MisD-MoE: A Multimodal Misinformation Detection Framework with Adaptive Feature Selection

Moyang Liu[1] **Kaiying Yan**[2] **Yukun Liu**[3*] **Ruibo Fu**[4]
Zhengqi Wen[5] **Xuefei Liu**[4] **Chenxing Li**[4]
[1]Beihang University
[2]School of Mathematics, Sun Yat-sen University
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences
[4] Institute of Automation, Chinese Academy of Sciences
[5]Beijing National Research Center for Information Science and Technology, Tsinghua University
{moyang_liu}@buaa.edu.cn {yanky}@mail2.sysu.edu.cn {yukunliu927}@gmail.com

Abstract

The rapid growth of social media has led to the widespread dissemination of misinformation across multiple content forms, including text, images, audio, and video. Compared to unimodal misinformation detection, multimodal misinformation detection benefits from the increased availability of information across multiple modalities. However, these additional features may introduce redundancy, where overlapping or irrelevant information is included, potentially disrupting the feature space and consequently impairing the model's performance. To address the issue, we propose a novel framework, Misinformation Detection Mixture of Experts (MisD-MoE), which employs distinct expert models for each modality and incorporates an adaptive feature selection mechanism using top-k gating and Gumbel-Sigmoid. This approach dynamically filters relevant features, reducing redundancy and improving detection accuracy. Extensive experiments on the FakeSV and FVC-2018 datasets demonstrate that MisD-MoE significantly outperforms state-of-the-art methods, with accuracy improvements of 3.45% and 3.71% on the respective datasets compared to baseline models.

1 Introduction

In recent years, with the rapid growth of social media platforms, the spread of misinformation has emerged as a significant global societal issue. On social media, users can rapidly disseminate information through likes, shares, and comments, regardless of its veracity[1]. Such misinformation can not only mislead public perception and influence social opinion[2], but it can also have severe impacts on various domains, including politics, economics, and public health[3].

As a result, accurately detecting misinformation becomes crucial in helping individuals identify and differentiate between authentic and false content[4]. Nowadays, misinformation on social media appears in various forms, such as text, images, audio, and video, making unimodal detection methods insufficient[5]. Text can be manipulated to bypass keyword filters[6], images and videos can be altered, and audio can be synthesized or edited[7]. Thus, relying on unimodal detection often fails to deliver ideal results[8].

To address these challenges, multimodal misinformation detection[9] has emerged. Multimodal detection methods integrate information from multiple sources, such as text, images, audio, and video, utilizing cross-modal feature fusion and correlation analysis to achieve a more comprehensive

*Corresponding Author

4th NeurIPS Efficient Natural Language and Speech Processing Workshop (ENLSP-IV 2024).

identification of misinformation[10]. Song et al.[11] proposed a multimodal misinformation detection framework based on Cross-modal Attention Residual and Multi-channel convolutional neural Network (CARMN). However, these feature fusion methods often encounter the issue of losing information at the shallow layers[12]. In response to this issue, Jing et al.[13] proposed a network that can capture the representational information of each modality at different levels. Nevertheless, most existing multimodal models have neglected certain modalities[14], which hinders their performance improvement in misinformation detection[15]. Qi et al.[16] constructed China's largest fake news short video dataset FakeSV, which includes various contents such as titles, videos, keyframes, audios, metadata, and user comments.

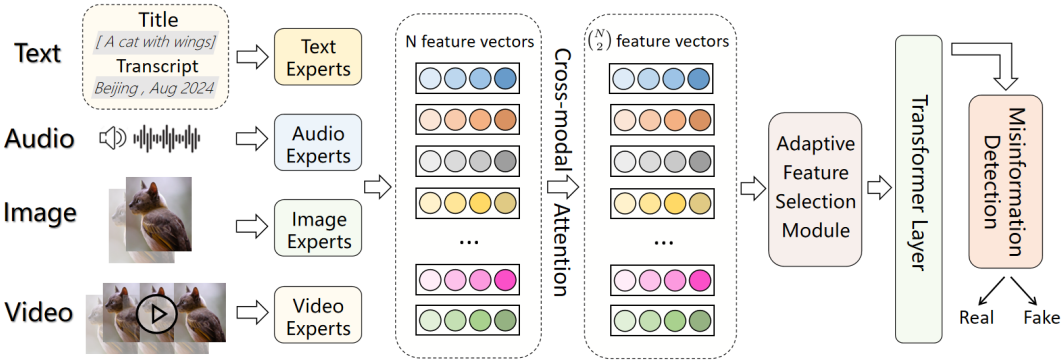

Figure 1: Architecture of the multimodal misinformation detection framework MisD-MoE

