Asking Diversified Reasonable Questions with External Commonsense Knowledge to Infer Inconsistency for Multi-modal Clickbait Detection

Jianxing Yu, Shiqi Wang, Qi Chen, Huaijie Zhu, Libin Zheng, Wenqing Chen, Jian Yin(✉)

School of Artificial Intelligence, Sun Yat-sen University, Zhuhai, 519082, China
Key Laboratory of Sustainable Tourism Smart Assessment Technology, Ministry of Culture and Tourism of China, Sun Yat-sen University, Zhuhai, 519082, China
Pazhou Lab, Guangzhou, 510330, China
{yujx26, wangshq25, chenq539, zhuhuaijie, zhenglb6, chenwq95, issjyin}@mail.sysu.edu.cn

Abstract. Clickbait posts have become rampant on social media platforms, causing a multitude of detrimental impacts, e.g., the propagation of disinformation. Most of the existing studies focus on text-centric clickbait. They often make judgments based on shallow single-modality features of the text content, ignoring the deep content and commonsense clues contained in the images. It is difficult to explain where the inconsistencies are, and these shallow features cannot indicate complex clickbait well. To address these problems, we propose a new approach to infer inconsistencies from a suspect-then-verify perspective. It can suspect each potential clue of the multi-modal content by asking diverse and incisive questions, and then verify each clue by commonsense reasoning. That can well find evolving bait tricks with good interpretability. Specifically, we first analyze and represent the multi-modal content and context of the post. We then question the authenticity and consistency of the content. We take into account knowledge such as external commonsense to generate reasonable and diverse questions, capturing various bait types better. To facilitate the generation process, we learn the questions' reasoning structures and expressive patterns from open-source data. Next, we answer these questions to infer potential inconsistencies. The verification results are combined with typical clickbait features to derive the final prediction. We conduct a comprehensive set of experiments on three popular datasets to demonstrate the effectiveness of our approach.

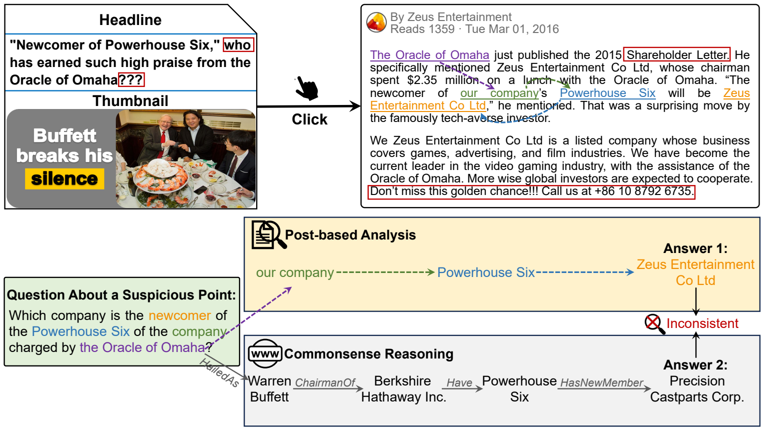
Keywords: Clickbait detection · Reasoning question · Common sense.

1 Introduction

On social media platforms, an overwhelming volume of posts flood in daily. This torrent of data poses a formidable challenge for users in processing. To obtain information efficiently, users would click on a post only when the headline and thumbnail trigger their interest. More clicks can bring more advertising revenue to the posts' creators, giving them a greater competitive advantage in the market. That prompts some unethical creators to create sensational baits, tricking users into clicking [22]. The content in baits is attractive and easily spread, which has detrimental consequences to users, media platforms, and society. Thus, there is a compelling need to detect such bait posts [29].

However, it is inadequate for the manual review method to curb the rampant spread of clickbait posts [41]. Thus, researchers are turning to automated detection methods. Their methods can be summarized into two categories. One is based on social propagation characteristics [11]. For example, a bait post usually spreads quickly, where user satisfaction is low and complaints are high because of its low quality. That requires sustained user engagement over time to accumulate historical data. This delay results in clickbait posts often being identified by machines only after they have been widely disseminated. To detect clickbait in time, another direction is to analyze the post's quality [36]. Existing work focuses on finding textual baits [3]. However, baits might exist in content of various modalities, such as text and images [6]. The cross-modal inconsistencies are difficult to identify via a single-modal method. That requires multi-modal reasoning to find out, which will encounter the heterogeneous gap problem. There are only a few multi-modal studies, which often collect features on each modality, and then detect baits based on a classifier built by fusing these features [1]. The fusion is often simple, such as concatenation and addition. That is too rough and does not take into account the importance of each modality. It is hard to accurately identify the subtle inconsistencies that require complex reasoning and deep exploration of the internal correlation among features in various modalities. In addition, the classifier is usually based on the neural network which is a black box. The intermediate decision-making process is invisible, resulting in poor interpretability [4].

To fill this research gap, we focus on the task of multi-modal clickbait detection and propose a new question-guided framework. It can infer baits with disinformation and inconsistencies hidden in posts by asking a series of incisive questions. As shown in Fig.(1), the post creator used a photo featuring his leader and *Buffett* as a gimmick to make false propaganda about the company for which he works. To discover this bait, we can raise a question to explore the newcomer of the *Powerhouse Six*. Based on this question, we can conduct cross-validation to find inconsistency. For example, one plausible answer *Zeus Entertainment Co Ltd* can be found based on the post content, but another correct answer *Precision Castparts Corp* can be derived by resorting to open-domain commonsense knowledge. This answer discrepancy indicates the bait in the post. Concretely, we first extract doubtful clues by analyzing the post's content and related commonsense knowledge. To facilitate cross-modal reasoning, the multi-modal content of the post is unified into the textual form. We then ask questions to suspect the clues. Considering that a fixed asking pattern is insufficient to cover a wide range of bait types, our questions are diverse, covering a variety of inquiry contents and expression forms. The diverse questions



Headline

"Newcomer of Powerhouse Six," who has earned such high praise from the Oracle of Omaha???

By Zeus Entertainment
Reads 1359 · Tue Mar 01, 2016

The Oracle of Omaha just published the 2015 Shareholder Letter. He specifically mentioned Zeus Entertainment Co Ltd, whose chairman spent $2.35 million on a lunch with the Oracle of Omaha. "The newcomer of our company's Powerhouse Six will be Zeus Entertainment Co Ltd," he mentioned. That was a surprising move by the famously tech-averse investor.

