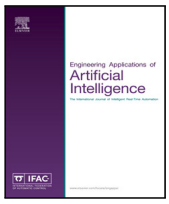
Engineering Applications of Artificial Intelligence 142 (2025) 109931

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Cross-modal augmentation for few-shot multimodal fake news detection

Ye Jiang [a], Taihang Wang [a], Xiaoman Xu [a], Yimin Wang [b,*], Xingyi Song [c], Diana Maynard [c]

[a] College of Information Science and Technology, Qingdao University of Science and Technology, China
[b] College of Data Science, Qingdao University of Science and Technology, China
[c] Department of Computer Science, University of Sheffield, UK

ARTICLE INFO

Keywords:
Fake news detection
Multimodal fusion
Few-shot learning
Natural language processing

ABSTRACT

The nascent topic of fake news requires automatic detection methods to quickly learn from limited annotated samples. Therefore, the capacity to rapidly acquire proficiency in a new task with limited guidance, also known as few-shot learning, is critical for detecting fake news in its early stages. Existing approaches either involve fine-tuning pre-trained language models which come with a large number of parameters, or training a complex neural network from scratch with large-scale annotated datasets. This paper presents a multimodal fake news detection model which augments multimodal features using unimodal features. For this purpose, we introduce Cross-Modal Augmentation (CMA), a simple approach for enhancing few-shot multimodal fake news detection by transforming n-shot classification into a more robust (n × z)-shot problem, where z represents the number of supplementary features. The proposed CMA achieves state-of-the-art (SOTA) results over three benchmark datasets, utilizing a surprisingly simple linear probing method to classify multimodal fake news with only a few training samples. Furthermore, our method is significantly more lightweight than prior approaches, particularly in terms of the number of trainable parameters and epoch times. The code is available here: https://github.com/zgjiangtoby/FND_fewshot

1. Introduction

The recent proliferation of social media has not only transformed the landscape of information exchange, but also led to the pernicious spread of fake news. The detection and mitigation of fake news have consequently become pivotal areas of research (Conroy et al., 2015; Long et al., 2017). Traditional approaches, primarily relying on textual analysis, have shown limitations due to the sophisticated and multi-faceted nature of fake news (Wang et al., 2018; Lao et al., 2021). In response, many studies have incorporated multimodal methods that consider both text and accompanying images, yielding a more comprehensive and effective framework for identifying and debunking fake news (Chen et al., 2022; Zhou et al., 2023).

To explore the inconsistent semantics between text and image in fake news, many studies have either incorporated contrastive learning to achieve better alignment between image-text pairs (Wang et al., 2023b), or designed complex neural networks to strengthen the deep-level fusion of multimodal features (Wu et al., 2023b; Qu et al., 2024). The former relies on contrastive loss to align image-text pairs, but most image-text pairs in fake news are inherently not matched (Gao et al., 2022), and different image-text pairs may also have potential correlations (Li et al., 2021), which can consequently confuse the model. The latter typically needs to be trained from scratch, which is fundamentally bounded by the availability of large-scale annotated data (Rashkin et al., 2017; Shu et al., 2020).

In contrast to machines, the process of concept learning in humans involves integrating multimodal signals and representations (Meltzoff and Borton, 1979; Nanay, 2018). When processing uncertain information, people inherently seek help from other modalities. This capability enables humans to learn from a limited number of samples by incorporating cross-modal information, as shown in Fig. 1. Meanwhile, the efficacy of fake news detection (FND) in the context of nascent topics, such as COVID-19, remains a significant challenge for prevailing strategies. This difficulty is compounded by the lack of extensive data and annotations in the target domain, underscoring the critical role of few-shot learning in mitigating the spread of early-stage fake news (Wu et al., 2023a).

In the context of emerging topics with limited training samples, prompt learning, through its few-shot learning capacity, encapsulates news articles in task-specific textual prompts for direct knowledge extraction from pre-trained language models (PLMs), achieving comparable performance across different tasks (Gao et al., 2021; Ding et al.,

* Corresponding author.
E-mail address: yimin.wang@qust.edu.cn (Y. Wang).

https://doi.org/10.1016/j.engappai.2024.109931
Received 25 July 2024; Received in revised form 26 November 2024; Accepted 23 December 2024
Available online 1 January 2025

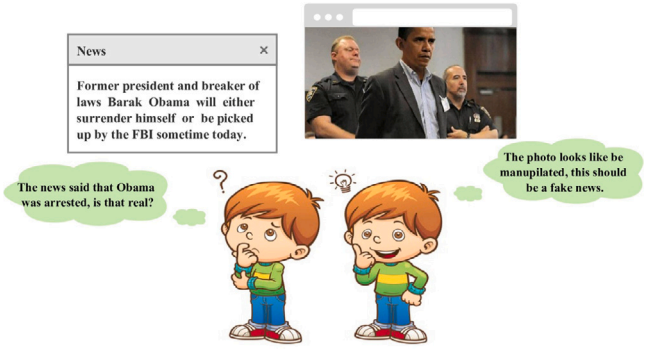


Fig. 1. Information from different modalities assists humans in decision-making, especially when faced with uncertainty.

2021). However, most prompt-based methods primarily tune the PLM with unimodal textual information from fake news (Jiang et al., 2022; Wu et al., 2023a), thus once again ignoring the multimodal nature of fake news.

Even though the previous method (Jiang et al., 2023) attempts to integrate different prompt templates with image features extracted from the pre-trained vision model, the fusion strategy still utilized the multimodal features only, potentially struggling to address spatial discrepancies between visual and textual semantics (Wu et al., 2023b; Guo et al., 2023). Meanwhile, few-shot FND approaches often suffer from sample quality issues, where the quality of unimodal sample (Jiang et al., 2023) might potentially harm the performance of few-shot learners.

In this paper, we propose a **C**ross-**M**odal **A**ugmentation (CMA) method to explore how to enhance multimodal few-shot FND tasks by augmenting unimodal feature representations. The unimodal feature representation of each text-image pair can be extended to z unimodal augmentations, enhancing sample quality for the FND task.

Specifically, we leverage the foundational multimodal model Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) to extract textual and visual features from fake news simultaneously. Utilizing class labels as supplementary one-shot training instances, the n-shot classification can then be converted to an $(n \times z)$-shot problem. Meanwhile, we also fuse unimodal features by utilizing the cross-attention mechanism (Wang et al., 2023b) as another supplementary. Finally, we employ a simple linear probing (He et al., 2022) for each modality as well as for the fused multimodal features. The experimental results indicate that CMA achieves SOTA results across three datasets.

The main contributions of this paper are:

- Introduction of a Cross-Modal Augmentation method for few-shot multimodal fake news detection, utilizing unimodal features to enhance multimodal fusion.
- Leveraging a pre-trained multimodal model to extract unimodal features, and repurposing class labels as additional one-shot training samples, transforming the n-shot classification into a more robust $(n \times z)$-shot problem.
- By freezing the pre-trained multimodal model and training only with a simple linear classifier, the proposed CMA achieves SOTA results over three datasets, outperforming 11 baseline models and surpassing previous methods in efficiency.

