MAFT: Multimodal Automated Fact-Checking via Textualization

Kazuya Kakizaki, Yuto Matsunaga, Ryo Furukawa

NEC Corporation
7-1, Shiba 5-chome Minato-ku, Tokyo 108-8001 Japan
{kazuya1210, yuto-matsunaga, rfurukawa}@nec.com

Abstract

This paper proposes MAFT, a novel multimodal automated fact-checking system capable of handling content in any combination of text, images, videos, and audio. The core idea behind our system is the textualization of multimodal content using various machine learning techniques. MAFT comprehensively analyzes this textualized content along with external information collected via web APIs by large language models (LLMs). MAFT generates interpretable fact-checking reports that include not only verification results but also a detailed verification process. With its adaptability and ability to automatically verify multimodal content, MAFT contributes to the fight against the spread of multimodal misinformation.

Introduction

In recent years, the spread of misinformation has become increasingly severe due to the advancement of generative AI and the proliferation of social media. Especially, misinformation encompassing multimodal content, such as text, images, videos, and audio, is posing a significant threat due to its high credibility and spreadability (Akhtar et al. 2023; Newman and Zhang 2020; Li and Xie 2020). Moreover, this type of misinformation may include deceptive and sophisticated content, like deepfakes (Pei et al. 2024; Khanjani, Watson, and Janeja 2023), making manual verification extremely challenging. Thus, there is a rapidly growing demand for automated fact-checking systems that can verify the factuality of multimodal misinformation.

Related Work. Much effort has been devoted to automated fact-checking to combat misinformation consisting of unimodal and multimodal content. In unimodal automated fact-checking, each type of data is analyzed within its own modality to verify its factuality. Significant attention has been given to detecting manipulated images, videos, and audio, including deepfakes (Dong et al. 2022; Kawa et al. 2023), synthetic data (Ojha, Li, and Lee 2023; Cazenavette et al. 2024), and forged data (Yu et al. 2024; Li et al. 2024). Furthermore, recent works have focused on verifying the factuality of text by using large language models (LLMs) to check the consistency between the target text and external information (Chern et al. 2023; Min et al. 2023). On the other hand, multimodal automated fact-checking methods verify the factuality of multimodal content by checking inconsistencies among different modalities. The inconsistency between text and images has been examined with entity recognition (Müller-Budack et al. 2020), knowledge graphs (Fung et al. 2021), and vision-language models (VLMs) (Qi et al. 2024; Abdelnabi, Hasan, and Fritz 2022; Yao et al. 2023; Liu et al. 2024). Some studies have proposed methods for detecting deepfakes that include both generated video and audio (Oorloff et al. 2024; Yang et al. 2023) based on audio-visual inconsistency. Despite the progress in automated fact-checking, the current methods are often limited to specific modalities. This poses a significant challenge when dealing with the diverse combinations of text, images, video, and audio that are commonly found in multimodal misinformation in real-world scenarios.

Our Approach. In this paper, we propose **M**ultimodal **A**utomated **F**act-checking via **T**extualization (MAFT), a novel multimodal automated fact-checking system capable of handling content in any combination of text, images, videos, and audio. The core idea behind our system is the textualization of multimodal content through various machine learning techniques. Specifically, images, videos, and audio are converted into text representing their respective content using image caption generation, video caption generation, and speech recognition. Moreover, we also employ deepfake detectors to generate text indicating whether the multimodal content has been tampered with. Once all input data are converted into text, they are integrated with external information obtained via web APIs and then comprehensively analyzed using LLMs. Finally, our system provides interpretable fact-checking reports, including not only the verification results but also a detailed verification process. With its adaptability and ability to automatically verify multimodal content, MAFT contributes to the fight against the spread of multimodal misinformation.

System Overview

Given any combination of text, images, videos, and audio, MAFT outputs a fact-checking report. MAFT consists of four processes: (i) multimodal textualization, (ii) claim extraction, (iii) evidence collection, and (iv) report generation. Fig. 1 (a) shows an overview of MAFT.

