Interpretable Multimodal Misinformation Detection with Logic Reasoning

Hui Liu[1] **Wenya Wang**[2,3] **Haoliang Li**[1]
[1]City University of Hong Kong
[2]Nanyang Technological University
[3]University of Washington
liuhui3-c@my.cityu.edu.hk, wangwy@ntu.edu.sg, haoliang.li@cityu.edu.hk

Abstract

Multimodal misinformation on online social platforms is becoming a critical concern due to increasing credibility and easier dissemination brought by multimedia content, compared to traditional text-only information. While existing multimodal detection approaches have achieved high performance, the lack of interpretability hinders these systems' reliability and practical deployment. Inspired by Neural-Symbolic AI which combines the learning ability of neural networks with the explainability of symbolic learning, we propose a novel logic-based neural model for multimodal misinformation detection which integrates interpretable logic clauses to express the reasoning process of the target task. To make learning effective, we parameterize symbolic logical elements using neural representations, which facilitate the automatic generation and evaluation of meaningful logic clauses. Additionally, to make our framework generalizable across diverse misinformation sources, we introduce five meta-predicates that can be instantiated with different correlations. Results on three public datasets (Twitter, Weibo, and Sarcasm) demonstrate the feasibility and versatility of our model. The implementation of our work can be found in this link [1].

1 Introduction

Misinformation refers to incorrect or misleading information[2] which includes fake news, rumors, satire, etc. The enormous amount of misinformation emerged on online social platforms is attributed to users' reliability on the information provided by the internet and the inability to discern fact from fiction (Spinney, 2017). Moreover, widespread misinformation can have negative consequences for both societies and individuals. Therefore, there is an urgent need to identify misinformation automatically. While numerous posts are in multimodal style (i.e., text and image) on social media, this work concentrates on multimodal misinformation detection.

Multimodal approaches, which either fuse text and image features (Wang et al., 2018; Khattar et al., 2019; Xue et al., 2021; Chen et al., 2022b) or investigate discrepancies between the two modalities (Li et al., 2022a; Qi et al., 2021), have been used for misinformation detection with some success. However, these methods often lack interpretability because of the black-box nature of the neural network. Some frameworks have been proposed to solve this challenge. As depicted in Fig. 1, methods based on attention maps, such as those outlined in (Liang et al., 2021) and (Liu et al., 2022a), have been employed to identify highly correlated text or image content (referred to here as "where") according to attention weights, while multi-view based methods, such as those described in (Zhu et al., 2022b) and (Ying et al., 2022), have been utilized to highlight the most contributive perspectives[3] (referred to here as "how"). However, the explainability of the fusion of such attention or views has yet to be fully established (Liu et al., 2022b), and these methods cannot concurrently illustrate both the "where" and "how" of the reasoning process. Such interpretability is crucial for ensuring trust, reliability, and adoption of deep learning systems in real-world applications (Linardatos et al., 2021; Sun et al., 2021; Cui et al., 2022), particularly when it comes to detecting misinformation (Cui et al., 2019).

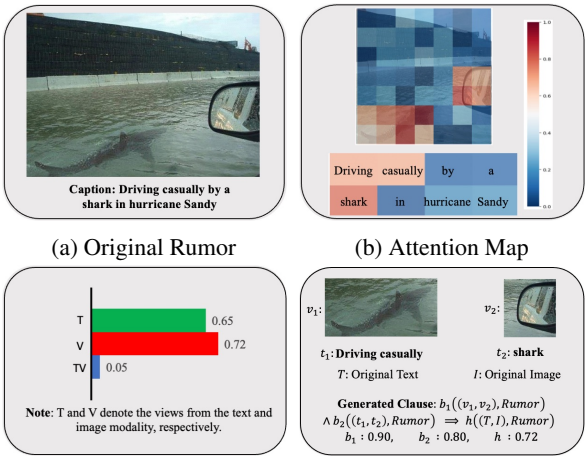
To address the aforementioned limitations, owing to Neural-Symbolic learning (Raedt et al., 2020; Hamilton et al., 2022), we propose to incorporate

[1] https://github.com/less-and-less-bugs/LogicMD

[2] https://www.merriam-webster.com/dictionary/misinformation

[3] Perspective is defined as a particular aspect to identify misinformation. In our work, it involves different types of assembly of different modalities, following a popular classification method of existing misinformation detection approaches (Alam et al., 2022).

Findings of the Association for Computational Linguistics: ACL 2023, pages 9781–9796
July 9-14, 2023



(a) Original Rumor (b) Attention Map

(c) Multi-View (d) Neural-Symbolic

Figure 1: Examples of explanations generated by attention map, multi-view, and our proposed Neural-Symbolic-based method for a rumor sample in Twitter dataset. For (c) and (d), a higher value indicates a higher probability of being detected as a rumor.

logic reasoning into the misinformation detection framework to derive human-readable clauses. As shown in Fig. 1d, the clause $b_1((v_1, v_2), Rumor) \wedge b_2((t_1, t_2), Rumor) \Rightarrow h((T, I)), Rumor)$ is induced from the text-image pair where constants v_1, v_2, t_1, t_2 are crucial visual patches and textual tokens for predication, corresponding to "where". Body predicates b_1 and b_2 indicate relationships between patches and tokens for misinformation identification, corresponding to "how". We propose to automatically learn these logic clauses which explicitly express evident features and their interactions to promote interpretability and improve the final performance, which has not been explored by previous work.

However, given the intrinsic complexity and diversity of multimodal context, it is hard to explicitly predefine the exact relationships as logic predicates. To this end, we introduce five general perspectives relevant to the task of misinformation detection as meta-predicates for clause formulation. These perspectives include suspicious atomic textual content, visual content, relationships between text tokens, visual patches and both modalities. Each meta-predicate can be instantiated with different correlations between contents of the text-image pair and target labels (e.g., (t_1, t_2) and $Rumor$ in Fig. 1d), aiming to cover a wide range of aspects leading to misinformation. For instance, the fifth perspective implicates exploiting cross-modal contents to debunk misinformation while cross-modal ambiguity learning (Chen et al., 2022b), inconsistency between news contents and background knowledge (Abdelnabi et al., 2022) and entities misalignment (Li et al., 2022a) are candidate correlations to achieve this goal.

Building upon these definitions, we propose a logic-based multimodal misinformation detection model (**LogicDM**). LogicDM first extracts embeddings for text tokens and image patches using corresponding encoders and then generates cross-modal object embeddings for different predicates using a multi-layer graph convolutional network (GCN). We then propose to parameterize meta-predicates by weighing the importance of each correlation. When combined with different object constants, these meta-predicates are softly selected to produce interpretable logic clauses defining the target predicate. The whole framework can be trained end-to-end with differentiable logic operators and probabilistic logic evaluations. To summarize, the contributions of this work include: 1) We propose an explainable neural-symbolic approach capable of automatically generating logic clauses instantiated with multimodal objects via differentiable neural components. 2) We define five meta-predicates building upon existing misinformation detection perspectives and introduce an adaptive mechanism to represent these predicates using soft selections over multiple pre-defined correlations. 3) We provide comprehensive evaluations of our model on three benchmark datasets.