2. Related work

2.1. Unimodal fake news detection

Unimodal fake news detection aims to extract significant semantics from either news texts or images. Given the precision of semantics in text, previous approaches have concentrated on the task of text-based unimodal fake news detection. Early works focused on analyzing statistical characteristics of text (e.g., length, punctuation, exclamation marks) (Castillo et al., 2011) and metadata (e.g., likes, shares) (Tabibian et al., 2017; Geeng et al., 2020) for manual fake news detection. However, these manual feature engineering approaches are time-consuming and struggle with processing large-scale, real-time data (Liu et al., 2016; Fedoryszak et al., 2019).

The advent of deep learning has significantly advanced automated fake news detection. These methods primarily utilize deep learning models like Bidirectional Long Short-Term Memory (BiLSTM) (Bahad et al., 2019; Sridhar and Sanagavarapu, 2021), Graph Neural Network (GNN) (Phan et al., 2023), pre-trained models (Song et al., 2021; Jiang et al., 2021, 2020) to analyze text features, extracting various attributes such as emotional (Ghanem et al., 2020), stance-based (Jiang, 2023), and stylistic elements (Wu et al., 2021). Recently, graph-based studies have demonstrated advancements in FND tasks. For instance, GAMC (Yin et al., 2024) utilizes a Graph Autoencoder with Masking and Contrastive Learning to leverage both content and contextual propagation signals as self-supervised cues. Similarly, RF-RSL (Hou et al., 2024) combines a greedy sensor deployment strategy with a limited-information-oriented inference approach. However, the recent proliferation of multimodal information (text, images, videos) in social networks has shifted the propagation of fake news from solely text-based to multimodal formats.

2.2. Multimodal fake news detection

Multimodal methods employing cross-modal discriminative patterns have been introduced, aiming to enhance performance in fake news detection. For example, Multimodal Co-Attention Networks (Wu et al., 2021) employs multiple co-attention layers to effectively integrate textual and visual features in detecting fake news. Ambiguityaware multimodal fake news detection (CAFE) (Chen et al., 2022) quantifies cross-modal ambiguity through the assessment of the Kullback–Leibler (KL) divergence among the distributions of unimodal features. Singhal et al. (2022) determines the modality that exhibits greater confidence in the context of fake news detection. Crossmodal contrastive learning framework (COOLANT) (Wang et al., 2023b) focuses on improving the alignment between image and text representations, utilizing contrastive learning for finer semantic alignment and cross-modal fusion to learn inter-modality correlations. However, these approaches are limited by the need for extensive annotated data in the context of emerging topics.

2.3. Cross-modal few-shot fake news detection

Few-shot learning is designed to master new tasks using a limited number of labeled examples (Wang et al., 2020). Current few-shot learning methodologies, such as prototypical networks, acquire class-specific features in metric spaces for swift adaptation to novel tasks (Vinyals et al., 2016; Snell et al., 2017). Within computer vision, the concept of few-shot domain adaptation is explored in image classification for transferring knowledge to novel target domains (Motiian et al., 2017; Zhao et al., 2021). In natural language processing, meta-learning is suggested as a means to enhance few-shot learning performance in tasks like language modeling (Sharaf et al., 2020; Han et al., 2021) and misinformation detection (Yue et al., 2023; Zhang et al., 2021b). To our knowledge, the application of few-shot multimodal fake news detection through cross-modal augmentation remains unexplored in existing literature.

Meanwhile, previous multimodal learning approaches have sought to enhance unimodal tasks by leveraging data from various modalities (Schwartz et al., 2022; Zhang et al., 2021c). With multimodal pre-trained models achieving notable success in classic vision tasks (Radford et al., 2021; Zhang et al., 2022), there is a growing interest in formulating more efficient cross-modal augmentation techniques.

However, the prevailing techniques are based on successful strategies originally designed for multimodal foundational models. For example, CLIP utilizes linear probing (He et al., 2022, 2020) and comprehensive fine-tuning (Girdhar and Ramanan, 2017) in its application to downstream tasks. CLIP-Adapter (Gao et al., 2023) and Tip-Adapter (Zhang et al., 2021a) draw inspiration from parameter-efficient finetuning approaches (Houlsby et al., 2019) that focus on optimizing lightweight multilayer perceptrons (MLPs) while maintaining a fixed encoder. However, all the aforementioned methods, including ensemble the weights of the zero-shot and fine-tuned models (WiSE-FT) (Wortsman et al., 2022), employ an alternative modality, such as textual labels, as classifier weights, and continue to compute a unimodal Softmax loss on few-shot tasks. In contrast, this paper demonstrates the enhanced effectiveness of incorporating additional modalities as training samples.

3. Methodology

The proposed CMA enhances few-shot fake news detection by integrating samples from different modalities, and extends traditional unimodal few-shot classification to leverage the richness of cross-modal data. This section starts with a standard unimodal few-shot FND framework, and the loss function is discussed. Then, it extends this to multiple modalities, assuming each training example is a combination of five different modalities. The modality-specific features are passed through linear classifiers to obtain their inferences. Finally, we combine the inferences and train a meta-linear classifier to compute the final prediction.

3.1. Unimodal few-shot FND

Initially, unimodal few-shot FND learns from a labeled dataset of $(x, y) \in X$, where x is either the text or image passing to a pre-trained feature encoder. The ultimate goal is to allocate a binary classification label of $y \in \{0, 1\}$, in which 0 denotes real news and 1 denotes fake news. We assume only an n-shot subset (x_i, y_i) from X is provided for training, where $i \in [1, n]$ (i.e., n samples per class); the rest of X is used as the test set.

Therefore, the standard unimodal FND can be denoted as minimizing the cross-entropy loss L:

$$L = -\frac{1}{2n}\sum_{i=1}^{2n}(y_i log(y_i') + (1 - y_i) log(1 - y_i')), \tag{1}$$

where y_i' represents the output of the model for a given input sample, derived from the inference of the linear classifier after applying the softmax function.

$$y_i' = \text{softmax}(\text{linear probing}(f)) = \frac{e^{w_j * f}}{\sum_k e^{w_k * f}}, \tag{2}$$

where f denotes the feature representation extracted using a pre-trained unimodal feature encoder, followed by an linear layer. w_j is the weight of the linear layer assigned to the correct class. The denominator, $\sum_k e^{w_k * f}$, is the sum of exponentials over all elements in the output of the linear layer. This normalization ensures that the outputs form a valid probability distribution, with their sum constrained to 1. The optimization process involves updating w_j and w_k for the predicted classes through gradient descent (LeCun et al., 2015), with the objective of minimizing the cross-entropy loss (Rubinstein and Kroese, 2004).

3.2. Multimodal few-shot FND

The architecture of the CMA model is illustrated in Fig. 2. To extend the approach to multimodal FND, we consider each training sample x_i as comprising an image-text pair, denoted as $[x_{mi}, x_{ti}]$, where x_{mi} represents the visual modality and x_{ti} represents the textual modality. We employ linear probing to adapt pre-trained CLIP models to downstream tasks. Linear probing involves: (1) leveraging features extracted from the pre-trained model, and (2) training a lightweight linear classifier to tailor the pre-trained features for the few-shot FND task. The feature representation f is a combination of five components, each processed by linear probing: (1) a text-only feature f_t; (2) an image-only feature f_m; (3) concatenation of L2 normalized $f_c = [f_t \oplus f_m]$, where $\oplus$ is the concatenation operation; (4) an image-text cross-attended feature f_{mt}; (5) a text-image cross-attended feature f_{tm}.

