Report on architecture and design of final model for identification of organised campaigns

Deliverable D5.1

Authors: Anselmo Peñas[1]

Affiliation: (1) UNED

Version FINAL
February 28, 2025

Project Acronym	HAMiSoN
Project full title	Holistic Analysis of Organised Misinformation Activity in Social Networks
Funding Scheme	CHIST-ERA 2021
Project website	http://nlp.uned.es/hamison-project
Project Coordinator	Prof. Dr. Anselmo Peñas UNED NLP group Universidad Nacional de Educación a Distancia Email: anselmo@lsi.uned.es
Document number	D5.1
Status & version	FINAL, version 1
Contractual date of delivery	February 2025
Actual date of delivery	February 28, 2025
Type	Report
Security (distribution level)	Public
Number of pages	13
WP contributing to the deliverable	WP5
Authors: Anselmo Peñas	
Keywords: Identification of disinformation campaigns	
Abstract: This report presents a comprehensive methodology for analyzing disinformation campaigns on social networks by integrating content analysis, network analysis, and social-communicative context. It highlights key signals, both direct and derived, that can be used to detect coordinated disinformation efforts. The report explains how claims, narratives, sentiment, network structures, and actor intentions can be analyzed to uncover hidden patterns and actors behind disinformation campaigns. This framework provides a foundation for building effective tools to counter online disinformation.	

Table of Revisions

Version	Date	Descriptions and reason	By	Affected sections
0.1	Jan 31, 2025	First draft of the document	Anselmo Peñas	All
0.2	Feb 7, 2025	Second draft of the document	Anselmo Peñas	All
1.0	Feb 7, 2025	First version presented to the consortium		
1.1	Feb 14, 2025	Review	All	All
1.2	Feb 21, 2025	Final corrections	All	All
FINAL	Feb 28, 2025	Approval by project manager	Anselmo Peñas	All

Executive Summary

This report outlines a methodology to analyze disinformation campaigns and the associated behavior within social networks. These campaigns are typically characterized by the presence of coordination among multiple actors, biased narratives, propaganda techniques, and intentional efforts to influence the sentiment around particular topics.

The primary focus of this analysis is to identify disinformation networks, their behaviors, and the narratives they propagate. By examining the messages, interactions, and users within social networks over a given period, the analysis produces insights into the structure of the network, detecting clusters, nodes with negative sentiment, and propaganda techniques. The methodology also provides insights into the narratives involved, their goals, and the actors promoting them.

Key elements of the model include:

1. **Content Analysis:** This aspect of the model involves assessing messages, posts, or multimedia content to identify claims, topics, sentiments, and propaganda techniques. The analysis also uncovers the narratives driving these claims, showing how they promote specific stances over topics.

2. **Network Analysis:** This involves mapping out users and their interactions (followers, retweets, quotes) to identify core influencers, followers, bots, and personas. By analyzing network topology and communication patterns, clusters of nodes promoting similar narratives or having similar goals are identified. Connections between users are weighted by interaction frequency and analyzed for signs of coordination indicative of disinformation efforts.

3. **Social-Communicative Context:** Here, the actors involved, their goals, and relationships are examined. This aspect also identifies the narratives used by actors within different scenarios, including the creation of new narratives.

The methodology involves identifying the key areas of the network that merit attention through network and content analysis. It then proceeds to identify disinformation goals, the narratives being used, and the factions or communities promoting the same stances and goals. Visualizations, including graphs of nodes and clusters, help illustrate how disinformation campaigns are organized and propagated across social networks.

The purpose of this document is to serve as a guide for the development of assistant tools that enable analysis, reveal underlying disinformation intentions, detect coordinated behavior, and uncover factions promoting the same narratives or stances, offering a comprehensive understanding of disinformation tactics and their execution in the digital domain.

Contents

1. Introduction

Disinformation campaigns have become a significant threat in online ecosystems, influencing public perception and decision-making through coordinated, misleading, or false information. This article presents a structured methodology for analyzing disinformation campaigns within social networks, using a combination of content analysis, network analysis, and social-communicative context. By examining coordinated tweets, we can uncover patterns of disinformation and identify the actors behind these campaigns.

2. Defining Disinformation Campaigns

What is Disinformation?
Disinformation refers to the intentional spread of false or misleading information to deceive or manipulate an audience. Unlike misinformation (which is inaccurate but not intentionally deceptive), disinformation is goal-driven and often part of a broader strategy.

Characteristics of Disinformation Campaigns

- **Coordinated Behavior**: Multiple users or accounts post similar or identical tweets within a short timeframe, suggesting premeditated efforts.
- **Spread of Biased Narratives**: Tweets contain one-sided claims that favor a particular viewpoint, often omitting critical context or facts.
- **Intentionality**: The tweets aim to influence public opinion or undermine trust in credible sources.

Example: During a political election, dozens of seemingly independent accounts tweet identical messages accusing a candidate of voter fraud using the hashtag #ElectionScam within minutes of each other.

