the rampant spread of fake news, particularly in the form of multi-modal content, poses a great social threat. While automatic multi-modal fake news detection methods have shown promising results, the lack of explainability remains a significant challenge. Existing approaches provide superficial explainability by displaying learned important components or views from well-trained networks, but they often fail to uncover the implicit deceptive patterns that reveal how fake news is fabricated. To address this limitation, we begin by predefining three typical deceptive patterns, namely image manipulation, cross-modal inconsistency, and image repurposing, which shed light on the mechanisms underlying fake news fabrication. Then, we propose a novel Neuro-Symbolic Latent Model called NSLM, that not only derives accurate judgments on the veracity of news but also uncovers the implicit deceptive patterns as explanations. Specifically, the existence of each deceptive pattern is expressed as a two-valued learnable latent variable, which is acquired through amortized variational inference and weak supervision based on symbolic logic rules. Additionally, we devise pseudo-siamese networks to capture distinct deceptive patterns effectively. Experimental results on two real-world datasets demonstrate that our NSLM achieves the best performance in fake news detection while providing insightful explanations of deceptive patterns.

1 Introduction

Nowadays, the Internet's rapid expansion has greatly facilitated the dissemination and acquisition of information. However, this also provides an avenue for malicious actors to fabricate and spread fake news with ulterior motives. The ubiquity of fake news makes it challenging for individuals to discern reliable information online and significantly threatens the modern media ecosystem (Allcott and Gentzkow 2017; Wang et al. 2023). This hazard becomes even more evident against the backdrop of Large Language Models (LLMs) such as ChatGPT (OpenAI 2023), which inadvertently generates and propagates fake information due to AI hallucination (Goldstein et al. 2023). On the other hand, the Internet is increasingly flooded with multi-modal (e.g., text and image) online posts, renowned for their heightened allure and deceptive attributes (Cao et al. 2020). Consequently, developing automatic detection systems to verify and combat multi-modal fake news has become an urgent necessity.

(a) Image manipulation (b) Image repurposing

Figure 1: Typical examples of fake news manifesting different deceptive patterns.

Existing efforts utilizing Deep Neural Networks (DNN) have been made to tackle the multi-modal fake news detection problem by integrating various features (Dhawan et al. 2022) by constructing graph (Jin et al. 2022a,b) or exploring cross-modal correlations (Qi et al. 2021; Dong et al. 2023). While achieving promising results, such methods often lack explainability and are commonly referred to as "black boxes", as they focus on learning unclear latent features (Mishima and Yamana 2022). Poor explainability not only extremely undermines user trust but also impedes system debugging and upgrading. Recently, several approaches have attempted to provide explanations by highlighting the contributive semantics components within text description and image region (Wu, Liu, and Zhang 2023), exhibiting coarse prediction scores from each view that includes individual modality and cross-modality correlation (Ying et al. 2023), or jointly locating evident contents and their logic

*Corresponding author

interactions (Liu, Wang, and Li 2023). These explanations display the input components or views most relevant to the predictions in some way. However, they overlook a different route to explainability, one that involves uncovering how fake news is fabricated, which we term as deceptive patterns implicit in the news. Our starting point is that tracing back to the root, the diverse and unique features manifested within fake news articles stem from various deceptive patterns employed during their creation. We posit that unveiling these patterns could enhance the detection of fake news and provide succinct explanations behind the news being fake.

Accordingly, inspired by common visual patterns prevalent in fake news (Cao et al. 2020), we explore three primary deceptive patterns frequently utilized to forge fake news: image manipulation, cross-modal inconsistency, and image repurposing. Among them, cross-modal inconsistency refers to the semantic inconsistency between text and image, which is a readily understandable pattern. Hence, we present two imperceptible fake news examples related to the other two patterns, sourced from Snopes[1] in Figure 1. At first glance, both of them do not appear to be fake. However, in the original image shown in Figure 1 (a), Trump was holding a pen, not a commemorative bill, clearly indicating image manipulation. As for the image in Figure 1 (b), it actually depicts an unnamed mother, daughter, and maid in Johannesburg, South Africa, during apartheid, which conflicts with the textual description of the news, revealing image repurposing.

In practice, predicting news authenticity and mining deceptive patterns as explanations jointly are challenging due to the lack of deceptive pattern labels for news samples in the dataset. Furthermore, deceptive patterns within fake news, as exemplified in Figure 1, are often not easily recognizable even to human annotators, rendering manual labeling unfeasible and augmenting the intricacy of our task. Thus, this study attempts to answer the question: can we unveil those unlabeled deceptive patterns in multi-modal news as an insightful and concise explanation?

