

Real-Time Detection of Social Media Disinformation Using DISARM: Twitter Case Study

Daniele Granata¹, Roberto Nardone¹

¹University of Naples "Parthenope", Isola C4, Centro Direzionale, 80143 Naples, Italy

Abstract

Social networks have become a major source of information, but their popularity brings significant challenges, including disinformation and security threats. This paper presents an architecture leveraging the DISARM framework, a real-time analytics system capable of integrating data from multiple sources, detecting anomalies, and assessing risks. The proposal combines advanced analytical techniques with a dedicated Policy Engine to promptly monitor, detect, and report suspicious activities, assessing their potential risks. The validation of the proposal was conducted through a case study focused on Twitter, which was chosen for its extensive usage and susceptibility to disinformation campaigns. Using different datasets, we demonstrate that the architecture effectively detects anomalies such as bot-driven activities, promotional campaigns, and suspicious behaviors in real time.

Keywords

Disinformation detection, Real-time analytics, Twitter analysis, SIEM, DISARM framework

1. Introduction

The dissemination of information is a fundamental process in modern societies, but the problem of disinformation often accompanies it. The rapid circulation of content, facilitated by digital media, makes it difficult to distinguish between reliable and manipulated information, with significant consequences for public perception and policy decisions.

Since social networks have expanded significantly in the past decade, communication and information sharing have become more accessible. However, this has also led to a rise in disinformation, making it more difficult to distinguish reliable sources from false or misleading content. The rapid spread of inaccurate information leads to different challenges, influencing public opinion, decision-making, and trust in institutions. To address this issue, policy frameworks have been developed to ensure the accuracy and reliability of information. Among them, the ABCDE framework proposed by Pamment [1] provides a structured approach to counter disinformation through analysis and response strategies.

Among the existing policy tools, the DISARM framework [2] stands out as an initiative from the EU StratCom Task Force that standardizes how disinformation incidents are described and handled. It defines a taxonomy covering tactics, techniques, actors, and countermeasures involved in disinformation campaigns. DISARM is organized into two main parts: the Red Framework, which documents offensive tactics used to spread disinformation, and the Blue Framework, which provides defensive measures and recommended responses. This structured vocabulary enables interoperability between organizations and systems when analyzing or mitigating disinformation threats.

However, despite these policies, concrete techniques are lacking to implement them effectively. The absence of operational tools to integrate these frameworks into communication systems and media allows misleading content to continue spreading. Therefore, studying and developing adequate technical solutions is essential to improve the quality of available information. Our contribution aims at filling this gap: we have designed an architecture to systematically monitor disinformation, particularly on social networks, and respond accordingly. The DISARM framework has also been embedded into

ITADATA2025: The 4th Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy

[†]These authors contributed equally.

 daniele.granata@uniparthenope.it (D. Granata); roberto.nardone@uniparthenope.it (R. Nardone)

 0000-0002-6776-9485 (D. Granata); 0000-0003-4938-9216 (R. Nardone)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

the proposed real-time monitoring architecture, mapping its tactics and countermeasures to concrete detection rules and automated responses within a Security Information and Event Management (SIEM) system and a Security Orchestration Automation and Response (SOAR) platform. This allows us to detect suspicious patterns—such as rapid message amplification or coordinated low-profile account activity—and trigger counteractions consistent with DISARM’s response definitions. Moreover, we have conducted different experiments in different scenarios, using Twitter as the social media platform for analysis.

The rest of the paper is organized as follows. Section 2 presents the state-of-the-art contributions, describing various datasets used for detecting disinformation and fake news. Section 3 introduces the proposed architecture based on DISARM, detailing the detection rules and corresponding responses. Section 4 explores real-world scenarios using different datasets and, finally, Section 5 summarizes the conclusions and outlines future work.