The cross-attention mechanism, which swaps the text query Q_t with the image query Q_m, to obtain the cross-attended feature f_{mt} is denoted as follows:

$$f_{mt} = CrossAtt_{m\to t}(Q_m, K_t, V_t) = softmax(\frac{Q_m K_t^T}{\sqrt{d}})V_t. \tag{3}$$

In contrast, by swapping the image query Q_m with the text query Q_t, the cross-attended feature f_{tm} can be obtained:

$$f_{tm} = CrossAtt_{t\to m}(Q_t, K_m, V_m) = softmax(\frac{Q_t K_m^T}{\sqrt{d}})V_m, \tag{4}$$

where K_t and K_m represent the key vectors for text and image features respectively, V_t and V_m denote the corresponding value vectors, and d refers to the dimensionality of the model.

In this study, five distinct types of features are treated as separate modalities. Each modality is processed through its respective linear probing, implemented as a linear layer, to generate five corresponding inferred probabilities. The Representer Theorem (Schölkopf et al., 2001) suggests that optimally trained classifiers can be expressed as linear combinations of their training samples. Inspired by this theorem, we concatenate the five inferred probabilities to construct a new input, which is then processed by a meta-linear classifier to produce the final prediction:

$$\hat{y} = \text{softmax}(\text{linear probing}(f_t \oplus f_m \oplus f_c \oplus f_{mt} \oplus f_{tm})). \tag{5}$$

The weights of all the five modality-specific linear classifiers, the final meta-linear classifier, and the cross attentions layers are updated simultaneously through a unified back-propagation process. Consequently, we convert the standard n-shot classification to an $(n \times z)$-shot problem. The training procedure for the proposed CMA framework is outlined in Algorithm 1.

Algorithm 1 Cross-modal Augmentation Algorithm

Input: source data X, number of seeds S, number of shots n
Initialize pre-trained multimodal model;
for seed $\in \{1, 2, \ldots, S\}$ **do**
  for x_i in $\{x_1, \ldots, x_n\}$, where x_i comprises an image-text pair, $[x_{mi}, x_{ti}] \in x_i$ **do**
    Apply the pre-trained vision model to x_{mi} to extract image feature f_m;
    Apply the pre-trained language model to x_{ti} to extract text feature f_t;
    Concatenate f_m with f_t and apply L2-normalization to obtain f_c;
    Compute cross-attended features f_{mt} and f_{tm} with Equations (3) and (4);
    Obtain the inferences for the above features' using their respective linear classifiers;
    Concatenate the inferences and compute the final prediction with Equation (5);
    Compute the cross-entropy loss with Equation (1);
    Perform back-propagation to update the weights of all the five modality-specific linear classifiers, the final meta-linear classifier, and the cross attentions layers;
  end for
end for

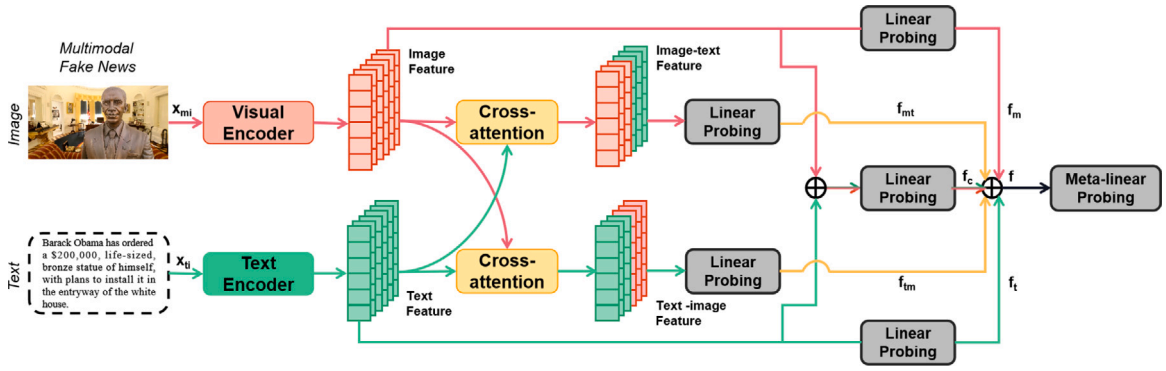


Fig. 2. The overall architecture of the CMA model.

4. Experiment

This section details experiments conducted to validate the effectiveness of the proposed approach. Initially, benchmark datasets are introduced, followed by the implementation details for experiments. The experimental results are analyzed in comparison to unimodal, multimodal, and few-shot FND methods. Finally, detailed analyses are provided to enhance the understanding of the proposed methods.

4.1. Data setup

Three publicly available datasets are utilized for evaluation.

PolitiFact (Shu et al., 2020) comprises a dataset of political news categorized as either fake or real by expert evaluators and is part of the benchmark FakeNewsNet project. Using the provided data crawling scripts, news with no images or invalid image URLs are removed, resulting in 198 multimodal news articles.

GossipCop (Shu et al., 2020) features entertainment stories rated on a scale from 0 to 10, with stories scoring less than five classified as fake news by the author of FakeNewsNet. Using the same retrieval strategies as PolitiFact, 6,805 multimodal news articles are collected.

Weibo (Jin et al., 2017), a dataset sourced from Chinese social media platforms, comprises a multimodal fake news collection featuring both text and images. Authentic news items were crawled from a reputable source (Xinhua News), and fake news was obtained from Weibopiyao, an official rumor refutation platform of Weibo, that aggregates content either through crowdsourcing or official rumor refutation efforts. The same pre-processing methods as in previous work (Wang et al., 2023b) are followed, resulting in 7,853 Chinese news articles.

Notably, the original text data in the FakeNewsNet and Weibo datasets often corresponds to multiple images. Rather than randomly selecting an image to pair with the text input, we follow the approach outlined in Jiang et al. (2023) by calculating the semantic similarity between the feature representations of text and images. Such calculations are commonly employed to pair the most semantically aligned image-text samples in cross-modal retrieval. For example, Wang et al. (2023a) analyzes semantic similarity by computing the cosine similarity between text-image pairs in the MediaEval Benchmark 2023. To find the most relevant image, the cosine similarity between each image and its corresponding text is calculated, and the image-text pair with the highest similarity, as determined by the pre-trained CLIP, is retained. The resulting dataset statistics are presented in Table 1.

Table 1
The statistics of the pre-processed multimodal fake news datasets.

Statistics	PolitiFact	GossipCop	Weibo
Total news	198	6805	7853
Fake news	96	1877	4211
Real news	102	4928	3642
Avg tokens	2148	728	67

Avg tokens denotes the average tokens per article.