2 Related Work

2.1 Misinformation Detection

Misinformation detection has gained significant attention in recent years due to the proliferation of content on online social media (Alam et al., 2022). To identify misinformation, the text modality can be used with clues such as semantics (Zhu et al., 2022b; Ma et al., 2019), writing style (Zhou et al., 2019), emotion (Zhu et al., 2022b), special word usage (Zhu et al., 2022a), and punctuation (Pérez-Rosas et al., 2018; Rubin et al., 2016). In addition, image features can help detect misinformation, with fake and real news often having distinct image distribution patterns, including differences in image semantics and compression trace (Jin et al., 2017a,b). Intra-modal inconsistency and incongruity within the text or image (Tay et al., 2018; Huh et al., 2018) can also serve as indicators of misinformation. Cross-modal interaction and fusion, used by many recent multimodality-based methods,

can assist in detecting misinformation. For example, (Li et al., 2022a; Qi et al., 2021) compared the characteristics of entities across the textual and visual modalities, while Ying et al. (2022) measured cross-modal inconsistency through Kullback-Leibler divergence between unimodal distributions.

2.2 Neural-Symbolic Reasoning

Deep learning has achieved impressive results, but its limitations in interpretability and logical reasoning have been noted by (Hamilton et al., 2022). To address these limitations, the integration of symbolic reasoning and neural networks, known as Neural-Symbolic AI, has gained attention as a potential solution (Raedt et al., 2020). One approach enhances neural networks with structured logic rules, such as first-order logic, that act as external constraints during model training (Hu et al., 2016; Manhaeve et al., 2018; Wang and Pan, 2021; Chen et al., 2022a). The other approach, Inductive Logic Programming (ILP), aims to automatically construct first-order logic rules from noisy data (Cropper et al., 2022). There have been various proposed ILP architectures, including NeuralLP (Yang et al., 2017), LNN (Sen et al., 2022), δILP (Evans and Grefenstette, 2018), and RNNLogic (Qu et al., 2021). ILP has been applied in a range of areas including knowledge-base completion (Qu et al., 2021), question answering (Li et al., 2022b), and multi-hop reading comprehension (Wang and Pan, 2022). However, multimodal misinformation detection, unlike these previous applications, faces the challenge of lacking well-defined predicates and constants due to the unstructured and modality-different text-image input.

3 Preliminaries

3.1 Task Definition

In this paper, we aim to address the problem of multimodal misinformation detection. Given a text-image pair (T, I), we seek to predict its label. To incorporate logic reasoning into the neural network, we define a candidate label set $\mathcal{Y} = \{\text{NonRumor}, \text{Rumor}\}$ for rumor detection task while $\mathcal{Y} = \{\text{NonSarcasm}, \text{Sarcasm}\}$ for sarcasm detection task. We also define a 2-ary predicate h that takes as input a text-image pair and a label, with the implicit meaning that the text-image pair satisfies the label. Our goal can then be reformulated as selecting a label $y \in \mathcal{Y}$ such that $h((T, I), y)$ holds. It is worth noting that this definition allows for the extension of our framework to multi-class classification tasks by increasing the size of the set of labels $\mathcal{Y}$.

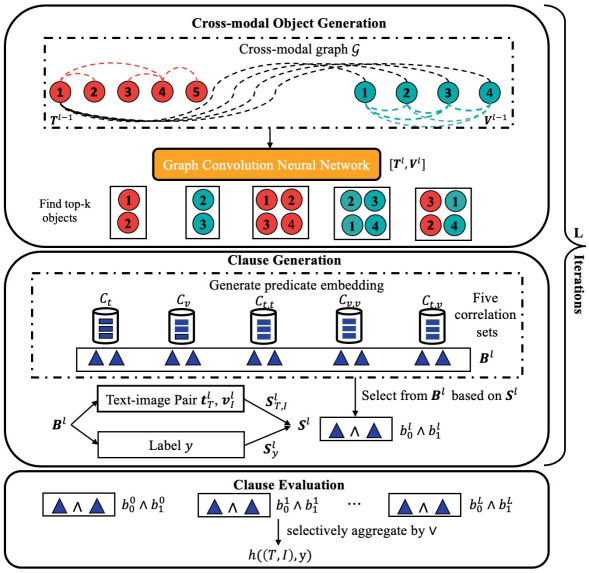


Figure 2: The core architecture of the proposed interpretable multimodal misinformation detection framework based on logic reasoning (**LogicDM**). Textual nodes are fully connected to visual nodes but we only visualize edges between one textual node and visual nodes for ease of illustration.

3.2 Inductive logic programming

To address the interpretability challenge in misinformation detection, we propose a framework that induces rules or clauses of the form $b_1 \wedge \ldots \wedge b_q \Rightarrow h$, where $b_1, \ldots, b_q$ are predicates in the body, h is the head predicate, and $\wedge$ denotes the conjunction operation. The body predicates are 2-ary, defined over object variable O (i.e., combinations of text tokens, image patches, or both) and label variable Y (i.e., labels in the set $\mathcal{Y}$). These predicates with associated variables, such as $b(O, Y)$, are referred to as logic atoms. By instantiating variables in body atoms with constants (e.g., $b(o, y)$, where o is an object and y is a label), we can obtain truth values of these body atoms and subsequently derive the value of the head atom $h((T, I), y)$ using logic operators (e.g., conjunction $\wedge$ and disjunction $\vee$), where the truth value indicates the probability of the atom or clause being true and is in the range of 0 to 1, denoted as $\mu(\cdot) \in [0, 1]$.

4 Methodology

This section introduces the proposed logic-based multimodal misinformation detection model (**LogicDM**), which offers a more explicit reason-

ing process and better performance than existing approaches. The model consists of four main components: Feature Extraction, Cross-modal Object Generation, Clause Generation, and Clause Evaluation. Feature Extraction generates representations for text tokens and image patches using encoders. Cross-modal Object Generation constructs a cross-modal graph and applies a multi-layer graph convolutional neural network to generate multi-grained representations that constitute cross-modal objects as logic constants. Clause Generation produces dynamic embeddings for predicates (see Table 1) by weighing the importance of different correlations and considers the logic relationship among all predicates to adaptively derive probable logic clauses. These clauses, when instantiated with object constants, can be evaluated to determine the truth value as Clause Evaluation. The overview of this model is shown in Fig. 2 and a running example is depicted in Fig. 6.

4.1 Feature Extraction

Given text-image pair (T, I) as input, we first tokenize T into m tokens, denoted as $X_T = \{w_1, w_2, \ldots, w_m\}$. Then we use BERT (Devlin et al., 2019) with a one-layer LSTM (Hochreiter and Schmidhuber, 1997) as the textual encoder to obtain d-dimension representations for all tokens in X_T, given as $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_m]$, where $\mathbf{T} \in \mathbb{R}^{m \times d}$.

For image modality, we first resize the image to the size 224×224 and divide each image into $r = z^2$ patches, where the size of each patch is $224/z \times 224/z$. Similar to text modality, these patches are reshaped to a sequence, denoted as $X_I = \{p_1, p_2, \ldots, p_r\}$. Then we exploit the pre-trained visual backbone neural network (e.g., ResNet34 (He et al., 2016) and ViT (Dosovitskiy et al., 2021)) to extract visual features and map these features to d-dimension using a two-layer MLP as $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r]$, where $\mathbf{V} \in \mathbb{R}^{r \times d}$.