2. Related Work

Disinformation detection systems are part of a broader effort to ensure security and integrity in various domains. Research in securing smart grids and health information sharing provides insights into securing digital ecosystems, which could be adapted to safeguard social media platforms against malicious content and disinformation campaigns [3], [4]. Social media offers easy access and rapid dissemination of news, but it also facilitates the spread of fake news and disinformation, which can have harmful societal impacts. Detecting disinformation is challenging due to its misleading nature and the need to analyze social engagement data, which is often large, unstructured, and noisy. Some authors [5] reviewed detection methods, challenges, and future research directions to improve fake news identification on social media. In literature, there are some datasets enumerating (in an anonymised way) data from the social network (e.g., Twitter, Facebook, etc) [6]. The same authors propose a dataset¹ [7], which includes an overview of social media, news, and spatiotemporal information. This data set helps analysts study how and when fake news evolves on the Web using events and timestamps.

Other examples involve fake news incidents, like FakeNewsIndia [8], which comprises 4,803 records reported by six fact-checking websites in India from June 2016 to December 2019. It includes associated data such as 5,031 tweets and 866 YouTube videos linked to these incidents. The dataset enables impact evaluation on Twitter and YouTube using engagement-based metrics, with machine learning models predicting content popularity more effectively on YouTube than on Twitter. Another relevant dataset is *BuzzFace* [9], a comprehensive collection of Facebook data refined into four categories: mostly true, mostly false, a mixture of true and false, and no factual content. It incorporates Facebook comments, reactions, and content accessible via the platform’s Graph API, along with additional features such as article body text, images, links, and plugin comments from Facebook and Disqus. Embedded tweets included in the dataset provide opportunities for expansion across other social media platforms. With over 1.6 million text items, it is significantly larger than other datasets.

Most researchers are focusing on detecting fake news and disinformation threats using a machine learning approach and sentiment analysis. As an example, Khalil et al. [10] provide a detailed analysis of the social media-related dataset and, therefore, the associated detection model. The authors underline how the text representation impacts the accuracy of deep learning models and how the hand-crafted features are important for obtaining accurate results.

Another complementary direction is leveraging Security Information and Event Management (SIEM) systems to monitor disinformation campaigns. Various platforms and technologies have been developed to secure digital systems from potential vulnerabilities. For example, effective SIEMs have been designed with the support of Digital Twins of the system [11]. While SIEM is traditionally used for security threats, its capability to analyze logs and network traffic and detect anomalies could be extended to track patterns of coordinated disinformation. Existing studies exploring this intersection are limited, but this area holds potential for future research and practical applications. To address this gap, we