Although multimodal misinformation detection has made significant progress, existing frameworks still face numerous challenges[17]. As the number of modalities increases, some modality features may become redundant, failing to complement other modalities and thereby disrupting the feature space[18]. Moreover, the issue of feature alignment between different modalities during the fusion process has become increasingly complex and challenging. Ensuring the consistency and coordination of modality features in fusion remains a pressing issue[19]. Additionally, current frameworks typically employ only a single encoder for each modality, which limits the comprehensive extraction and utilization of multi-level information within each modality[20]. This further hampers the framework's ability to capture complex and diverse information effectively. And simply increasing the number of encoders would only lead to information redundancy, without ensuring the complementarity of the features.

In response to these challenges, we have developed a novel framework for multimodal misinformation detection, called Misinformation Detection Mixture of Experts (MisD-MoE). This framework incorporates an adaptive feature selection mechanism that minimizes feature redundancy during the fusion phase, ensuring the complementarity between modality features. Additionally, by dynamically adjusting the contribution weights of each expert model, our approach effectively captures the unique characteristics of each modality, while also ensuring proper feature alignment and effective cross-modal feature integration. Furthermore, the framework introduces distinct expert models for each modality to fully capture the complete information within each, further improving detection accuracy and robustness.

2 MisD-MoE Framework

2.1 Framework

To address the task of multimodal misinformation detection, we propose a novel framework MisD-MoE, as shown in the Fig. 1.

In this framework, to tackle the widespread occurrence of various types of misinformation encountered in real-world scenarios, we leverage multiple modalities within short video content—including text, audio, image, and video to comprehensively extract relevant information. For each modality, we employ distinct expert models to extract features and integrate the extracted modal information for misinformation detection.

Next, we apply cross-attention[21] between the features extracted by the expert models for each modality pairwise, aiming to capture the complementary information across different modalities. For example, given the audio features $F_a = (a_1, a_2, \ldots, a_n)$ extracted by an expert model for the audio modality and the text features $F_t = (x_1, x_2, \ldots, x_n)$ extracted by an expert model for the text modality, we can calculate the attention weights and compute the attention-weighted sum of audio-enhanced text features and text-enhanced audio features as follows.

$$F_{t \leftarrow a} = \text{Attention}(Q_t, K_a, V_a) = \text{softmax}\left(\frac{Q_a K_t^T}{\sqrt{d_k}}\right) V_t \tag{1}$$

$$F_{a \leftarrow t} = \text{Attention}(Q_a, K_t, V_t) = \text{softmax}\left(\frac{Q_t K_a^T}{\sqrt{d_k}}\right) V_a \tag{2}$$

where d_k is the dimensionality of the feature vectors. Assuming there are N modality experts, performing pairwise cross-attention on the N extracted features would yield $\binom{N}{2}$ modality-enhanced vectors. Subsequently, these modality-enhanced features undergo an adaptive feature selection mechanism. This adaptive feature selection mechanism comprises an attention-based top-k gating filter[22] and a Gumbel-Sigmoid-based[23] dynamic selection mechanism. This approach enables the model to dynamically focus on the most relevant features, improving overall detection performance. Finally, the selected features F_{final} are fed into a transformer layer, and the output of this layer is classified into true or false categories by a classifier:

$$\hat{y} = \text{Classifier}(\text{Transformer}(F_{\text{final}})) \tag{3}$$

Table 1: The experimental results of the multimodal disinformation detection framework using different models and combinations of modalities (text, audio, image, video). E1-E7 represent a series of experiments conducted within the SV-FEND framework, each utilizing different modalities for feature extraction and fusion. E1 and E2 used different audio encoders, VGG and Wav2Vec2.0, respectively. E7 used VGG and Wav2Vec2.0 encoders simultaneously.