4.2. Implementation details

The pre-trained OpenAI CLIP (ViT-B-32) (Radford et al., 2021) and Chinese CLIP (ViT-B-16) (Yang et al., 2022) models are utilized to respectively extract text and image features for different languages. The hidden size for the cross-attention projection layer is 512, which is the same as the output dimension of CLIP encoders. The AdamW optimizer is employed with a learning rate of 1e−3 and a decay parameter of 1e−2. The model is trained for 20 epochs, with the optimal checkpoint being determined by peak validation performance. Early stopping is utilized with a patience of three epochs.

In the few-shot context, the model is trained using a restricted set of samples, selected from the dataset to form an n-shot scenario. Here, $n \in [2, 8, 16, 32]$ represents the number of samples for each class, while the remainder of the samples are reserved for testing purposes. Given that the data quality of the sampled training set might significantly impact the model's performance, data sampling is repeated 10 times with random seeds, and the average score is reported after excluding the highest and lowest scores.

4.3. Benchmarked models

The proposed CMA is benchmarked against 11 representative models. Specifically, we extensively compare the proposed method with unimodal approaches (1)-(3), multimodal approaches (4)-(6), and the few-shot approaches (7)-(13).

(1) Explainable fake news detection (**dEFEND**) (Shu et al., 2019) utilizes the hierarchical attention network for FND. In this study, we remove the user comments from the original model.

(2) Latent Dirichlet Allocation Hierarchical Attention Network (**LDA-HAN**) (Jiang et al., 2020) integrates pre-calculated topic distributions from Latent Dirichlet Allocation into a hierarchical attention network for text classification.

(3) Fine-tuned RoBERTa (**FT-RoBERTa**) is a standard, fine-tuned version of the pre-trained language model RoBERTa; we use Huggingface Trainer to conduct the fine-tuning experiment.

(4) Multi-modal Framework for Fake News Detection (**SpotFake**) (Singhal et al., 2019) employs the pre-trained VGG and BERT for extracting image and text features, respectively, and then concatenating them for final classification.

(5) Similarity-Aware Multi-Modal Fake News Detection (**SAFE**) (Zhou et al., 2020) transforms images into textual descriptions and utilizes the correlation between text and visual information for FND.

(6) Cross-modal Ambiguity Learning for Multimodal Fake News Detection (**CAFE**) (Chen et al., 2022) employs an ambiguity-aware multimodal strategy to adaptively aggregate unimodal features and their correlations.

Table 2
Performance comparison between CMA and baseline models in accuracy (%).

Method	PolitiFact				GossipCop				Weibo				AVG
	2	8	16	32	2	8	16	32	2	8	16	32	
dEFEND	21.3	39.7	37.5	54.1	25.6	26.0	44.1	47.8	31.9	33.0	40.1	44.5	37.1
LDA-HAN	39.4	47.3	52.2	54.9	21.2	30.4	39.5	41.3	40.3	41.8	44.4	50.9	42.0
FT-RoBERTa	52.0	63.1	70.0	72.5	41.3	60.4	62.6	65.9	39.7	58.1	64.3	66.3	59.7
SAFE	19.0	27.3	48.7	52.1	31.3	45.2	45.4	47.1	21.1	19.3	39.4	41.1	36.4
SpotFake	49.3	53.7	58.5	63.4	28.3	28.4	34.4	36.1	36.9	41.3	40.4	53.7	43.7
CAFE	38.6	46.4	48.9	51.0	42.3	48.1	55.9	59.3	44.4	40.6	47.5	51.3	47.9
KPL	55.1	60.7	65.5	66.3	53.3	54.8	58.6	61.3	45.4	49.3	50.2	59.9	56.7
M-SAMPLE	56.2	66.1	69.5	73.4	53.4	54.1	59.7	66.0	49.7	52.1	59.8	65.7	60.5
KPT	68.1	74.8	80.0	83.2	52.5	56.5	58.1	67.0	56.9	69.4	69.9	71.2	67.3
PET	73.2	68.4	68.3	70.1	65.7	66.9	68.3	71.1	65.4	66.6	70.3	71.5	68.8
P&A	71.9	**80.7**	81.7	83.5	54.9	58.4	**75.6**	69.3	–	–	–	–	72.0
AMPLE	67.0♣	71.0♣	78.0♣	71.3	55.0♣	64.0♣	65.0♣	65.1	67.6	**71.5**	73.0	71.3	68.3
MPL	69.0♣	73.0♣	82.0♣	–	52.0♣	64.0♣	65.0♣	–	–	–	–	–	67.5
CMA(Ours)	**73.5**	75.8	**82.5**	**87.3**	**71.9**	**69.0**	71.7	**77.0**	**74.5**	69.9	**73.8**	**76.5**	**75.3**

#Bold indicates the best performance. Underline is the second-best performance. AVG denotes the average accuracy per model across all n-shot settings and datasets. ♣ denotes the scores are directly taken from the paper. Notably, the experimental results of P&A in Weibo are not accessible since it would require constructing the news proximity graph from the raw social context, which is not provided in the Weibo dataset.

(7) Knowledgeable Prompt Learning (**KPL**) (Jiang et al., 2022) employs prompt learning in RoBERTa by enhancing it with external knowledge representations.

(8) Mixed Template in Similarity-Aware Multimodal Prompt Learning for Fake News Detection (**M-SAMPLE**) (Jiang et al., 2023) incorporates prompt learning with multimodal FND. It also applies a similarity-aware fusing to adaptively combine the intensity of multimodal representation for FND.

(9) Pattern-Exploiting Training (**PET**) (Schick and Schütze, 2021) employs PLMs with task descriptions for supervised training, employing task-related cloze questions and verbalizers.

(10) Knowledgeable Prompt Tuning (**KPT**) (Hu et al., 2022) enhances the label word space by incorporating class-related tokens that exhibit diverse granularities and perspectives.

(11) Prompt&Align (**P&A**) (Wu et al., 2023a) combines prompt-based learning with social alignment techniques and addresses label scarcity by using task-specific prompts in PLMs to elicit relevant knowledge.

(12) Emotion-Aware Multimodal Fusion Prompt Learning (**AMPLE**) (Xu et al., 2024) integrates emotion-aware sentiment analysis and multimodal data fusion using hybrid prompt learning templates.

(13) Multi-Modal Prompt Learning (**MPL**) (Hu et al., 2024) leverages multi-modal pre-trained models and learnable prompts for early fake news detection.

4.4. Results

The FND accuracy comparison between the proposed CMA and all the baselines at various few-shot settings over the three datasets are shown in Table 2.

Comparing with unimodal baselines. First, we assess the accuracy of both unimodal approaches and the proposed CMA to evaluate their performances. Overall, CMA outperforms the best unimodal approach, FT-RoBERTa, achieving a 15.6% enhancement in average accuracy across all datasets, demonstrating its superiority in few-shot scenarios.

Surprisingly, FT-RoBERTa emerges as the most accurate model among both unimodal and multimodal approaches, suggesting that conventional fine-tuning methods can reach competitive levels of performance solely through the analysis of textual information from fake news. However, this method necessitates increased epoch time due to the adjustment of numerous parameters in the pre-trained language model (as shown in Table 4), making it impractical for real-world few-shot FND applications.