¹<https://github.com/KaiDMML/FakeNewsNet>

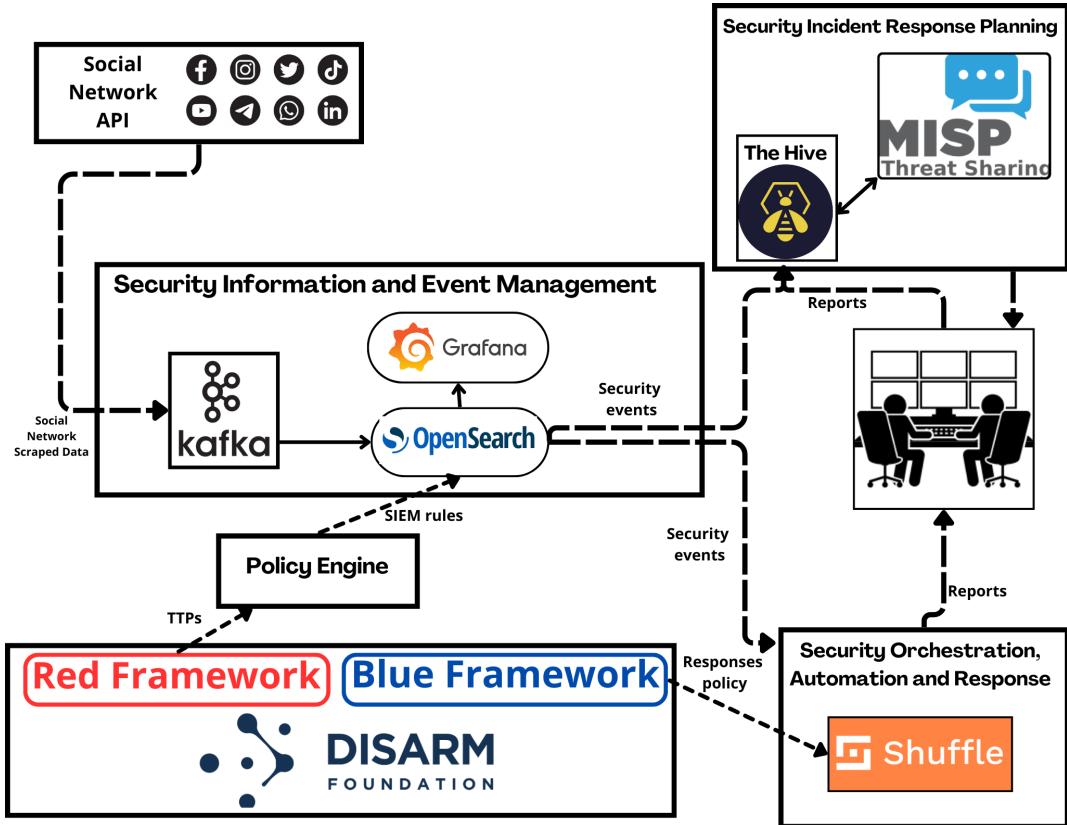


Figure 1: The Proposed Architecture

propose an architecture that integrates multiple information sources, detects anomalies (e.g., fake news), and implements efficient response mechanisms. The following section provides a detailed explanation of this proposed architecture.

3. Proposed Architecture

Our architecture is designed to detect disinformation-related anomalies on social networks in real time by directly leveraging the DISARM framework. While DISARM was initially proposed as a structured vocabulary for describing disinformation incidents, we extend its use to guide both detection and response mechanisms within our system. By leveraging the structured approach of DISARM, the architecture effectively identifies irregular patterns and potential security threats through different analytical techniques. The resulting framework seamlessly integrates with existing data streams, enabling continuous monitoring and real-time analysis of incoming information.

As illustrated in Figure 1, data is sourced from social network APIs (e.g., YouTube API, Twitter). A custom script collects structured information from these APIs, such as posts, replies, and video metadata, and publishes the social media data to Kafka. As stated above, the DISARM framework [2] is divided into Red and Blue frameworks. The red framework collects all the techniques, tactics, and procedures aimed at compromising disinformation, while the Blue Framework focuses on security countermeasures and response policies. In this case, the Red Framework is used to collect all the relevant techniques used to identify related threats and, therefore, Security Information and Event Management (SIEM) rules that mitigate these threats. The Open Search consumes the messages stored in Kafka and checks which rules are verified on the scraped data, publishing the data. On the other hand, Security Orchestration Automation and Response (SOAR) stores the related response policies based on the Blue DISARM Framework and suggests some automated responses triggered on the security events. This mechanism is based on the fact that DISARM is the common knowledge base between SIEM rules and security

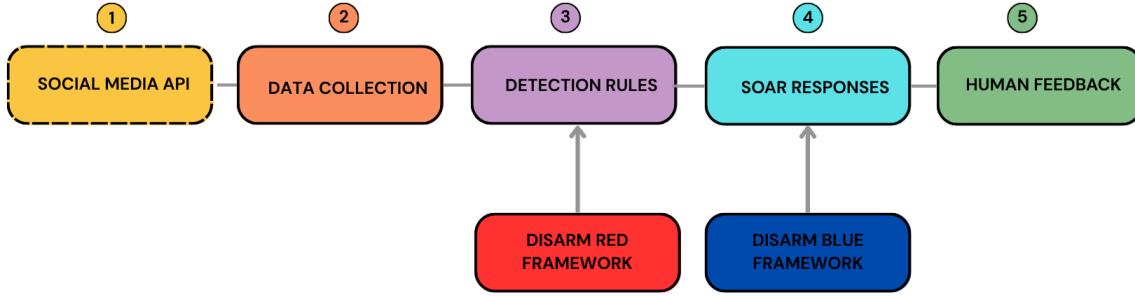


Figure 2: The data flow process

responses. Accordingly, when a specific security event is verified, the SOAR platform (i.e., Shuffle) can apply the appropriate countermeasures to mitigate the threats. In this part, security operators check the reports produced from SIEM and SOAR components. The highest abstraction level is the *Security Incident Response Planning (SIRP)* that collects the incidents from security operators (even belonging to different infrastructures) and sends feedback back to security operators, considering different data sources. The figure 2 summarizes the overall process.