3. Methodology for Analyzing Disinformation Campaigns

This methodology focuses on detecting disinformation by analyzing both content and behavior in social networks, particularly around coordinated tweets.

3.1. Content Analysis

Content analysis examines the text, images, and multimedia content of disinformation messages.

- **Sentiment Analysis**: Disinformation tweets often convey highly negative or inflammatory sentiments aimed at discrediting individuals or institutions. For

example, tweets claiming “The election is a fraud! They are stealing your vote!” indicate high negative sentiment.

- **Topic Identification**: The methodology extracts key topics or themes within the tweets, such as "election fraud" or "vaccine danger."

- **Propaganda Techniques**: Tweets may use fear-mongering, exaggerated claims, or emotionally charged language. For example, “If they win, your country is doomed!” is a clear example of fear-based propaganda.

- **Narrative Examination**: Disinformation campaigns often revolve around specific narratives. In the example of coordinated tweets accusing a candidate of fraud, the narrative is designed to delegitimize the election process.

Tools: Natural Language Processing (NLP) techniques can automate sentiment and topic analysis, identifying frequent words, hashtags, and phrases in the tweets.

3.2. Network Analysis

Network analysis uncovers the relationships between users and detects patterns of coordination in disinformation.

- **User Mapping**: Disinformation campaigns often involve a mix of real users, bots, and personas that push the same message. In the election fraud example, users tweeting identical messages can be visualized as a cluster of interconnected accounts.

- **Interaction Analysis**: By tracking retweets, likes, and mentions, network analysis shows how tweets propagate across the network. For instance, if multiple accounts retweet the same message at the same time, it indicates coordinated action.

- **Detecting Coordination**: Signs of coordination include identical tweets, synchronized posting times, and high-frequency interactions between specific accounts.

- **Network Topology**: Visualizations, such as node-link diagrams, illustrate clusters of users posting similar content, with influential users (large nodes) directing traffic and smaller nodes (followers) amplifying their messages.

Example: A network analysis shows that 50 accounts retweeted the same election fraud claim within two minutes of its original posting, revealing a pattern of orchestrated dissemination.

3.3. Social-Communicative Context

This step looks beyond the content and network to understand the actors, goals, and broader social context behind disinformation.

- **Understanding Actors**: Disinformation campaigns often involve different types of actors, from influencers to bot farms. For instance, in the election example, the main actors could include politically motivated influencers and anonymous troll accounts amplifying their message.

- **Factions and Communities**: Factions in disinformation networks are often united by a shared goal, such as discrediting a political opponent. Identifying these communities helps to map out the narratives they are promoting.

- **Narrative Evolution**: Narratives often evolve in response to external events. For instance, if an official dismisses fraud claims, disinformation actors may shift to questioning the legitimacy of that official, reinforcing distrust.

4. Input signals to be consider by the analysis assistant

To create an effective assistant that can detect and analyze disinformation campaigns, we need to establish clear categories of signals—both direct inputs and derived data. The assistant will leverage these signals to analyze content, networks, and social-communicative context. Here's an organized breakdown, along with some additional signals to enrich the model:

Social-Communicative Context Signals

These signals provide context for the messages and network data, focusing on the scenario, actors, and their goals.

- **A priori Sets of Communicative Intentions**: Preset communicative goals (e.g., to mislead, to persuade, to discredit) based on the type of disinformation campaign being analyzed.

- **Type of Scenario / Topic**: The broader scenario in which the disinformation is occurring (e.g., political election, public health crisis).

- **Actors of Current Scenario**: The key players participating in the narrative (e.g., political figures, media outlets, anonymous accounts).

    - **Entities Involved**: Information about the entities (people, organizations, governments) being referenced in the narratives, including their roles (e.g., supporters, opponents).

- **Names, Nicknames, Camouflage**: Variations in how entities are referred to (e.g., nicknames, derogatory terms) or attempts to hide their true identity (opponents using aliases).

- **Relationships Between Entities**: Analyzing interactions or alliances between entities (e.g., political figures retweeting media outlets that support their stance).

- **Narratives**

  - **(Intention/Goal) + (Scenario/Topic) → Narratives**: Understanding how an actor's goal and the current scenario lead to the generation of specific narratives (e.g., if the goal is to undermine trust in vaccines, narratives may emerge about vaccine dangers).

  - **Types of Narratives Seen in the Past**: Refers to past patterns of disinformation narratives that match the current scenario. For example, in past elections, disinformation about vote tampering may resurface with variations.

  - **Instances of Narratives in the Current Scenario**: Identifying real-time occurrences of specific narratives within the current topic, along with their stance (e.g., "Vaccines cause infertility" narrative being pushed by anti-vaccine actors).

  - **New Narratives**: Detection of emerging narratives that have not been seen before in the current scenario. These are generated as new claims and topics arise.

Content Signals

Input Signals

These are the raw, unprocessed data that the assistant receives directly.

- **Message, Post, Tweet (Text + Image)**: This includes the full text, hashtags, mentions, and attached multimedia (images, GIFs).