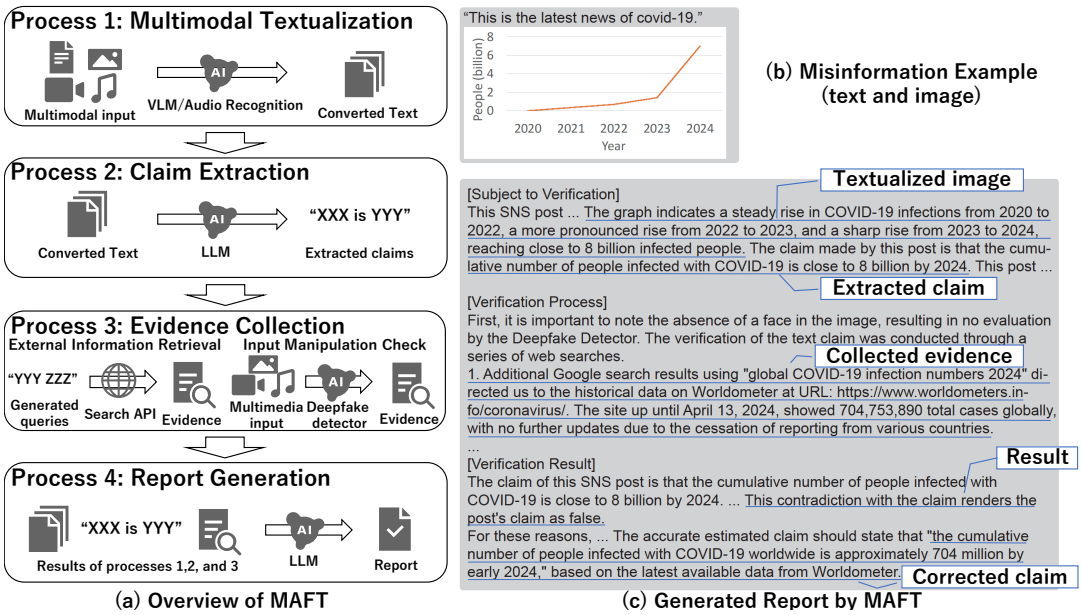

Figure 1: (a) Overview of our proposed multimodal automated fact-checking system, MAFT. (b) An example of misinformation composed of text and an image. (c) The report generated by MAFT for the misinformation shown in (b).

Multimodal Textualization. This process involves converting images, video, and audio into text, which allows for subsequent fact-checking by the LLMs. MAFT utilizes a vision language model (VLM) and a speech recognition model for the textualization. The VLM is employed to convert images and videos into text. For images, the summary generated by the VLM is directly used as the converted text. In the video case, the VLM produces a summary of the selected frames in the video and then integrates the frame summaries to make the converted text. MAFT uses speech recognition models to extract spoken content for audio, which is then used as the converted text. MAFT uses GPT-4o (OpenAI 2024) and Whisper (Radford et al. 2023) as the VLM and the speech recognition model, respectively.

Claim Extraction. In this process, claims to be verified are extracted from the multimodal inputs. To achieve this, MAFT has the LLM comprehensively analyze the textualized input through the multimodal textualization process. Specifically, the original modality of each text is made explicit in the prompt, and the LLM synthesizes its content to extract the claims. In all processes, MAFT uses GPT-4 (Achiam et al. 2023) as the LLM.

Evidence Collection. MAFT collects evidence through *external information retrieval* and *input manipulation check* to check the factuality of the extracted claims. In external information retrieval, the system retrieves information related to the extracted claims using the Google Custom Search API (Developers 2024). As in the existing study of automatic fact-checking (Chern et al. 2023), search terms are generated using a LLM. MAFT summarizes the retrieved web pages by the LLM to fewer than 1000 characters. In input manipulation check, MAFT checks whether the input images, videos, and audio have been manipulated or generated by deepfake detectors for face images/videos and audio. to ensure comprehensive analysis. The outputs from the detectors are aggregated as text; for instance, if the deepfake detector for face returns a "True", a statement such as "An evaluation using a detector revealed that this person in this image is a deepfake." is collected as evidence. MAFT uses ICT (Dong et al. 2022) and Whisper-based detector (Kawa et al. 2023) as the detector for face and audio, respectively.

