

A Literature Review on Multimodal Misinformation Detection with External Knowledge Augmentation

Ding Zhang, Yuejin Zhang *

Beijing Union University, Beijing, China

* Corresponding author: Yuejin Zhang

Abstract: Social media misinformation is increasingly presented through combinations of headlines, body text, images, comments, interaction traces, and source metadata. This shift complicates detection because a misleading post may arise not from an explicitly false sentence, but from an authentic image reused in another event, a weakly related headline, a satirical publishing context, or missing background information. Recent multimodal and vision-language models improve the joint representation of textual and visual signals, yet most of them still operate mainly on evidence contained in the sample. They therefore have limited access to factual context, entity relations, or information about the platform and source. External knowledge augmentation offers one way to address this gap by adding knowledge-graph facts, retrieved evidence, event descriptions, entity links, or source-context information. This review examines how such knowledge is represented and combined with multimodal features, with attention to commonly used datasets, fusion strategies, and evaluation practices. Particular emphasis is placed on three unresolved issues: noisy or irrelevant knowledge, leakage caused by label-bearing descriptions, and the gap between semantic alignment and genuine fact verification. The discussion is intended to clarify when external knowledge is useful and what controls are required for reliable and interpretable multimodal misinformation detection.

Keywords: Misinformation Detection; Multimodal Learning; External Knowledge; Semantic Enhancement; Text-image Consistency; Knowledge Fusion.

1. Introduction

Social media is now a routine route for obtaining news and discussing public events. Its low publication threshold and rapid sharing mechanisms allow text, images, short videos, and comments to circulate almost immediately. The same mechanisms also accelerate the spread of misinformation, especially when platform recommendation, repeated reposting, and emotional engagement amplify uncertain claims. The consequences extend beyond individual misunderstanding, affecting public trust and the quality of online discourse. Shu et al. were among the first to frame social-media misinformation detection as a data-mining problem and argued that news content should be analyzed together with signals such as user behavior and social context[1]. Zhou and Zafarani later organized the field around false knowledge, writing style, propagation patterns, and source credibility[2]. Taken together, these studies show that the task is inherently multi-source rather than a conventional text-classification problem.

Early systems nevertheless treated misinformation detection largely as a textual classification task. Handcrafted statistics, sentiment and topic features, syntactic patterns, sequence models, and later pre-trained language models were used to represent headlines, claims, or short posts. These approaches remain useful when the misleading cue is expressed directly in language. Their coverage is weaker, however, when a post relies on an unrelated photograph, recycles an older image, exploits a sensational headline, or derives its meaning from satire or source identity. In such cases, the text itself may appear plausible, while the misleading relation is created by the combination of content and context. A text-only model can therefore miss whether the image supports the claim and whether the post is being interpreted in the correct event or platform setting.

These limitations motivated a move toward multimodal detection. Current systems combine text, images, comments, interaction records, and metadata through feature concatenation, attention-based fusion, graph models, or vision-language pre-training. Fakeddit has been influential in this transition because it provides Reddit text, images, metadata, and comments with 2-way, 3-way, and 6-way labels[3]. Its scale and label granularity make it suitable for both binary and fine-grained evaluation. The emergence of CLIP has also changed the available feature space. By learning text and image encoders through contrastive training on large collections of paired data, CLIP maps both modalities into comparable semantic spaces[4]. For misinformation detection, this offers a practical way to measure whether a caption and an image are semantically related and to fuse their representations.

Cross-modal alignment, however, should not be mistaken for verification. A photograph can match the words in a headline while depicting the same type of event at another time or place. A headline can also be visually plausible even when the underlying claim is unsupported. Source context introduces a further complication: satire communities, image-editing forums, news outlets, and general-interest communities establish different expectations for similar content. A model restricted to the headline and image may therefore confuse topical relevance with factual support or overlook the role of the publishing environment.

External knowledge is introduced to supply information that is absent from the sample itself. Depending on the design, it may consist of retrieved articles, encyclopedic descriptions, entity and relation data, event history, platform-community descriptions, domain attributes, or knowledge-graph substructures. These signals can be encoded alongside the textual and visual features and used during classification. The intended benefit is not simply a larger feature vector; it is a

broader basis for judging whether the observed content is compatible with known facts and with the context in which it was published.

That additional evidence also creates new failure modes. Retrieved material may be outdated, incomplete, irrelevant, or drawn from an unreliable source. More seriously, a knowledge description that contains terms such as 'fake', 'false', 'misleading', 'fabricated', or 'manipulated' can reveal the target label directly. A classifier may then learn a shortcut instead of the intended semantic relation. Fusion design and interpretability are also unresolved: the model should be able to discount noisy knowledge and make clear when an external item influences a prediction. Consequently, evaluation must consider knowledge construction, leakage control, and the transparency of the decision process, not only the final accuracy.