- **Video Shorts**: Clips from platforms like TikTok, Instagram, or YouTube. Can include captions, speech transcription (via speech-to-text tools), and visual elements.

Derived Signals

These are signals extracted or inferred using natural language processing (NLP), machine learning, and other state-of-the-art tools.

- **Claim**: Specific assertions or statements made in the content, which can be fact-checked (e.g., "The election was rigged"). This claims can be classified and focus on Opinions (OPN) and claims Worth it to Check (FCW).

- **Topic**: General subjects or themes discussed (e.g., "elections," "vaccines"). Tools like topic modeling or keyword extraction can be used.

- **Stance**: The position or attitude toward the topic (e.g., pro-vaccine, anti-election fraud). Sentiment and context-based NLP models can help identify this.

- **Sentiment**: The emotional tone of the message (positive, negative, neutral), derived using sentiment analysis models.

- **Propaganda Techniques**: Identification of manipulative techniques such as fear mongering, appeals to authority, or bandwagon effects. This can be detected using NLP models trained on propaganda categories.

- **Narrative**: The broader storyline or message the content is promoting. Narratives often include multiple claims, and the same claim may be used in different or even opposing narratives.

    - **Example**: "The election is fraudulent" might be part of a pro-democracy narrative (“defend the election”) or an anti-democracy narrative (“undermine the election process”).

- **Claim-Narrative Relationship**:

    - **Narratives contain claims**: A narrative is built around claims.

    - **Narratives promote a stance over a topic**: Each narrative pushes a specific stance (positive or negative) about a given topic (e.g., "The vaccine is dangerous" vs. "The vaccine is safe and necessary").

Network Signals

Input Signals

These are data that represent the structure and activity within the social network.

- **Users / accounts / channels**: Accounts participating in discussions or retweeting content with all associated metadata like, for example, time of creation.

- **Links**: The topology of the social network, mapping how users are connected (followers, following, retweets, replies mentions).

- **Associated content**: Links between users and the messages they interact with.

Derived Signals

These are insights generated from network analysis tools and models.

- **Accounts categorization**: Users can be categorized based on their role in the network.
    - **Core / Regular / Casual**: Core accounts are defined as the minimum number of accounts that originate 80% of all communications. They introduce narratives or claims, and usually act also as influencers. Regular corresponds to the following 15% and Casual to the last 5% of all communications.
    - **Newbies / Inbetweeners / Oldies:** According to the account creation time. New accounts created closer to an event often indicate bot or fake accounts than longer established users.
    - **Influencers / Followers**: Accounts that amplify narratives and influence others versus accounts that mostly amplify but rarely create original content.
    - **Bots**: Accounts identified as automated or fake, detected via bot-detection algorithms.
- **Connection strength**: Describes the nature and strength of connections between users based on interaction history, such as the frequency of retweets or shared users, revealing strong or weak ties between accounts.
- **Topology analysis:** Centrality metrics (e.g., degree centrality or betweenness centrality) to analyze the network structure, such as which nodes are most influential.

5. Content signals aggregated at network level

Both input and derived signals coming from all levels explained above can be aggregated at network level to provide a view of the role of the disinformation campaign inside the network.

The starting point is to construct heterogeneous graphs with the following definition:

- Nodes:
    - Accounts
    - Messages
    - Narratives
    - Communicative goals

- Edges:
  - Links among accounts
  - Links between accounts and messages
  - Links between messages and narratives
  - Links between messages and goals

Therefore, the categorization of messages can help to detect the hottest portions of the social graphs through community detection techniques.

Communities around messages

The classification of messages content allows finding:

- **Cliques of Negative Sentiment**: Community clusters of accounts and messages spreading negative or hostile sentiment.
- **Cliques of Propaganda**: Community clusters of messages and linked accounts that repeatedly use manipulative messaging.
- **Time coordination**: Timing patterns of accounts (e.g., sudden spikes of activity within minutes) sending related messages can indicate coordination.

Communities around goals and narratives

Once the community clusters are detected, the assistant can analyse the messages to identify (or generate via LLMs):

- Communication goals of the community
- Topics and stances toward those topics
- Narratives they promote

This information can be shown over the graph for human analysis and operation in order to identify disinformation goals, narratives used and accounts involved.

6. Challenges and Limitations

- **Complexity of Disinformation**: It is difficult to separate legitimate content from coordinated disinformation, especially when both use similar rhetoric or hashtags.
- **Evolving Tactics**: Disinformation actors frequently adjust their strategies, making it challenging to develop static detection methods. For example, accounts may use different variations of a hashtag to avoid detection.
- **False Positives**: Algorithms can mistakenly flag organic behavior as coordinated, particularly in high-traffic events where similar topics are trending naturally.

7. Conclusion

This report presented a comprehensive methodology for detecting and analyzing disinformation campaigns in social networks through coordinated tweets. By combining content analysis, network mapping, and social-communicative analysis, researchers can uncover hidden narratives, identify coordinated actions, and expose the actors behind these efforts. As disinformation tactics evolve, continuous refinement of analytical tools is essential to stay ahead of emerging threats.