It is necessary to underline that the proposed architecture is structured into two layers. The upper layer handles the API component, SIEM, SOAR, and SIRP in an agnostic way, while the lower layer focuses on specific technologies like DISARM, OpenSearch, and Shuffle. The next subsection will explain the rules identified through the DISARM framework, covering both detection mechanisms in SIEM and security responses in SOAR.

3.1. DiSARM-based detection rules

The creation of rules for anomaly detection is based on the DISARM framework. This process, integrated within the Policy Engine component, involves: i) Analyzing DISARM Tactics for specific objectives (e.g., detecting disinformation sources on social networks); ii) Identifying different threats (i.e. threat modeling [12]) affecting disinformation of the platform; iii) Correlating each threat with the DISARM tactic; iv) Identifying SIEM Rules that mitigate the related threats. The table 1 identifies a part of the rule list.

For example, the *Identification of Trending Topics or Hashtags* (T0080.003 DISARM ID) is a tactic that can lead to security issues such as the *rapid spread of suspicious content*. To detect this phenomenon, we can define a threshold for identifying anomalies in sharing behavior. Formally, let $S(C, t)$ represent the number of shares of content C within a time interval of t minutes. An anomaly is flagged if:

$$S(C, t) > X$$

where X is the predefined sharing threshold. In other words, if a content C is shared more than X times within $t = Y$ minutes, it is considered anomalous behavior that may indicate the rapid dissemination of suspicious content.

This approach clearly defines the detection criteria and facilitates the implementation of automated monitoring algorithms to promptly identify potential security threats. Another example can be related to a low-profile user. Let a low-profile user be defined as a user with fewer than T_F followers, where T_F is a predefined threshold for the number of followers (e.g., $T_F = 100$). Additionally, let the engagement rate of the user be defined by E_u , where E_u represents the number of interactions per hour. An alert should be triggered when the following condition is met:

$$F_u < T_F \quad \text{and} \quad E_u > T_E$$

#	DISARM ID	DISARM Tactic	Associated Threat	Implementation in SIEM
1	T0080.003	Identify Trending Topics/Hashtags	Rapid spread of suspicious content	Set thresholds for rapid spread (e.g., content shared more than 100 times in 10 minutes).
2	T0090.001	Create Anonymous Accounts	Abnormal engagement from low-profile users	Configure alerts when low-profile users (e.g., less than 100 followers) receive high engagement (e.g., more than 100 interactions per hour).
3	T0081.005	Identify Existing Conspiracy Narratives/Suspicious	Presence of suspicious content (keywords, hashtags, links)	Create a list of known fake news keywords and trigger alerts when they are detected in posts.
4	T0084	Reuse Existing Content	Frequent content modifications	Trigger alerts when a post is modified multiple times (e.g., more than 3 times in 24 hours).
5	T0019	Generate Information Pollution	Flooding of irrelevant or misleading data	Monitor for high volumes of posts with repeated irrelevant keywords and set alerts for excessive similar content (e.g., more than 50 posts with identical patterns in 1 hour).

Table 1
Some Tactics, the Associated Threats and Siem Rules

where T_E is a predefined threshold for engagement (e.g., $T_E = 100$).

Note that threshold values can be refined through behavioral analysis of the platform, but fixed values are used here for illustration.

3.2. DiSARM-based related responses

Based on alerts detected by the SIEM system, a corresponding set of related responses has been compiled using information from the DISARM framework. The table 2 maps identified threats to their associated DISARM IDs and provides the appropriate countermeasures, along with their corresponding DISARM countermeasure IDs, to mitigate or respond to these threats effectively.