We Zeus Entertainment Co Ltd is a listed company whose business covers games, advertising, and film industries. We have become the current leader in the video gaming industry, with the assistance of the Oracle of Omaha. More wise global investors are expected to cooperate. Don't miss this golden chance!!! Call us at +86 10 8792 6735.

Question About a Suspicious Point:
Which company is the newcomer of the Powerhouse Six of the company charged by the Oracle of Omaha?

Fig. 1. Clickbait sample that requires complex commonsense inference. Red boxes mark the bait content.

can explore inconsistencies in the post from various perspectives, thus expanding the scope of detection. To alleviate the labeled data scarcity problem, we learn reasoning and expressive patterns from external open-source data that is easily accessible. Next, we answer each question from two sources, i.e., the post context and external commonsense knowledge. Based on the matching of these answers, along with six other typical bait-aware features, we build a detector to infer inconsistency. The questions and answers corresponding to the inconsistencies found can be used to explain the detection process.

We summarize the main contributions of this paper as follows:

- We focus on reliably identifying multi-modal inconsistencies in clickbait posts. This is a promising new direction for this task.
- We propose a novel detector in a suspect-then-verify way. Suspect by asking questions and then verify them by answering from multiple sources. The questions are created by considering the diversity of asking patterns on inquiry content and expressive forms. That helps to cover more clickbait types.
- We conducted a series of experiments. The results showed that the proposed method outperformed other baselines. We can reliably detect various types of clickbait posts, accompanied by explainable decision-making processes.

2 Approach

As illustrated in Fig.(2), our method comprises three main components. Initially, we analyze the multi-modal content of the post and extract potential clues. We then generate questions based on these clues and relevant commonsense knowledge. These questions are deep and can interrogate potential disinformation and inconsistencies within the post by commonsense reasoning. Also, they have a

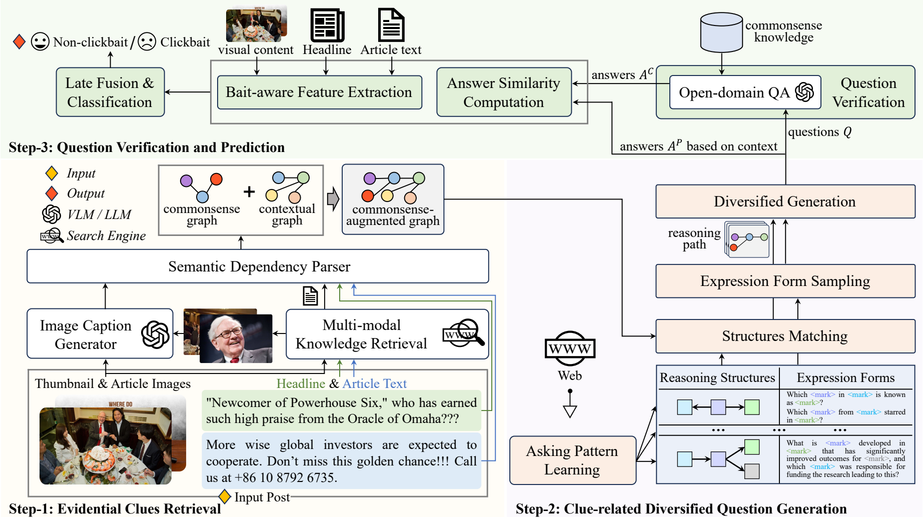

Non-clickbait / Clickbait
visual content
Headline
Article text
commonsense knowledge
Late Fusion & Classification
Bait-aware Feature Extraction
Answer Similarity Computation
answers A^c
Open-domain QA
Question Verification
answers A^P based on context
questions Q
Step-3: Question Verification and Prediction
Input
Output
VLM / LLM
Search Engine
commonsense graph
+
contextual graph
commonsense-augmented graph
Diversified Generation
reasoning path
Semantic Dependency Parser
Expression Form Sampling
Image Caption Generator
Multi-modal Knowledge Retrieval
Structures Matching
WWW
Web
Thumbnail & Article Images
Headline & Article Text
"Newcomer of Powerhouse Six," who has earned such high praise from the Oracle of Omaha???
More wise global investors are expected to cooperate. Don't miss this golden chance!!! Call us at +86 10 8792 6735.
Input Post
Asking Pattern Learning
Reasoning Structures
Expression Forms
Which <mark> in <mark> is known as <mark>?
Which <mark> from <mark> starred in <mark>?
What is <mark> developed in <mark> that has significantly improved outcomes for <mark>, and which <mark> was responsible for funding the research leading to this?
Step-1: Evidential Clues Retrieval
Step-2: Clue-related Diversified Question Generation

Fig. 2. Overview of our question-guided clickbait detector.

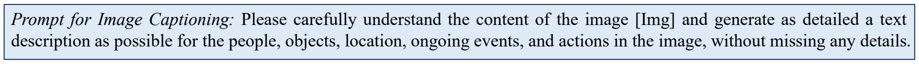
good diversity that can cover various types of baits. We verify consistency by answering these questions from multiple sources. Finally, we fuse the verified result with other classic multi-modal features extracted from the posts to build a clickbait detector. Next, we elaborate on each module of our method.

2.1 Problem Formulation

Our task aims to learn a detector $\mathcal{H}(l|po)$ to predict whether a given post po is clickbait ($l = 1$) or not ($l = 0$). po is composed of the headline, thumbnail, and linked article with mixed-modal content. Clickbait posts contain various disinformation and inconsistencies, which require complex multi-modal reasoning to detect. Traditional methods make predictions by analyzing typical deceptive features from past posts. However, there are vast types of baits. And malicious creators continually develop new tricks to craft increasingly covert clickbait posts. It is difficult to exhaustively enumerate them. To detect baits in an explainable way, we propose a question-guided detector which suspects and verifies the details in the post by asking and answering questions. In this QG+QA way, we can find the inconsistent bait and further explain where the bait is located.

2.2 Evidential Clues Retrieval

For clickbait posts, we can often find hidden contradictions within and between its multi-modal content. To systematically expose these hidden baits, we need to comprehend this content. Thus, we first establish a semantic graph $G^P = \{(h, r, t)\}$ to capture the fine-grained entities and relations in the posts, where h



Prompt for Image Captioning: Please carefully understand the content of the image [Img] and generate as detailed a text description as possible for the people, objects, location, ongoing events, and actions in the image, without missing any details.

