On Effective Ways to Warn People Against Misleading Contents

Authors: Luigi Catuogno, Emanuel Di Nardo – Università Parthenope
Clemente Galdi, **Gennaro Ragucci** – Università di Salerno

Conference: ITADATA2025

Date: September 9 - 11, 2025, Turin, Italy

Research objectives

1) To categorize systems to alert users regarding the presence of misleading content.

2) To identify criticalities and provide guidelines for future research areas about alert systems.

Introduction: The Challenge of Information Disorder

- Rapid proliferation of online information sources.
- Traditional channels are no longer the primary source of information.
- Malicious actors deliberately spread misleading content to manipulate public opinion.
- "Fake news" is just one part of a much larger problem called **Information Disorder**.

Types of Information Disorder

Categorizing misleading contents

- **Misinformation**: Inaccurate information shared without malicious intent. It's often due to a misunderstanding or lack of verification.
- **Disinformation**: False information intentionally created to cause harm and deceive. This includes fabricated narratives and distorted events.
- **Mal-information**: Genuine, verifiable information shared with malicious intent to cause harm. Examples include *doxxing* or leaking confidential data.

Note: These definitions are more precise than the general term "fake news."

Attack Target Classification

Two Main Attack Categories

- **Individual-Targeted Social Engineering**: Focuses on a single person or small group
  - **Objective**: Direct financial gain, credential theft, malware installation.
  - **Psychological Factor**: Exploits urgency, fear, curiosity, or false trust.
  - **Examples**: Phishing, pretexting, baiting, scareware.
- **Public Opinion-Targeted Mass Manipulation**: Aims at a broad audience.
  - **Objective**: Political influence, narrative shaping, belief change.
  - **Psychological Factor**: Leverages cognitive biases, emotional amplification, and social proof.
  - **Examples**: Disinformation campaigns, astroturfing, deepfakes, hate speech

Warning Systems for Individuals

Countering Individual-Targeted Attacks

- **Early Measures**: Security toolbars and SSL/HTTPS indicators were introduced in browsers
  - **Effectiveness**: These were found to have limited effectiveness.
- **Active Warnings**: Modern browsers use interstitial warnings that interrupt navigation when a risk is detected, forcing a user action.
  - **Effectiveness**: More effective, but performance varies based on user type and interface design.
  - **Our Observation**: Because individual attacks often have distinct technical signatures (e.g., IP spoofing), warning systems can take more decisive, even preventive, actions.

Warning Systems for Public Opinion

Countering Mass Manipulation

- **Social Media Measures**: Platforms like Facebook, X, and YouTube have implemented various measures.
  - **Warning Labels**: "Disputed" or "Rated false" tags on posts.
  - **Content Moderation**: Removing content that violates policies, such as hate speech or conspiracy theories.
  - **Community Notes**: User-driven fact-checking systems
- **Effectiveness Issues:**
  - Warning labels don't always stop users from sharing misinformation.
  - Algorithms can still amplify fake news and contribute to political polarization.
  - The sheer volume of content makes moderation challenging.

Warning Systems Components

Key Features of Warning Systems

1) **How they warn**:
   - **Contextual**: Flags or marks placed near the content, without interrupting the user's flow.
   - **Interstitial**: Pop-up windows or dialog boxes that block access until the user takes an action. Interstitial warnings are generally more effective but can be perceived by the user as restrictive.
2) **The content of the warning:**
   - Can be simple icons, labels, or detailed pop-up windows with explanatory text and links to reliable sources.
3) **The subject of the notification:**
   - The warning can be based on the content itself, its emotional tone, the sender's identity, or the trustworthiness of the source website.

Factors Limiting Effectiveness

Why Warnings Still Fail

- **Platform Dependency**: Many warning systems are built on proprietary technologies and policies of individual platforms, limiting their reach and consistency.
- **Lack of Standards**: There is no comprehensive framework for measuring the effectiveness of warnings across different use cases.
- **Human Factors**: User characteristics like age and education level can impact how they react to warnings.
- **Transparency Issues**: Users are more likely to accept warnings if the criteria for flagging content are transparent and explainable.

Future Directions

Moving Forward: Our Proposals

- **Platform-Independent Systems**: We must research and develop warning systems that can operate independently of a specific platform's API or policies. This is a rare area of research, but it's crucial for achieving stability and usability.
- **Comprehensive Evaluation Framework:**
  There is a need to design a standardized framework for evaluating warning systems, allowing for a meaningful comparison of different approaches.

Conclusion

Key Takeaways

- Misleading content attacks can be categorized into individual-targeted and public-opinion targeted attacks.
- Warning systems for each category must be designed differently. Individual attacks allow for more decisive, preventative measures, while mass manipulation requires greater user involvement.
- The effectiveness of current warning systems is limited by platform dependency, the lack of a standardized evaluation framework, and user-acceptance issues.
- Future research should focus on building platform-independent tools and creating a comprehensive framework for evaluation.

Questions

Contacts:

Luigi Catuogno: luigi.catuogno@uniparthenope.it*
Emanuel Di Nardo: emanuel.dinardo@uniparthenope.it
Clemente Galdi: clgaldi@unisa.it
Gennaro Ragucci: gragucci@unisa.it

**Corresponding author*