LDA-HAN yields the second best in accuracy among unimodal models, with dEFEND coming in next. This could be attributed to two factors: firstly, the vanilla LDA model struggles to effectively generate topics from short texts, a characteristic of the datasets from GossipCop and Weibo (as detailed in Table 1) used in LDA-HAN; secondly, the employment of GloVe embeddings for initializing LDA-HAN and dEFEND may not perform as effectively as the contextualized embeddings generated by the BERT family.

Comparing with multimodal baselines. We evaluate the performance of CMA in comparison with multimodal approaches. CMA outperforms the best multimodal baseline, CAFE, with a 27.4% improvement in average accuracy across all datasets. The reason might be that the complex architecture of multimodal approaches inherently comes with a large number of trainable parameters, which might easily lead to overfitting in few-shot scenarios.

Excluding FT-RoBERTa, all multimodal baselines outperform unimodal models on average, showing that the inclusion of the image modality can significantly affect model accuracy. While these multimodal approaches excel in scenarios with abundant data, their effectiveness heavily relies on the availability of high-quality annotated training samples, which may not be readily accessible during the initial stages of FND. Moreover, all multimodal approaches utilize pre-trained unimodal models, such as VGG, ResNet, and BERT, to independently extract features from images and text. Yet, since these unimodal models are trained separately, merging their extracted features during the multimodal fusion process could potentially introduce noise (Jiang et al., 2023).

Comparing with few-shot baselines. The effectiveness of the proposed CMA is evaluated in comparison with the latest prompt-based few-shot models. CMA outperforms the best few-shot baseline, P&A, with a 3.3% improvement in average accuracy, showing that using unimodal features to assist multimodal probing without prompting the pre-trained language model could also benefit the FND task.

While P&A demonstrates performance on par with CMA, it requires the pre-calculation of a news proximity graph. However, such social context data may not always be accessible, particularly in datasets not sourced from Twitter, like Weibo. After analyzing PET and KPT, it is evident that these methods yield comparable outcomes, likely due to variations introduced by the manually crafted verbalizers used in prompting. This underscores the significance of hand-designed discrete templates in prompt-based learning. Concurrently, M-SAMPLE, a multimodal adaptation of KPL, demonstrates superior performance, suggesting that incorporating image modality can significantly enhance FND effectiveness.

Additionally, we compared CMA with two recent few-shot FND approaches, AMPLE and MPL. For PolitiFact and GossipCop, the $2, 8, 16$-shot accuracies were directly taken from their published papers, while

Table 3
Ablation experiments of CMA.

Method	PolitiFact				GossipCop				Weibo				AVG
	2	8	16	32	2	8	16	32	2	8	16	32	
CMA	73.5	75.8	82.5	87.3	71.9	69.0	71.7	77.0	74.5	69.9	73.8	76.5	75.3
-cross	67.6	76.7	81.2	84.0	71.8	71.8	71.6	71.1	58.4	65.2	68.4	75.2	71.6
-meta	72.2	74.1	74.7	78.4	49.0	53.2	56.8	56.1	50.0	50.9	57.4	61.7	61.2
-img	59.6	61.7	68.5	71.4	48.3	48.4	54.3	56.1	46.9	47.3	50.4	52.1	55.4
-txt	39.0	37.4	45.6	52.1	41.3	43.3	45.1	47.6	39.1	39.3	39.4	45.1	42.9

#-cross denotes the cross-attention is removed from the CMA. -meta means the meta-linear layer is removed. -img means the image features are removed and only text features are used. -txt denotes the text features are removed and only image features are used.

the 32-shot setting for AMPLE is re-constructed based on the same implementation, as described in Section 4.2. Consequently, AMPLE and MPL demonstrate comparable performance; however, the inclusion of emotional elements (e.g., sentiment and subjectivity) in AMPLE may contribute to its slightly higher average accuracy compared to MPL.

5. Analysis

5.1. Distributional diversity analysis

We conducted a Kolmogorov–Smirnov (KS) test to analyze the distributional diversity. The results indicate that the original and augmented instances belong to different distributions, with a test statistic of 0.8763 and a p-value of 3.002e−59. Given that the p-value is less than 0.05, the KS test rejects the null hypothesis, indicating that the sample data does not originate from the same distribution.

To obtain these results, we employed the ks_2samp function from SciPy,[1] calculating the test statistic and p-value for CLIP-encoded original and cross-modal augmented instances across all shots and seeds, then averaging the results.

5.2. Ablation study

We investigate the impact of key components in CMA by assessing the framework's performance in a range of complete and partial configurations. In each experiment, CMA is selectively utilized by removing different components, followed by training the framework from scratch. The results are averaged over five random seeds in each shot, and indicate the performance decay of CMA in the absence of each component in most configurations, underscoring the significance of each key module within CMA, as shown in Table 3.

Specifically, removing the cross-attention from the CMA (i.e., **-cross**) results in a slight decrease in accuracy, showing that the cross-attended features from text and image capture semantic correlations and contribute to improved performance. Further removal of the meta-linear layer from the CMA (i.e., **-meta**) transforms the model into a standard n-shot classification, where it simply classifies concatenated multimodal features. This leads to a significant decrease in accuracy, emphasizing the importance of jointly updating all modality-specific weights in a meta-linear classifier for cross-modal adaptation and accuracy improvement. The meta-linear layer integrates modality-specific features, resembling an ensemble that transforms n-shot classification into a more robust $(n \times z)$-shot problem, enhancing cross-modal adaptation in few-shot classification.

Additionally, experiments are performed by excluding either the image features (**-img**) or the text features (**-txt**), relying solely on the remaining modality for classification. Such setups led to additional reductions in accuracy, underscoring the comparative importance of text over image features in FND. This highlights the complexities in multimodal FND tasks, where the spatial discrepancies between visual and textual semantics tend to be more subtle than in broader multimodal datasets.

[1] https://docs.scipy.org

5.3. Stablility test

Given the selection of few-shot examples can significantly affect the model performance, we assess the stability of the CMA and other prompt-based baselines by measuring the standard deviation of accuracies in the few-shot settings, as shown in Fig. 3.

Overall, the standard deviation for all models decreases in tandem with an increase in the number of n-shot settings, underscoring the importance of augmenting training examples in few-shot scenarios. This augmentation can be further observed that the standard deviation of the CMA tends to be the most stable among the few-shot approaches, indicating that the ensemble of unimodal features in the meta-linear layer can enhance the robustness of multimodal fusion in classification. Additionally, the GossipCop dataset exhibits greater instability compared to the PolitiFact dataset. This instability may be attributed to the semantic complexity in GossipCop, which is responsible for the lower accuracy across all models.

