SEMANTIC DISTILLATION AND STRUCTURAL ALIGNMENT NETWORK FOR FAKE NEWS DETECTION

Shangdong Liu, Xiaofan Yue, Fei Wu, Jing Sun, Yujian Feng, Yimu Ji

Nanjing University of Posts and Telecommunications

ABSTRACT

In recent years, the rapid proliferation of multi-modal fake news has posed potential harm across various sectors of society, making the detection of multi-modal fake news crucial. Most existing methods can not effectively reduce the redundant information and preserve both semantic and structural information. To address these problems, this paper proposes a semantic distillation and structural alignment (SDSA) network. We design an semantic distillation module for modality-specific features to preserve task-relevant semantic information and eliminate redundant information. Then, we propose a triple similarity alignment module to preserve structural information. Specifically, intra-modal similarity alignment mines intra-modal consistency by preserving the neighborhood structure within each modality, inter-modal similarity alignment explores cross-modality consistency by bringing the cross-modality feature neighborhood structures, and joint similarity alignment aims to preserve the structural information of fused features. Experiments conducted on two widely used fake news datasets demonstrate that the SDSA method outperforms state-of-the-art approaches.

Index Terms— Information bottleneck, structural similarity alignment, fake news detection

1. INTRODUCTION

Nowadays, a piece of fake news may consist of various modalities, which presents significant challenges for fake news detection tasks [1]. Therefore, establishing an effective multi-modal fake news detection system is of paramount importance [2, 3]. Recently, various methods have been proposed to utilize multi-modal information to improve the accuracy of fake news detection. EANN [4] introduces an event discriminator to reduce event differences. SpotFake [5] performs true/fake news classification directly without relying on subtasks. BDANN [6] employs domain-adaptive neural networks with a domain classifier to reduce event disparities. MVAE [7] introduces a variational autoencoder to obtain multi-modal features. MEAN [8] introduces adversarial mechanisms in event discriminator and classifier. CAFE [9] proposes a cross-modal ambiguity learning problem for multi-modal fake news detection. MRML [10] obtains semantic relationships within modalities through metric learning. Another set of methods, as described in references [11–13], employ semantic similarity between modalities to aid in feature learning. For instance, CAFE [11] determines whether text and images exhibit similarity, categorizing most non-matching cases as fake news. However, these fake news detection methods cannot effectively remove the redundant information [14–16] and ignore the structural information [17] in the news.

To solve these problems, in this paper, we introduce a method called the semantic distillation and structural alignment network (SDSA) for fake news detection. Specifically, we first employ a semantic distillation encoder based on the information bottleneck theory to encode input features for each modality, thereby retaining modality-specific semantic information relevant to the fake news detection task. Second, we propose a triple similarity alignment module including intra-modal similarity alignment, inter-modal similarity alignment, and joint similarity alignment to preserve structural information. Intra-modal similarity alignment utilizes the similarity matrix of samples within each modality for similarity alignment, *i.e.*, maintaining the consistency of the similarity matrix before and after encoding. Inter-modal similarity alignment uses the similarity relationship between the two modal similarity matrices after encoding to mine the consistency between the modalities, thereby reducing the differences between the modalities. Joint similarity alignment pulls the low-dimensional and high-dimensional similarity matrices before and after mapping to maintain the structural consistency of cross-modal fusion features.

The main contributions of our study can be summarized as follows:

(1) We employ the information bottleneck theory to distill the input features, utilizing structural alignment to preserve structural information.

(2) We propose a triple similarity alignment module to mine intra-modal consistency, inter-modal consistency, and the structural information of the fusion feature.

(3) Experimental results demonstrate that our method out-

This work is supported by National Natural Science Foundation of China (No. 62076139), 1311 Talent Program of Nanjing University of Posts and Telecommunications.