Report Generation. MAFT produces a report on the fact-checking results and the process by having the LLM integrally analyze all the results obtained in the previous processes. The generated report consists of three sections: subject to verification, verification process, and verification result. The subject to verification section explains the content of the input and the claim extracted from the inputs. The verification process provides the evidence used for fact-checking and how that evidence was obtained (e.g., URL), which improves the interpretability and acceptability of the fact-check results. The verification result shows the result of the fact-checking in three stages: factual, false, or undetermined. Additionally, if the claim can be corrected, the correct content is provided based on the evidence. Fig. 1 (c) shows the report generated by MAFT for the input in Fig. 1 (b), which is partially abridged due to page limitations.

Conclusion

This paper proposes MAFT, a multimodal automated fact-checking system capable of handling content in any combination of text, images, videos, and audio. For future work, we could focus on implementing detectors for synthetic and forged data, as well as utilizing textualized externally retrieved multimodal information as evidence.

References

Abdelnabi, S.; Hasan, R.; and Fritz, M. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14940–14949.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akhtar, M.; Schlichtkrull, M. S.; Guo, Z.; Cocarascu, O.; Simperl, E.; and Vlachos, A. 2023. Multimodal Automated Fact-Checking: A Survey. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Cazenavette, G.; Sud, A.; Leung, T.; and Usman, B. 2024. FakeInversion: Learning to Detect Images from Unseen Text-to-Image Models by Inverting Stable Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10759–10769.

Chern, I.; Chern, S.; Chen, S.; Yuan, W.; Feng, K.; Zhou, C.; He, J.; Neubig, G.; Liu, P.; et al. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv preprint arXiv:2307.13528*.

Developers, G. 2024. Programmable search engine. https://developers.google.com/custom-search/v1/overview. Accessed: 2024-10-02.

Dong, X.; Bao, J.; Chen, D.; Zhang, T.; Zhang, W.; Yu, N.; Chen, D.; Wen, F.; and Guo, B. 2022. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9468–9478.

Fung, Y.; Thomas, C.; Reddy, R. G.; Polisetty, S.; Ji, H.; Chang, S.-F.; McKeown, K.; Bansal, M.; and Sil, A. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1683–1698.

Kawa, P.; Plata, M.; Czuba, M.; Szymański, P.; and Syga, P. 2023. Improved deepfake detection using whisper features. *INTERSPEECH*.

Khanjani, Z.; Watson, G.; and Janeja, V. P. 2023. Audio deepfakes: A survey. *Frontiers in Big Data*, 5: 1001063.

Li, S.; Ma, W.; Guo, J.; Xu, S.; Li, B.; and Zhang, X. 2024. UnionFormer: Unified-Learning Transformer with Multi-View Representation for Image Manipulation Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12523–12533.

Li, Y.; and Xie, Y. 2020. Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of marketing research*, 57(1): 1–19.

Liu, X.; Li, P.; Huang, H.; Li, Z.; Cui, X.; Liang, J.; Qin, L.; Deng, W.; and He, Z. 2024. FKA-Owl: Advancing Multimodal Fake News Detection through Knowledge-Augmented LVLMs. In *ACM MM*.

Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.-t.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Müller-Budack, E.; Theiner, J.; Diering, S.; Idahl, M.; and Ewerth, R. 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 international conference on multimedia retrieval*, 16–25.

Newman, E. J.; and Zhang, L. 2020. Truthiness: How non-probative photos shape belief. In *The Psychology of Fake News*, 90–114. Routledge.

Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24480–24489.

Oorloff, T.; Koppisetti, S.; Bonettini, N.; Solanki, D.; Colman, B.; Yacoob, Y.; Shahriyari, A.; and Bharaj, G. 2024. AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27102–27112.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-10-02.

Pei, G.; Zhang, J.; Hu, M.; Zhai, G.; Wang, C.; Zhang, Z.; Yang, J.; Shen, C.; and Tao, D. 2024. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*.

Qi, P.; Yan, Z.; Hsu, W.; and Lee, M. L. 2024. SNIFFER: Multimodal Large Language Model for Explainable Out-of-Context Misinformation Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13052–13062.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Yang, W.; Zhou, X.; Chen, Z.; Guo, B.; Ba, Z.; Xia, Z.; Cao, X.; and Ren, K. 2023. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 18: 2015–2029.

Yao, B. M.; Shah, A.; Sun, L.; Cho, J.-H.; and Huang, L. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2733–2743.

Yu, Z.; Ni, J.; Lin, Y.; Deng, H.; and Li, B. 2024. DiffForensics: Leveraging Diffusion Prior to Image Forgery Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12765–12774.