4.2 Cross-modal Object Generation

Cross-modal Object Generation aims to produce representations for constants (e.g., (v_1, v_2), (t_1, t_2) in Fig. 1) to instantiate logic clauses. Different from the common definition of constants as single objects (in images or texts), we define constants according to our newly introduced meta-predicates. Specifically, we define meta-predicates as higher-level perspectives pertinent to discriminating misinformation. For this task, we use five meta-predicates, namely b_t for single-token perspective, b_v for single-image-patch perspective, $b_{t,t}$ for intra-text interactions, $b_{v,v}$ for intra-image interactions and $b_{t,v}$ for inter-modal interactions. The detailed explanations are shown in Table 1. The constants for these meta-predicates include a single token t_i, a single image patch v_i, a pair of tokens (t_i, t_j), a pair of image patches (v_i, v_j), and a pair consisting of both modalities (t_i, v_j). The representations, denoted by $\mathbf{o}$, for these constants are computed according to the formula in Table 1 and will be illustrated next.

The atoms, defined in Table 1, necessitate disparate uni-modal and cross-modal inputs, thus, requiring our model to capture intricate intra-modal and inter-modal representations concurrently. Inspired by recent work on multimodal task (Liang et al., 2021; Liu et al., 2020), we propose to construct a cross-modal graph $\mathcal{G}$ for (T, I) to leverage the relations among text tokens X_T, image patches X_I as well as those units between both modalities for computing representations of cross-modal constants.

Concretely, we take textual tokens X_T and visual patches X_I as nodes of graph $\mathcal{G}$, i.e., the node matrix is the concatenation of X_T and X_I, denoted as $[X_T, X_I]$ and the initial node embedding matrix is the concatenation of text-modality and image-modality representations, denoted as $\mathbf{H} = [\mathbf{T}, \mathbf{V}]$, where $\mathbf{H} \in \mathbb{R}^{(m+r) \times d}$. For edges, the semantic dependencies among textual tokens are first extracted by Spacy[4]. And if there exits a dependency between any two tokens, there will be an edge between them in $\mathcal{G}$. Then visual patches are connected according to their geometrical adjacency in the image, following (Liu et al., 2022a). Additionally, we assume the text nodes and visual nodes are fully connected to each other to increase interactions between two modalities, thus reducing the modality gap. Finally, the adjacency matrix $\mathbf{A} \in \mathbb{R}^{(m+r) \times (m+r)}$ can be represented as

$$\mathbf{A}_{ij} = \begin{cases} 1, & if\ i, j \leq m \text{ and a dependency exists in } w_i, w_j \\ 1, & if\ i \leq m, j > m \text{ or } i > m, j \leq m \\ 1, & if\ i, j > m \text{ and } p_{i-m}, p_{j-m} \text{ are adjacent,} \end{cases} \tag{1}$$

where p_{i-m} and p_{j-m} are determined as adjacent when $|(i-m) \bmod z - (j-m) \bmod z| \leq 1$ and $|(i-m)/z - (j-m)/z| \leq 1$. Subsequently, a L-layer GCN (Kipf and Welling, 2017) is used to update each node embedding after fus-

[4] https://spacy.io/

Logic Atom	Predicate Meaning	Formula of Objects
$b_t(t, y)$	token t is related to label y	$\mathbf{o}_t = \mathbf{t}\mathbf{W}_t$, $\mathbf{W}_t \in \mathbb{R}^{d\times d}$
$b_v(v, y)$	image patch v is related to label y	$\mathbf{o}_v = \mathbf{v}\mathbf{W}_v$, $\mathbf{W}_v \in \mathbb{R}^{d\times d}$
$b_{t,t}((t_i, t_j), y)$	the pair of tokens (t_i, t_j) is related to label y	$\mathbf{o}_{t_i,t_j} = [\mathbf{t}_i, \mathbf{t}_j, \mathbf{t}_i - \mathbf{t}_j, \mathbf{t}_i \circ \mathbf{t}_j]\mathbf{W}_{t,t}$, $\mathbf{W}_{t,t} \in \mathbb{R}^{4d\times d}$
$b_{v,v}((v_i, v_j), y)$	the pair of patches (v_i, v_j) is related to label y	$\mathbf{o}_{v_i,v_j} = [\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_i \circ \mathbf{v}_j]\mathbf{W}_{v,v}$, $\mathbf{W}_{v,v} \in \mathbb{R}^{4d\times d}$
$b_{t,v}((t_i, v_j), y)$	the pair of token and patch (t_i, v_j) is related to label y	$\mathbf{o}_{t_i,v_j} = [\mathbf{t}_i, \mathbf{v}_j, \mathbf{t}_i - \mathbf{v}_j, \mathbf{t}_i \circ \mathbf{v}_j]\mathbf{W}_{t,v}$, $\mathbf{W}_{t,v} \in \mathbb{R}^{4d\times d}$

Table 1: The meaning of proposed five meta-predicates and formulas to produce cross-modal objects for each predicate. $\mathbf{t}^l \in \mathbb{R}^d$ and $\mathbf{v}^l \in \mathbb{R}^d$ denote textual and visual features obtained in the l-th iteration of GCN, and the subscripts i and j represents two different features. The bold symbol $\mathbf{o} \in \mathbb{R}^d$ represents the embedding of corresponding constant. And $\mathbf{W}_t$, $\mathbf{W}_v$, $\mathbf{W}_{t,t}$, $\mathbf{W}_{v,v}$ and $\mathbf{W}_{t,v}$ are trainable parameters.

ing the information from its neighbor nodes via $\mathbf{H}^l = \text{ReLU}(\tilde{\mathbf{A}}\mathbf{H}^{l-1}\mathbf{W}^l)$, where $l \in \{0, 1, \ldots, L\}$ represents the l-th iteration of GCN, $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, $\mathbf{D}$ is the degree matrix of $\mathbf{A}$, and $\mathbf{W}^l \in \mathbb{R}^{d\times d}$ is a layer-specific trainable weight matrix. $\mathbf{H}^l \in \mathbb{R}^{(m+r)\times d}$ denotes the output of l-th GCN where $\mathbf{H}^l = [\mathbf{T}^l, \mathbf{V}^l]$ and $\mathbf{H}^0 = \mathbf{H}$. Especially, $\mathbf{T}^l \in \mathbb{R}^{m\times d}$ and $\mathbf{V}^l \in \mathbb{R}^{r\times d}$ are updated textual and visual representations at the l-th layer.

With $\mathbf{T}^l$ and $\mathbf{V}^l$, we compute representations of the cross-modal objects $\mathbf{O}_t^l \in \mathbb{R}^{m\times d}$, $\mathbf{O}_v^l \in \mathbb{R}^{r\times d}$, $\mathbf{O}_{t,t}^l \in \mathbb{R}^{(m\times m)\times d}$, $\mathbf{O}_{v,v}^l \in \mathbb{R}^{(r\times r)\times d}$ and $\mathbf{O}_{t,v}^l \in \mathbb{R}^{(m\times r)\times d}$ as constants for those meta-predicates, according to formulas in Table 1. In subsequent illustrations, we omit the layer index l for ease of illustration. Intuitively, different objects have different importance for multimodal misinformation detection task. As such, we feed the embedding of each object to a separate MLP (one linear layer with a ReLU as the activation function) to compute its importance score corresponding to a specific meta-predicate. Then k objects are chosen for each meta-predicate based on their importance scores for clause generations and evaluations. We denote their representations as $\hat{\mathbf{O}}_t$, $\hat{\mathbf{O}}_v$, $\hat{\mathbf{O}}_{t,t}$, $\hat{\mathbf{O}}_{v,v}$ and $\hat{\mathbf{O}}_{t,v}$, each of which belongs to $\mathbb{R}^{k\times d}$.