Fig. 3. Prompt for converting post's images into textual descriptions.

and t denote nodes of the head and tail, representing potential clues, r denotes the relation. To fill semantic gaps, we further capture their context by referring to external commonsense knowledge. This knowledge is not explicitly mentioned in the post but shared by most humans, including facts about the physical world, history, celebrities, events, etc. It can help detect hidden inconsistencies in the bait posts. Following the creation process of G^P, we make another one G^C. We then merge identical nodes in these two graphs to obtain G.

Construction of Post Contextual Graph. Considering that there are great differences in representation between the visual and textual modalities, it is difficult to directly make cross-modal correlations and reasoning. To bridge this heterogeneity gap, we first transform visual content into text descriptions. Specifically, we prompt the *Vision-Language Model* (*VLM*) *LLaVA* [21] to obtain the caption cap_{img} which describes the image img in the thumbnail and article, that is, $cap_{img} = LLaVA(img, pt_1)$. The prompt pt_1 is shown in Fig.(3). It can draw *LLaVA*'s attention to the salient content of the image, such as individuals, locations, and events. Since the baits may be present in the text inside the image, we use the optical character recognition (*OCR*) model *DTrOCR* [12] to obtain its embedded text cap_{img}^{emb} for each image img. Afterward, we utilize a semantic dependency parser [17] to extract entities and relations from the image descriptions, headline, and article texts, thereby constructing a graph G^P. This provides structural and uniform representations of the post to facilitate deriving reasoning paths to ask to-the-point questions.

Construction of Commonsense-augmented Graph. Commonsense knowledge is hidden inside the posts and useful for identifying clickbait, since it can help to recognize deliberately fabricated disinformation. For example, malicious creators may use a headline like "*Shocking! Eating too many cherries can lead to cyanide poisoning*" to attract readers' attention. Using commonsense knowledge, we can infer that "*cyanide poisoning is only likely to occur when a large number of cherry pits are chewed and swallowed*". To retrieve post-related commonsense clues, traditional methods usually leverage knowledge graphs (*KGs*) with entity alignment. However, posts often involve the latest events. *KGs* only cover a limited amount of knowledge, some of which is even outdated. In addition, the retrieval often only uses the text content as a query, ignoring visual content. This will cause commonsense omissions, resulting in performance degradation. Therefore, we propose to retrieve related text and image data from open source like the Web and use them as augmented commonsense knowledge. Compared with *KGs*, the retrieved content encompasses vast and updated knowledge, which is more suitable for our task. We obtain the commonsense clues in three steps:
(Step-1) Multi-modal Knowledge Retrieval. Both textual and visual commonsense clues are conducive to finding inconsistent baits. For each image img in

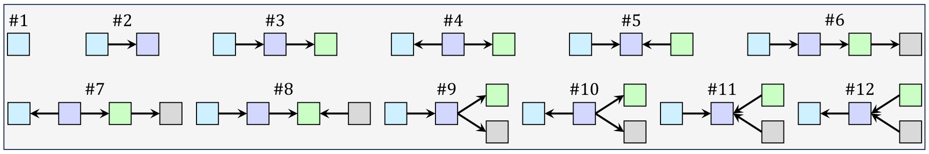


Fig. 4. Different types of reasoning structures that contain at most four nodes.

the post, we use *Google Lens API* to retrieve similar images. We also extract images of entities mentioned in the post's text by searching with *Google*. Taking the post's text and image captions as queries, we collect top-k related text from the *Google* search engine. Afterward, we use *LLaVA* to yield descriptions for the collected images, making retrieval results in a unified modality.
(Step-2) Trustworthy Knowledge Screening. The results retrieved from the Web may contain false content. That would introduce noise in the detector. We thus analyze the posts, and observe that the number of trustworthy posts far exceeds that of clickbait ones. Based on this observation, we design a screening trick to select reliable clues from the retrieval results. Specifically, for each query, we first use a text encoder to obtain the features of its retrieval results. We then calculate the center point of these features in the semantic space, and derive the distance between each feature and the center point. The retrieval results whose distances exceed the threshold β are viewed as noise and discarded.
(Step-3) Commonsense Graph Construction. The remaining retrieval results are fed into the parser to obtain a commonsense graph G^C. It captures the potential correlations among multi-modal commonsense knowledge. That helps to fill the semantic gaps and provide an indispensable detection basis.

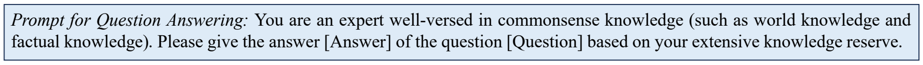
2.3 Clue-related Diversified Question Generation

Based on G, we raise questions to find the potential inconsistent details of the post po. Considering G contains multiple aspects of the post content, it is necessary to cover all aspects when questioning, since diverse questions help better detect the bait from more perspectives. We thus collect subgraphs in G and they can be viewed as reasoning paths for certain questions. Based on the subgraph, we then yield a question. We further impose asking patterns to ensure the logical and expressive correctness of the results. The patterns are obtained from a large amount of easily accessible open-source data, constraining reasoning structures, and expression forms. The structures are the results of removing the node and edge labels from the questions' reasoning paths. We observe there are up to four nodes in most structures of popular QA datasets, such as *SQuAD 2.0* [30], *GrailQA* [13], and *CosmosQA* [16]. We summarize and list them in Fig.(4). Next, we classify reasoning structures for questions from the open-source data, and learn their expression forms to help produce diversified results.

Recognition of Reasoning Structures. Given a question q, we can classify its reasoning structure, thus collecting 12 sets $\{D_i^Q\}_{i=0}^{12}$. By analyzing the questions under each structure, we can learn abundant expression forms. Specifically, we input q into the encoding model *TextCNN* [18] to obtain its representation $\mathbf{x}_q = TextCNN_{\Theta}(q)$, where Θ is the parameter set. Subsequently, $\mathbf{x}_q$ is fed into a multi-layer perceptron classifier to predict its reasoning structure. Compared with sequential networks such as *RNN*, *LSTM*, and *GRU*, which are good at capturing semantic content, *TextCNN* focuses more on structural features. Thus, it can better learn the mappings for reasoning structure types. Learning this classifier requires vast labeled data of *(question, reasoning structure label)* pairs, but this data is expensive to annotate. To handle this issue, we resort to the QA datasets *KQA Pro* [5] and *LC-QuAD 2.0* [10]. Their questions come with *SPARQL* query labels, which can be associated with one of the 12 reasoning structures by matching. In the way of *question → SPARQL query → reasoning structure*, we can collect sufficient annotated structure resources for training.