This review concentrates on the intersection between multimodal detection and external knowledge. Rather than covering every branch of misinformation research, it asks three narrower questions: what information is missing from text-image pairs, how external evidence can be represented and fused, and what methodological risks accompany knowledge-enhanced systems. Section 2 defines the main tasks and summarizes commonly used datasets. Section 3 reviews multimodal fusion and text-image consistency modeling. Section 4 discusses knowledge sources, representations, fusion strategies, and leakage-free source-context construction. Section 5 compares representative approaches and uses a preliminary Fakeddit experiment to illustrate both the benefit of neutral source context and the inflation caused by label-bearing knowledge. Section 6 concludes the review.

2. Misinformation Detection Tasks and Common Datasets

The goal of misinformation detection is to determine the veracity of a given piece of information. Depending on the format of the input information and the granularity of the labels, this task can be categorized into text-based misinformation detection, multimodal misinformation detection, rumor detection, fact-checking, and fine-grained misinformation classification, among others. Early research primarily focused on text-based inputs, such as news headlines, news articles, or political statements; however, as the nature of social media content has evolved, images, videos,

comments, user interactions, dissemination patterns, and source-context information have gradually been incorporated into the detection process.

2.1. Task Definition and Evaluation Metrics

Given a sample to be detected, it may contain text content, image content, social context, and external knowledge. Misinformation detection can be regarded as a classification task that maps these input components to a veracity label. In binary classification, the model usually distinguishes between real information and misinformation. In multi-class classification, the label space can be further extended to finer-grained categories, such as satire, misleading content, manipulated content, and false context.

In multimodal misinformation detection, models no longer make judgments based solely on text content but instead utilize text, images, and other contextual information simultaneously. Key questions include whether the text contains misleading statements, whether the image content is semantically consistent with the text, whether the image has been misused, whether the source or community context affects content interpretation, and whether external knowledge is needed to supplement the factual background.

Misinformation detection is typically evaluated using metrics such as Accuracy, Precision, Recall, F1-score, and Macro-F1. Among these, Accuracy reflects the model's overall classification accuracy; Precision and Recall measure the accuracy and coverage of positive class predictions, respectively; and the F1-score comprehensively considers both Precision and Recall. In cases of class imbalance, Macro-F1 better reflects the model's overall ability to identify different classes; therefore, research on multimodal misinformation detection typically reports both Accuracy and Macro-F1. In the subsequent preliminary case study of this paper, we also adopt Accuracy and Macro-F1 as the primary evaluation metrics.

2.2. Common Datasets and Dataset Characteristics

The development of misinformation detection research is closely tied to publicly available datasets. Different datasets vary in terms of data sources, modalities, labeling granularity, and applicable tasks. Table 1 lists some of the most commonly used representative datasets in this field.

Table 1. Comparison of Commonly Used Misinformation Detection Datasets

Dataset	Main Modalities	Data Source	Label Type	Applicable Task
LIAR	Text	PolitiFact political statements	Six-class labels	Political statement veracity detection
FakeNewsNet	Text, social context, temporal information	PolitiFact, GossipCop, social platforms	Binary classification	Misinformation detection, propagation analysis
PHEME	Text, conversational threads, propagation information	Twitter discussions on breaking events	Rumor/non-rumor, stance labels	Rumor detection, stance classification
Weibo	Text, images, social features	Weibo platform	Rumor/non-rumor	Chinese multimodal rumor detection
MediaEval VMU	Text, images/videos, metadata	Twitter and other social media platforms	Real/misleading	Multimedia content verification
Fakeddit	Text, images, metadata, comments	Reddit platform	2-way, 3-way, and 6-way labels	Fine-grained multimodal misinformation detection

LIAR is one of the earliest widely used datasets for text-based misinformation detection. It was constructed by Wang using political statements from PolitiFact and annotated with

a six-level veracity label[5]. This dataset is suitable for researching the veracity classification of short text statements, but since its primary modality is text, it is difficult to use for

modeling images or social context.

FakeNewsNet is a data resource designed for social media misinformation research, containing news content, social context, and temporal propagation information [6]. Compared to datasets that include only news text, FakeNewsNet is better suited for studying the relationships among user engagement, propagation dynamics, and news content. However, as this dataset primarily focuses on news websites and social media propagation scenarios, its coverage of semantic consistency issues between text and images is relatively limited.

The PHEME dataset focuses primarily on rumor detection and stance analysis in Twitter-based breaking news events, and includes event-related tweets, reply threads, and rumor annotations[7]. This dataset emphasizes the event propagation process within a social media context and is suitable for studying the evolutionary characteristics of rumors within conversational threads; however, its multimodal visual information is relatively limited.

Weibo datasets are commonly used for rumor detection on Chinese social media. Jin et al. studied multimodal rumor detection based on Weibo and Twitter data, combining text, images, and social contextual information to demonstrate that rumor detection in Chinese social media scenarios also requires the use of multi-source information[8]. Since misinformation on the Weibo platform is often accompanied by images, comments, and user interactions, such datasets are valuable for research on Chinese multimodal misinformation detection.

The “Verifying Multimedia Use” task in MediaEval focuses on the misuse and tampering of online multimedia content; its objective is to detect whether images or videos on social media are being used in a misleading manner[9]. Such datasets are important for research on the reuse of old images, image tampering, image misuse, and multimedia verification, reflecting the evolution of misinformation detection from text-based veracity assessment to multimedia content verification.