	Modality				FakeSV				FVC-2018			
Method	Text	Audio	Image	Video	Acc	F1	Pre	Rec	Acc	F1	Pre	Rec
VGGish+SVM		✓			61.25	61.31	61.24	61.33	58.44	58.61	58.48	58.63
VGG19+Att			✓		68.53	68.51	68.53	68.50	65.79	65.81	65.49	66.08
C3D+Att				✓	70.26	70.24	70.25	70.25	71.81	71.72	71.89	71.85
Bert+Att	✓				74.31	74.35	74.30	74.39	76.37	76.35	76.39	76.33
TikTec	✓	✓	✓	✓	75.07	75.04	75.18	75.07	77.02	73.95	74.24	73.67
FANVN	✓	✓	✓	✓	75.04	75.02	75.11	75.04	85.81	85.32	85.20	85.44
SV-FEND	✓	✓	✓	✓	79.31	79.24	79.62	79.31	84.71	85.37	84.25	86.53
E1	✓	✓			81.55	81.40	81.34	81.73	78.61	78.20	78.21	78.54
E2	✓	✓			82.29	81.03	83.44	81.21	79.77	79.50	79.36	79.43
E3	✓	✓	✓		78.60	78.46	78.40	78.78	84.39	84.01	84.32	84.13
E4	✓	✓		✓	78.60	78.96	78.80	79.57	83.82	84.16	83.27	84.50
E5	✓		✓	✓	79.84	79.12	78.53	78.83	86.12	84.77	85.58	84.25
E6		✓	✓	✓	78.60	78.03	78.21	77.58	81.50	81.57	81.60	81.63
E7	✓	✓	✓	✓	78.12	78.83	78.60	79.33	83.82	83.98	83.72	84.89
MisD-MoE	✓	✓	✓	✓	**82.84**	**82.22**	**83.76**	81.60	**88.44**	**89.02**	**88.56**	**89.62**

2.2 Expert Models

To extract valuable information from each modality, it is crucial to select appropriate expert models based on the current modality.

In the text modality, BERT[24] is a mature and highly versatile text encoder. Therefore, we utilize BERT as the expert model to extract textual features. For the audio modality, we employed two expert models: the CNN-based VGG[25] and the transformer-based Wav2Vec2.0[26] for feature extraction. This choice is motivated by VGG's strength in capturing background audio features, while Wav2Vec2.0[27] excels at extracting semantic information. By leveraging these complementary models, we aim to achieve a comprehensive understanding of the audio content across different levels. To detect the spatiotemporal and multi-granularity information in the video modality, we used a pre-trained C3D[28] model to extract motion features. For the image modality, we extracted a specific number of frames from each video and fed them into a pre-trained VGG19[25] model to learn static visual features.

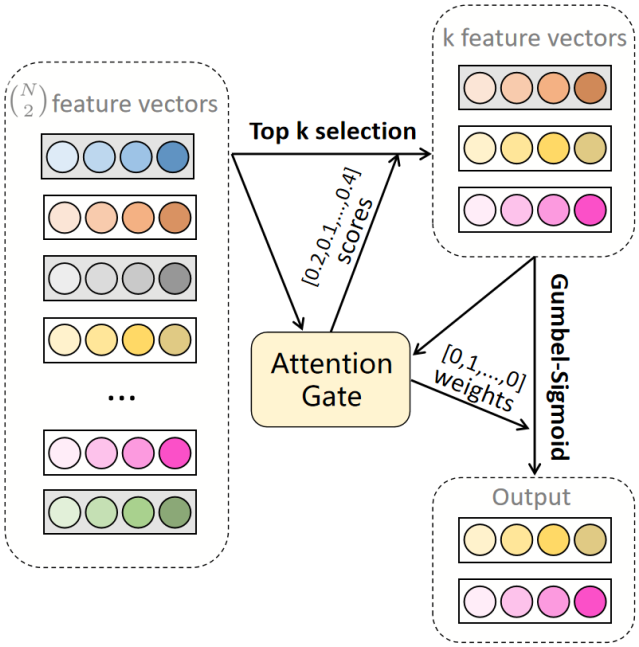

Figure 2: Architecture of adaptive feature selection mechanism. The input set of $\binom{N}{2}$ feature vectors is processed through an attention gate, from which the top k vectors are selected based on their scores. These k feature vectors are then passed through another attention gate, where the Gumbel-Sigmoid function is applied to generate the final output.

3 Adaptive Feature Selection Mechanism

3.1 Gumbel-Sigmoid Technique

Gumbel-Sigmoid is a stochastic technique designed for differentiable sampling from discrete distributions. It merges the Gumbel-Max trick, which is commonly used for categorical sampling, with the Sigmoid function to enable continuous and differentiable approximations of binary decisions. This makes it particularly well-suited for tasks such as feature selection, where we need to decide whether to retain or discard a feature in a way that still allows backpropagation for model optimization during training.

