

MEWS: Real-time Social Media Manipulation Detection and Analysis

Trenton W. Ford
Michael Yankoski
William Theisen
Tom Henry
Farah Khashman
Katherine R. Dearstyne
Tim Weninger

*Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46530*

TFORD5@ND.EDU
MYANKOSK@ND.EDU
WTHEISEN@ND.EDU
THENRY3@ND.EDU
FKHASHMA@ND.EDU
KDEARSTY@ND.EDU
TWENINGER@ND.EDU

Pamela Bilo Thomas

*Department of Computer Science and Engineering
University of Louisville
Louisville, KY 40292*

PAMELA.THOMAS.1@LOUISVILLE.EDU

Abstract

This article presents a beta-version of MEWS (Misinformation Early Warning System). It describes the various aspects of the ingestion, manipulation detection, and graphing algorithms employed to determine—in near real-time—the relationships between social media images as they emerge and spread on social media platforms. By combining these various technologies into a single processing pipeline, MEWS can identify manipulated media items as they arise and identify when these particular items begin trending on individual social media platforms or even across multiple platforms. The emergence of a novel manipulation followed by rapid diffusion of the manipulated content suggests a disinformation campaign.

Keywords: Social media, misinformation, graph theory, near real-time detection

1. Introduction

One of the most challenging aspects of online disinformation is the overwhelming volume of content published on social media platforms. Typical approaches to combating this problem rely on tracking headlines, news stories, hashtags, and accounts Ferrara (2017); Glenski et al. (2018). However, disinformation campaigns are becoming increasingly visual Theisen et al. (2020), but organizing and analyzing this volume of content in the hope of detecting disinformation campaigns in near real-time is impossible for humans without the assistance of automated tools. This problem is especially pertinent in young and struggling democracies whose traditional media organizations lack the ability to keep pace with the explosion of deep-fake, manipulated, altered, or plainly-fake online media Yankoski et al. (2021). To provide such capacity, we have developed a real-time social media manipulation detection and analysis system called MEWS (Misinformation Early Warning System) Yankoski et al.

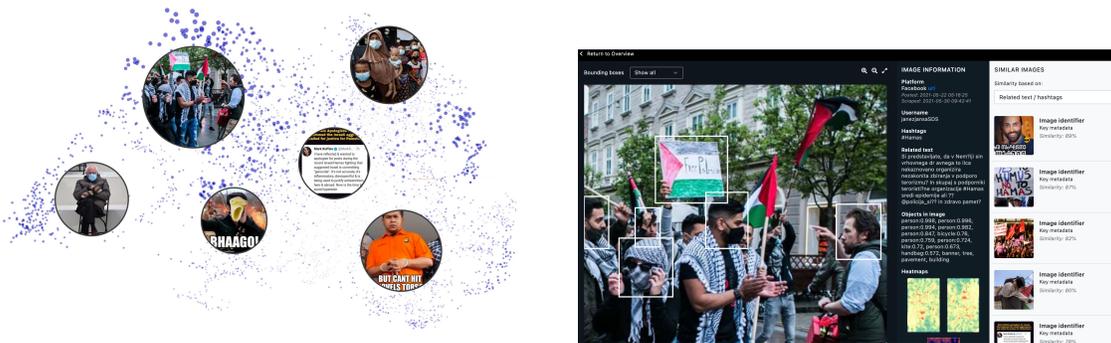


Figure 1: (Left) The macro-view of MEWS displaying image clusters and image-representatives of each cluster. (Right) The image detail page of MEWS including embedded results of image forensics, facial recognition, meme-text, related images and other pertinent information.

(2020). Figure 1 shows MEWS’ similarity graph with top images present (on the left) and a single image’s detail page (on the right).

This system combines work in digital forensics, computer vision, graph analysis, and media studies to accomplish three specific tasks:

1. MEWS ingests enormous amounts of images and video from various social media platforms (*e.g.*, Facebook, Instagram, Twitter, Telegram) using keyword targets provided by partner media organizations with context-specific knowledge domains from across the world;
2. MEWS employs state-of-the-art AI systems to detect and extract faces [Albiero et al. \(2021\)](#); [Guo et al. \(2021\)](#), objects [Carion et al. \(2020\)](#), text (including meme-text) [Lee and Smith \(2012\)](#), image features [Bay et al. \(2006\)](#), and any potential manipulations [Bianchi et al. \(2011\)](#) from the visual content; and
3. MEWS constructs a media-graph which pairs similar sub-images, objects, and manipulations for display in an interactive, easily-navigable, and searchable user interface.

For NeurIPS 2021 we demonstrated MEWS’ organizational and analytic capabilities using tens of millions of images (and other media) collected from several social media platforms (Facebook, Instagram, and Twitter) from within the Indonesian social-media context. In particular, this demonstration highlighted MEWS’ ability to:

1. Interactively reveal emergent trends in social media images in (near) real-time.
2. Identify media manipulations and alterations that recur across media items and platforms.
3. Represent the relationships between social media posts on a variety of axes, including: meme-text, ancillary post-text (*e.g.*, hashtags), detected objects, faces, and their identities, etc.