Learning of Expression Forms. We analyze question set D_i^Q of each reasoning structure i to learn the applicable expression forms. Traditional work uses a dependency tree parser [20] to obtain the syntactic structure of the question. They then retain the syntactic tags and relations within the parsing result to acquire the expression form. However, the parser may introduce extra errors and noises. We propose to tackle this issue based on the longest common subsequence (LCS). We differentiate between tokens of expression form and content according to frequencies. The former appears more frequently in the question set than the latter. For each question $q_j \in D_i^Q$, we record its LCS with other questions $q_k \in D_i^Q - \{q_j\}$, and replace the content tokens with the placeholder *<mark>*, obtaining the LCS set $\hat{E}_{q_j}$. For example, an element in $\hat{E}_{q_j}$ is "*Where does <mark> come from?*". We retain a subsequence $e \in \hat{E}_{q_j}$ that satisfies two pre-defined rules simultaneously as an expression form. The rules include (1) it should contain at least one interrogative term from the set *When, Where, Which, Why, How, Who, What*; and (2) its frequency should exceed 15 times. We collect subsequences that can serve as expression forms for each question, and aggregate them into the expression form set E_i for the i-th reasoning structure.

Diversified Reasonable Generation. We integrate two guidances into the generator, i.e., inquiry content and expression forms. For the inquiry content, we obtain subgraphs from G according to the 12 reasoning structures. Based on the i-th structure, we match and collect all matching subgraphs to construct inquiry content set SG_i. When generating questions for a subgraph $sg \in SG_i$, we sample λ expression forms from E_i to guide the generation based on a copy-based *GRU* decoder. It takes the sampled expression form e as the input. Its hidden state $\mathbf{s}_0$ is initialized by the feature $\mathbf{x}_{sg}$, which is obtained by feeding sg into the *GCN*. It can decompose the generation into a series of sub-generation tasks. In each task, it decides whether to copy content from e or fill in content on its own. We define these two states with special symbols $< C >$ and $< I >$. During decoding, we set a pointer r to mark the current position in e. The state h_j at the j-th step of decoding is determined by the output result of the previous step. When the result z_{j-1} of the $j-1$-th step is $< C >$, the copy state is activated, and h_j is set to 1. We take the token e_r pointed to by the pointer r as the output and move r to the next token. If z_{j-1} is $< I >$, the filling state is activated, and h_j is set



Prompt for Question Answering: You are an expert well-versed in commonsense knowledge (such as world knowledge and factual knowledge). Please give the answer [Answer] of the question [Question] based on your extensive knowledge reserve.

Fig. 5. Prompt for answering questions with open-domain commonsense knowledge.

to 0. In this state, the decoder combines the hidden state $\mathbf{s}_{j-1}$ of the $j-1$-th step to yield the filling content. That is, $z_j = Softmax(GRU(z_{j-1}, \mathbf{s}_{j-1}))$.

2.4 Question Verification and Prediction

Each question checks a certain content of the post. When the content is contrary to commonsense facts, the post is likely to contain bait. We find these contradictions by cross-verifying the answer of each question from multiple sources.

Question Verification. When generating questions Q, we can simultaneously obtain their answers A^P from the reasoning paths. These answers are derived from post content which may be incorrect due to the baits. The answers do not necessarily match the real facts. Therefore, questions with incorrect answers can indicate clickbait. To find these questions, we utilize *GPT-4* [26] to deduce answers A^C for Q, as shown in Fig.(5). *GPT-4* has a vast amount of knowledge. Compared with traditional search engines, it has a powerful reasoning ability and can flexibly apply commonsense knowledge to get the answer. Subsequently, we extract each answer pair (a_i^P, a_i^C) from A^P and A^C, and calculate the cosine similarity as $sim_i =< \mathbf{x}_{a_i^P}, \mathbf{x}_{a_i^C}, >$, where $\mathbf{x}_{a_i^P}$ and $\mathbf{x}_{a_i^C}$ are the *Word2Vec* [23] embeddings of a_i^P and a_i^C, respectively. A large sim_i means a high similarity between a_i^P and a_i^C. Conversely, a small sim_i may indicate the post's content is inconsistent with real facts. We average the similarities of all answer pairs as a verification feature. That is, $s = \frac{1}{|Q|}\sum_{i=1}^{|Q|} sim_i$.

Bait-aware Feature Extraction. In addition to the QA verified feature, we employ other six features to build a detector, so as to improve its robustness. *(1) Feature of visual disparity.* We analyze the consistency of visual content in the posts' thumbnails and articles' images by weighted directed graphs. The thumbnail is viewed as a root node and image as a child node. The edge is weighted by their similarity which is calculated by two metrics. The first one is the entity overlap degree m_1 of people and objects. They are recognized by *QMagFace* [37] and *RepVL-PAN* [8] respectively. We list the entity sets detected in the thumbnail and the image as E^{thu} and E^{img}. As shown in Eq. (1), we count the number of common entities CE, and use its ratio to the total number of entities to define m_1. The second metric is the pixel-level similarity m_2. To intuitively analyze whether the topics of the images match, we represent these two visual contents as color histograms and compare them. We extract the summation, mean and standard deviation of the bins to build the low-level visual features. m_2 is calculated by the *Euclidean* distance between these features. It is normalized in the range of $[0, 1]$. We take the average of metrics as $s_2 = \frac{m_1+m_2}{2}$. Besides weighting the edges, we also represent the features of each node image using *ResNet* [15]. Finally, the graph is fed into the *GCN* network [19] to obtain a feature $\mathbf{v}_1$ representing the visual disparity.

$$m_1 = \begin{cases} \dfrac{2CE}{E^{thu} + E^{img}}, E^{thu} + E^{img} > 0 \\ None, otherwise \end{cases}. \tag{1}$$