In this mechanism, a Gumbel noise is sampled from a Gumbel distribution, which is known for modeling the maximum of a set of variables. By applying the noise to a logit (log-odds), followed by a temperature-scaled Sigmoid function, the mechanism approximates a discrete decision in a differentiable manner. Mathematically, the Gumbel noise is generated as:

$$g_i = -\log(-\log(U_i)), \quad U_i \sim \text{Uniform}(0, 1) \tag{4}$$

Then, the final decision for feature selection is computed as:

$$z_i = \sigma\left(\frac{1}{\tau}\left(\log(\alpha_i) + g_i\right)\right) \tag{5}$$

where σ is the Sigmoid function, and τ is the temperature parameter. By adjusting τ, we can fine-tune the balance between exploration and exploitation in feature selection.

3.2 Adaptive Feature Selection Mechanism

Compared to unimodal misinformation detection, multimodal misinformation detection benefits from an increased amount of available information due to the addition of multiple modalities. However, these features may introduce redundancy, which can disrupt the feature space and potentially hinder the model's performance.

Therefore, we designed an adaptive feature selection mechanism, which incorporates an attention-based top-k gating filter and a Gumbel-Sigmoid-based dynamic selection mechanism.

First, we concatenate the N feature vectors obtained through the cross-attention mechanism: $F_{\text{concat}} = [F_{1\to2}, F_{1\to3}, \ldots, F_{N-1\to N}]$. Subsequently, these feature vectors are fed into a self-attention-based feedforward layer, where each vector is assigned a score:

$$\alpha_i = \frac{\exp(\text{score}(F_i))}{\sum_{j=1}^{N} \exp(\text{score}(F_j)} \tag{6}$$

The top k vectors, based on their scores, are then selected for retention

$$F_{\text{top-k}} = \text{TopK}([F_{\text{concat}}]) \tag{7}$$

ensuring that only the most informative features are preserved for further processing.

At this point, the Gumbel-Sigmoid technique is applied to the selected top-k features. For each feature, the Gumbel noise is added to the logit score, and the Sigmoid function is used to generate a probability for retaining or discarding the feature. The outcome of this process is the final set of features:

$$F_{\text{final}} = \{f_i \mid z_i \geq 0.5\} \tag{8}$$

This allows us to preserve the features that are most useful for misinformation detection while excluding redundant features, thereby minimizing interference and improving the overall detection accuracy.

4 Experiments

4.1 Datasets and Experiment Details

We conducted extensive experiments on two multimodal datasets, FakeSV and FVC-2018. The details of the datasets are described as follows:

- FakeSV dataset, constructed by Qi et al.[16], comprises a large collection of Chinese news short videos. This dataset includes multiple modalities such as text, video, audio, and social context, which can cover various data in social media scenarios.
- The FVC-2018 dataset[29] contains real and fake videos on topics from YouTube like politics, sports, and entertainment. It includes multiple modalities such as titles, videos, comments, and URLs, making it valuable for misinformation detection.

In the experiments, the dataset was divided into training, validation, and test sets in a 70:15:15 ratio following a chronological order. The model utilized the cross-entropy loss function and AdamW optimizer, with a batch size of 64. The final results were obtained by evaluating this best model on the test set.

Qi et al.[16] proposed a multimodal fake information detection framework, SV-FEND, which integrates text, audio, image, and video modalities based on an attention mechanism. We conducted a series of experiments on feature extraction and fusion based on SVFEND, investigating the model's performance when different modal information is included separately.

4.2 Performance Results

In this study, we first explore the impact of modal information on multimodal fake information detection. Extensive experiments conducted on the baseline models using the FakeSV dataset and the FVC-2018 dataset indicate that incorporating additional modal information is not always beneficial. As the number of modalities increases, the information from these modalities can become redundant, potentially disrupting the feature space.

First, it can be observed from Table 1 that multimodal fake information detection generally outperforms single-modal approaches by a significant margin. This indicates that the complementarity between different modalities can, to a certain extent, enhance the model's performance, as integrating

multiple modalities enables a more comprehensive capture of the characteristics of multimodal information.

The results also showed that, on the FakeSV dataset, the fusion of text features and audio features extracted via the wav2vec encoder achieved the best results. Meanwhile, on the FVC-2018 dataset, the optimal model performance was observed when text, video, and image features were fused. It is worth noting that both of these optimal modality feature fusion methods only incorporate a subset of the four available modalities. Other experiments presented in Table 1 further show that adding more modal information to these fusion strategies results in a decline in performance. This phenomenon highlights the redundancy of modal information and the lack of complementarity among the modalities.