Threat	Threat DISARM ID	Countermeasure DISARM ID	Countermeasure Specification
Rapid spread of suspicious content	T0080.003	C00126	Send the triggered alert for rapid spread of content to security operators.
Abnormal engagement from low-profile users	T0090.001	C00070	Block Access to Disinformation Resources
Presence of suspicious content (keywords, hashtags, links)	T0081.005	C00126	Use automated systems to detect and flag suspicious keywords and phrases to security operators and send alert to security operators
Frequent content modifications	T0084	C00074	Monitor and track frequency of content modifications to detect disinformation attempts
Flooding of irrelevant or misleading data	T0019	C00074	Identify and delete or rate limit identical content

Table 2
Some Threats and Associated DISARM IDs with Countermeasures

As evidence, a way to respond to the detection of *Rapid spread of suspicious content* is to send the

triggered alert to security operators who can monitor the presence of threats and, accordingly, block malicious behaviors. Similarly, identifying abnormal engagement from low-profile users helps detect suspicious activities. For example, when accounts with fewer than 100 followers receive significant interactions, such as over 100 in an hour, the system intervenes by blocking access to disinformation resources. This approach disrupts potential coordinated or automated efforts, minimizing the impact of orchestrated attacks. Another example is related to the flooding of irrelevant or misleading data undermines the reliability of information on a platform. Systems identify instances of repetitive posts, such as 50 or more identical messages shared within an hour, and take steps to delete redundant content or impose rate limits. This ensures that the platform remains focused on credible information, preventing the overshadowing of legitimate content by disinformation or spam. It is important to note that, for brevity, only a selection of threats and their corresponding countermeasures are presented here; however, the complete list is available upon request.

4. Twitter Case Study

To evaluate the effectiveness of the rules and implement the anomaly detection workflow within a specific architecture, we simulated the social network data using different datasets. The usage of a dataset simplifies data collection and aligns with the article's focus: proposing rules and response policies to combat disinformation. It is worth noticing that the same results can be obtained by scraping data using a common API provided by a social network platform. To do this, we conducted two experiments using datasets with different characteristics (as discussed in Section 2):

- Tweet-Level Analysis: Using a dataset of raw tweets, requiring more complex processing.
- High-Level Analysis: Using aggregated data from social network APIs, which simplifies implementation.

Each approach has advantages and trade-offs, as discussed in the following paragraphs. It is important to note that all social media data has been anonymised by the dataset authors.

4.1. Experiment 1: Tweet-Level Analysis (Raw Data)

Twitter News Dataset [13] contains 5,234 news events collected from Twitter in 2014. The table below summarizes the dataset:

File Name	Description
events.csv	Contains details about events, including: <ul style="list-style-type: none"> • event ID: Numeric identifier of the event (1 to 5,234). • date: Event date in YYYY-MM-DD format. • total keywords: Number of keywords associated with the event. • total tweets: Number of tweets related to the event. • keywords: Keywords for the event, separated by semicolons.
tweets.csv	Contains tweet IDs and their corresponding event IDs: <ul style="list-style-type: none"> • tweet ID: Numeric identifier of the tweet (usable via Twitter's REST API). • event ID: Identifier linking the tweet to its event.
cluster_labels.txt	Provides cluster labels for events, ranging from 0 to 19.
time_resolutions.txt	Provides temporal resolutions for events, expressed in minutes.

Table 3

Description of files in the Twitter News Dataset.

This data set is useful for analyzing news events and understanding the dynamics of Twitter dissemination. For each profile, the data set provides valuable information, such as the number of followers and tweets. Some rules (described above in table 1) have been implemented in OpenSearch to show the applicability of our approach.

4.1.1. Rule1: Rapid spread of suspicious content

As an example, considering the chosen dataset, we implemented Rule 1, which detects if any tweet has been shared at least 10 times within a 30-minute window.