4.3 Clause Generation

In Clause Generation, we derive logic clauses consisting of meta-predicates that deduce the head atom $h((T, I), y)$, e.g., $b_v(v, y) \wedge b_t(t, y) \Rightarrow h((T, I), y)$. For each meta-predicate, we predefine a set of g fine-grained correlations (parameterized with embeddings) between objects and labels, denoted by $\mathbf{C} \in \mathbb{R}^{g\times d}$ (i.e., $\mathbf{C}_t$, $\mathbf{C}_v$, $\mathbf{C}_{t,t}$, $\mathbf{C}_{v,v}$, $\mathbf{C}_{t,v}$ corresponding to b_t, b_v, $b_{t,t}$, $b_{v,v}$ and $b_{t,v}$, respectively). For example, $\mathbf{C}_t$ stores g correlations between text tokens and labels relevant to meta-predicate $b_t(t, y)$. These correlations can be flexibly combined to form an embedding for each meta-predicate with different instantiations.

Concretely, taking meta-predicate $b_t(t, y)$ as an example, the embedding $\mathbf{B}_t$ for $b_t(t, y)$ with all instantiations t (i.e., $\hat{\mathbf{O}}_t$) is computed as

$$\mathbf{B}_t = \text{sparsemax}([\hat{\mathbf{O}}_t, \mathbf{y}]\mathbf{W}_t^e\mathbf{C}_t^\top)\mathbf{C}_t. \tag{2}$$

Here $\mathbf{B}_t \in \mathbb{R}^{k\times d}$ consists of k embeddings corresponding to k different objects extracted in $\hat{\mathbf{O}}_t$. $\mathbf{y}$ is the d-dimension embedding of label y and is broadcasted to $k \times d$ for concatenation. $\mathbf{W}_t^e \in \mathbb{R}^{2d\times d}$ is a learnable matrix. In addition, we utilize sparsemax, a sparse version of softmax, to select only a small number of correlations, which has been proven effective in multi-label classification tasks (Martins and Astudillo, 2016). The intuition of Eq. 2 is to softly select correlations to form the meta-predicate embedding when the input constants are t and y. By adapting Eq. 2 to other meta-predicates, we obtain a complete set of predicate embeddings $\mathbf{B} \in \mathbb{R}^{5k\times d}$ where $\mathbf{B} = [\mathbf{B}_t, \mathbf{B}_v, \mathbf{B}_{t,t}, \mathbf{B}_{v,v}, \mathbf{B}_{t,v}]$.

Furthermore, we obtain the embedding of the entire text input $\mathbf{t}_T \in \mathbb{R}^d$ and image $\mathbf{v}_I \in \mathbb{R}^d$ via weighed summations of all tokens and patches, respectively: $\mathbf{t}_T = \mathbf{T}^\top\text{softmax}(\mathbf{T}\mathbf{W}_T)$ and $\mathbf{v}_I = \mathbf{V}^\top\text{softmax}(\mathbf{V}\mathbf{W}_I)$, where $\mathbf{W}_T \in \mathbb{R}^{d\times 1}$ and $\mathbf{W}_I \in \mathbb{R}^{d\times 1}$ are trainable parameters to compute importance scores of tokens and patches.

To generate valid clauses, given the predicate embeddings $\mathbf{B}$, textual representation $\mathbf{t}_T$ and image representation $\mathbf{v}_I$, we use two sparse attention networks to select relevant predicates pertinent to the image-text input, as well as the given label, to form the body of a clause. Formally, we have two attention scores $\mathbf{S}_{T,I}$ and $\mathbf{S}_y$ indicative of the input text-image pair and label respectively, given as

$$\begin{aligned} \mathbf{S}_{T,I} &= \text{sparsemax}(\mathbf{B}\mathbf{W}_{T,I}[\mathbf{t}_T, \mathbf{v}_I]), \\ \mathbf{S}_y &= \text{sparsemax}([\mathbf{B}, \mathbf{y}, \mathbf{B} - \mathbf{y}, \mathbf{B} \circ \mathbf{y}]\mathbf{W}_y), \end{aligned} \tag{3}$$

where $\mathbf{W}_{T,I} \in \mathbb{R}^{d\times 2d}$ and $\mathbf{W}_y \in \mathbb{R}^{4d\times 1}$ are learnable parameters. The final score $\mathbf{S} \in \mathbb{R}^{5k}$ is obtained via

$$\mathbf{S} = \text{sparsemax}(\mathbf{S}_{T,I} \circ \mathbf{S}_y). \tag{4}$$

Each score in $\mathbf{S}$ indicates the probability of its corresponding predicate being selected to deduce the

head atom $h((T, I), y)$. Then $\lfloor 5k \times \beta \rfloor$ atoms ranking at the top of $\mathbf{S}$ are selected to complete the clause generation, where $\beta \in (0, 1)$ is a hyper-parameter. For instance, if $b_v(v, y)$ and $b_t(t, y)$ are selected, the clause will become $b_v(v, y) \wedge b_t(t, y) \Rightarrow h((T, I), y)$.

4.4 Clause Evaluation

In Clause Evaluation, we aim to derive the truth value of the head atom for each clause, given body atoms which are instantiated with constants. Specially, given an atom $b_t(t, y)$, its truth value $\mu(b_t(t, y))$ is computed as

$$\mu(b_t(t, y)) = \text{sigmoid}([\mathbf{b}_t, \mathbf{p}, \mathbf{b}_t - \mathbf{p}, \mathbf{b}_t \circ \mathbf{p}]\mathbf{W}_\mu), \quad (5)$$

where $\mathbf{p} \in \mathbb{R}^d$, $\mathbf{p} = \mathbf{o}_t \circ \mathbf{y}$, and $\mathbf{W}_\mu = \mathbf{W}^{4d \times 1}$ is a trainable parameter. Note that $\mathbf{b}_t \in \mathbb{R}^d$, $\mathbf{o}_t \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^d$ are representations of b_t, t, y, respectively, and $\mathbf{b}_t$ is taken from $\mathbf{B}$.

To obtain the truth value of the head atom, we approximate logic operators $\wedge$ and $\vee$ using product t-norm, an example of T-Norm (i.e., $T : [0, 1] \times [0, 1] \rightarrow [0, 1]$) (Klement et al., 2000). Product t-norm defines $T_\wedge(\mu_1, \mu_2) = \mu_1\mu_2$ and $T_\vee(\mu_1, \mu_2) = 1 - (1 - \mu_1)(1 - \mu_2)$, with $\mu_1, \mu_2 \in [0, 1]$ referring to truth values of atoms. With Product t-norm, the truth value of the head atom $\mu(h((T, I), y))$ can be derived as long as the value for each body atom is given. Recall that our GCN model generates representations for each layer $l \in \{0, ..., L\}$. Therefore, with logic clauses $b_1^l \wedge ... \wedge b_n^l \Rightarrow h((T, I), y)$ generated for each layer l, we use disjunctive operators to combine clauses across all the layers as $(b_1^0 \wedge ...) \vee (b_1^1 \wedge ...) \vee ... \vee (b_1^L \wedge ...) \Rightarrow h((T, I), y)$.