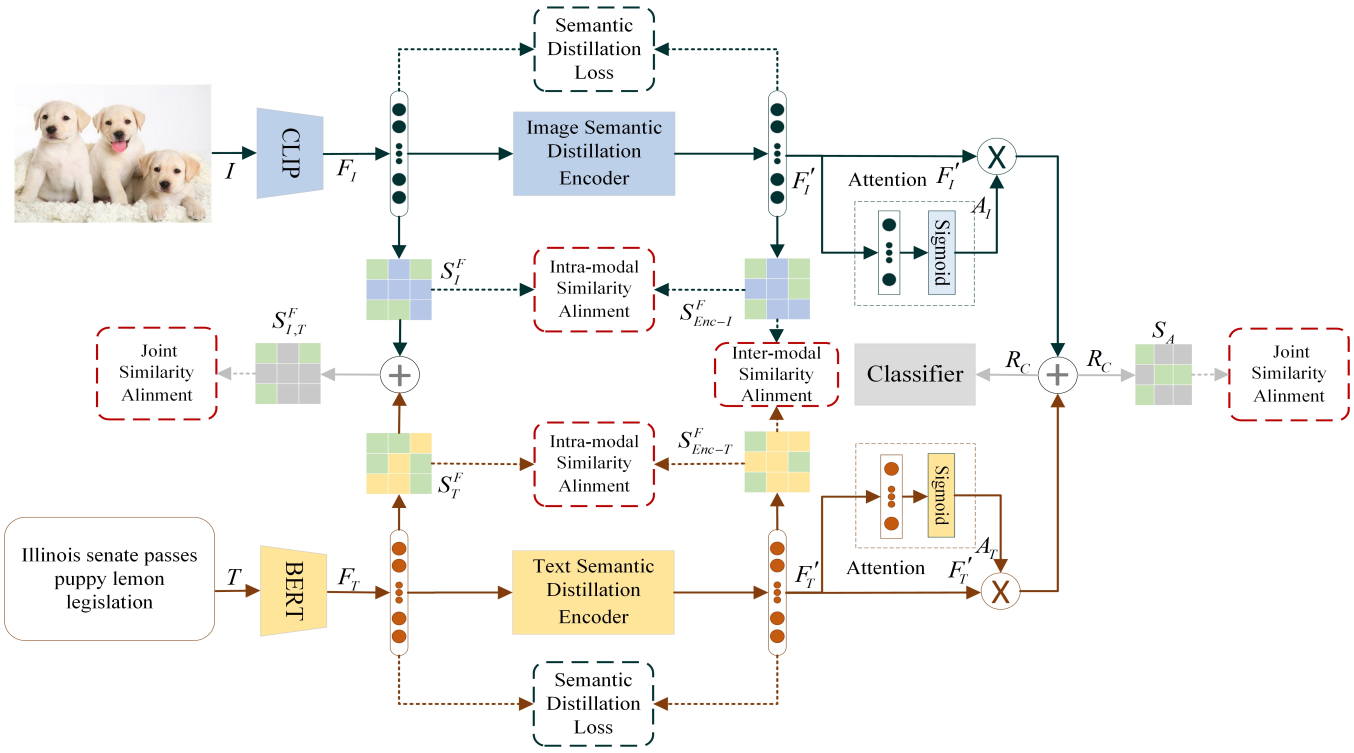


Fig. 1: The overall framework of our SDSA approach

performs relevant state-of-the-art approaches.

2. METHODOLOGY

2.1. Problem Definition

Let $O = \{o_n, l_n\}_{n=1}^{N}$ represent a set of image-text pairs, where each instance $o_n = (i_n, t_n)$ contains an image feature vector $i_n \in \mathbb{R}^{d_i}$ and a text feature vector $t_n \in \mathbb{R}^{d_t}$, d_i, d_t representing the dimensions of the image and text features, respectively. The feature matrices for the image and text modalities are denoted as $I = [i_1, \cdots, i_N]$ and $T = [t_1, \cdots, t_N]$, respectively. l_n represents the semantic label vector associated with o_n, and the label vectors are denoted as $L = [l_1, \cdots, l_N]$. Fig. 1 illustrates the overall architecture of our SDSA approach.