Fortunately, from the perspective of human cognition, there is at least one deceptive pattern if the news is fake, while no deceptive pattern if the news is real. Inspired by the powerful expressive capabilities of first-order logic language in capturing complex relationships (Enderton 2001), our mind starts by formalizing these rules using first-order logic as a form of weak supervision inspired by (Chen et al. 2022a). By doing so, we establish a correlation between the available labels for news authenticity and the presence of unsupervised deceptive patterns, enabling the underlying deceptive patterns to be automatically learned. Building upon these insights, we propose a Neuro-Symbolic Latent Model (NSLM) that concurrently predicts the veracity of news and reveals deceptive patterns as explanations. Central to our NSLM is the modeling of each deceptive pattern's existence as a corresponding two-valued learnable latent variable, learned through weak supervision from logic rules. Specifically, the presence prediction of each deceptive pattern is treated as an atomic predicate in the logic rules, and the final prediction is aggregated using the conjunction of these individual predicates. This design effectively captures that the presence of one or more deceptive patterns indicates fake news, whereas the absence of all deceptive patterns confirms the news as real. Overall, we formulate the problem as a probabilistic maximum likelihood estimation with latent variables and adopt variational auto-encoding (Kingma and Welling 2014) to address it. For effectively capturing different deception patterns, we design a pseudo-siamese network within the encoder. In addition, we employ a distill-based strategy to influence the learning of latent variables subject to the pre-specified logic rules.

To sum up, the contributions of our work are three-folded:

- We propose a novel fake news detection approach named NSLM, capable of revealing the unlabeled deceptive patterns within multi-modal news data as illuminating explanations.
- Each deceptive pattern is treated as a two-valued learnable latent variable, and we introduce logic rules based on human cognition to provide weak supervision for the existence of the proposed three deceptive patterns.
- Experimental results on two benchmark datasets demonstrate that our NSLM achieves state-of-the-art performance in fake news detection and provides clear explanations for its predictions.

2 Preliminaries

2.1 Task Definition

Given a news article x with text x_t, an attached image x_v, and image contexts x_r retrieved by the image inverse search (Zlatkova, Nakov, and Koychev 2019), this work aims at predicting its label $y \in \{\textit{Real, Fake}\}$ by modeling the probability distribution $p(y \mid x)$, while at the same time mining its deceptive patterns acting as explanations. Here, we associate the presence of each proposed deceptive pattern with a two-values learnable latent variable $z_k \in \{\textit{Not Exist, Exist}\}, k \in \{\mathrm{IM}, \mathrm{CI}, \mathrm{IR}\}$, where $z_{\mathbf{IM}}$ for image manipulation, $z_{\mathbf{CI}}$ for cross-modal inconsistency, and $z_{\mathbf{IR}}$ for image repurposing. Note that we assume the independence of z_k. We further define $z = (z_{\mathbf{IM}}, z_{\mathbf{CI}}, z_{\mathbf{IR}})$. Formally, our objective function based on maximum likelihood estimation is given as follows:

$$\max \mathcal{O} = \mathbb{E}_{(x,y^*)\sim p_{\text{train}}} \log p\left(y^* \mid x\right), \tag{1}$$

where y^* is the ground truth label of news article x, and p_{train} denotes the distribution of the training data.

2.2 Logic Rules

To introduce weak supervision signals for imperceptible deceptive patterns and subsequently unveil these patterns as explanations, our model incorporates logic rules based on human intuition. We empirically observe that fake news typically involves at least one deceptive pattern, whereas true news lacks any deceptive patterns. These logical intuitions are regarded as a crucial link connecting the veracity of news and the presence of deceptive patterns. Moreover, they are well-suited to be represented using first-order logic language

[1] https://www.snopes.com

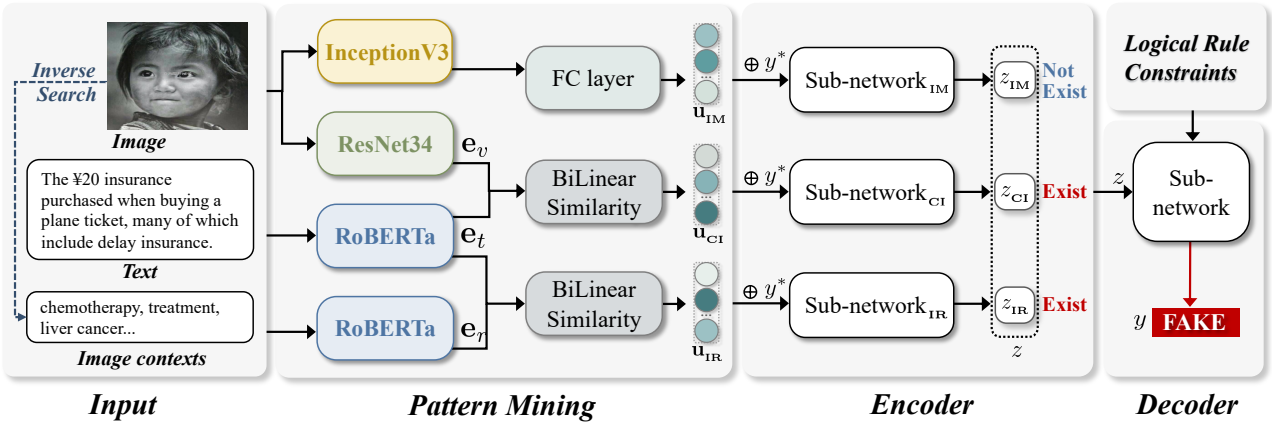


Figure 2: Architecture of the proposed NSLM. The main modules of our model include Pattern Mining, Encoder, Decoder, and Logical Rule Constraints. The learning of deceptive patterns in NSLM is constrained by symbolic logic rules. Here $\oplus$ denotes the concatenation operation, both y^* and z are converted into vectors in continuous space.

with strong expressive capabilities, which can be formulated as follows:

$$z_{\text{IM}} \wedge z_{\text{CI}} \wedge z_{\text{IR}} \Rightarrow y, \tag{2}$$

where z_k serves as the unary body predicate, y serves as the head predicate, and the conjunction operator $\wedge$ shows the relationship between body predicates. Then the detailed reasoning rules derived by Eq. (2) can be defined as:

$$\begin{aligned} y &= \textit{Fake}, \text{ iff } \exists\ z_k = \textit{Exist}, \\ y &= \textit{Real}, \text{ iff } \forall\ z_k = \textit{Not Exist}. \end{aligned} \tag{3}$$

With the above definitions in place, we will subsequently introduce the proposed latent model NSLM and outline how the logic rules are employed to supervise it.