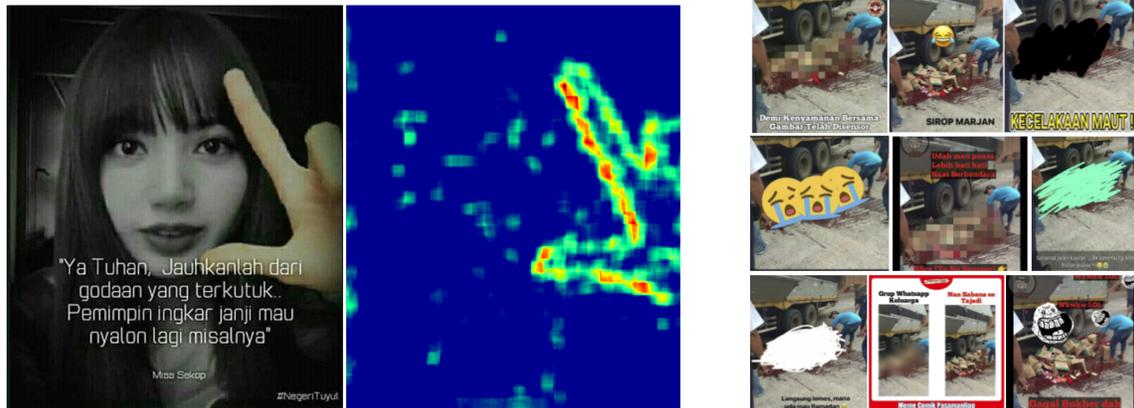


Figure 2: (Left) Example of image manipulation detection and text recognition. (Right) Cluster of self-similar images that represent a possible social media influence campaign.

4. Provide a searchable interface that users (*i.e.*, newsrooms, civil society, government agencies, and others) can use to understand the way in which disinformation is spreading within online social media.
5. Provide authenticated end-users the ability to upload media and visualize their relationships to other media contained in the MEWS dataset.

2. Examples of MEWS’s Capabilities

For example, MEWS extracted the manipulated/inserted finger-and-thumb motif from the image in Fig. 2 (left) as well as the overlaid text and several other image features. In this particular instance, we find that the finger-and-thumb motif was frequently inserted to show support for a political candidate in Indonesia inauthentically. Figure 2 (right) shows another example of a cluster of altered images of an industrial accident that MEWS detected [Theisen et al. \(2020\)](#).

MEWS also identified an image manipulation aimed at political influence: Fig. 3. A 2019 Presidential Candidate is pictured standing with another person whose shirt has been manipulated to display a hammer and sickle. The flag at the left in the image has also been manipulated with similar symbols. The heat map on the right is the algorithmic detection of the image’s manipulated portion(s) as identified by the MEWS system.

3. Interactive Virtual Demonstration

MEWS provided NeurIPS attendees an opportunity to witness several state-of-the-art technologies applied to a pertinent social problem. During the virtual conference, users were able to browse the existing collection of images through a Google Maps-like interface. They explored the faces, objects, alterations, text extractions, and information MEWS collected from each image.



Figure 3: (Left) The image as it appeared on Indonesian social media. (Center) Facial recognition identified the person on the left as Prabowo Subianto, a presidential candidate. (Right) Manipulation detection algorithms identified areas of high probability image manipulation artifacts.

In addition to demonstrating the combined power of several AI technologies, MEWS is one of the first image-based social listening and detection services, filling a wide gap in the study of social media, communications, and international security.

4. The Path Ahead

Disinformation campaigns will likely continue to influence social, political, and economic processes for the foreseeable future. We believe that these disinformation campaigns will become increasingly visual – taking the style of low-effort memes and out-of-context or cropped photography rather than sophisticated Deep-fakes.

MEWS provides an early-stage example of the application of AI technologies in the service of helping human users better navigate their social media networks. Despite its performance capabilities, MEWS is not intended to be a standalone solution. Rather, we envision MEWS as a tool for use by a robust partner network of fact-checkers, journalists, human rights watchers, and potentially government representatives who will use the information MEWS surfaces to identify and respond to disinformation threats more efficiently as they emerge on social media.

Acknowledgments

This work was supported by the US Agency for International Development (USAID) Cooperative Agreement number 7200AA18CA00059 and by the Defense Advanced Research Projects Agency and the Air Force Research Laboratory under agreement number FA8750-16-2-0173.

References

- Vítor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *CVPR*, 2021. URL <https://arxiv.org/abs/2012.07791>.
- Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.
- Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. Improved dct coefficient analysis for forgery localization in jpeg images. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2444–2447. IEEE, 2011.
- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.
- Emilio Ferrara. Disinformation and social bot operations in the run up to the 2017 french presidential election. *arXiv preprint arXiv:1707.00086*, 2017.
- Maria Glenski, Tim Wenering, and Svitlana Volkova. Propagation from deceptive news sources who shares, how much, how evenly, and how quickly? *IEEE Transactions on Computational Social Systems*, 5(4):1071–1082, 2018.
- Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021.
- Dar-Shyang Lee and Ray Smith. Improving book ocr by adaptive language and image models. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 115–119. IEEE, 2012.
- William Theisen, Joel Brogan, Pamela Bilo Thomas, Daniel Moreira, Pascal Phoa, Tim Wenering, and Walter Scheirer. Automatic discovery of political meme genres with diverse appearances. *Proceedings of the International AAAI Conference on Web and Social Media*, 15, 2020.
- Michael Yankoski, Tim Wenering, and Walter Scheirer. An ai early warning system to monitor online disinformation, stop violence, and protect elections. *Bulletin of the Atomic Scientists*, 76(2):85–90, 2020.
- Michael Yankoski, Walter Scheirer, and Tim Wenering. Meme warfare: Ai countermeasures to disinformation should focus on popular, not perfect, fakes. *Bulletin of the Atomic Scientists*, 77(3):119–123, 2021.