Given input I and T,$* \in I, T$, to obtain modality-specific features, we use two modality-specific base feature extractors CLIP [18] and BERT [19] to learn the base feature representations $F_* = F_Q(*; \theta_*) = [f_n^*]_{n=1}^{N} \in \mathbb{R}^{d_{fi} \times N}$ for the image and text modalities, where $F_Q(*; \theta_*)$ is the mapping functions for the two modalities. To obtain task-relevant semantic information, we encode the features F_* using an information distillation encoder $F_*^{'} = Enc(F_*; \delta_*) = [f_n^{*'}]_{n=1}^{N} \in \mathbb{E}^{d'_{fi} \times N}$, where $Enc(F_*; \delta_*)$ represents the information distillation encoder. This encoding process is aimed at maximizing the retention of semantic information relevant to the fake news detection task while eliminating redundant information unrelated to the task for the encoded features. We obtain the fused features $R_C = A\left(F_I', F_T'; \theta_A\right)$ by using an attention subnetwork to fuse the encoded image features and text features.

2.2. Model

2.2.1. *Semantic Distillation*

For the fake news detection task, semantic information is composed of two parts: task-relevant semantic information and task-irrelevant semantic information (redundant information). For F_*, we can represent it as follows:

$$F_* = f(y; l) + f(z; y \mid l) \tag{1}$$

where $f(y; l)$ represents task-relevant information and $f(z; y \mid l)$ denotes redundant information [20]. Our objective is to retain only the task-relevant semantic information and eliminate the redundant information from F_*. This implies that $F_*^{'} = f(m; l)$ is a sufficient representation of $f(y; l)$. Thus, our aim is to minimize the difference between $F_*^{'} = f(m; l)$ and $f(y; l)$:

$$\min f(m; l) - f(y; l) \tag{2}$$

For ease of computation, we equivalently transform Equation (2) into minimizing the conditional entropy $H(m; l)$ and $H(x; l)$:

$$\min f(m; l) - f(y; l) \longleftrightarrow \min H(m; l) - H(y; l) \tag{3}$$

where $H(y;l) = -\int p(l)dl \int p(y \mid l) \log p(y \mid l)dy$. If the KL divergence between $f(m;l)$ and $f(y;l)$ is equal to 0, then $f(m;l)$ is a sufficient representation of $f(y;l)$:

$$\mathcal{D}_{KL}[\mathbb{P}_{F'_*}||\mathbb{P}_{F_*}] = 0 \Rightarrow H(m;l) - H(y;l) = 0 \quad (4)$$

where $P_{F_*} = p(y;l)$ and $P_{F'_*} = p(m;l)$ represent the predicted distribution, D_{KL} represents the KL divergence. Therefore, by minimizing the KL divergence between P_{F_*} and $P_{F'_*}$, we can ensure that only semantic information related to the task is retained.

$$\mathcal{L}_{VSD} = \min_{\theta} \mathbb{E}_{F_* \sim E_\theta(F_*|*)} \left[\mathbb{E}_{F'_* \sim E_\phi(F'_*|F_*)} \left[D_{KL} \left[\mathbb{P}_{F*} \| \mathbb{P}_{F'_*} \right] \right] \right] \quad (5)$$

where θ represents the parameters of the information distillation encoder.

2.2.2. Triple Similarity Alignment

To retain the structural information of the original data and reduce cross-modal disparities, we use a similarity matrix to represent the neighborhood structure of the dataset [21], and we propose a triple similarity alignment module to exploit the similarity information within modalities and across modalities' neighborhood structures, which includes intra-modal similarity alignment, inter-modal similarity alignment, and joint similarity alignment.

Specifically, given a batch of input instances F_*, we utilize l_2-norm to normalize F_* to obtain $\hat{F}_*$, and calculate the cosine similarity matrix $S^F_* = \hat{F}_*(\hat{F}_*)^T$ for F_*, which is used to describe the original neighborhood structures in different modalities. Similarly, by normalizing the encoded features F'_* through l_2-norm, we can compute the cosine similarity matrix $S^F_{Enc-*} = \hat{F}'_*(\hat{F}'_*)^T$, which is used to describe the encoded neighborhood structures in different modalities.