3 Methodology

Figure 2 illustrates the framework of the proposed NSLM, which aims to uncover implicit deceptive patterns in fake news acting as explanations when giving authenticity predictions of news. To achieve this, we formulate a neuro-symbolic latent model and represent each deceptive pattern as a two-valued learnable latent variable z_k that requires inference. As shown in Figure 2, our NSLM consists of a pattern mining module, encoder, and decoder, while also integrating a logical constraint component for guided learning. Given a multi-modal news article as input, the pattern mining module initially extracts coarse-grained features linked to three deception modes using pre-trained models. Subsequently, the encoder employs pseudo-siamese networks to process features from the pattern mining module, producing distinct latent variables, which are then fed into the decoder for final news credibility predictions. Besides, taking inspiration from (Hu et al. 2016), we apply knowledge distillation to incorporate information from the logic rules into variables y and z. In practice, we optimize the NSLM through a variational inference-based algorithm, where both the encoder and decoder are jointly optimized to train the model.

3.1 Probabilistic Formalization

We begin by formulating fake news detection from a probabilistic standpoint, where the underlying deceptive patterns are treated as latent variables. Assuming that news articles are independent of each other, the objective function in Eq. (1) could be equivalently decomposed into maximizing the logarithmic likelihood function for each news article. Hence, we next delve into the details of our NSLM from the perspective of an individual news article. For a piece of news x, our objective is to compute the target distribution, considering the incorporation of latent variables, as follows:

$$p_\theta(y \mid x) = \sum_z p_\theta(y \mid z, x)\, p(z \mid x), \tag{4}$$

where $p_\theta(y \mid z, x)$ defines the conditional probability of y given input x and latent variables z parameterized by θ, and $p(z \mid x)$ denotes the prior distribution of the latent variables z conditioned on the input x.

Nevertheless, due to latent variables introducing additional dimensions to the parameter space, direct optimization using the EM algorithm becomes computationally intractable. To address this, we adopt recent advancements in variational inference, i.e., the amortization of the variational posterior distribution using neural networks (Kingma and Welling 2014). Specifically, a variational posterior distribution $q_\omega(z \mid x, y)$ is introduced to approximate the true posterior distribution $p_\theta(z \mid x, y)$, which makes the objective function for news x into maximizing the well-known Evidence Lower BOund (ELBO). The ELBO is defined as:

$$\mathbb{E}_{q_\omega(z|x,y)}\left[\log p_\theta\left(y \mid z, x\right)\right] - D_{\text{KL}}\left[q_\omega(z \mid x, y) \| p(z \mid x)\right], \tag{5}$$

where $D_{\text{KL}}[\cdot]$ denotes the Kullback-Leibler divergence. Here we treat the Eq. (5) with a negative sign as one term of the overall loss function to minimize:

$$\mathcal{L}_{\text{elbo}\,(\theta,\omega)} = -\text{ELBO}. \tag{6}$$

3.2 Parameterization

Pattern Mining In pursuit of capturing the three underlying deceptive patterns within fake news, we devise three branches to extract pertinent features $\mathbf{u}_k, k \in \{\mathrm{IM}, \mathrm{CI}, \mathrm{IR}\}$ corresponding to three patterns, i.e., Image Manipulation, Cross-modal Inconsistency, and Image Repurposing.

To capture the image manipulation, we leverage InceptionV3 (Szegedy et al. 2016) coupled with a fully connected layer to extract coarse features $\mathbf{u}_{\mathbf{IM}} \in \mathbb{R}^d$ (d is the fixed feature dimension) from the image's frequency domain. This choice stems from the fact that recompressed or tampered images often exhibit periodicity in the frequency domain (Qi et al. 2019), which can be effectively discerned by InceptionV3. To find out the cross-modal inconsistency, the pre-trained ResNet34 (He et al. 2016) and RoBERTa (Liu et al. 2019) with fully connected layers are employed to extract semantic features $\mathbf{e}_v, \mathbf{e}_t \in \mathbb{R}^d$ from image and text respectively. Leveraging the advantages of BiLinear Similarity (Kim et al. 2017) in capturing intricate relationships between two features, we apply it to uncover inconsistencies between $\mathbf{e}_v$ and $\mathbf{e}_t$, which can be computed as:

$$\mathbf{u}_{\mathbf{CI}}[i] = \mathbf{e}_v^\top \mathbf{W}_{\mathbf{CI}}^i \mathbf{e}_t + b_{\mathbf{CI}}^i, \tag{7}$$