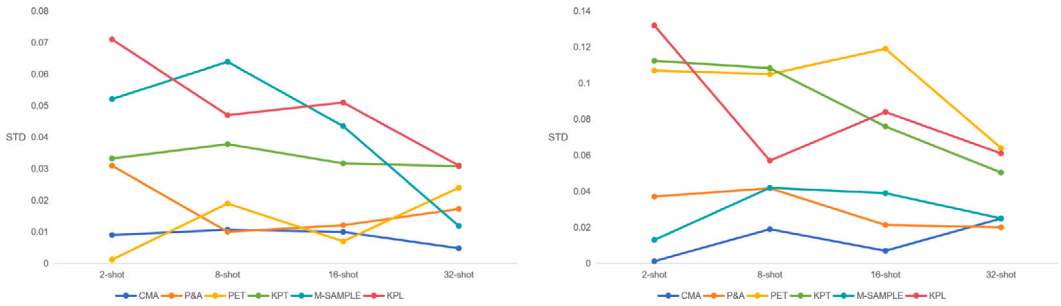
5.4. Model efficiency

Given the CMA achieves the best performance with a surprisingly simple augmentation, we further explore its efficiency in comparison to other baseline models. Table 4 showcases a comparison of the accuracies and epoch times between baselines and the CMA. The average accuracy of each model is determined in a 16-shot setting as shown in Table 2, along with the recording of average epoch times for each model. All experiments are tested with batch size 32 on a single RTX 4090 GPU in the GossipCop dataset for a fair comparison.

Among unimodal models, dEFEND and LDA-HAN exhibit comparable accuracy and epoch times, attributed to their analogous hierarchical architectural design. While FT-RoBERTa exceeds the performance of various unimodal (e.g., 18% higher than dEFEND) and multimodal methods (e.g., 6.9% higher than CAFE), it requires modifying a significant number of trainable parameters, thus extending epoch durations (on average, four minutes per epoch) relative to other unimodal baselines.

In the multimodal models, SAFE yields the lengthiest epoch durations owing to its prerequisite for independently pre-generating image descriptions. Although Spotfake achieves the fastest epoch duration due to its simple concatenation of the image and text features from the BERT and VGG respectively, it achieves the worst performance compared with other models. CAFE achieves the best multimodal FND outcomes by integrating a degree of ambiguity in the similarity across text and image features, albeit at the cost of marginally increased model complexity and consequently, slightly extended epoch durations.

All few-shot baselines demonstrate significant improvements over both unimodal and multimodal counterparts, indicating the suboptimality of traditional methods in contexts with limited annotated data. Specifically, the integration of external knowledge into the prompt-tuning phase by both KPL and KPT results in comparable epoch durations. However, KPL's design of an FND-specific prompt may underlie its superior performance over KPT. PET records the lengthiest epoch duration among the few-shot baselines, potentially due to the repeated fine-tuning of the PLM for reconfiguring input examples with the task



(a) Standard deviation comparisons in the PolitiFact.

(b) Standard deviation comparisons in the GossipCop.

Fig. 3. The standard deviations of accuracies for both PolitiFact and GossipCop datasets among the few-shot baselines and the proposed CMA.

Table 4
Comparisons of model efficiency.

Model	Accuracy	Time	Gain
dEFEND	40.9	2 min	0
LDA-HAN	38.7	2 min	−2.2
FT-RoBERTa	58.9	4 min	+18.0
SAFE	41.1	7 min	+0.2
Spotfake	33.9	2 min	−7.0
CAFE	52.0	3 min	+11.1
KPL	57.5	3 min	+16.6
M-SAMPLE	58.1	5 min	+17.2
KPT	54.3	3 min	+13.4
PET	69.9	6 min	+29.0
P&A	71.5	2 min	+30.6
CMA	74.1	<1 min	+33.2

#Both Accuracy (%) and Time represent averages derived from five random seeds. Times displayed in green signify an average duration of less than 3 min, whereas those in red indicate an average exceeding 3 min. Gain denotes notable improvements in accuracy relative to the dEFEND model.

Table 5
Domain shift performance comparison.

Method	Poli→Goss				Goss→Poli				AVG
	2	8	16	32	2	8	16	32	
KPT	40.1	31.7	31.4	31.1	**56.3**	**55.3**	54.1	55.8	44.5
PET	51.0	51.3	51.5	51.6	53.1	54.1	**54.5**	54.1	52.6
P&A	**53.2**	53.4	53.2	54.5	50.1	50.4	50.3	50.5	51.9
CMA	48.7	**53.5**	**56.1**	**58.6**	51.4	**55.3**	53.0	**55.9**	**54.1**

Poli→Goss refers to utilze few-shot samples from the Politifact as training and the Gossipcop for testing. Goss→Poli denotes the Gossipcop is utilized as training set and the Politifact is the test set. Bold and Underline denote the best and the second best accuracy (%) in that n-shot setting. AVG is the mean accuracy across all n-shot settings.

description. P&A not only achieves the second-best performance but also the second-shortest epoch durations, benefiting from the integration of user engagements. However, it incorporates an external alignment module to correlate user engagement with the PLM's predictions, consequently increasing epoch times relative to CMA. Finally, CMA is more efficient and precise as it avoids the need for extensive parameter fine-tuning and does not depend on intensive image augmentation processes. Additionally, the inclusion of linear probing layers atop the image and text features presents a more streamlined approach than extensive fine-tuning and precise-crafted complex model designs.

5.5. Domain shift analysis

Real-world fake news demonstrates significant distribution discrepancies, which is also referred to as domain shift (Zhu et al., 2022). Consequently, automatic FND methods are required to rapidly adapt to emerging topics by using limited resources.

To address this, we investigate the cross-domain capability of the proposed CMA against three strong few-shot FND baselines (i.e., P&A, PET and KPT). Considering Politifact's focus on political news using formal language and Gossipcop's emphasis on entertainment and celebrity narratives in a more casual tone, we first utilize Politifact for training and Gossipcop for testing, later inverting this arrangement.

The outcomes following domain shift are presented in Table 5. Notably,while the CMA model records the highest average accuracy among the few-shot baselines, the performance of each model markedly differs from that observed in the comparison experiments (as shown in Table 2). For example, KPT exhibits the strongest performance in both 2- and 8-shot scenarios in Goss→Poli. PET and P%A also achieve the highest performance in Goss→Poli and Poli→Goss respectively, highlighting the disparity between present few-shot FND methodologies and their adaptability to domain adaptation.

5.6. Feature visualization

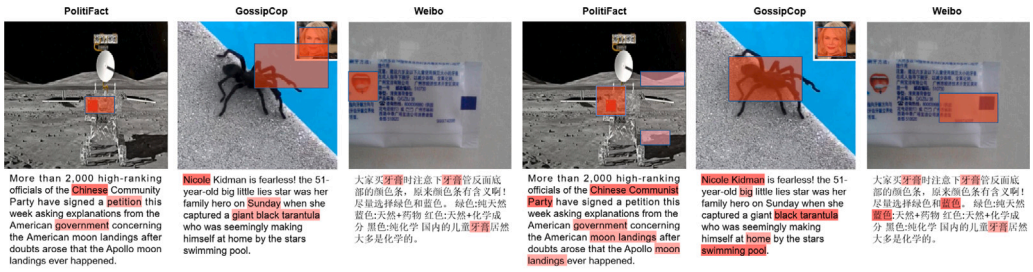
At last, we present a visual comparison of the features extracted by M-SAMPLE and CMA, both of which are multimodal few-shot approaches. This involves the visualization of multimodal features alongside an assessment of their semantic correlations. For each dataset, a specific sample is chosen, with the corresponding multimodal features depicted in Fig. 4.