The above experiments demonstrate that in multimodal information fusion, the selection of appropriate modalities and fusion strategies is crucial. Therefore, we incorporated an adaptive feature selection mechanism while enriching the modal information, aiming to eliminate redundant information and retain useful features. Extensive experiments have also demonstrated the effectiveness of our proposed MisD-MoE model. Specifically, compared to the corresponding state-of-the-art methods, SV-FEND, our MisD-MoE model achieved accuracy improvements of 3.45% on the FakeSV dataset and 3.71% on the FVC-2018 dataset.

4.3 Ablation Study

We designed a series of ablation experiments to validate the effectiveness of the model. Specifically, we compared the model's performance under several different gating mechanisms.

First, we examined the differences in model performance and modality selectivity when using sigmoid, softmax, and gumbel-sigmoid for the final output weighting. It was observed that the model's performance with softmax was inferior to that with Gumbel-Sigmoid. This is because the softmax mechanism exhibited stronger continuity in modality selection, resulting in a smoother weight distribution across modalities and a lack of clear selectivity. In contrast, when employing the Gumbel-Sigmoid mechanism, the model demonstrated more discrete modality selection, effectively distinguishing the contributions of different modalities, and thereby enhancing overall model performance. This suggests that the Gumbel-Sigmoid mechanism is more advantageous for processing multimodal information, as it strengthens the model's sensitivity to critical modalities while suppressing irrelevant or redundant information, ultimately improving decision-making performance.

Table 2: The performance comparison of the model when employing sigmoid, softmax, and Gumbel-Sigmoid mechanisms

Mechanism	Acc	F1	Pre	Rec
sigmoid	81.37	81.30	81.56	81.23
softmax	81.89	81.67	81.95	81.43
Gumbel-Sigmoid	**82.48**	**82.04**	**82.30**	**81.60**

To validate this modality selection capability, we specifically examined the weight values output by the model when using these two mechanisms. As shown in the Fig.3, with the softmax mechanism, the model's output weights tend to follow a fixed pattern. In contrast, when using the Gumbel-Sigmoid mechanism, the selected modalities exhibit no clear regularity in the model's output. This is precisely due to the more discrete nature of the Gumbel-Sigmoid mechanism, which allows the model to dynamically adjust its reliance on different modalities under varying inputs, thus preventing any single modality from disproportionately influencing the model's decisions.

Next, we compared three selection mechanisms—top k, Gumbel-Sigmoid, and a combination of top k and Gumbel-Sigmoid, based on attention gating. It is evident that the combination of top k + Gumbel-Sigmoid yields the best performance. This is because the combination of top-k and Gumbel-Sigmoid effectively leverages the strengths of both mechanisms. The top-k mechanism ensures that only the most relevant modalities are considered by focusing on the top-ranked ones, thereby reducing noise and irrelevant information. Meanwhile, the Gumbel-Sigmoid mechanism introduces a level of stochasticity and discrete selection, allowing the model to explore different modality combinations and prevent over-reliance on specific modalities. By combining these two

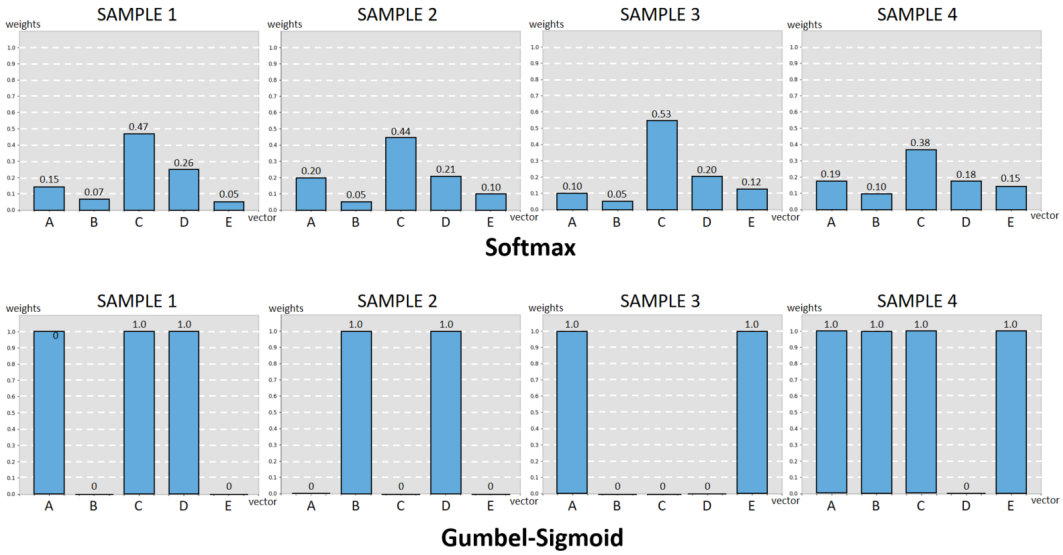

Figure 3: The weight values output by the model when using two mechanisms.

mechanisms, the model not only filters out less important modalities but also dynamically adjusts its modality selection, leading to improved robustness and overall performance.