where $\mathbf{u}_{\mathbf{CI}} \in \mathbb{R}^d$ denotes the pattern features for the cross-modal inconsistency, and $\mathbf{u}_{\mathbf{CI}}[i], i \in (1, 2, ..., d)$ is the component value of the i-th dimension of $\mathbf{u}_{\mathbf{CI}}$, $\mathbf{W}_{\mathbf{CI}}^i \in \mathbb{R}^{d \times d}$ is a learnable parameter matrix, $b_{\mathbf{CI}}^i$ is a bias for $\mathbf{u}_{\mathbf{CI}}[i]$. As for image repurposing, it is difficult to be detected solely by the news content since it does not contain any contexts where the original image appeared. Therefore, we employ image reverse search to retrieve the contextual information of images from the Web. This process can be efficiently automated and scaled to a large number of images using Google's Vision API [2], which returns a list of pages and entities related to the image. We gather the concatenated entities as image contexts. They are fed into the RoBERTa to obtain its representation $\mathbf{e}_r \in \mathbb{R}^d$. Similarly, another BiLinear Similarity is applied between text embedding $\mathbf{e}_t$ and image contexts embedding $\mathbf{e}_r$ to learn their differences:

$$\mathbf{u}_{\mathbf{IR}}[i] = \mathbf{e}_t^\top \mathbf{W}_{\mathbf{IR}}^i \mathbf{e}_r + b_{\mathbf{IR}}^i, \tag{8}$$

where $\mathbf{u}_{\mathbf{IR}} \in \mathbb{R}^d$ represents the pattern features indicating the image repurposing. The parameters dimension is consistent with that in Eq.(7).

Encoder & **Decoder** After calculating the above representations, we parameterize the variational distribution $q_\omega(z \mid x, y)$ and target distribution $p_\theta(y \mid z, x)$ with neural networks, which corresponds to encoder and decoder in the variational autoencoder, respectively.

The encoder is designed as a pseudo-siamese structure, whose goal is to generate a set of latent variables z that represent diverse deception patterns. More precisely, due to $q_\omega(z \mid x, y) = \prod_k q_{\omega,k}(z_k \mid x, y)$, we employed three structurally consistent but weight-disjoint sub-networks to model the three distinct distribution $q_{\omega,k}(z_k \mid x, y)$, and each sub-network consists of two fully connected layers with a softmax function. For each sub-network$_k$, it utilizes the concatenation of $\mathbf{u}_k$ and embeddings of y as input to generate the probability distribution of z_k.

The decoder mirrors the encoder's sub-network structure. It takes the concatenation of the probability distribution of $z_{\mathbf{IM}}$, $z_{\mathbf{CI}}$, $z_{\mathbf{IR}}$, along with $\mathbf{e}_v$ and $\mathbf{e}_t$, as input to predict the distribution of the news credibility label y.

Logical Rule Constraints We adopt the knowledge distillation strategy with a teacher model and a student model to integrate logic rules into latent variables, providing weak supervision inspired by (Chen et al. 2022a). The teacher model projects the variational distribution $q_\omega(z \mid x, y)$ into a subspace $q_\omega^\star(y_z \mid x, y)$ adhering to the logic rules, with $y_z \in \{\textit{Real, Fake}\}$ representing the logical aggregation of z. This allows us to transfer logical knowledge to the student model $p_\theta(y \mid z, x)$ that we aim to optimize. The whole process can be understood analogously to human education, where a knowledgeable teacher possesses systematic general rules and guides students by offering her solutions to specific questions (Hu et al. 2016). The following distillation loss is defined to guide this process:

$$\mathcal{L}_{\text{logic}}(\theta, \omega) = D_{\mathrm{KL}}\left(p_\theta(y \mid z, x) \,\|\, q_\omega^\star(y_z \mid x, y)\right). \tag{9}$$

A pivotal aspect here pertains to how to get the logical aggregation label y_z, we transfer hard logic defined in Section 2.2 into soft logic with product t-norms (Li et al. 2019) to ensure differentiability. Then the projected distribution $q_\omega^\star(y_z \mid x, y)$ is given by:

$$\begin{aligned} q_\omega^\star(y_z = \textit{Real} \mid x, y) &= \prod_k q_{\omega,k}(z_k = \textit{Not Exist} \mid x, y), \\ q_\omega^\star(y_z = \textit{Fake} \mid x, y) &= 1 - q_\omega^\star(y_z = \textit{Real} \mid x, y). \end{aligned} \tag{10}$$

3.3 Model Learning

Next, we introduce the optimization strategy to achieve the objective in Eq. (1). Combining the ELBO loss $\mathcal{L}_{\text{elbo}}$ and logic loss $\mathcal{L}_{\text{logic}}$, our final loss function $\mathcal{L}_{\text{all}}$ for the news x is defined as:

$$\mathcal{L}_{\text{all}}(\theta, \omega) = (1 - \mu)\mathcal{L}_{\text{elbo}}(\theta, \omega) + \mu\mathcal{L}_{\text{logic}}(\theta, \omega), \tag{11}$$

where $\mu \in (0, 1)$ is placed to balance between the two terms.