Observations indicate that: (1) CMA can capture more consistent features from the image-text pair of fake news than those of M-SAMPLE. For example, although both M-SAMPLE and CMA successfully correlate the flag in the image with the word "Chinese" in the text, CMA can also identify the semantic meaning of "moon landing" between the text and image in the PolitiFact example; (2) The proposed CMA is more accurate in capturing important features from the image than M-SAMPLE. For example, although both models can identify the person "Nicole Kidman" and "black tarantula" in both the text and the image in the GossipCop example, the image region of the tarantula slightly overlaps with that of Nicole Kidman provided by M-SAMPLE. This is even more obvious in the Weibo example, as CMA successfully captures the "blue" color bar in the toothpaste, but M-SAMPLE fails to do so.

6. Conclusion and future work

This paper presented Cross-Modal Augmentation (CMA), a novel approach to enhance few-shot multimodal fake news detection by leveraging unimodal features to improve multimodal fusion. CMA utilizes a pre-trained multimodal model for unimodal feature extraction and transforms n-shot classification into a robust (n × z)-shot problem. With simple linear classifiers, CMA outperforms several FND baselines on three datasets under few-shot settings while demonstrating superior efficiency compared to existing methods.

We acknowledge several limitations in this study: (1) CMA's few-shot capabilities were evaluated exclusively using CLIP. Future research will explore how other multimodal models affect its performance. (2) In some datasets, a single text may correspond to multiple images. To address this, cosine similarity was employed to select images from among



(a) Feature visualization from M-SAMPLE. (b) Feature visualization from CMA.

Fig. 4. Feature visualization comparisons between M-SAMPLE and CMA. English translation of the Weibo example: "*When you buy toothpaste, pay attention to the color bar on the bottom of the toothpaste tube, the color bar has meaning! Try to choose greens and blues. Green: natural, blue: natural + medicine, Red: natural + chemical composition, Black: pure chemical. Surprisingly, most children's toothpaste brands on the domestic market contain chemical ingredients.*".

the available options. However, this approach may lead to performance variations depending on the text-image pairing strategy utilized. (3) CMA demonstrates limited performance under domain shift scenarios, leaving room for future improvements through the incorporation of knowledge distillation or domain adaptation techniques.

CRediT authorship contribution statement

Ye Jiang: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Taihang Wang:** Writing – review & editing, Methodology. **Xiaoman Xu:** Writing – review & editing, Data curation. **Yimin Wang:** Writing – review & editing, Methodology, Funding acquisition. **Xingyi Song:** Writing – review & editing, Supervision. **Diana Maynard:** Writing – review & editing, Supervision, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is funded by the Natural Science Foundation of Shandong Province under grant ZR2023QF151 and the Natural Science Foundation of China under grant 12303103.

Data availability

Data will be made available on request.

References

Bahad, P., Saxena, P., Kamal, R., 2019. Fake news detection using bi-directional LSTM-recurrent neural network. Procedia Comput. Sci. 165, 74–82.
Castillo, C., Mendoza, M., Poblete, B., 2011. Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web. pp. 675–684.
Chen, Y., Li, D., Zhang, P., Sui, J., Lv, Q., Tun, L., Shang, L., 2022. Cross-modal ambiguity learning for multimodal fake news detection. In: Proceedings of the ACM Web Conference 2022. pp. 2897–2905.
Conroy, N.K., Rubin, V.L., Chen, Y., 2015. Automatic deception detection: Methods for finding fake news. Proc. Assoc. Inform. Sci. Technol. 52 (1), 1–4.
Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H.-T., Sun, M., 2021. OpenPrompt: An open-source framework for prompt-learning. arXiv preprint arXiv:2111.01998.
Fedoryszak, M., Frederick, B., Rajaram, V., Zhong, C., 2019. Real-time event detection on social data streams. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2774–2782.
Gao, T., Fisch, A., Chen, D., 2021. Making pre-trained language models better few-shot learners. In: Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. ACL-IJCNLP 2021, Association for Computational Linguistics (ACL), pp. 3816–3830.
Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y., 2023. Clip-adapter: Better vision-language models with feature adapters. Int. J. Comput. Vis. 1–15.
Gao, Y., Liu, J., Xu, Z., Zhang, J., Li, K., Ji, R., Shen, C., 2022. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. In: Advances in Neural Information Processing Systems, vol. 35, pp. 35959–35970.
Geeng, C., Yee, S., Roesner, F., 2020. Fake news on facebook and Twitter: Investigating how people (don't) investigate. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–14.
Ghanem, B., Rosso, P., Rangel, F., 2020. An emotional analysis of false information in social media and news articles. ACM Trans. Internet Technol. (TOIT) 20 (2), 1–18.
Girdhar, R., Ramanan, D., 2017. Attentional pooling for action recognition. In: Advances in Neural Information Processing Systems, vol. 30.
Guo, Q., Kang, Z., Tian, L., Chen, Z., 2023. TieFake: Title-text similarity and emotion-aware fake news detection. arXiv preprint arXiv:2304.09421.
Han, C., Fan, Z., Zhang, D., Qiu, M., Gao, M., Zhou, A., 2021. Meta-learning adversarial domain adaptation network for few-shot text classification. In: Findings of the Association for Computational Linguistics. ACL-IJCNLP 2021, pp. 1664–1673.
He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009.
He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738.
Hou, D., Gao, C., Wang, Z., Li, X., Li, X., 2024. Random full-order-coverage based rapid source localization with limited observations for large-scale networks. IEEE Trans. Netw. Sci. Eng..
Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S., 2019. Parameter-efficient transfer learning for NLP. In: International Conference on Machine Learning. PMLR, pp. 2790–2799.
Hu, S., Ding, N., Wang, H., Liu, Z., Wang, J., Li, J., Wu, W., Sun, M., 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2225–2240.
Hu, W., Wang, Y., Jia, Y., Liao, Q., Zhou, B., 2024. A multi-modal prompt learning framework for early detection of fake news. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 18, pp. 651–662.
Jiang, Y., 2023. Team QUST at SemEval-2023 task 3: A comprehensive study of monolingual and multilingual approaches for detecting online news genre, framing and Persuasion techniques. In: Proceedings of the 17th International Workshop on Semantic Evaluation. SemEval-2023, Association for Computational Linguistics, Toronto, Canada, pp. 300–306.
Jiang, G., Liu, S., Zhao, Y., Sun, Y., Zhang, M., 2022. Fake news detection via knowledgeable prompt learning. Inf. Process. Manage. 59 (5), 103029.
Jiang, Y., Song, X., Scarton, C., Aker, A., Bontcheva, K., 2021. Categorising fine-to-coarse grained misinformation: An empirical study of COVID-19 infodemic. arXiv preprint arXiv:2106.11702.
Jiang, Y., Wang, Y., Song, X., Maynard, D., 2020. Comparing topic-aware neural networks for bias detection of news. In: ECAI 2020. IOS Press, pp. 2054–2061.
Jiang, Y., Yu, X., Wang, Y., Xu, X., Song, X., Maynard, D., 2023. Similarity-aware multimodal prompt learning for fake news detection. Inform. Sci. 647, 119446.
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.
Lao, A., Shi, C., Yang, Y., 2021. Rumor detection with field of linear and non-linear propagation. In: Proceedings of the Web Conference 2021. pp. 3178–3187.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521 (7553), 436–444.
Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H., 2021. Align before fuse: Vision and language representation learning with momentum distillation. In: Advances in Neural Information Processing Systems, vol. 34, pp. 9694–9705.