For the target task of multimodal misinformation detection, given (T, I), we derive truth values $\mu(h((T, I), y))$ for different candidate labels y, e.g., $y \in \{NonRumor, Rumor\}$. Then a cross-entropy loss is adopted to train our model in an end-to-end manner which maximizes the truth values for gold labels. During inference, we compare the truth values for both labels and pick the one corresponding to a larger value as the final prediction.

5 Experiment

5.1 Experiment Setup

We verify the effectiveness of our approach on two public misinformation datasets (*Twitter* and *Weibo*) and further demonstrate its versatility on a sarcasm detection dataset (*Sarcasm*). Three datasets are described as follows: 1) *Twitter* (Boididou et al., 2018) contains 7334 rumors and 5599 non-rumors for training and 564 rumors and 427 non-rumors for testing. 2) *Weibo* (Jin et al., 2017a) includes 3749 rumors and 3783 non-rumors for training and 1000 rumors and 996 non-rumors for testing. 3) *Sarcasm* (Cai et al., 2019) comprises 8642 sarcasm posts and 11174 non-sarcasm posts for training, 959 sarcasm posts and 1451 non-sarcasm posts for validating and 959 sarcasm posts and 1450 non-sarcasm posts for testing. Furthermore, for *Twitter* and *Weibo*, only samples with both text and image are kept, following previous work (Boididou et al., 2018; Chen et al., 2022b). The data pre-processing of *Sarcasm* follows Cai et al. (2019). For all experiments, we set $k = 5, g = 10$ and $\beta = 0.1$. Other details of the implementation and baselines can be found in the appendix.

5.2 Overall Performance

Table 2 and Table 3 present comparison results for multimodal misinformation detection and sarcasm detection tasks against popular baselines. Despite well-recognized tradeoffs between performance and model interpretability (Raedt et al., 2020), both tables indicate our proposed **LogicDM** consistently surpasses existing state-of-art methods in terms of both Accuracy and F1 Score. Especially our model brings 3.9% and 1.2% improvements based on accuracy over state-of-art **BMR** on *Twitter* and **CAFE** on *Weibo*. Moreover, our model demonstrates superior Precision than other baselines on *Sarcasm*. Such results verify the advantage of the integration of logical reasoning and neural network. We conjecture that logic components may motivate our model to learn useful rules instead of overfitting to noise. In addition, it is also worth mentioning that there is a difference in performance between Rumor and Non Rumor on *Twiiter*, which may be due to unbalanced proportions within the training set.

Furthermore, it is observed that multi-modality based methods generally outperform uni-modality based methods, suggesting that text and image can provide complementary information to enhance detection performance. In addition, **CAFE** and **BMR** can estimate the importance of different modalities to adaptively aggregate unimodal representations by ambiguity measure component and multi-view learning, thus, showing better performance than simple fusion or concatenation. In contrast,

Dataset		Method	Acc	Rumor			Non Rumor		
				Precision	Recall	F1 Score	Precision	Recall	F1 Score
Twitter	Uni-Modal	Bert (Devlin et al., 2019)	0.733	0.571	0.754	0.650	0.857	0.722	0.784
		ResNet (He et al., 2016)	0.644	0.473	0.712	0.568	0.812	0.610	0.697
	Multi-Modal	Vanilla	0.784	0.669	0.683	0.676	0.843	0.834	0.838
		EANN (Wang et al., 2018)	0.648	0.810	0.498	0.617	0.584	0.759	0.660
		MAVE (Khattar et al., 2019)	0.745	0.745	0.719	0.758	0.689	0.777	0.730
		SAFE (Zhou et al., 2020)	0.762	0.831	0.724	0.774	0.695	0.811	0.748
		MVNN (Xue et al., 2021)	0.784	0.778	0.781	0.779	0.790	0.787	0.788
		CAFE (Chen et al., 2022b)	0.806	0.807	0.799	0.803	0.805	0.805	0.809
		BMR (Ying et al., 2022)	0.872	0.842	0.751	0.794	0.885	0.931	0.907
		LogicDM	**0.911**	**0.909**	**0.816**	**0.859**	**0.913**	**0.958**	**0.935**
Weibo	Uni-Modal	Bert (Devlin et al., 2019)	0.716	0.671	0.671	0.671	0.692	0.762	0.725
		ResNet (He et al., 2016)	0.678	0.701	0.638	0.668	0.658	0.720	0.688
	Multi-Modal	Vanilla	0.746	0.610	0.622	0.616	0.814	0.806	0.810
		EANN (Wang et al., 2018)	0.795	0.806	0.795	0.800	0.752	0.793	0.804
		MAVE (Khattar et al., 2019)	0.824	0.854	0.769	0.722	0.720	0.740	0.730
		SAFE (Zhou et al., 2020)	0.816	0.818	0.818	0.817	0.816	0.818	0.817
		MVNN (Xue et al., 2021)	0.823	0.858	0.801	0.828	0.787	0.848	0.816
		CAFE (Chen et al., 2022b)	0.840	0.855	0.830	0.842	0.825	0.851	0.837
		BMR (Ying et al., 2022)	0.831	0.831	0.838	0.834	0.831	0.824	0.827
		LogicDM	**0.852**	**0.862**	**0.845**	**0.853**	**0.843**	**0.859**	**0.851**

Table 2: Comparison results for multimdoal misinformation detection on Twitter and Weibo datasets.

Model		Acc	P	R	F1
Uni-Modal	BERT (Devlin et al., 2019)	0.839	0.787	0.823	0.802
	ViT (Dosovitskiy et al., 2021)	0.678	0.579	0.701	0.634
multimodal	HFM (Cai et al., 2019)	0.834	0.766	0.842	0.802
	D&R Net (Xu et al., 2020)	0.840	0.780	0.834	0.806
	Att-BERT (Pan et al., 2020)	0.861	0.809	0.851	0.829
	InCrossMGs (Liang et al., 2021)	0.861	0.814	0.844	0.828
	HCM (Liu et al., 2022a)	0.874	0.818	**0.865**	0.841
	LogicDM	**0.881**	**0.857**	0.850	**0.853**

Table 3: Comparison results for mutlimodal sarcasm detection on Sarcasm dataset.

our model achieves this goal by softly choosing predicates to induce logic clauses when taking into consideration the logic relationship among these predicates.

5.3 Interpretation Study

To illustrate the interpretability of our proposed framework **LogicDM**, we visualize the learned rules in Fig. 3. Despite the complicated text-image input, it is evident that our model can explicitly locate highly correlated content as constants for "where" and softly choose suitable meta-predicates for "how". For example, as shown in Fig. 3c, objects "*a city*" and "*my baby*" are selected to instantiate b_1 (i.e., $b_{t,t}$) and b_2 (i.e., b_t) where both predicates implicate that samples with indefinite pronouns are more likely to be rumors. By comparison, samples of proper nouns can usually be detected as non-rumors because of their more realistic description, as seen in Fig. 3d. Moreover, the derived explanation can provide supplementary insights and knowledge previously unknown to practitioners. For example, as seen from Fig. 3a, the logic reasoning based on two visual patches, b_1, b_2 (i.e., both are b_v) implies that these areas are hand-crafted[5] (i.e., produced by Photoshop), which is difficult to be discriminated by human-beings.