These similarity matrices are then utilized to optimize F'_*. Additionally, the cross-modal fused feature R_c is normalized using l_2-norm to obtain $\hat{R}_c$, and we calculate the cosine similarity matrix $S_A = \hat{R}_C(\hat{R}_C)^T$ to describe the neighborhood structure of the fused feature.

Intra-modal Similarity Alignment: To preserve the neighborhood structure from the original features within each modality. We align S^F_* with S^F_{Enc-*}, and calculate the mean squared error losses between S^F_* and S^F_{Enc-*}. By minimizing these intra-modal mean squared error losses, we can retain the intra-modal neighborhood structures as much as possible. We define the intra-modal similarity loss as follows:

$$L_{intra} = \left\| S^F_I - S^F_{Enc-I} \right\|^2 + \left\| S^F_T - S^F_{Enc-T} \right\|^2 \quad (6)$$

Inter-modal Similarity Alignment: To effectively reduce the differences between modalities, we align S^F_{Enc-I} and S^F_{Enc-T}, calculating the mean squared error loss between S^F_{Enc-I} and S^F_{Enc-T}. By minimizing this inter-modal mean squared error loss, we aim to bring the cross-modal feature neighborhood structures closer together, thus diminishing inter-modal disparities. The inter-modal similarity loss function is as follows:

$$L_{intra} = \left\| S^F_{Enc-I} - S^F_{Enc-T} \right\|^2 \quad (7)$$

Joint Similarity Alignment: We utilize the complementary relationships among inter-modal neighborhood structures to construct a joint similarity matrix $S^F_{I,T}$. We align $S^F_{I,T}$ and S_A to enable fused features to preserve neighborhood structures from various modalities. Due to varying levels of semantic richness among different modalities, we introduce a hyperparameter α to adjust the weighting of different modality structural information in the joint similarity matrix. This facilitates better utilization of neighborhood information from different modalities. The construction method is defined as follows:

$$S^F_{I,T} = \alpha S^F_I + (1-\alpha) S^F_T \quad (8)$$

Consequently, we have devised a triple similarity alignment approach, encompassing intra-modal, inter-modal, and joint similarity alignment. We define the overall similarity alignment loss as follows:

$$L_S = L_{intra} + L_{inter} + L_{joint} \quad (9)$$

2.2.3. Classification

We introduce classification loss L_C, which is based on cross-entropy loss, to enhance the performance of real and fake news classification.

Therefore, the total loss of the entire model is:

$$L_{total} = L_C + L_S + L_{ib} \quad (10)$$

3. EXPERIMENT

3.1. Experimental Datasets and Details

We conduct comprehensive experiments on two Weibo [22] and Fakeddit [23] datasets, to validate the effectiveness of our approach. In the Weibo dataset, the training set comprises 7,532 news items, with 3,749 being fake news and 3,783 being non-fake news, the testing set consists of 1,996 news articles. In the Fakeddit dataset, we randomly select 30,000 image-text pairs in the Fakeddit training set as our training set and randomly select 10,000 image-text pairs from the test set as our test set.

The experiments are deployed on an NVIDIA GeForce 1080Ti GPU with PyTorch. The batch size for both datasets is set to 32. We employed a grid search method to fine-tune the hyperparameters in equation (8). The search range for α was [0, 1], with a step size of 0.1. Ultimately, we set $\alpha = 0.4$ for the Weibo dataset and $\alpha = 0.5$ for the Fakeddit dataset. We employed Accuracy and F1-score as evaluation metrics for model performance assessment.