(2) Feature of headline-thumbnail disparity. Setting inconsistent headlines and thumbnails to create the curiosity gap is a common bait trick. To detect this inconsistency, we compare the thumbnail description cap_{thu} with the headline t_{hd}. We calculate the cosine similarity between the *Word2Vec* embeddings $\mathbf{x}_{cap_{thu}}$ and $\mathbf{x}_{t_{hd}}$ as a disparity feature $\mathbf{v}_2 = [< \mathbf{x}_{cap_{thu}}, \mathbf{x}_{t_{hd}} >]$.
(3) Feature of textual disparity. The bait post may use an attractive headline t_{hd} that is unrelated to the article text T. t_{hd} is a short text, while T is long with more details. Due to the different densities of the knowledge contained, it is unfair to compare their semantic features directly. We thus build a graph, where the root node is the headline, and a child node denotes article text $t \in T$. We use *BiGRU* [9] to encode the node text. Edges are weighted by the cosine similarity of nodes. We use *GCN* to encode this graph as a disparity feature $\mathbf{v}_3$.
(4) Feature of cross-modal sentiment. Clickbait posts often use strong feeling headlines and thumbnails to stir up the readers' emotions. We thus capture these emotions by a multi-modal sentiment analyzer *ITIN* [47], with three kinds of emotions, *{Positive, Neutral, Negative}*. We adopt the feature output from the layer before *Softmax*, as $\mathbf{v}_4$.
(5) Feature of headline lexical. Malicious creators tend to use exclamation marks, emojis, capital letters in the headlines. We capture the lexical properties with a statistical feature $\mathbf{v}_5$, which includes the units of the count of digits present; number of punctuation marks, such as question marks, exclamation marks, quotation marks, etc.; count of emojis; and ratio of capital letters to lowercase letters.
(6) Feature of headline bait. Bait posts often use some deceptive phrases and words in the headlines, such as "*You must know...*", "*Incredible!*", "*Get rich quick!*", etc. We capture these common phrases and words by the statistical feature $\mathbf{v}_6$, which includes the units of the number of personal names mentioned; a number of slangs used, such as "*LOL*" (*Laughing Out Loud*), "*YOLO*" (*You Only Live Once*), "*WTH*" (*What The Heck*); count of obscene and violent words, like "*nude*", "*prostitute*", "*necrophilia*"; count of typical bait phrases and tokens, such as "*You must know*", "*What happened next will shock you*", "*Shocking*", etc.

Joint Prediction. To combine the QA-verified feature s and bait-aware features $\mathbf{v}_{1\sim 6}$ for prediction, a straightforward way is to concatenate them into a single feature. However, these features have distinct structural characteristics. Rough fusion would disrupt their original structures, leading to the loss of salient knowledge and even causing semantic confusion. Therefore, we adopt late fusion to integrate these judgment bases. Specifically, we set up basic multi-layer perceptron (*MLP*) classifiers for s and $\mathbf{v}_{1\sim 6}$, respectively. The classification results are then fed into a meta-classifier to generate the final prediction $\hat{l}$, with a loss function as Eq. (2). The basic classifiers are learned on the training set, and the

Table 1. Comparisons in terms of all evaluated metrics with corresponding variances.

Datasets	*CLDInst*				*Clickbait17*				*FakeNewsNet*			
Methods	ACC ↑	PRE ↑	REC ↑	F1 ↑	ACC ↑	PRE ↑	REC ↑	F1 ↑	ACC ↑	PRE ↑	REC ↑	F1 ↑
PSGT	79.47± 0.23	78.46± 0.13	77.92± 0.24	78.19± 0.09	81.43± 0.11	81.06± 0.25	80.84± 0.18	80.95± 0.13	80.67± 0.32	80.33± 0.10	80.18± 0.13	80.25± 0.30
Varifocal	77.65± 0.08	78.29± 0.16	78.36± 0.09	78.32± 0.31	78.80± 0.25	78.94± 0.12	78.83± 0.33	78.88± 0.07	78.48± 0.19	78.62± 0.26	78.44± 0.14	78.53± 0.27
CAFE	82.41± 0.32	83.10± 0.12	82.77± 0.08	82.93± 0.21	86.15± 0.22	85.42± 0.15	85.39± 0.20	85.40± 0.19	84.22± 0.31	85.02± 0.33	84.29± 0.14	84.65± 0.22
PT-CD	80.15± 0.15	79.92± 0.07	79.74± 0.34	79.83± 0.32	81.22± 0.20	80.84± 0.14	80.53± 0.31	80.68± 0.22	80.85± 0.15	80.14± 0.21	80.06± 0.08	80.10± 0.14
MINIGPT-4	85.72± 0.12	84.33± 0.04	83.97± 0.32	84.15± 0.10	89.02± 0.08	87.42± 0.11	86.23± 0.29	86.82± 0.14	88.14± 0.22	87.40± 0.33	86.27± 0.35	86.83± 0.28
Ours	**91.47**± 0.24	**90.69**± 0.17	**90.03**± 0.16	**90.36**± 0.24	**93.76**± 0.17	**92.97**± 0.33	**92.51**± 0.13	**92.74**± 0.31	**93.17**± 0.09	**92.46**± 0.19	**91.84**± 0.34	**92.15**± 0.22

meta-classifier is trained on the output features of these basic classifiers.

$$\mathcal{L} = -[l \cdot \log(\hat{l}) + (1 - l) \cdot \log(1 - \hat{l})]. \tag{2}$$

3 Evaluations

We fully evaluated our method with qualitative and quantitative analyses.

3.1 Data and Experimental Settings

We performed evaluations on three classic datasets, including *CLDInst* [14], *Clickbait17* [28], and *FakeNewsNet* [35]. Sources of these datasets were popular in social media platforms, such as *Twitter*, *PolitiFact* and *GossipCop*. Through manual annotation, these datasets collected a total of 4,260, 9,276, and 5,755 clickbait posts, as well as 3,509, 29,241, and 17,441 non-clickbait posts. Their clickbait samples exhibit diverse deception types. A significant portion of the baits can only be identified by reasoning over multi-modal context and relevant commonsense knowledge. We employed four evaluation metrics that were typical in the field of bait detection, including accuracy (*ACC*), precision (*PRE*), recall (*REC*), and F1-score (*F1*). The larger the metric values, the stronger the model's ability to identify clickbait/non-clickbait. We trained all evaluated models with the oversampling technique to address the class imbalance problem. To avoid bias, we repeated running 15 times and reported the average performance.