Furthermore, our model can mitigate the trust problem of AI systems according to further analyzing derived clauses. For instance, although the non-rumor in Fig. 3b is identified accurately, it may not be sufficiently convincing based on only "*tower*", "*landmark*" and relevant predicates b_1, b_2 (i.e., both belongs to $b_{t,t}$). In other words, the decision result may not be reliable in this case. The interpretability of the model allows for further understanding of the decision-making process, thus increasing the reliability and trustworthiness of the system.

5.4 Ablation Study

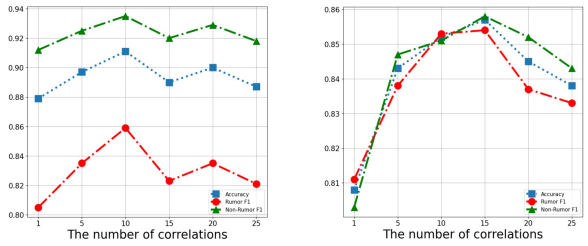
In the ablation study, we conduct experiments to analyze the impact of different parameters for performance, including the number of correlations g and rate β in Sec. 4.3 as well as selected iterations l in Sec. 4.4. For illustration, we report the precision, recall, F1 Score of rumor and accuracy on *Twitter* and *Weibo* datasets.

Impact of Number of Correlations. In order to effectively deal with the diverse online misinformation, we propose to adaptively represent predicates through their corresponding correlation sets in Clause Generation. As seen in Fig. 4, the influence of varying numbers of correlations (i.e., g) on performance reveals that the results dramatically increase as g increases and then gradually decrease after reaching a peak (e.g., 10 for the *Twitter* dataset and 15 for the *Weibo* dataset). These results validate the effectiveness of dynamic predicate em-

[5] https://phogotraphy.com/2015/03/20/iss-fake-photo/

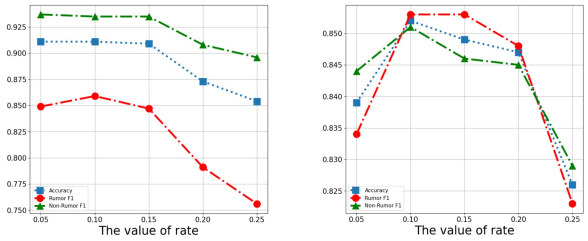
(a) Rumor in Twitter (b) Non Rumor in Twitter (c) Rumor in Weibo (d) Non Rumor in Weibo

Figure 3: Examples of derived clauses and related constants. For (c) and (d), we translate the text from Chinese to English.



(a) Twitter (b) Weibo

Figure 4: The influence of the number of correlations g for dynamic predicate representation.



(a) Twitter (b) Weibo

Figure 5: The influence of rate β for logic clause generation.

bedding mechanism and suggest that the optimal number of correlations depends on the complexity of specific scenarios. However, it should be noted that our model can be tolerant of an excessive number of correlations without significantly impacting performance.

Impact of Logic Clause Length. In Clause Generation, we deduce the logic clause of a fixed length by adjusting rate β. As illustrated in Fig. 5, it is evident that the performance drops significantly as β increases from 0.15. This observation can be attributed to two possible reasons: 1) Product t-norm may result in exponential decay when the number of atoms in the clause grows, leading to decreased stability, as previously reported in literature (Wang and Pan, 2022). 2) Including redundant logic atoms may inevitably introduce noise and negatively impact performance. These findings suggest that a moderate β is optimal for clause generation.

Impact of Selected Iterations. In Clause Evaluation, we obtain the final truth value of head atom $h((T, I), a)$ by selectively aggregating clauses produced at different iterations of GCN based on disjunction operator $\vee$. Table 4 compares various ways for computing $\mu(h((T, I), a))$, revealing that our model achieves the best performance when $l = 2$ while yielding the worst performance when $l = 0$. Such results highlight the importance of capturing intra-modal and inter-modal interactions of multimodal input through multi-layer GCN for our task.

selected iteration l	Twitter			Weibo		
	Accuracy	Rumor F1	Non-Romor F1	Accuracy	Rumor F1	Non-Romor F1
$l \in \{0\}$	0.745	0.638	0.799	0.825	0.826	0.824
$l \in \{1\}$	0.882	0.821	0.912	0.840	0.837	0.843
$l \in \{2\}$	**0.911**	**0.859**	**0.935**	**0.852**	**0.853**	**0.851**
$l \in \{0, 1\}$	0.847	0.762	0.887	0.847	0.848	0.846
$l \in \{1, 2\}$	0.902	0.842	0.928	0.841	0.832	0.849
$l \in \{0, 1, 2\}$	0.842	0.742	0.886	0.847	0.843	0.850

Table 4: The influence of selected iterations for clause evaluation. $l \in \{0\}$, $l \in \{1\}$, $l \in \{2\}$ are non-disjunctive combination clauses and the others are disjunctive combination clauses. For example, when $l \in \{0\}$, $h((T, I), a) = (b_1^0 \wedge ...)$ and when $l \in \{0, 1\}$, $h((T, I), a) = (b_1^0 \wedge ...) \vee (b_1^1 \wedge ...)$.

Furthermore, it is observed that disjunctive combination clauses perform more robustly than non-disjunctive combination clauses on *Weibo*, potentially due to the logic-based fusion of information at different iterations. These results provide insights into the importance of incorporating multiple iterations in clauses for better performance in some cases.

6 Conclusion

We propose an interpretable multimodal misinformation detection model **LogicDM** based on neural-symbolic AI. We predefine five meta-predicates and relevant variables evolved from corresponding misinformation detection perspectives. And we propose to dynamically represent these predicates by fusion of multiple correlations to cover diversified online information. Moreover, we differentiate reasoning process to smoothly select predicates

and cross-modal objects to derive and evaluate explainable logic clauses automatically. Extensive experiments on misinformation detection task demonstrate the effectiveness of our approach and external experiments on sarcasm detection task reveal the versatility.

Limitations

Our work has two limitations that may impact the generalization ability of our proposed framework. Firstly, in the Clause Generation section (Sec. 4.3), we deduce logic clauses involving a fixed number of atoms, represented by $\lfloor 5k \times \beta \rfloor$, rather than variable length for each iteration of GCN. While this approach has demonstrated superior performance on the multimodal misinformation detection and sarcasm detection tasks, it may harm the generalization of our framework to more complex multimodal misinformation tasks, such as the detection of fake news that involves various modalities, including social networks, text, user responses, images and videos, as discussed in (Zhou and Zafarani, 2021; Alam et al., 2022). Secondly, in our work, the incorporation of logic into the neural network relies on the use of product t-norm to differentiate logic operators (i.e., $\wedge$ and $\vee$). However, as shown in the Ablation Study (Sec. 5.4), product t-norm may lead to vanishing gradients with the increase of logic atoms during the training stage, which may limit the ability of our proposed framework to handle more sophisticated scenarios. We plan to address these limitations in future research.

Ethics Statement

This paper complies with the ACM Code of Ethics and Professional Conduct. Firstly, our adopted datasets do not contain sensitive private information and will not harm society. Secondly, we especially cite relevant papers and sources of pretrained models and toolkits exploited by this work as detailed as possible. Moreover, our code will be released based on the licenses of any used artifacts. At last, our proposed multimodal misinformation detection approach will contribute to protecting human beings from the detrimental and unordered online environment with more trustworthy interpretations.