Fakeddit is a representative dataset in recent years for multimodal misinformation detection. Sourced from Reddit, this dataset includes text, images, metadata, comment information, as well as 2-way, 3-way, and 6-way multi-granularity labels. Compared to traditional datasets, Fakeddit is advantageous because of its large sample size, rich modal information, and the provision of finer-grained categories of misinformation. Because it includes source information such as subreddits and domains, this dataset is also well-suited for researching how contextual knowledge from external sources can enhance multimodal misinformation detection.

Table 1 shows that misinformation detection datasets have evolved from text-based political statements to multimodal social media data containing images, metadata, social context, and fine-grained labels. This evolution provides a stronger basis for multimodal detection, but it also increases the complexity of modality alignment, label design, and cross-dataset comparison.

However, existing datasets still have some shortcomings. First, annotation standards and labeling granularity vary across different datasets, making it difficult to directly compare model results across them. Second, the relationship between images and text in multimodal datasets is complex; simply providing both text and images does not mean that models can truly understand the relationship between them. Furthermore, many datasets lack explicit external knowledge annotations, so models can typically only make judgments

based on information within the samples themselves. For samples involving event contexts, entity relationships, source communities, and factual evidence, relying solely on text and images often makes it difficult to reach accurate conclusions.

Therefore, multimodal misinformation detection enhanced with external knowledge requires the development of additional knowledge-supplementation mechanisms based on existing datasets. On the one hand, background information can be obtained from news search engines, encyclopedias, knowledge graphs, and fact-checking websites; on the other hand, neutral source-context knowledge can be constructed from platform sources, community types, domain attributes, and publication contexts. For datasets like Fakeddit, which include subreddit and domain fields, source-context knowledge can serve as a form of external knowledge to supplement the contextual information that text and images alone cannot provide.

3. Multimodal Misinformation Detection Methods

Misinformation on social media often contains text, images, comments, and source information simultaneously. Compared to detection methods that rely solely on text content, multimodal misinformation detection can leverage visual cues in images, event descriptions in text, and the semantic relationships between the two to assess information veracity more comprehensively. Particularly in scenarios such as the reuse of old images, mismatched text and images, misleading thumbnails, and visual tampering, text-only models struggle to identify such issues, whereas multimodal methods can enhance detection capabilities through the joint analysis of text and images.

3.1. From Text-Based Detection to Text-Image Feature Fusion

Early misinformation detection primarily relied on text content, such as news headlines, body text, or social media posts. Text-based methods can extract semantic features through techniques such as word frequency statistics, sequence modeling, and pre-trained language models; however, their input data is relatively limited. When misinformation spreads through the misuse of images, inconsistencies between text and images, or visual misdirection, analyzing text content alone can easily lead to misjudgments. Consequently, researchers began incorporating modal information—such as images, user interactions, and social context—to develop multimodal misinformation detection methods.

Jin et al. proposed a multimodal rumor detection method based on attention-based recurrent neural networks, combining text, images, and social context information for rumor detection tasks on Weibo and Twitter. This demonstrated that the fusion of multimodal information can improve the effectiveness of social media rumor detection. Building on this foundation, related research has expanded from simple feature concatenation to methods such as event adversarial learning, variational representation learning, text-image consistency modeling, and pre-trained vision-language models.

In multimodal misinformation detection, text and images are usually encoded separately. The text encoder extracts semantic features from headlines, claims, or post content, while the image encoder extracts visual features from scenes,

objects, people, and visual style. These two types of features are then fused through concatenation, attention mechanisms, or other fusion modules, and the fused representation is fed

into a classifier for veracity prediction. This process forms the basic framework of multimodal misinformation detection, as shown in Figure 1.

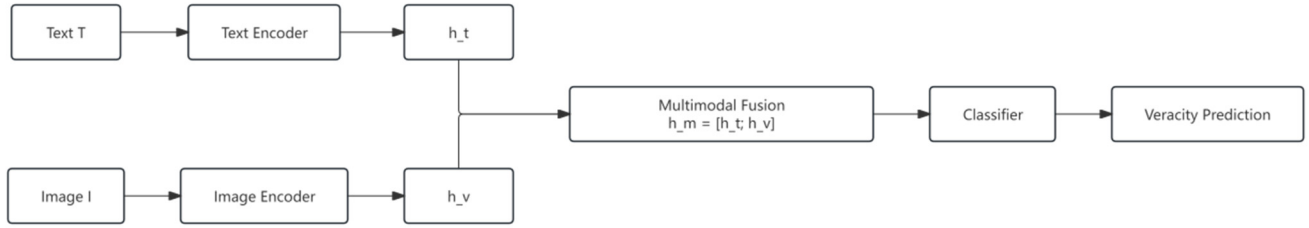


Figure 1. Basic Framework of Multimodal Misinformation Detection

Wang et al. proposed the EANN model, or Event Adversarial Neural Network. This method is designed for multimodal misinformation detection; it learns event-independent features from text and images and uses adversarial learning to reduce the model’s dependence on specific events, thereby improving the model’s generalization ability to new events [10]. The significance of EANN lies in the fact that it not only focuses on the fusion of text and images but also addresses the issue of event transfer in misinformation detection—namely, that models may perform well on events they have encountered but struggle to adapt to newly emerging events.