Liu, X., Li, Q., Nourbakhsh, A., Fang, R., Thomas, M., Anderson, K., Kociuba, R., Vedder, M., Pomerville, S., Wudali, R., et al., 2016. Reuters tracer: A large scale system of detecting & verifying real-time news events from Twitter. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 207–216.

Long, Y., Lu, Q., Xiang, R., Li, M., Huang, C.-R., 2017. Fake news detection through multi-perspective speaker profiles. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 252–256.

Meltzoff, A.N., Borton, R.W., 1979. Intermodal matching by human neonates. Nature 282 (5737), 403–404.

Motiian, S., Jones, Q., Iranmanesh, S., Doretto, G., 2017. Few-shot adversarial domain adaptation. In: Advances in Neural Information Processing Systems, vol. 30.

Nanay, B., 2018. Multimodal mental imagery. Cortex 105, 125–134.

Phan, H.T., Nguyen, N.T., Hwang, D., 2023. Fake news detection: A survey of graph neural network methods. Appl. Soft Comput. 110235.

Qu, Z., Meng, Y., Muhammad, G., Tiwari, P., 2024. QMFND: A quantum multimodal fusion-based fake news detection model for social media. Inf. Fusion 104, 102172.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y., 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2931–2937.

Rubinstein, R.Y., Kroese, D.P., 2004. The Cross-Entropy Method: a Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning, vol. 133, Springer.

Schick, T., Schütze, H., 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 255–269.

Schölkopf, B., Herbrich, R., Smola, A.J., 2001. A generalized representer theorem. In: International Conference on Computational Learning Theory. Springer, pp. 416–426.

Schwartz, E., Karlinsky, L., Feris, R., Giryes, R., Bronstein, A., 2022. Baby steps towards few-shot learning with multiple semantics. Pattern Recognit. Lett. 160, 142–147.

Sharaf, A., Awadalla, H.H., Daumé III, H., 2020. Meta-learning for few-shot NMT adaptation. In: Proceedings of the Fourth Workshop on Neural Generation and Translation. pp. 43–53.

Shu, K., Cui, L., Wang, S., Lee, D., Liu, H., 2019. dEFEND: Explainable fake news detection. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '19, Association for Computing Machinery, New York, NY, USA, pp. 395–405. http://dx.doi.org/10.1145/3292500.3330935.

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.

Singhal, S., Pandey, T., Mrig, S., Shah, R.R., Kumaraguru, P., 2022. Leveraging intra and inter modality relationship for multimodal fake news detection. In: Companion Proceedings of the Web Conference 2022. pp. 726–734.

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. BigMM, IEEE, pp. 39–47.

Snell, J., Swersky, K., Zemel, R., 2017. Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems, vol. 30.

Song, X., Petrak, J., Jiang, Y., Singh, I., Maynard, D., Bontcheva, K., 2021. Classification aware neural topic model for COVID-19 disinformation categorisation. PLoS One 16 (2), e0247086.

Sridhar, S., Sanagavarapu, S., 2021. Fake news detection and analysis using multitask learning with BiLSTM CapsNet model. In: 2021 11th International Conference on Cloud Computing, Data Science & Engineering. Confluence, IEEE, pp. 905–911.

Tabibian, B., Valera, I., Farajtabar, M., Song, L., Schölkopf, B., Gomez-Rodriguez, M., 2017. Distilling information reliability and source trustworthiness from digital traces. In: Proceedings of the 26th International Conference on World Wide Web. pp. 847–855.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al., 2016. Matching networks for one shot learning. In: Advances in Neural Information Processing Systems, vol. 29.

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.

Wang, T., Tian, J., Li, X., Xu, X., Jiang, Y., 2023a. Ensemble pre-trained multimodal models for image-text retrieval in the NewsImages MediaEval. NewsImages in MediaEval 2023.

Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M., 2020. Generalizing from a few examples: A survey on few-shot learning. ACM Comput. Surv. (csur) 53 (3), 1–34.

Wang, L., Zhang, C., Xu, H., Xu, Y., Xu, X., Wang, S., 2023b. Cross-modal contrastive learning for multimodal fake news detection. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5696–5704.

Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al., 2022. Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7959–7971.

Wu, J., Li, S., Deng, A., Xiong, M., Hooi, B., 2023a. Prompt-and-align: Prompt-based social alignment for few-shot fake news detection. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 2726–2736.

Wu, L., Long, Y., Gao, C., Wang, Z., Zhang, Y., 2023b. MFIR: Multimodal fusion and inconsistency reasoning for explainable fake news detection. Inf. Fusion 100, 101944.

Wu, Y., Zhan, P., Zhang, Y., Wang, L., Xu, Z., 2021. Multimodal fusion with co-attention networks for fake news detection. In: Findings of the Association for Computational Linguistics. ACL-IJCNLP 2021, pp. 2560–2569.

Xu, X., Li, X., Wang, T., Jiang, Y., 2024. AMPLE: Emotion-aware multimodal fusion prompt learning for fake news detection. arXiv preprint arXiv:2410.15591.

Yang, A., Pan, J., Lin, J., Men, R., Zhang, Y., Zhou, J., Zhou, C., 2022. Chinese clip: Contrastive vision-language pretraining in chinese. arXiv preprint arXiv:2211.01335.

Yin, S., Zhu, P., Wu, L., Gao, C., Wang, Z., 2024. GAMC: an unsupervised method for fake news detection using graph autoencoder with masking. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, (no. 1), pp. 347–355.

Yue, Z., Zeng, H., Zhang, Y., Shang, L., Wang, D., 2023. MetaAdapt: Domain adaptive few-shot misinformation detection via meta learning. In: Association for Computational Linguistics. pp. 5223–5239.

Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H., 2021a. Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930.

Zhang, Q., Huang, H., Liang, S., Meng, Z., Yilmaz, E., 2021b. Learning to detect few-shot-few-clue misinformation. arXiv preprint arXiv:2108.03805.

Zhang, H., Koh, J.Y., Baldridge, J., Lee, H., Yang, Y., 2021c. Cross-modal contrastive learning for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 833–842.

Zhang, H., Zhang, P., Hu, X., Chen, Y.-C., Li, L., Dai, X., Wang, L., Yuan, L., Hwang, J.-N., Gao, J., 2022. Glipv2: Unifying localization and vision-language understanding. Adv. Neural Inf. Process. Syst. 35, 36067–36080.

Zhao, A., Ding, M., Lu, Z., Xiang, T., Niu, Y., Guan, J., Wen, J.-R., 2021. Domain-adaptive few-shot learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1390–1399.

Zhou, X., Wu, J., Zafarani, R., 2020. Similarity-aware multi-modal fake news detection. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, pp. 354–367.

Zhou, Y., Yang, Y., Ying, Q., Qian, Z., Zhang, X., 2023. Multimodal fake news detection via clip-guided learning. In: 2023 IEEE International Conference on Multimedia and Expo. ICME, IEEE, pp. 2825–2830.

Zhu, Y., Sheng, Q., Cao, J., Li, S., Wang, D., Zhuang, F., 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2120–2125.