During the training process, all training news samples are sequentially processed through the pattern mining module, encoder, and decoder, which are jointly optimized using Eq. (11). It's crucial to highlight that in the variational distribution $q_\omega(z \mid x, y)$, y actually is the ground-truth label y^* for each x during training. In our encoder, y^* is converted into one-hot encoding and then used to derive embeddings.

3.4 Model Inference

During the testing phase, the input news samples are first processed through the pattern mining module. Then we randomly initialize the probability of z from a standard Gaussian distribution and use it as input for the decoder. The output distribution of news authenticity y generated by the decoder is then passed through the encoder. This process, involving passing through the decoder and encoder, continues

[2] http://cloud.google.com/vision/

to update the distributions of y and z iteratively until convergence. As a result, we obtain both the final news veracity and the latent variables with respect to deceptive patterns, providing valuable insights into how a piece of news is forged. This end-to-end training and decoding approach contributes to a more reliable and transparent explanation mechanism.

4 Experiments

4.1 Datasets and Experimental Setup

Dateasets We evaluate the proposed NSLM on two real-world datasets called *Fakeddit* (Nakamura, Levy, and Wang 2020) and *Weibo* (Jin et al. 2017), respectively. The Fakeddit dataset is derived from diverse subreddits on the Reddit platform, comprising comments and metadata. Notably, due to the abundance of short-text samples in Fakeddit, extracting their internal semantic information poses challenges. To this end, we create a subset of the dataset by selecting samples with a token count greater than 15 for further evaluation. In the Weibo dataset, the real news samples are gathered from Xinhua News Agency, a reputable news source in China, while the fake news samples are verified using Weibo's official rumor debunking system.

In our study, we exclude samples for which the corresponding image or Google inverse search results are unavailable. Statistically, Fakeddit comprises 31,011 news samples for training and 6,181 for testing, whereas Weibo consists of 5,455 news samples for training and 1,493 for testing.

Comparison Models To validate the performance of the proposed NSLM, we compare it against 11 baselines, including two categories of models: 1) **Uni-modal methods**, consisting of the pre-trained ResNet34 (He et al. 2016), InceptionV3 (Szegedy et al. 2016), and RoBERTa (Liu et al. 2019) models combined with a fully connected layer; 2) **Multi-modal methods**, containing EANN (Wang et al. 2018), SpotFake (Singhal et al. 2019), BTIC (Zhang, Gui, and He 2021), HMCAN (Qian et al. 2021), CAFE (Chen et al. 2022b), CMC (Wei et al. 2022), BMR (Ying et al. 2023) and LogicDM (Liu, Wang, and Li 2023). These methods commonly utilize deep neural networks and well-designed strategies, such as cross-modal knowledge exploitation and contrastive learning. Although BMR and LogicDM offer a certain degree of explainability, they do not effectively identify the deceptive patterns that reveal how fake news is fabricated. In experiments, we employ the same pre-processed data to re-run the official code provided by the aforementioned papers for comparison.

Implementation Details In our NSLM, we adopt a randomly sampled Gaussian distribution as the prior distribution $p(z \mid x)$. We set the dimension d to 256 and the trade-off weight μ to 0.5. During training, we use a batch size of 8, while for testing, the batch size is set to 16. We employ a learning rate of 1e-5 for both datasets. The Fakeddit dataset allows a maximum text length of 45 and an image contexts length of 12, while for the Weibo dataset, the respective maximum lengths are 110 for text and 10 for image contexts. The whole model is trained with the Adaptive Moment Estimation (Adam) optimizer (Kingma and Ba 2014).

4.2 Experiment Results

Table 1 presents a comprehensive comparison of our NSLM against popular baseline methods in terms of Accuracy, Precision, Recall, and Macro F_1 score. The results consistently indicate NSLM's superior performance over other models across all four metrics on both datasets, especially NSLM brings 1.6% and 1.0% improvements in Accuracy over best-performing CMC on the Fakeddit and Weibo datasets, respectively. Such performance proves the efficacy of unraveling the mechanisms underpinning fake news fabrication.

Table 1 also reveals that the image uni-modal approaches yield quite inefficient performance, particularly evident in the Weibo dataset characterized by intricate semantic images. In stark contrast, the text uni-modal method exhibits much better performance, emphasizing the pivotal role of textual information in effective fake news detection. Moreover, the multi-modal methods generally achieve even more promising results, which demonstrates the potential for complementary effects of the two modalities to improve detection accuracy. Among the multi-modal models, we can observe the results of CAFE are suboptimal. This could be attributed to CAFE's consideration of cross-modal ambiguity, which can be regarded as a specific aspect of deceptive patterns and might not universally apply in real scenarios. On the other hand, we notice the exceptional performance of CMC, which may relate to its adeptness in leveraging feature correlations through a well-designed mutual learning strategy. It's important to note that CMC's two-stage nature introduces additional training time and complexity compared to others. Regarding the best results of our NSLM, we believe this benefits from our model's ability to reveal how fake news is fabricated, enabling the identification of common deceptive patterns shared among fake news.

4.3 Ablation Study

To thoroughly comprehend the impact of each suggested deceptive pattern and its collective significance, we systematically exclude each pattern (w/o z_k) individually and combinations of two patterns (w/o z_k, z_j, where $k \neq j$). The empirical results for model variants in Accuracy and Macro F1 scores are reported in Table 2.