Singhal et al. proposed the SpotFake framework, which uses news text and corresponding images together for misinformation detection. This method uses a pre-trained language model to extract text features and a pre-trained visual model to extract image features, then fuses these two types of features for classification[11]. Compared to monomodal methods that use only text or images, SpotFake demonstrates the complementary nature of text and visual information.

Khattar et al. proposed MVAE, a multimodal variational autoencoder method. This method uses a variational autoencoder to learn a joint latent representation of text and images, which is then applied to misinformation detection[12]. Compared to simple feature concatenation, MVAE places greater emphasis on learning a shared latent space from text and images, enabling the model to capture multimodal semantic relationships at a higher level.

Although text-image fusion provides a basic framework for multimodal detection, it still needs further modeling of cross-modal consistency and external context.

3.2. Text-Image Consistency and Vision-Language Pre-training Models

The complexity of multimodal misinformation lies not only in the falsity of text or images themselves, but also in semantic mismatches between the two modalities. For example, a genuine image may be used to describe a different event; a headline may intentionally exaggerate the content of an image; or an image may be relevant to the text’s topic but fail to support the factual claims made in the text. Therefore, semantic consistency between text and images has become a critical issue in multimodal misinformation detection.

Text-image consistency is usually measured by comparing the semantic representations of text and images. If the visual content strongly supports the textual description, the semantic similarity between the two modalities tends to be high. If the headline deviates from the image content or uses an irrelevant

image to attract attention, the similarity tends to be lower. Therefore, text-image similarity can serve as an important signal for identifying text-image mismatch, image misuse, and misleading visual presentation.

Zhou et al. proposed the SAFE method, which stands for Similarity-Aware Multimodal Fake News Detection. This method extracts features from both text and images separately and further models the similarity relationship between them to identify misinformation that uses irrelevant images to attract readers’ attention [13]. A key feature of SAFE is that it not only focuses on the individual representations of text and images but also incorporates cross-modal similarity as one of the detection criteria. Such methods are particularly effective at identifying text-image mismatches, image misuse, and misleading headlines.

However, text-image consistency does not equate to factual accuracy. Information with a high degree of text-image consistency may still be false, as the image could originate from a different time or place; conversely, information with low text-image similarity is not necessarily false, since news images sometimes serve merely as supplementary illustrations. Therefore, text-image consistency is better suited as an important clue in multimodal detection rather than as the ultimate basis for determining factual truth. To further enhance detection reliability, it is necessary to incorporate external knowledge such as event background, entity relationships, retrieved evidence, and source-context information.

In recent years, vision-language pre-training models have provided a new technical foundation for multimodal misinformation detection. Traditional multimodal methods typically involve training or invoking text encoders and image encoders separately, followed by fusion in downstream tasks. Vision-language pre-training models, on the other hand, learn cross-modal semantic alignment capabilities through large-scale text-image data, enabling text and images to be mapped to interrelated semantic spaces.

CLIP is a representative vision-language pre-training model. Radford et al. trained text and image encoders through large-scale image-text contrastive learning, enabling the model to learn transferable visual representations based on natural language supervision. In misinformation detection, CLIP can be used to extract semantic vectors for both text and images, as well as to compute the semantic relevance between images and text. Compared to traditional visual models, CLIP’s advantage lies in the fact that its training objective inherently incorporates text-image matching relationships, making it better suited for addressing multimodal semantic alignment problems.

For a sample consisting of a caption and an image, CLIP

can be used to obtain the semantic representation of the text and the visual representation of the image. These representations can then be fused for downstream classification or used to measure text-image relevance. This approach leverages CLIP’s cross-modal semantic alignment capability and provides useful foundational features for multimodal misinformation detection. However, CLIP primarily learns semantic alignment between text and images and does not directly perform fact-checking. In other words, CLIP can determine whether text and images are related in semantic space, but it cannot independently determine whether the events described in the text actually occurred. Therefore, in scenarios involving historical events, news facts, source-context information, and background knowledge, external knowledge is still required to further supplement the model.

4. Multimodal Misinformation Detection with External Knowledge Augmentation

Multimodal misinformation detection can simultaneously utilize text and image information; however, text and images still provide only in-sample information. For many complex cases of misinformation, relying solely on in-sample content is insufficient to make a reliable judgment. For example, text and images may be highly consistent semantically, yet the corresponding event may not have actually occurred; an image may be authentic but be used in connection with the wrong time, place, or event; a headline may not contain obviously false words, but its meaning can only be determined by considering the news context, entity relationships, or source-context information. Therefore, external knowledge augmentation has gradually become an important direction in multimodal misinformation detection.