3.2 Comparisons Against State-of-the-Arts

We compared our method against six classic baselines, including (1) *PSGT* [46], a propagation structure-aware detector based on the graph transformer; (2) *Varifocal* [27], which questioned text content to find bait; (3) *CAFE* [7], that used cross-modal alignment to fuse multiple features and then injected them into the classifier; (4) *PT-CD* [39], a prompt-tuning method which classified baits only based on the headline; (5) *MINIGPT-4* [45], a multi-modal pre-trained detector. We reimplemented these open-source baselines following their original settings.

As illustrated in Table 1, we presented the comparison results in terms of four metrics and their associated variances. The proposed method comprehensively outperformed the baselines on all datasets. In terms of the *F1* score, *Ours* surpassed the best baseline *MINIGPT-4*, by 7.38%, 6.82% and 6.12% on

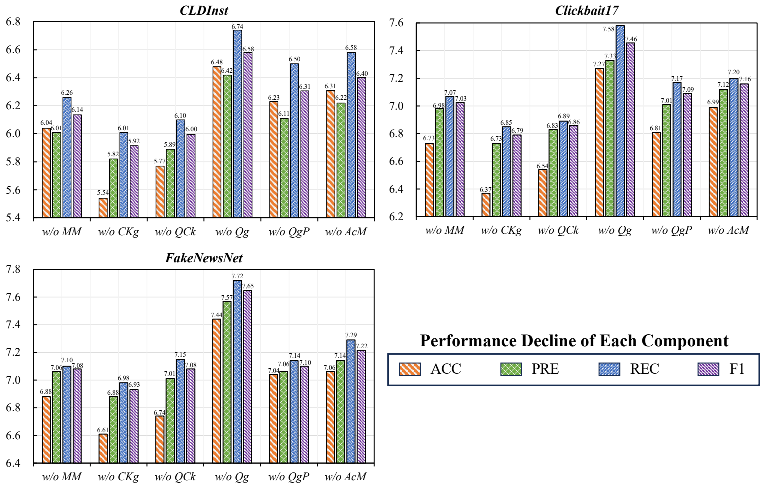
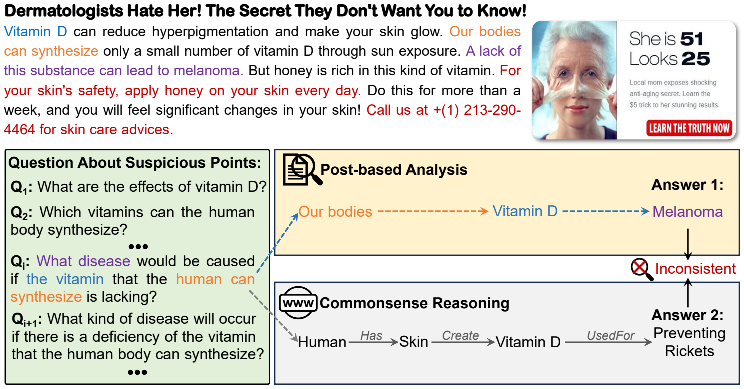


Fig. 6. Ablation studies, performance change ratios.

CLDInst, *Clickbait17*, and *FakeNewsNet*, respectively. That indicated the diverse questions generated by *Ours* could cover various types of deception, effectively supporting the clickbait detection task. Moreover, we observed that the multi-modal baselines generally performed better than single-modal ones. This verified that multi-modal content could provide strong clues, which helped to figure out inconsistencies and disinformation in bait posts. To evaluate the necessity of external knowledge, we randomly selected 600 samples from each dataset for manual analysis. The results showed that approximately 32%, 35%, and 36% of the posts need external knowledge to make predictions. In addition, we observed that our model outperformed the QA-based method *Varifocal*. *Varifocal* yielded questions only based on textual content, whereas our method doubted each clue in various modalities of the post. Different from traditional classification-based methods, we employed a QA+QG pipeline to analyze the multi-modal clues in the posts. This fine-grained cross-verification can better detect inconspicuous inconsistencies to infer clickbait well.

3.3 Ablation Studies

To analyze the usefulness of our proposed components, we conducted ablation studies on six key parts, including (1) *w/o MM*, which focused on the textual content only; (2) *w/o CKg*, retrieved commonsense from *KGs*; (3) *w/o QCk*, discarded commonsense clues and asked questions only based on the post's multi-modal content; (4) *w/o Qg* that removed the question generation module and predicted by prompting large language models (*LLMs*) with the collected clues; (5) *w/o QgP* that threw away the guidance of patterns for question generation; and (6) *AcM* which removed the question verification module and predicted only based on six kinds of bait-aware features.



Dermatologists Hate Her! The Secret They Don't Want You to Know!

Vitamin D can reduce hyperpigmentation and make your skin glow. Our bodies can synthesize only a small number of vitamin D through sun exposure. A lack of this substance can lead to melanoma. But honey is rich in this kind of vitamin. For your skin's safety, apply honey on your skin every day. Do this for more than a week, and you will feel significant changes in your skin! Call us at +(1) 213-290-4464 for skin care advices.

Question About Suspicious Points:

Q_1: What are the effects of vitamin D?

Q_2: Which vitamins can the human body synthesize?

•••

Q_i: What disease would be caused if the vitamin that the human can synthesize is lacking?

Q_{i+1}: What kind of disease will occur if there is a deficiency of the vitamin that the human body can synthesize?

•••

Fig. 7. Case study of our proposed method.

As displayed in Fig.(6), the ablation of all evaluated components resulted in a performance decline of more than 5.54%. That demonstrated each of our proposed components played a crucial role in detecting baits. Removing the *Qg* module had the greatest impact on performance. Without this module, it was difficult to perform cross-modal inference with commonsesne knowledge. Discarding the *AcM* module also led to substantial degradation, with more than 6.31% and 6.40% dropped in terms of the *ACC* and *F1* metrics, respectively. That indicated question verification could uncover multi-modal inconsistencies and disinformation hidden in posts. In addition, the *Ckg* and *QCk* modules provided commonsense knowledge for our model. Their ablations would cause a noticeable performance drop. That validated the rationality of our framework.