From Table 2, it is evident that the removal of two latent variables has a more pronounced adverse impact on the model's performance than the elimination of only one. This suggests that discarding more pattern features would lead to inferior results. Specifically, on the Fakeddit dataset, we found that the contribution of cross-modal inconsistency (z_{CI}) holds slightly higher significance among the three deceptive patterns. Conversely, on the Weibo dataset, image manipulation (z_{IM}) is the most influential. This divergence may arise from the variations in deceptive pattern distributions across datasets with different languages and platforms. The ablation results confirm the significance of capturing the three proposed deceptive patterns in enhancing performance, as removing any of these patterns results in decreased accuracy.

Categories	Models	Fakeddit				Weibo			
		Accuracy	Precision	Recall	Macro F_1	Accuracy	Precision	Recall	Macro F_1
Uni-modal (image)	ResNet34	0.721	0.722	0.630	0.632	0.561	0.556	0.556	0.555
	InceptionV3	0.737	0.726	0.665	0.674	0.584	0.583	0.583	0.583
Uni-modal (text)	RoBERTa	0.832	0.819	0.806	0.812	0.829	0.828	0.829	0.829
Multi-modal (text+image)	EANN	0.826	0.821	0.790	0.801	0.727	0.749	0.738	0.726
	SpotFake	0.891	0.901	0.859	0.875	0.839	0.840	0.842	0.838
	BTIC	0.897	0.888	0.885	0.886	0.835	0.838	0.838	0.835
	HMCAN	0.892	0.885	0.876	0.880	0.832	0.833	0.835	0.832
	CAFE	0.848	0.844	0.816	0.826	0.812	0.818	0.817	0.812
	CMC	0.909	0.906	0.892	0.898	0.875	0.875	0.877	0.875
Multi-modal (text+image) +Explainability	BMR	0.901	0.890	0.890	0.891	0.843	0.843	0.843	0.843
	LogicDM	0.873	0.867	0.850	0.858	0.852	0.852	0.852	0.852
	NSLM (Ours)	**0.925**	**0.919**	**0.915**	**0.917**	**0.885**	**0.884**	**0.885**	**0.884**

Table 1: Comparison with the considered uni-modal and multi-modal baselines on Fakeddit and Weibo datasets in terms of Accuracy, Precision, Recall, and Macro F_1 score. The best results are in bold.

Models	Fakeddit		Weibo	
	Accuracy	F_1	Accuracy	F_1
NSLM	**0.925**	**0.917**	**0.885**	**0.884**
w/o z_{IM}	0.923	0.914	0.864	0.863
w/o z_{CI}	0.922	0.913	0.874	0.873
w/o z_{IR}	0.923	0.914	0.874	0.874
w/o $z_{\text{IM}}, z_{\text{CI}}$	0.918	0.910	0.865	0.865
w/o $z_{\text{IM}}, z_{\text{IR}}$	0.921	0.912	0.865	0.865
w/o $z_{\text{CI}}, z_{\text{IR}}$	0.919	0.910	0.861	0.861

Table 2: Comparison with different variants of NSLM. The best results are in bold. The "w/o" is the abbreviation of "without". The "F_1" denotes "Macro F_1".

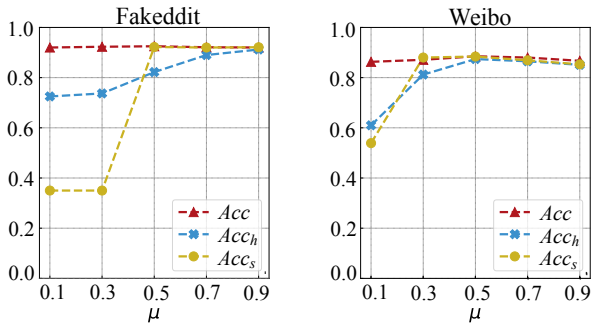


Figure 3: Evaluation of the overall accuracy and deceptive patterns quality by varying the trade-off weight μ in the loss function Eq. (11).

4.4 Overall Evaluation of Deceptive Patterns

The revelation of underlying deceptive patterns in fake news is a fundamental aspect of our model. To achieve this, we employ logical constraints to weakly supervise the learning of deceptive patterns z. By adjusting the trade-off weight μ in the overall loss function Eq. (11), we aim to investigate the impact of varying levels of logical supervision on the quality of learned latent variables z, and how it subsequently affects the model performance. The results depicted in Figure 3 show the influence of varying the weight μ from 0.1 to 0.9 on three key metrics: Acc evaluates the overall accuracy of the predicted label y; Acc_h and Acc_s indicate the accuracy of y_z obtained by logical aggregation of z through hard logic (Eq. (3)) and soft logic (Eq. (10)) respectively, which evaluate the overall quality of the learned z.

Starting with $\mu = 0.1$, Figure 3 illustrates that, with limited logical supervision, both Acc_h and Acc_s exhibit rather low values. This means that the latent variables z inadequately capture precise deceptive patterns in the absence of adequate logical guidance. As we increase this weight gradually, the quality of z improves significantly. Notably, when it hits 0.5, all the metrics achieve the best. However, as this value continues to increase, the model's performance tends to stabilize or even exhibit a slight decline. This observation highlights the significance of logic rules in shaping the quality of learned z and emphasizes that a moderate value of μ is crucial to achieving optimal model performance. In addition, the overall Acc exhibits remarkable robustness to variations in the weight, showing relatively minor fluctuations throughout the range. This finding verifies our NSLM's ability to discover deceptive patterns without compromising its overall predictive accuracy. In conclusion, the experimental analysis clarifies the effectiveness of deceptive patterns and the essential role of logical constraints.