The core idea behind external knowledge augmentation is to introduce additional factual background, entity relationships, retrieved evidence, source-context information, or common-sense information beyond the text and images, enabling the model to make judgments not only based on the surface-level semantics of the current sample but also by leveraging broader contextual information. Compared to simple multimodal fusion, external knowledge augmentation methods focus more on whether “the sample’s content is consistent with external facts or context,” and thus more closely resemble the fact-checking process in real-world scenarios.

4.1. Types and Representations of External Knowledge

In misinformation detection, external knowledge can be drawn from various sources and represented in different forms. According to its source and usage, it can generally be divided into factual knowledge, entity-relationship knowledge, retrieval-augmented knowledge, and source-context knowledge.

The first category is factual knowledge. Factual knowledge typically comes from encyclopedias, news databases, fact-checking websites, or reports from authoritative media outlets, and is used to provide background context for events and verify information veracity. For example, news stories involving people, places, organizations, time, and event relationships often require external facts to determine whether their descriptions are valid.

The second category is entity-relationship knowledge. This type of knowledge typically exists in the form of knowledge graphs, representing relationships between people, organizations, places, concepts, and events. In a knowledge graph, entity-relationship knowledge is usually represented through triples consisting of a head entity, a relation, and a tail entity. Through entity recognition and entity linking, entities in news texts can be mapped to external knowledge graphs, thereby obtaining entity attributes, neighboring nodes, and relationship paths. The KAN model proposed by Dun et al. utilizes entities and their contextual information from knowledge graphs to design a knowledge-aware attention mechanism that enhances misinformation detection[14].

The third category is retrieval-augmented knowledge. Methods in this category typically retrieve relevant evidence from search engines, news repositories, or document collections based on news headlines, statements, or entity keywords, and then use the retrieved text to assist in decision-making. Retrieval-augmented knowledge is advantageous because it can be updated rapidly and can therefore cover new events not yet included in knowledge graphs; however, its disadvantages are also evident, as search results may suffer from noise, off-topic content, or unreliable sources.

The fourth category is source-context knowledge. Source-context knowledge focuses on factors such as the information publishing platform, media domain, community type, topic background, and content style. For example, in a Reddit scenario, the subreddit and domain can reflect the content’s publishing environment; in a news scenario, media sources, site types, and section attributes also influence content understanding. Source-context knowledge is not directly equivalent to a veracity label; rather, it provides the model with contextual explanations regarding “where this information appears and in what context.”

External knowledge is not tied to a single representation. Some systems use natural-language passages, such as encyclopedia entries, retrieved news reports, source descriptions, or short evidence summaries. Others retain structure in the form of triples or local knowledge-graph subgraphs. Textual evidence can be passed through a language encoder and aligned with the post representation, whereas graph-based evidence is usually processed with graph embeddings, graph neural networks, or attention over entities and relations. The choice of representation therefore determines what relations are preserved and how directly the model can compare the evidence with the claim.

CompareNet, proposed by Hu et al., compares news content with external knowledge bases via entities and constructs a directed heterogeneous document graph containing topics and entities to capture consistency relationships between the news content and the knowledge base [15]. This method demonstrates that external knowledge can not only serve as additional features input to the model but can also be used to determine whether conflicts exist between news content and knowledge facts.

Source-context information can be handled without assigning a credibility label to a source. In the Fakeddit setting, subreddit names, domains, community types, or media categories can first be translated into short descriptive statements and then encoded as text. A news-oriented domain, for instance, may be described in terms of reporting public events, whereas an image-editing community may be characterized by its focus on visual creation and manipulation. These descriptions are intended to convey publishing context

and typical content form; they should not state or imply whether the associated samples are true or false.

4.2. Fusion Strategies and Effects of External Knowledge

Once the three inputs have been encoded, the simplest implementation joins the text, image, and knowledge vectors before classification. This baseline is useful because it tests whether external information adds signal under a controlled architecture. Its weakness is that every knowledge item receives the same opportunity to influence the decision.

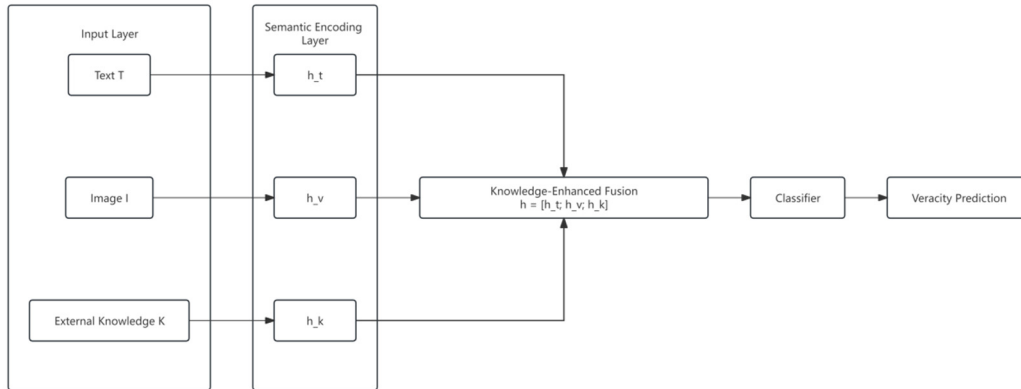


Figure 2. Framework of External Knowledge-Augmented Multimodal Misinformation Detection

Beyond simple concatenation, some studies employ attention mechanisms or graph structures to model the relationships between different modalities. The knowledge-aware multimodal adaptive graph convolutional network proposed by Qian et al. models text information, visual information, and knowledge concepts together as a unified graph structure and uses adaptive graph convolution to learn semantic relationships between different nodes[16]. Such methods can explicitly model the associations between text, images, and knowledge, rather than simply concatenating different features.