Table 1: Comparison Results of Different Models on Weibo and Fakeddit Datasets

Method	Weibo							Fakeddit						
	Acc	Rumor			Non-rumor			Acc	Rumor			Non-rumor		
		P	R	F1	P	R	F1		P	R	F1	P	R	F1
EANN [KDD-2018]	0.782	0.827	0.697	0.756	0.752	0.863	0.804	0.724	0.727	0.719	0.723	0.722	0.729	0.726
SpotFake [BigMM-2019]	0.892	0.902	0.964	0.932	0.847	0.656	0.739	0.819	0.801	0.848	0.824	0.839	0.790	0.813
BDANN [IJCNN-2020]	0.842	0.830	0.870	0.850	0.850	0.820	0.830	0.812	0.836	0.776	0.805	0.791	0.847	0.818
HMCAN [SIGIR-2021]	0.885	0.920	0.845	0.881	0.856	0.926	0.890	0.881	0.880	0.882	0.881	0.882	0.880	0.881
MEAN [SPR-2022]	0.894	0.900	0.870	0.890	0.890	0.910	0.90	0.910	0.930	0.890	0.910	0.890	0.930	0.910
CAFE [WWW-2022]	0.840	0.855	0.830	0.842	0.825	0.851	0.837	0.912	0.946	0.886	**0.959**	0.878	0.942	0.909
MRML [ICASSP-2023]	0.897	0.898	0.887	0.892	**0.896**	0.905	0.901	0.840	0.819	0.874	0.846	0.865	0.807	0.835
SDSA*	0.898	0.916	0.888	0.902	0.880	0.909	0.894	0.941	0.946	0.936	0.941	0.936	0.947	0.942
SDSA*+S	0.905	0.919	0.899	0.909	0.891	0.912	0.901	0.944	0.952	0.935	0.943	0.936	0.953	0.944
SDSA*+I	0.906	0.924	0.895	0.909	0.887	0.919	0.903	0.950	0.946	**0.954**	0.950	**0.954**	0.946	0.950
SDSA	**0.918**	**0.939**	**0.902**	**0.920**	**0.896**	**0.935**	**0.915**	**0.953**	**0.965**	0.939	0.952	0.940	**0.966**	**0.953**

3.2. Comparison With State-of-The-Art Methods

We compare our approach with state-of-the-art fake news detection methods, including EANN [4], SpotFake [5], BDANN [6], MVAE [7], MEAN [8], CAFE [9] and MRML [10].

The results of baselines and SPSSA are shown in Table 1. On the Weibo dataset, compared with MRML [10], the performance of our approach on Accuracy is 0.918, which significantly exceeds the MRML method by 2.1%. On the Fakeddit dataset, compared with CAFE [9], our approach leads by 4.1% on Accuracy. The leadership of the two data sets proves that our method can effectively reduce information redundancy and preserve both semantic and structural information.

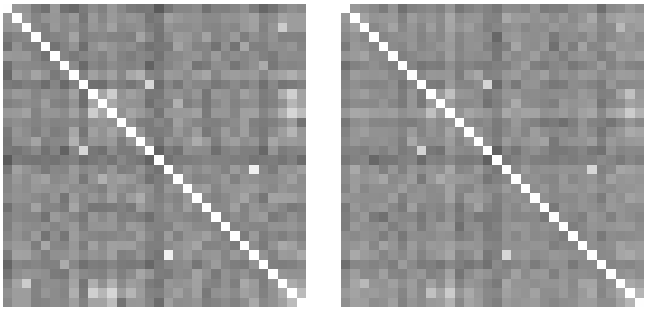
3.3. Ablation Study

Our baseline model comprises modality-specific feature extractors, an encoder with three fully connected layers, an attention network, and a classifier. We refer to this baseline model as SDSA*. SDSA*+S represents the addition of the triple similarity alignment module to the baseline model. SDSA*+I denotes replacing the baseline model's encoder with a semantic distillation encoder. SDSA indicates the simultaneous addition of both the triple similarity module and the semantic distillation encoder to the baseline model. We report the results in the bottom part of Table 1.

From Table 1, it can be observed that when compared to SDSA*, SDSA*+S has shown an improvement of 0.8% and 0.3% on Weibo dataset and the Fakeddit dataset, respectively. SDSA*+I has also improved by 0.9% and 0.9% on the two datasets. SDSA* improved by 2.2% and 1.3% on the two datasets, respectively. These experimental results indicate the effectiveness of our semantic distillation encoder and triple similarity alignment module.

3.4. Visualizations

To observe whether structural information is effectively preserved, we convert the similarity matrix of the dataset into grayscale images. The grayscale values represent the similarity between samples, with higher similarity resulting in brighter shades. The results are shown in Figure 2. By examining the distribution of the grayscale images, it is evident that the neighborhood structure for the samples have been well preserved before and after the mapping.