Table 3: The performance comparison of the model when employing top k, Gumbel-Sigmoid, and a combination of both mechanisms.

Mechanism	Acc	F1	Pre	Rec
Top k	80.44	80.39	80.25	80.36
Gumbel-Sigmoid	82.48	82.04	82.30	81.60
Top k + Gumbel-Sigmoid	**82.84**	**82.22**	**83.76**	**81.60**

5 Conclusion

In this paper, we proposed a novel framework, MisD-MoE, for multimodal misinformation detection that leverages expert models for each modality and an advanced feature fusion mechanism. Our approach dynamically selects the most relevant modal features through a combination of top-k gating and Gumbel-Sigmoid mechanisms, addressing the challenges posed by redundancy and misalignment in multimodal information. Extensive experiments conducted on the FakeSV and FVC-2018 datasets demonstrated that MisD-MoE outperforms state-of-the-art methods, achieving significant improvements in accuracy. Future work could focus on further aligning modal features[30] to enhance performance.

6 Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) (No.62101553).

References

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[2] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.

[3] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, vol. 57, no. 2, p. 102025, 2020.

[4] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019, pp. 312–320.

[5] A. Jain and A. Kasbe, "Fake news detection," in *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*. IEEE, 2018, pp. 1–5.

[6] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," *arXiv preprint arXiv:1708.07104*, 2017.

[7] J. C. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Supervised learning for fake news detection," *IEEE Intelligent Systems*, vol. 34, no. 2, pp. 76–81, 2019.

[8] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," *arXiv preprint arXiv:1811.00770*, 2018.

[9] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "Spotfake: A multi-modal framework for fake news detection," in *2019 IEEE fifth international conference on multimedia big data (BigMM)*. IEEE, 2019, pp. 39–47.

[10] K. Shu, S. Wang, and H. Liu, "Exploiting tri-relationship for fake news detection," *arXiv preprint arXiv:1712.07709*, vol. 8, 2017.

[11] N. Subramani and D. Rao, "Learning efficient representations for fake speech detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5859–5866.

[12] S. B. Parikh and P. K. Atrey, "Media-rich fake news detection: A survey," in *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 2018, pp. 436–441.

[13] X. Zhou, R. Zafarani, K. Shu, and H. Liu, "Fake news: Fundamental theories, detection strategies and challenges," in *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019, pp. 836–837.

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

[15] J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, "Multimodal fake news detection via progressive fusion networks," *Information processing & management*, vol. 60, no. 1, p. 103120, 2023.

[16] P. Qi, Y. Bu, J. Cao, W. Ji, R. Shui, J. Xiao, D. Wang, and T.-S. Chua, "Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, 2023, pp. 14 444–14 452.

[17] Y. Zhou and S.-N. Lim, "Joint audio-visual deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 800–14 809.

[18] W. Y. Wang, ""' liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint arXiv:1705.00648*, 2017.

[19] X. Shu, W. Wen, H. Wu, K. Chen, Y. Song, R. Qiao, B. Ren, and X. Wang, "See finer, see more: Implicit modality alignment for text-based person retrieval," in *European Conference on Computer Vision*. Springer, 2022, pp. 624–641.

[20] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu, "Learnable manifold alignment (lema): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS journal of photogrammetry and remote sensing*, vol. 147, pp. 193–205, 2019.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] M. Liu, Y. Liu, R. Fu, Z. Wen, J. Tao, X. Liu, and G. Li, "Exploring the role of audio in multimodal misinformation detection," *arXiv preprint arXiv:2408.12558*, 2024.

[24] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[25] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

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

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[29] X. Zhou and R. Zafarani, "Network-based fake news detection: A pattern-driven approach," *ACM SIGKDD explorations newsletter*, vol. 21, no. 2, pp. 48–60, 2019.

[30] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, and P. S. Yu, "Ti-cnn: Convolutional neural networks for fake news detection," *arXiv preprint arXiv:1806.00749*, 2018.