4.5 Case Study

To give an intuitive comprehension of our NSLM's explainability, we display the outputs of several fake news cases from the Weibo dataset in Figure 4. This illustration includes

Text: Pigmen were found in Jishan Pa goda, Shuangyang Town, Zhangping City, Fujian Province.
Image contexts: Pigman, Human-Beast, Hybrid

Learned deceptive patterns: **image manipulation**
Predicted label: **Fake**

(a) Good case

Text: The ¥20 insurance purchased when buying a plane ticket, many of which include delay insurance.
Image contexts: Chemotherapy, Treatment, Liver cancer.

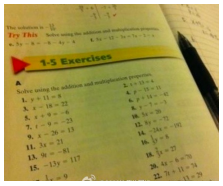
Learned deceptive patterns: **cross-modal inconsistency, image repurposing**
Predicted label: **Fake**

(b) Good case

Text: #Boston Explosion# The 8-year-old child died in today's explosion.
Image contexts: Fitness, Power t-shirts, French marathon

Learned deceptive patterns: **None**
Predicted label: **Fake**

(c) Bad case



Text: The 10th grade Math in the U.S. Please summarize your feelings in one sentence!
Image contexts: Mathematics Textbook, High School, USA

Learned deceptive patterns: **None**
Predicted label: **Real**

(d) Bad case

Figure 4: Several fake examples of learned deceptive patterns from the Weibo dataset. The texts are translated from Chinese to English. Good cases and bad cases showcase the successes and limitations of our NSLM, respectively.

the retrieved image contexts, learned deceptive patterns, and predicted authenticity labels for each news example.

For the first two cases, NSLM performs effectively. In case (a), it captures the presence of image manipulation, since the manipulated ears in its image provide clear evidence of fabrication through image tampering. In case (b), a stark incongruity between text and image is evident. This inconsistency also existed between text and the retrieved image contexts such as chemotherapy, therapy, and liver cancer that deviate from the textual description. The above observations substantiate the presence of both cross-modal inconsistency and image repurposing in this case.

We also present two bad cases (c) and (d) in Figure 4 to further analyze the limitations of our NSLM. In case (c), our model incorrectly identifies the absence of deceptive patterns, likely due to poor image contexts retrieved through reverse search, failing to recognize the actual content of the image depicting a girl wearing a race bib for the Chosun City Jogging 5K. However, as we can see, the final predicted label is correct, suggesting that the imposed logical constraints may not be effectively incorporated. In the last example, the learned results also indicate the absence of all three deceptive patterns, resulting in an erroneous judgment of the prediction y that tries to be consistent with the logical aggregated label of z. While the image indeed represents a genuine math book, determining whether it belongs to the mentioned American 10th-grade mathematics requires leveraging external knowledge.

It is worth mentioning that though several approaches achieve explainability by emphasizing specific content components or views in image and text of news, real-world scenarios may not always allow humans the time or expertise to carefully analyze every sample. Instead, they require clear and concise explanations. NSLM excels in providing such explanations directly, unveiling the deceptive patterns in fake news. For example, if people know that the case in Figure 4 (a) contains a deceptive pattern of image manipulation, they can quickly judge it as fake. This superiority becomes particularly valuable when dealing with large-scale datasets and time-sensitive situations, where quick and accurate decisions are paramount.

5 Related Works

Explainable fake news detection has become a prominent area of research. For instance, (Chen et al. 2022a) made notable contributions in the field of fact-checking by utilizing evidential information and combining phrase-level veracity reasoning to determine the veracity of entire claims. This approach provides a more clear explanation. (Ying et al. 2023) disentangles multi-modal features through single-view prediction and explains which view is critical to the final decision. (Liu, Wang, and Li 2023) integrated logical clauses to express the reasoning process of the target task, identifying the contributing factors and selecting appropriate perspectives for explanations. While the above models achieved certain explainability, none could reveal the deceptive patterns within multi-modal fake news as concise explanations. Our work uniquely bridges the gap by unveiling those patterns through the constraints of symbolic logic rules.

6 Conclusion

In this work, we blaze a novel path to explainability by elucidating unlabeled deceptive patterns within multi-modal news. In detail, we propose NSLM that converts the veracity of a news article into the presence of a set of deceptive patterns, thereby providing insightful explanations.

Deceptive practices are constantly evolving, potentially giving rise to new patterns. So in the future, we plan to extend our model into a dynamically adaptable framework to adapt to these evolving patterns through the incorporation of a versatile combined pattern mining module, which is an extension of the Pattern Mining module in Figure 2. This extended module operates by amalgamating various input sources, thereby enabling the selection of specific inputs and the extraction of implicit deceptive pattern characteristics.

7 Acknowledgments

This work was supported by the National Key Research and Development Program of China (2023YFC3304503), the National Natural Science Foundation of China (No. 92370111, 62272340, 62276187, 62302333), and the China Postdoctoral Science Foundation (No. 2023M732593). Carl Yang was not supported by any funds from China.