3.4 Case Studies and Discussions

To further explore the pros and cons of our proposed method, we performed case studies to analyze the actual effect. As shown in Fig.(7), this post misled readers through lure thumbnails and headlines, creating a wrong impression that the post might share some useful skincare tips to make one look younger. However, when readers click on the linked article, they would realize that it is an advertising post containing disinformation. To detect this clickbait, we first retrieved a set of suspicious clues from the post content and its context of external commonsense facts. We then inferred baits by asking diverse questions to suspect clues and verified them. Here, we showed one sampled question with the highest inconsistent score. To answer this complex question, we reason over multiple facts, like '*Human skin can produce vitamin D on its own.*,' and '*Vitamin D is beneficial for preventing rickets.*' However, based solely on the post content, what we actually get is another false conclusion '*Vitamin D is useful for preventing melanoma.*' This inconsistency can help the detector to judge correctly and give a reasonable explanation for the bait.

4 Related Work

Detecting clickbait posts is a crucial way to prevent the spread of disinformation. In the past, this task was reviewed by humans with expert knowledge. That is labor-intensive [31] and incapable of handling the massive volume of emerging posts. An automated detector is proposed to address this problem. Early work predicts clickbait based on the social behavior of users. Clickbait posts can receive lots of views and shares in a short period. Because of their low-quality content, user satisfaction is often low with negative feedback, such as short reading durations and scathing comments [24]. This property can be used to find baits, but there is a time lag in the user feedback. When the bait is identified, the post usually has spread widely [25]. To detect clickbait in time, some studies mined bait-aware features from the post content [34]. Traditional methods used rules to find linguistic features of clickbait posts, e.g., word choice, topic selection, emotional polarity, etc. These rules were hand-crafted, with poor scalability. To address this issue, *Wang et al.* [39] proposed to learn mappings between post content and the bait label by neural networks. Due to the lack of a deep understanding of the post content, this method can only perceive simple bait cases. *Bourgonje et al.* [2] evaluated the similarity between features of headline and article text to discover inconsistencies. The article text is long text while the headline is short. This mismatch results in poor similarity calculations. Some work proposed to summarize the article text into a short one [33], whose length approaches that of the headline. These methods mostly focus on baits in a single modality but neglect inconsistencies across various modalities, like text and image. To identify these multi-modal inconsistencies, we have to deduce across multiple pieces of detail and even resort to commonsense knowledge [43]. Existing multi-modal methods utilize techniques like concatenation, addition, and attention mechanism to fuse textual and visual features for classification [44]. The fusion is too rough, ignoring subtle relations between features in different modalities, which are crucial to identifying baits.

Another related work is question generation which can verify the authenticity of the content [38]. Prior work had primarily focused on simple one-hop questions by considering text only [32]. *Yu et al.* [42] studied complex questions with a reasoning chain, but the diversity is insufficient. To solve it, few studies employ technologies like beam search and reinforcement learning, yet the generated results often lack coherence and logical soundness. In addition, some work extracted entities from the text, and collected relevant triples from knowledge graphs (*KGs*) as common-sense knowledge [40]. Since *KGs* are manually constructed, it would suffer from low coverage and untimely updates. When the post involves new topics not covered by *KGs*, it is difficult to generate relevant questions. Unlike existing methods, we generate questions by using clues in multiple modalities. The generator is guided by asking patterns. The abundant patterns facilitate the diversity and logical correctness of the results. In addition, we obtain commonsense knowledge from the Web. Compared with *KGs*, it has a wider coverage and updated knowledge.

5 Conclusion

This paper studied clickbait detection of posts in social media. These posts often contain various inconsistencies among the multi-modal content. Existing neural models are black-box, which have poor interpretability to falsify the bait content. To address this problem, we proposed a question-guided framework. In detail, we first retrieved plausible clues related to the post content, and then examined clues by asking a series of questions. Different from shallow ones, our questions involved multi-hop reasoning and commonsense inference which can help to check complex and inconspicuous facts and relations flexibly. Our questions have a good diversity which can better identify the various types of baits. By cross-verifying the answers from multiple knowledge sources, we can find inconsistencies to infer bait content. Based on these inconsistencies and six extra features, we built a robust model that can well explain the bait points. Extensive evaluations conducted on three datasets showed our method's effectiveness.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62276279, 62472455, 62306344, U2001211, U22B2060), Guangdong Basic and Applied Basic Research Foundation (2024B1515020032, 2024A1515010253), and Research Foundation of Science and Technology Plan Project of Guangzhou City (2023B01J0001, 2024B01W0004).

References

1. Abdali, S., Shaham, S., Krishnamachari, B.: Multi-modal misinformation detection: Approaches, challenges and opportunities. ACM Computing Surveys **57**(3), 76:1–76:29 (2025)
2. Bourgonje, P., Moreno Schneider, J., Rehm, G.: From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles. In: EMNLP. pp. 84–89 (2017)
3. Bronakowski, M., Al-khassaweneh, M., Al Bataineh, A.: Automatic detection of clickbait headlines using semantic analysis and machine learning techniques. Applied Sciences **13**(4), 2456 (2023)
4. Caled, D., Carvalho, P., Sousa, F., Silva, M.: Domain: Explainable credibility assessment tools for empowering online readers coping with misinformation. ACM Transactions on the Web **19**(1), 1–31 (2024)
5. Cao, S., Shi, J., Pan, L., Nie, L., Xiang, Y., Hou, L., Li, J., He, B., Zhang, H.: KQA pro: A dataset with explicit compositional programs for complex question answering over knowledge base. In: ACL. pp. 6101–6119 (2022)
6. Chen, J., Jia, C., Zheng, H., Chen, R., Fu, C.: Is multi-modal necessarily better? robustness evaluation of multi-modal fake news detection. IEEE Transactions on Network Science and Engineering **10**(6), 3144–3158 (2023)
7. Chen, Y., Li, D., Zhang, P., Sui, J., Lv, Q., Lu, T., Shang, L.: Cross-modal ambiguity learning for multimodal fake news detection. In: WWW. pp. 2897–2905 (2022)

8. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: CVPR. pp. 16901–16911 (2024)
9. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP. pp. 1724–1734 (2014)
10. Dubey, M., Banerjee, D., Abdelkawi, A., Lehmann, J.: Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia. In: ISWC. pp. 69–78 (2019)
11. Elyashar, A., Bendahan, J., Puzis, R.: Detecting clickbait in online social media: You won't believe how we did it. In: CSCML. pp. 377–387 (2022)
12. Fujitake, M.: Dtrocr: Decoder-only transformer for optical character recognition. In: WACV. pp. 8010–8020 (2024)
13. Gu, Y., Kase, S., Vanni, M., Sadler, B.M., Liang, P., Yan, X., Su, Y.: Beyond I.I.D.: three levels of generalization for question answering on knowledge bases. In: WWW. pp. 3477–3488 (2021)
14. Ha, Y., Kim, J., Won, D., Cha, M., Joo, J.: Characterizing clickbaits on instagram. In: ICWSM. pp. 92–101 (2018)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
16. Huang, L., Bras, R.L., Bhagavatula, C., Choi, Y.: Cosmos QA: machine reading comprehension with contextual commonsense reasoning. In: EMNLP-IJCNLP. pp. 2391–2401 (2019)
17. Jiang, S., Li, Z., Zhao, H., Ding, W.: Entity-relation extraction as full shallow semantic dependency parsing. IEEE Transactions on Audio, Speech, and Language Processing **32**, 1088–1099 (2024)
18. Kim, Y.: Convolutional neural networks for sentence classification. In: EMNLP. pp. 1746–1751 (2014)
19. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
20. Li, J., Lu, W.: Contextual distortion reveals constituency: Masked language models are implicit parsers. In: ACL. pp. 5208–5222 (2023)
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
22. Liu, T., Yu, K., Wang, L., Zhang, X., Zhou, H., Wu, X.: Clickbait detection on wechat: A deep model integrating semantic and syntactic information. Knowledge-Based Systems **245**, 108605 (2022)
23. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR (2013)
24. Nan, Q., Sheng, Q., Cao, J., Hu, B., Wang, D., Li, J.: Let silence speak: Enhancing fake news detection with generated comments from large language models. In: CIKM. pp. 1732–1742 (2024)
25. Nan, Q., Sheng, Q., Cao, J., Zhu, Y., Wang, D., Yang, G., Li, J.: Exploiting user comments for early detection of fake news prior to users' commenting. Frontiers of Computer Science **19**(10), 1910354 (2025)
26. OpenAI: GPT-4 technical report. In: arXiv (2023)
27. Ousidhoum, N., Yuan, Z., Vlachos, A.: Varifocal question generation for fact-checking. In: EMNLP. pp. 2532–2544 (2022)
28. Potthast, M., Gollub, T., Komlossy, K., Schuster, S., Wiegmann, M., Fernandez, E.P.G., Hagen, M., Stein, B.: Crowdsourcing a large corpus of clickbait on twitter. In: COLING. pp. 1498–1507 (2018)
29. Raj, R., Sharma, C., Uttara, R., Animon, C.R.: A literature review on clickbait detection techniques for social media. In: ICRITO. pp. 1–5 (2024)

30. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for squad. In: Gurevych, I., Miyao, Y. (eds.) ACL. pp. 784–789 (2018)
31. Rastogi, S., Bansal, D.: A review on fake news detection 3t's: Typology, time of detection, taxonomies. International Journal of Information Security **22**(1), 177–212 (2023)
32. Sachan, D.S., Lewis, M., Joshi, M., Aghajanyan, A., Yih, W., Pineau, J., Zettlemoyer, L.: Improving passage retrieval with zero-shot question generation. In: EMNLP. pp. 3781–3797 (2022)
33. Sepúlveda-Torres, R., Vicente, M.E., Saquete, E., Lloret, E., Palomar, M.: Leveraging relevant summarized information and multi-layer classification to generalize the detection of misleading headlines. Data & Knowledge Engineering **145**, 102176 (2023)
34. Shi, C., Yin, Y., Zhang, Q., Xiao, L., Naseem, U., Wang, S., Hu, L.: Multiview clickbait detection via jointly modeling subjective and objective preference. In: EMNLP. pp. 11807–11816 (2023)
35. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data **8**(3), 171–188 (2020)
36. Supriya, Singh, J.P., Kumar, G.: Identification of clickbait news articles using sbert and correlation matrix. Social Network Analysis and Mining **13**(1), 153 (2023)
37. Terhörst, P., Ihlefeld, M., Huber, M., Damer, N., Kirchbuchner, F., Raja, K.B., Kuijper, A.: Qmagface: Simple and accurate quality-aware face recognition. In: WACV. pp. 3473–3483 (2023)
38. Uehara, K., Harada, T.: K-VQG: knowledge-aware visual question generation for common-sense acquisition. In: WACV. pp. 4390–4398 (2023)
39. Wang, Y., Zhu, Y., Li, Y., Qiang, J., Yuan, Y., Wu, X.: Clickbait detection via prompt-tuning with titles only. IEEE Transactions on Emerging Topics in Computational Intelligence **9**(1), 695–705 (2025)
40. Yu, J., Wang, S., Yin, H., Chen, Q., Liu, W., Rao, Y., Su, Q.: Diversified generation of commonsense reasoning questions. Expert Systems with Applications **263**, 125776 (2025)
41. Yu, J., Wang, S., Yin, H., Sun, Z., Xie, R., Zhang, B., Rao, Y.: Multimodal clickbait detection by de-confounding biases using causal representation inference. In: EMNLP. pp. 10300–10317 (2024)
42. Yu, J., Wang, S., Zheng, L., Su, Q., Liu, W., Zhao, B., Yin, J.: Generating deep questions with commonsense reasoning ability from the text by disentangled adversarial inference. In: ACL. pp. 470–486 (2023)
43. Zhang, H., Fang, Q., Qian, S., Xu, C.: Multi-modal knowledge-aware event memory network for social media rumor detection. In: ACM MM. pp. 1942–1951 (2019)
44. Zhang, L., Zhang, X., Zhou, Z., Huang, F., Li, C.: Reinforced adaptive knowledge learning for multimodal fake news detection. In: AAAI. pp. 16777–16785 (2024)
45. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: ICLR (2024)
46. Zhu, J., Gao, C., Yin, Z., Li, X., Kurths, J.: Propagation structure-aware graph transformer for robust and interpretable fake news detection. In: KDD. pp. 4652–4663 (2024)
47. Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H., Qian, J.: Multimodal sentiment analysis with image-text interaction network. IEEE Transactions on Multimedia **25**, 3375–3385 (2023)