In recent years, there have also been studies combining external knowledge with visual-language models. Gao et al. proposed a knowledge-enhanced vision-language model that incorporates large-scale open knowledge graphs into multimodal misinformation detection and constructs a graph structure containing text entities, relationships, and image objects to enhance the model’s ability to judge facts[17]. This demonstrates that external knowledge can complement vision-language pre-trained models: the vision-language model is responsible for semantic alignment between text and images, while external knowledge supplements entity relationships and factual context.

External knowledge mainly supplements information that is difficult to obtain from the sample itself, including event background, historical information, entity attributes, and source-context information. It can help the model determine whether the relationships among people, places, organizations, and events are consistent with external facts. In addition, external evidence and source-context knowledge can support the interpretation of text-image relationships, especially in cases involving image misuse, outdated images, or misleading source-context information.

Irrelevant evidence can therefore enter the representation, and fixed concatenation does not adapt when one modality is more informative than another for a particular sample.

Figure 2 summarizes the corresponding pipeline. Separate encoders produce representations for the post text, the accompanying image, and the external knowledge item. The fusion block combines these representations, after which a classifier estimates the veracity label. More elaborate variants replace the fixed fusion block with attention, gating, or graph reasoning, but the three-stage organization - encoding, cross-source fusion, and prediction - remains common.

4.3. Source-Context Knowledge and Leakage-Free Construction

Among the various types of external knowledge, source-context knowledge is closely tied to the data characteristics of social media platforms and is particularly well-suited for datasets that include fields such as subreddits and domains. It differs from traditional knowledge graphs and does not necessarily come from fact-checking websites; rather, it derives from the information-sharing environment itself. For example, subreddits, domains, media sources, section attributes, and community types can all provide contextual clues to help the model understand samples.

It is important to note that source-context knowledge must be constructed in a neutral manner. If the knowledge text directly contains strong labels such as “fake,” “false,” “misleading,” “fabricated,” or “manipulated,” the model may directly use these terms for classification, thereby causing knowledge leakage. In such cases, an improvement in model performance does not indicate that external knowledge has truly enhanced semantic understanding; rather, the model may have merely learned to follow the implications of the labels.

Therefore, when constructing source-context knowledge, one should avoid directly describing a source as “true” or “false,” and instead describe its subject matter, content format, community style, and media attributes. For example, an image-hosting domain could be described as “a platform for online image sharing or user-uploaded images”; news-related sources could be described as “sources related to news reporting and public events”; and visual creativity communities could be described as “communities focused primarily on image editing, visual creativity, or visual discussions.” Such neutral descriptions reduce the risk of label leakage while preserving the supportive role that source-context information plays in understanding the samples.

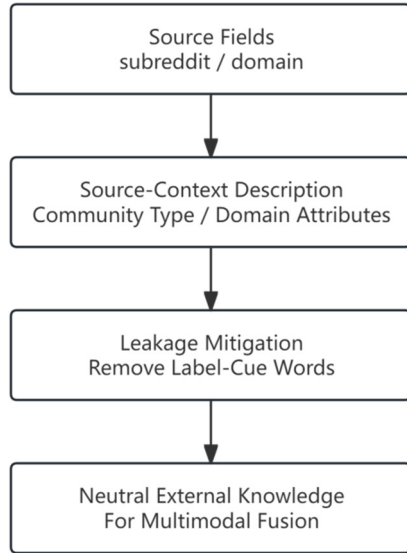


Figure 3. Construction Process of Leakage-Free Source-Context Knowledge

The value of source-context knowledge lies in its ability to supplement information about the publication environment that cannot be directly expressed by text or images. Particularly in social media datasets, the same text or image may carry different meanings across different communities, domains, and content contexts. Incorporating this type of

contextual information as external knowledge into multimodal detection helps models better understand the dissemination environment in which a sample is embedded.

Overall, external knowledge augmentation expands the model’s basis for judgment from in-sample text-image content to out-of-sample contextual information. Knowledge graphs, retrieved evidence, and source-context knowledge can respectively provide entity relations, event background, and platform context. However, the effectiveness of knowledge augmentation depends on knowledge quality, relevance, fusion strategy, and leakage control. Therefore, neutral construction, reliable retrieval, and interpretable fusion remain important issues for future research.