(a) (b)

Fig. 2: The grayscale images of the adjacency structures before and after mapping. (a) represents the grayscale image of the joint adjacency structure, (b) represents the grayscale image of the fused feature adjacency structure.

4. CONCLUSIONS

In this paper, we introduced a semantic distillation and structural alignment network (SDSA) for fake news detection. We utilize the semantic distillation module to distill features, retaining meaningful semantic information while eliminating redundant information. A triple similarity alignment module is proposed to preserve both structural and semantic information. Compared with the current state-of-the-art methods, our method shows higher accuracy on two widely used fake news detection datasets. Experimental results further validate the effectiveness of the key components of our proposed method.

5. REFERENCES

[1] X. Zhang, X. Chen, and G. Ali A, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, vol. 57, no. 2, pp. 102025, 2020.

[2] S. Mishra, P. Shukla, and R. Agarwal, "Analyzing machine learning enabled fake news detection techniques for diversified datasets," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–18.

[3] L. Hu, S. Wei, Z. Zhao, and B. Wu, "Deep learning for fake news detection: A comprehensive survey," *AI Open*, 2022.

[4] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "Eann: Event adversarial neural networks for multi-modal fake news detection," in *ACM Sigkdd International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 849–857.

[5] S. Singhal, R.R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "Spotfake: A multi-modal framework for fake news detection," in *International Conference on Multimedia Big Data*, 2019, pp. 39–47.

[6] T. Zhang, D. Wang, H. Chen, Z. Zeng, W. Guo, C. Miao, and L. Cui, "Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection," in *International Joint Conference on Neural Networks*, 2020, pp. 1–8.

[7] D. Khattar, J.S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *The world Wide Web Conference*, 2019, pp. 2915–2921.

[8] P. Wei, F. Wu, Y. Sun, H. Zhou, and X. Jing, "Modality and event adversarial networks for multi-modal fake news detection," *IEEE Signal Processing Letters*, vol. 29, pp. 1382–1386, 2022.

[9] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *ACM Web Conference*, 2022, pp. 2897–2905.

[10] L. Peng, S. Jian, D. Li, and S. Shen, "Mrml: Multimodal rumor detection by deep metric learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[11] X. Zhou, J. Wu, and R. Zafarani, "Safe: similarity-aware multi-modal fake news detection (2020)," *Preprint. arXiv*, vol. 200304981, 2020.

[12] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data," *Information Processing & Management*, vol. 58, no. 5, pp. 102610, 2021.

[13] Y. Shao, J. Sun, T. Zhang, Y. Jiang, J. Ma, and J. Li, "Fake news detection based on multi-modal classifier ensemble," in *International Workshop on Multimedia AI against Disinformation*, 2022, pp. 78–86.

[14] X. Tian, Z. Zhang, S. Lin, Y. Qu, Y. Xie, and L. Ma, "Farewell to mutual information: Variational distillation for cross-modal person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1522–1531.

[15] X. Tian, Z. Zhang, C. Wang, W. Zhang, Y. Qu, L. Ma, Z. Wu, Y. Xie, and D. Tao, "Variational distillation for multi-view learning," *arXiv preprint arXiv:2206.09548*, 2022.

[16] W. Xu, J. Wu, Q. Liu, S. Wu, and L. Wang, "Evidence-aware fake news detection with graph neural networks," in *ACM Web Conference*, 2022, pp. 2501–2510.

[17] J. Chen and G. Kou, "Attribute and structure preserving graph contrastive learning," in *AAAI Conference on Artificial Intelligence*, 2023, number 6, pp. 7024–7032.

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

[19] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[20] N. Tishby, F.C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[21] S. Su, Z. Zhong, and C. Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3027–3035.

[22] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *ACM International Conference on Multimedia*, 2017, pp. 795–816.

[23] K. Nakamura, S. Levy, and W. Wang, "r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," *arXiv preprint arXiv:1911.03854*, 2019.