ACKNOWLEDGEMENT

This work was supported in part by CityU Teaching Start-up Grant 6000801, CityU New Research Initiatives/Infrastructure Support from Central (APRC 9610528), the Research Grant Council (RGC) of Hong Kong through Early Career Scheme (ECS) under the Grant 21200522 and Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA).

References

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14920–14929. IEEE.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 6625–6643. International Committee on Computational Linguistics.

Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. 2018. Detection and visualization of misleading content on twitter. *Int. J. Multim. Inf. Retr.*, 7(1):71–86.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2506–2515. Association for Computational Linguistics.

Jiangjie Chen, Qiaoben Bao, Changzhi Sun, Xinbo Zhang, Jiaze Chen, Hao Zhou, Yanghua Xiao, and Lei Li. 2022a. LOREN: logic-regularized reasoning for interpretable fact verification. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10482–10491. AAAI Press.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022b. Cross-modal ambiguity learning for multimodal fake news detection. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2897–2905. ACM.

Andrew Cropper, Sebastijan Dumancic, Richard Evans, and Stephen H. Muggleton. 2022. Inductive logic programming at 30. *Mach. Learn.*, 111(1):147–172.

Limeng Cui, Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: A system for explainable fake news detection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 2961–2964. ACM.

Yue Cui, Zhuohang Li, Luyang Liu, Jiaxin Zhang, and Jian Liu. 2022. Privacy-preserving speech-based depression diagnosis via federated learning. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1371–1374. IEEE.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Richard Evans and Edward Grefenstette. 2018. Learning explanatory rules from noisy data (extended abstract). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5598–5602. ijcai.org.

Kyle Hamilton, Aparna Nayak, Bojan Bozic, and Luca Longo. 2022. Is neuro-symbolic AI meeting its promise in natural language processing? A structured review. *CoRR*, abs/2202.12205.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard H. Hovy, and Eric P. Xing. 2016. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, volume 11215 of *Lecture Notes in Computer Science*, pages 106–124. Springer.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017a. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 795–816. ACM.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017b. Novel visual and statistical image features for microblogs news verification. *IEEE Trans. Multim.*, 19(3):598–608.

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, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2915–2921. ACM.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Erich-Peter Klement, Radko Mesiar, and Endre Pap. 2000. *Triangular Norms*, volume 8 of *Trends in Logic*. Springer.

Peiguang Li, Xian Sun, Hongfeng Yu, Yu Tian, Fanglong Yao, and Guangluan Xu. 2022a. Entity-oriented multi-modal alignment and fusion network for fake news detection. *IEEE Trans. Multim.*, 24:3455–3468.

Xiao Li, Gong Cheng, Ziheng Chen, Yawei Sun, and Yuzhong Qu. 2022b. Adalogn: Adaptive logic graph network for reasoning-based machine reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7147–7161. Association for Computational Linguistics.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 4707–4715. ACM.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1):18.

Hui Liu, Wenya Wang, and Haoliang Li. 2022a. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. *CoRR*, abs/2210.03501.

Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, and Shiqi Wang. 2022b. Rethinking attention-model explainability through faithfulness violation test. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 13807–13824. PMLR.

Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. 2020. Learning cross-modal context graph for visual grounding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11645–11652. AAAI Press.

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, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3049–3055. ACM.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. 2018. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 3753–3763.

André F. T. Martins and Ramón Fernandez Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1614–1623. JMLR.org.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1383–1392. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3391–3401. Association for Computational Linguistics.

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 *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1212–1220. ACM.

Meng Qu, Junkun Chen, Louis-Pascal A. C. Xhonneux, Yoshua Bengio, and Jian Tang. 2021. Rnnlogic:

Learning logic rules for reasoning on knowledge graphs. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Luc De Raedt, Sebastijan Dumancic, Robin Manhaeve, and Giuseppe Marra. 2020. From statistical relational to neuro-symbolic artificial intelligence. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4943–4950. ijcai.org.

Victoria L Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17.

Prithviraj Sen, Breno W. S. R. de Carvalho, Ryan Riegel, and Alexander G. Gray. 2022. Neuro-symbolic inductive logic programming with logical neural networks. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 8212–8219. AAAI Press.

Laura Spinney. 2017. How facebook, fake news and friends are warping your memory. *Nature*, 543(7644).

Hao Sun, Zijian Wu, Yue Cui, Liwei Deng, Yan Zhao, and Kai Zheng. 2021. Personalized dynamic knowledge-aware recommendation with hybrid explanations. In *Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part III 26*, pages 148–164. Springer.

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1010–1020. Association for Computational Linguistics.

Wenya Wang and Sinno Jialin Pan. 2021. Variational deep logic network for joint inference of entities and relations. *Comput. Linguistics*, 47(4):775–812.

Wenya Wang and Sinno Jialin Pan. 2022. Deep inductive logic reasoning for multi-hop reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4999–5009. Association for Computational Linguistics.

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, KDD 2018, London, UK, August 19-23, 2018*, pages 849–857. ACM.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3777–3786. Association for Computational Linguistics.

Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Inf. Process. Manag.*, 58(5):102610.

Fan Yang, Zhilin Yang, and William W. Cohen. 2017. Differentiable learning of logical rules for knowledge base reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2319–2328.

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2022. Bootstrapping multi-view representations for fake news detection.

Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani. 2019. Fake news early detection: A theory-driven model. *CoRR*, abs/1904.11679.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: similarity-aware multi-modal fake news detection. In *Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part II*, volume 12085 of *Lecture Notes in Computer Science*, pages 354–367. Springer.

Xinyi Zhou and Reza Zafarani. 2021. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5):109:1–109:40.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022a. Generalizing to the future: Mitigating entity bias in fake news detection. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2120–2125. ACM.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022b. Memory-guided multi-view multi-domain fake news detection. *CoRR*, abs/2206.12808.

A Implementation

In Feature Extraction, we set $d = 200$ and employ pretrained Bert (i.e., bert-base-uncased[6] for *Twit-*

[6] https://huggingface.co/bert-base-uncased

ter and *Sarcasm* and bert-base-chinese[7] for *Weibo*) with one-layer LSTM as textual encoder to extract 200-dimension textual features. For visual modality, we divide the 224×224 image into 32×32 patches (i.e., $r = 49$, $z = 7$). We utilize ResNet34[8] as visual backbone for *Twitter* and *Weibo*, following (Chen et al., 2022b) and ViT[9] for *Sarcasm*, following (Liu et al., 2022a). The extracted visual features are subsequently mapped to the same dimension as textual features. In Cross-modal Objects Generation, we apply two-layer GCN (i.e., $L = 2$) to generate high-level representations of textual tokens and visual patches and then $k = 5$ to filter out five candidate objects for each meta-predicate. In Clause Generation, we set the number of correlations $g = 10$ and $\beta = 0.1$ to derive explainable logic clauses of length $\lfloor 5k \times \beta \rfloor$. At last, we set $h((T, I), a) = b_0^2 \wedge ... \wedge b^2_{\lfloor 5k \times \beta \rfloor - 1}$ (i.e., $l \in \{2\}$) to obtain the truth value of the target atom in Clause Evaluation. The number of parameters of our model is 4601019 without taking parameters of Bert and the visual backbone neural network (i.e., ResNet and ViT) into account.