5. Comparison and Analysis of Methods

Sections 3 and 4 show that the main approaches differ less in their final classification objective than in the evidence made available to the model. Text-only systems operate on linguistic cues; multimodal systems add visual or social signals; knowledge-enhanced systems extend the evidence beyond the sample. The practical trade-offs concern coverage, computational cost, dependence on external resources, and the extent to which a prediction can be traced to a meaningful relation. This section brings those differences together before examining the Fakeddit results.

5.1. Comparison of Different Detection Methods

Table 2. Comparison of Different Misinformation Detection Methods

Method Category	Representative Methods	Main Information Sources	Advantages	Limitations
Text-based detection methods	LIAR-related text classification methods, pre-trained language models	News titles, article bodies, claim texts	Simple structure, easy implementation, suitable for short-text and news-claim detection	Only use internal textual information and are difficult to identify image misuse and text-image mismatch
Multimodal fusion methods	Jin et al.’s method, EANN, SpotFake, MVAE	Text, images, social context	Can utilize complementary text-image information and are suitable for social media scenarios	Simple fusion makes it difficult to distinguish the contribution of different modalities and lacks external factual background
Text-image consistency methods	SAFE, CLIP feature similarity methods	Text, images, text-image similarity	Can identify text-image mismatch, misleading images, and weakly related text-image combinations	Text-image consistency does not equal factual truth and cannot independently perform fact verification
Knowledge graph-enhanced methods	KAN, CompareNet	Entities, relations, knowledge graph subgraphs	Can introduce entity relations and structured knowledge, with relatively strong interpretability	Entity linking and knowledge graph construction are costly, and coverage of emerging events is limited
Multimodal knowledge graph modeling methods	KMAGCN, knowledge-enhanced vision-language models	Text, images, entities, relations, visual objects	Can jointly model textual content, visual content, and external knowledge relations, producing richer semantic representations	Model structures are complex and highly dependent on knowledge quality and data annotation
Adaptive knowledge learning methods	AKA-Fake [18]	Multimodal content, candidate knowledge subgraphs	Can dynamically select relevant knowledge for each sample and reduce interference from irrelevant knowledge	Training and inference are complex, and the reliability of knowledge selection still needs further validation
Source-context knowledge-enhanced methods	Source-context knowledge based on subreddit, domain, and media sources	Platform communities, source domains, content styles	Suitable for social media data and can supplement publishing context and content background	Label words and category cues must be avoided; otherwise, knowledge leakage may occur

For comparison, the literature can be organized into six broad families: text-based classifiers, general multimodal

fusion models, text-image consistency models, knowledge-graph-enhanced systems, multimodal knowledge models, and source-context approaches. The categories are not mutually exclusive - for example, a knowledge-aware vision-language model belongs to more than one group - but they provide a useful way to compare the principal information source, expected benefit, and characteristic limitation of each design. Table 2 presents this comparison.

Table 2 reveals a progression in the evidence used for prediction. Text classifiers are comparatively lightweight and remain appropriate for short claims, yet they cannot observe visual misuse. Multimodal fusion adds complementary image and social signals, while consistency-oriented models ask whether the visual content supports the accompanying text. Neither step, by itself, establishes that the described event is true: both still reason mainly from the material contained in the post.

Knowledge-based methods move the decision boundary beyond the post. Graph and retrieval approaches contribute entity relations or event evidence, and multimodal knowledge models connect those signals with textual and visual representations through graph reasoning, attention, or vision-language encoders. Source-context knowledge serves a different purpose. It does not verify an event directly, but describes the community, domain, or publication setting in which the content appears. This information is useful only when it is written neutrally; label-bearing descriptions can turn contextual assistance into leakage and produce misleadingly high scores.

5.2. Preliminary Case Study on Fakeddit

The Fakeddit experiment is used here as an illustrative case rather than as a claim of state-of-the-art performance. Its purpose is to separate two possible explanations for a gain from external knowledge: genuine contextual assistance and direct exposure to category cues. CLIP provides the text and image representations, and logistic regression is retained as a deliberately simple classifier so that changes in performance can be attributed more clearly to the added features. Source-context descriptions are generated from the subreddit and domain fields. Before encoding, terms that explicitly signal a class - fake, false, misleading, fabricated, and manipulated - are removed from the neutral version. Table 3 compares the multimodal baseline with strong and neutral knowledge. The CLIP text-image model obtains 0.8250 for both Accuracy and Macro-F1. Adding neutral source-context knowledge raises both measures to 0.9350, indicating that the source and community descriptions contain useful information beyond the paired text and image.

Table 3. Preliminary Case Study Results on External Knowledge Augmentation

Method	Accuracy	Macro-F1
CLIP Text + Image	0.8250	0.8250
CLIP Text + Image + Strong Knowledge	0.9800	0.9800
CLIP Text + Image + Neutral Knowledge	0.9350	0.9350

The strongest score in Table 3 must be interpreted cautiously. With label-bearing knowledge, Accuracy and Macro-F1 reach 0.9800, but this increase is compatible with a shortcut explanation: the classifier may associate explicit category words with the target instead of learning how the knowledge relates to the post. The large gap between the

strong-knowledge setting and the baseline is therefore not sufficient evidence of improved reasoning. It is also a warning that the wording of external knowledge is part of the experimental protocol and should be reported and audited.