References

Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2): 211–36.

Cao, J.; Qi, P.; Sheng, Q.; Yang, T.; Guo, J.; and Li, J. 2020. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, 141–161.

Chen, J.; Bao, Q.; Sun, C.; Zhang, X.; Chen, J.; Zhou, H.; Xiao, Y.; and Li, L. 2022a. Loren: Logic-regularized reasoning for interpretable fact verification. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, volume 36, 10482–10491.

Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; and Shang, L. 2022b. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the 14th ACM Web Conference*, 2897–2905.

Dhawan, M.; Sharma, S.; Kadam, A.; Sharma, R.; and Kumaraguru, P. 2022. Game-on: Graph attention network based multimodal fusion for fake news detection. *arXiv preprint arXiv:2202.12478*.

Dong, Y.; He, D.; Wang, X.; Li, Y.; Su, X.; and Jin, D. 2023. A generalized deep markov random fields framework for fake news detection. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, 4758–4765.

Enderton, H. B. 2001. *A mathematical introduction to logic*. Elsevier.

Goldstein, J. A.; Sastry, G.; Musser, M.; DiResta, R.; Gentzel, M.; and Sedova, K. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hu, Z.; Ma, X.; Liu, Z.; Hovy, E.; and Xing, E. 2016. Harnessing Deep Neural Networks with Logic Rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2410–2420.

Jin, D.; Wang, L.; Zheng, Y.; Li, X.; Jiang, F.; Lin, W.; and Pan, S. 2022a. CGMN: A Contrastive Graph Matching Network for Self-Supervised Graph Similarity Learning. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, 2101–2107.

Jin, D.; Wang, R.; Ge, M.; He, D.; Li, X.; Lin, W.; and Zhang, W. 2022b. RAW-GNN: RAndom Walk Aggregation based Graph Neural Network. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, 2108–2114.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, 795–816.

Kim, J.-H.; On, K.-W.; Lim, W.; Kim, J.; Ha, J.-W.; and Zhang, B.-T. 2017. Hadamard Product for Low-rank Bilinear Pooling. In *Proceedings of the 5th International Conference on Learning Representations*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.

Li, T.; Gupta, V.; Mehta, M.; and Srikumar, V. 2019. A Logic-Driven Framework for Consistency of Neural Models. In *Proceedings of the 24th Conference on Empirical Methods in Natural Language Processing*, 3924–3935.

Liu, H.; Wang, W.; and Li, H. 2023. Interpretable Multimodal Misinformation Detection with Logic Reasoning. *arXiv preprint arXiv:2305.05964*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mishima, K.; and Yamana, H. 2022. A survey on explainable fake news detection. *IEICE TRANSACTIONS on Information and Systems*, 105(7): 1249–1257.

Nakamura, K.; Levy, S.; and Wang, W. Y. 2020. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 6149–6157.

OpenAI. 2023. ChatGPT. *https://openai.com/blog/chatgpt.*

Qi, P.; Cao, J.; Li, X.; Liu, H.; Sheng, Q.; Mi, X.; He, Q.; Lv, Y.; Guo, C.; and Yu, Y. 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.

Qi, P.; Cao, J.; Yang, T.; Guo, J.; and Li, J. 2019. Exploiting multi-domain visual information for fake news detection. In *Proceedings of the 19th IEEE International Conference on Data Mining*, 518–527.

Qian, S.; Wang, J.; Hu, J.; Fang, Q.; and Xu, C. 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.

Singhal, S.; Shah, R. R.; Chakraborty, T.; Kumaraguru, P.; and Satoh, S. 2019. Spotfake: A multi-modal framework for fake news detection. In *Proceedings of the 5th IEEE International Conference on Multimedia Big Data*, 39–47.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.

Wang, X.; Dong, Y.; Jin, D.; Li, Y.; Wang, L.; and Dang, J. 2023. Augmenting affective dependency graph via iterative incongruity graph learning for sarcasm detection. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, volume 37, 4702–4710.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM Sigkdd International Conference on Knowledge Discovery & Data Mining*, 849–857.

Wei, Z.; Pan, H.; Qiao, L.; Niu, X.; Dong, P.; and Li, D. 2022. Cross-modal knowledge distillation in multi-modal fake news detection. In *Proceedings of the 48th IEEE International Conference on Acoustics, Speech and Signal Processing*, 4733–4737.

Wu, L.; Liu, P.; and Zhang, Y. 2023. See How You Read? Multi-Reading Habits Fusion Reasoning for Multi-Modal Fake News Detection. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, volume 37, 13736–13744.

Ying, Q.; Hu, X.; Zhou, Y.; Qian, Z.; Zeng, D.; and Ge, S. 2023. Bootstrapping Multi-view Representations for Fake News Detection. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*.

Zhang, W.; Gui, L.; and He, Y. 2021. Supervised Contrastive Learning for Multimodal Unreliable News Detection in COVID-19 Pandemic. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3637–3641.

Zlatkova, D.; Nakov, P.; and Koychev, I. 2019. Fact-Checking Meets Fauxtography: Verifying Claims About Images. In *Proceedings of the 24th Conference on Empirical Methods in Natural Language Processing*, 2099–2108.