During model training, we set the batch size to 32, the epoch number to 20 and exploit Adam optimizer for three sets. Additionally, we adopt an initial learning rate of 0.0001 and a weight decay of 0.0005 for *Twitter* and *Weibo* and 0.00002 and 0.0005 for *Sarcasm*. Moreover, early stopping strategy is used to avoid overfitting. And we run our experiments on four NVIDIA 3090Ti GPUs.

For model evaluation, in accordance with prior research (Chen et al., 2022b), we report Accuracy, and Precision, Recall, F1 Score for rumor and non rumor on *Twitter* and *Weibo*, while Accuracy, and Precision, Recall, F1 Score for sarcasm posts on *Sarcasm*.

B Baseline Models

To comprehensively evaluate our proposed method **LogicDM**, we divide the baseline models into two categories: Uni-Modal and Multi-Modal methods. For Uni-Modal baselines, we adopt **Bert** (Devlin et al., 2019) where the mean embedding of all tokens is utilized for classification and pretrained visual backbone networks where the feature representation after the final pooling layer is used. Specifically, for the visual backbone model, we adopt **ResNet** (He et al., 2016) for *Twitter* and *Weibo* datasets as suggested by Chen et al. (2022b), and adopt **ViT** (Dosovitskiy et al., 2021) for sarcasm detection dataset by following Liu et al. (2022a).

For Multi-Modal baselines, we utilize different approaches for multimodal misinformation detection and sarcasm detection due to the discrepancy between both tasks. Concretely, for *Twiter* and *Weibo*, we adopt **Vanilla**, **EANN** (Wang et al., 2018), **MAVE** (Khattar et al., 2019), **SAFE** (Zhou et al., 2020), **MVNN** (Xue et al., 2021), **CAFE** (Chen et al., 2022b), **BMR** (Ying et al., 2022). Especially, **Vanilla** fuses the textual and visual features extracted by corresponding encoders of our proposed **LogicDM** for classification and we re-implement **BMR** by using the same Feature Extraction component as our method and removing image pattern branch for a fair comparison. For *Sarcasm*, we utilize **HFM** (Cai et al., 2019), **D&R Net** (Xu et al., 2020), **Att-BERT** (Pan et al., 2020), **InCrossMGs** (Liang et al., 2021) and **HCM** (Liu et al., 2022a).

C Running Example

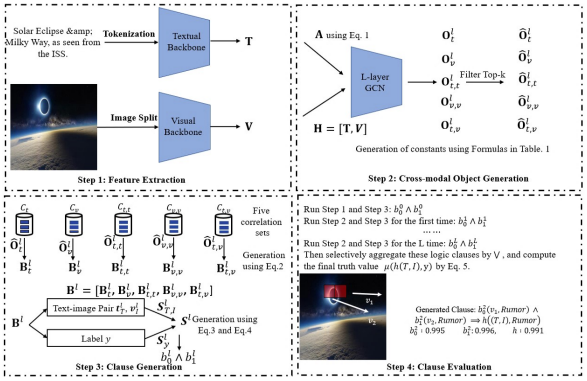


Figure 6: The running sample of our proposed **LogicDM**. In this example, we set $\lfloor 5k \times \beta \rfloor = 2$, implying that the derived clauses at each iteration are constituted of two logic atoms and the number of GCN layers is $L = 2$.

To facilitate understanding of the integral reasoning process, we provide an external running example as depicted in Fig. 6. The integral reasoning process can be summarized as follows: 1) Given the text-image pair as input, our model first extracts textual features $\mathbf{T}$ and visual features $\mathbf{V}$ using corresponding encoders. 2) These features are exploited to construct a cross-modal graph, denoted by the adjacency matrix $\mathbf{A}$ in Eq. 1 and node matrix $\mathbf{H} = [\mathbf{T}, \mathbf{V}]$. This graph is fed into

[7] https://huggingface.co/bert-base-chinese

[8] https://pytorch.org/vision/main/models/generated/torchvision.models.resnet34

[9] https://github.com/lukemelas/PyTorch-Pretrained-ViT

an L-layer GCN to conduct cross-modal reasoning. Especially at the iteration *l* of GCN, the output of GCN $\mathbf{H}^l$ is taken to construct cross-modal objects $\mathbf{O}_t^l$, $\mathbf{O}_v^l$, $\mathbf{O}_{t,t}^l$, $\mathbf{O}_{v,v}^l$ and $\mathbf{O}_{t,v}^l$, corresponding to each predicate, using formulas in Table 1. These objects are then refined through a purification process to retain only the most salient ones, denoted as $\hat{\mathbf{O}}_t^l$, $\hat{\mathbf{O}}_v^l$, $\hat{\mathbf{O}}_{t,t}^l$, $\hat{\mathbf{O}}_{v,v}^l$ and $\hat{\mathbf{O}}_{t,v}^l$, serve as constants to instantiate logic clauses. 3) To derive logic clauses at the iteration *l*, we obtain the predicate representations by weighting the importance of each clue in the corresponding clue set $\mathbf{C}$ for each pair of objects and label y using Eq. 2. Then two atoms from $\mathbf{B}^l$ are selected to derive logic clauses $b_0^l \wedge b_1^l$ based on the importance score $\mathbf{S}^l$ in Eq. 4. 4) As each iteration will produce one logic clause, the final logic clause can be deduced by $(b_0^0 \wedge b_1^0) \vee (b_0^1 \wedge b_1^1) \vee ... \vee (b_0^L \wedge b_1^L) \Rightarrow h((T, I), y)$, of where the truth value can be computed based on Eq. 5 and product t-norm. In this example, we only choose $b_0^2(v_1, Rumor) \wedge b_1^2(v_2, Rumor)$ as the final clause.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In Section Limitations.

☑ A2. Did you discuss any potential risks of your work?
Our work focuses on the multimodal misinformation task and will not harm society. We illustrate it in Section Ethics Statement in detail.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
In Section Abstract and Sec.1 Introduction.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Yes, we specify the pre-trained models (Bert, ViT, ResNet) and the toolkit (Spacy) we used in Sec. A Implementation. Moreover, we use the original Twitter, Weibo, and Sarcasm datasets, and the pre-processing follows existing work. For our code, we obey MIT License.

☑ B1. Did you cite the creators of artifacts you used?
Yes, in Sec. 5.1 Experiment Setup and Sec. A Implementation in Appendix.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Spacy, Our proposed framework LogicDM, ViT : MIT License Bert, ResNet, Twitter dataset: Apache License 2.0 For the other two public datasets, we cannot find related licenses.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
In Sec. Ethics Statement.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Related papers on three datasets (i.e., Weibo, Sarcasm, Twitter) have been published MM, ACL, International Journal of Multimedia Information Retrieval, and these datasets are popular datasets on misinformation and sarcasm detection tasks. We believe there is no offensive and private information in these datasets.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes, in Sec. A Implementation in Appendix and more detail can be found in the URLs provided in the footnotes.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In Sec. 5.1 Experiment Setup.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

In Sec. 5 Experiment.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In Sec. A Implementation in the Appendix.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Sec. A Implementation in the Appendix. In addition, for hyperparameter search, we search optimal learning rate from [0.001, 0.0001, 0.0005, 0.00002] and weight decay from [0.0005, 0.0001, 0.001]. However, we fix hyperparameters during model training for reported results.

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We report the average results of three runs to avoid randomness.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In Sec. A Implementation in the Appendix.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.