The neutral setting provides a more informative comparison. It retains descriptions of source type, community focus, content format, and media attributes while excluding direct class cues. Performance remains above the CLIP baseline, which suggests that publication context contributes predictive information even when the answer is not embedded in the knowledge text. The result should be read as contextual enrichment rather than fact verification: the added features help characterize the environment around a post, but they do not independently establish the truth of the underlying event.

Taken together, the method comparison and the case study point to a distinction that is easy to overlook. Multimodal fusion reduces the information loss of text-only models, whereas external knowledge addresses context that is not present in either modality. The benefit of the latter depends on what the knowledge actually contains. Label hints can produce a large numerical gain without a corresponding improvement in semantic analysis; neutral descriptions provide a stricter test of whether context itself is useful. For source-context experiments, construction rules should therefore emphasize topic, format, platform, and media properties while excluding words or categories that reveal the target.

The experiment also has clear limits. It uses one dataset and a lightweight classifier built on fixed CLIP features, so the reported values are not a comprehensive benchmark. No state-of-the-art comparison, repeated-run analysis, or statistical significance test is included. In addition, subreddit- and domain-based descriptions are better viewed as contextual metadata expressed in language than as independently retrieved factual evidence. The case study consequently supports a methodological point about leakage control; it does not establish that source-context knowledge will generalize across datasets, events, or model families.

6. Conclusion

The review indicates that the move from text-only detection to multimodal analysis has broadened what models can observe, but it has not removed the need for context. Many misleading posts depend on the relation between components: an authentic image is reused, a caption changes the event meaning, or a source setting encourages a satirical or edited interpretation. Text and image features can expose some of these relations, yet they remain insufficient when verification depends on time, place, entity history, or source background.

Accordingly, this paper examined three connected layers of the problem. The dataset discussion showed how current benchmarks differ in modality, label granularity, and social-context coverage. The method review then traced the transition from feature-level fusion to consistency modeling and vision-language pre-training. Finally, the knowledge section distinguished factual, entity-relation, retrieval-based, and source-context resources and considered how each can be represented and integrated.

The Fakeddit case study provides a concrete, though limited, illustration. A CLIP text-image baseline achieved 0.8250 Accuracy and Macro-F1; adding neutral descriptions derived from subreddit and domain information increased both measures to 0.9350. Because explicit label cues were

removed, this result is more consistent with a contribution from publication context than with direct label exposure. At the same time, the 0.9800 score obtained with strong knowledge demonstrates how easily an apparently superior result can be produced by leakage.

The central implication is that external knowledge should be evaluated as evidence, not merely as additional input. Its usefulness depends on relevance, reliability, temporal validity, and the absence of label cues. Future work should make knowledge construction reproducible, compare neutral and unfiltered variants, report retrieval or selection errors, and provide mechanisms that reveal which knowledge items affected the prediction. These controls are necessary if knowledge-enhanced multimodal models are to improve factual reasoning rather than exploit dataset-specific shortcuts.

References

- [1] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36.
- [2] Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), Article 109.
- [3] Nakamura, K., Levy, S., & Wang, W. Y. (2020). r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 6149–6157). European Language Resources Association.
- [4] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8748–8763). PMLR.
- [5] Wang, W. Y. (2017). “Liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 422–426). Association for Computational Linguistics.
- [6] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). 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.
- [7] Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3), Article e0150989.
- [8] Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia* (pp. 795–816). ACM.
- [9] Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D. T., Boato, G., Riegler, M., & Kompatsiaris, Y. (2015). Verifying multimedia use at MediaEval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop* (Vol. 1436). CEUR Workshop Proceedings.
- [10] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 849–857). ACM.
- [11] Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. (2019). SpotFake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data* (pp. 39–47). IEEE.
- [12] Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference* (pp. 2915–2921). ACM.
- [13] Zhou, X., Wu, J., & Zafarani, R. (2020). SAFE: Similarity-aware multi-modal fake news detection. In *Advances in Knowledge Discovery and Data Mining* (pp. 354–367). Springer.
- [14] Dun, Y., Tu, K., Chen, C., Hou, C., & Yuan, X. (2021). KAN: Knowledge-aware attention network for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 1, pp. 81–89).
- [15] Hu, L., Yang, T., Zhang, L., Zhong, W., Tang, D., Shi, C., Duan, N., & Zhou, M. (2021). Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 754–763). Association for Computational Linguistics.
- [16] Qian, S., Hu, J., Fang, Q., & Xu, C. (2021). Knowledge-aware multi-modal adaptive graph convolutional networks for fake news detection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17(3), Article 98.
- [17] Gao, X., Wang, X., Chen, Z., Zhou, W., & Hoi, S. C. H. (2024). Knowledge enhanced vision and language model for multi-modal fake news detection. *IEEE Transactions on Multimedia*, 26, 8312–8322.
- [18] Zhang, L., Zhang, X., Zhou, Z., Huang, F., & Li, C. (2024). Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 15, pp. 16777–16785).