It is important to note that this rule, for simplicity, uses the predefined time window provided by the dataset as input. However, in a production system, the analysis would be performed in real-time, dynamically determining the best time windows for detection as well as the optimal thresholds.

Date	Author	Tweet	Repetitions
2009-05-01 21:30:00	Wolverine811	You guys HAVE to go to this site! - UNLIMITED FREE RINGTONES!!	13
2009-05-03 00:30:00	007wisdom	"All that we are is the result of what we have thought" Buddah ... so think positive fabulous twitterverse	10
2009-05-03 17:30:00	TheOrigin953	you guys are going to LOVE me! DVD QUALITY of wolverine streaming online! no need to download or pay to watch http://tinyurl.com/cp5yhr	10
2009-05-03 20:30:00	bplusgrl445	Hey Twitter'ers! Im new to this but ive seen my friends do it Plz Follow me n I'll follow u back	13
2009-05-09 22:00:00	ukdjgrl210	FREE UNLIMITED RINGTONES!!! - http://tinyurl.com/freeringring - USA ONLY - Awesome 4 iphone	13
2009-05-16 20:30:00	JennE669	Twitter!!! Finally joined Follow me i'll follow u :d	13
2009-05-18 04:00:00	1nBpun	just a really really boring day	16
2009-05-21 23:30:00	wowlew	isPlayer Has Died! Sorry	10
2009-05-22 02:30:00	wowlew	isPlayer Has Died! Sorry	10

Table 4

Most repeated tweets in the dataset

As shown by the table 4, some tweets exhibited typical spam characteristics, including promotional language, urgency, and an external link. This pattern suggests automated posting or a coordinated campaign aimed at mass distribution.

We also implemented a response strategy (as described in Table 2). In this case, we developed a trigger in OpenSearch that uses Shuffle to schedule email alerts for the social network security administrator.

4.1.2. Rule2: Abnormal engagement from low-profile users

Another example relates to Rule 2, which uses OpenSearch to identify users with few followers (e.g., less than 100 followers) who exhibit high interaction activity on the social media platform within a short period.

The rule cannot be fully implemented since the dataset does not include additional information, such as followers or user details. In this case, we can only identify the most active users based on 30-minute time slots. Without follower counts, we do not have enough data to distinguish between influencers and potential malicious activities. Another issue is that evaluating user engagement requires analyzing interactions, not just tweets but also replies, likes, and retweets. The dataset does not provide the necessary data to trigger the rule.

4.2. Experiment 2: High-Level Analysis (Aggregated Data)

For the second experiment, we used the Twitter News Dataset 2020 [14] that provides pre-aggregated data (e.g., as Twitter API offers). Unlike the raw dataset, these APIs provide: i) Engagement metrics (likes, shares, comments); ii) Influence scores (e.g., number of verified interactions); iii) User metadata (account creation date, verification status). This pre-aggregation minimizes the need for custom data processing, enabling direct use of the API statistics available in the dataset.

This dataset is useful for different research purposes, particularly in understanding user engagement, sentiment analysis, and content trends on Twitter. Researchers can leverage tweet content to identify prevalent topics, assess public sentiment, and explore user interactions through retweets, replies, and likes.

Additionally, the temporal aspect of this dataset allows for studies of how tweets evolve over time and how conversations spread within particular time intervals. The inclusion of tweet URLs enables the retrieval of original posts, which is valuable for further context or multimedia content analysis.

This dataset is especially beneficial for analyzing social media behavior, misinformation detection, sentiment analysis, and trend forecasting. By examining tweet text and engagement features, researchers can gain insights into how information spreads across platforms, how user profiles interact with different content, and how digital discourse unfolds in real-time.

Attribute	Description
tweet_id	Unique identifier for each tweet. Useful for referencing and retrieval.
tweet_url	Direct link to the tweet, enabling further verification and context.
content	Full text of the tweet.
retweet_count	Number of retweets, indicating tweet popularity and potential virality.
reply_count	Number of replies to the tweet, reflecting user engagement and interaction.
like_count	Number of likes, showing user approval and tweet popularity.
created_at	Timestamp of when the tweet was posted.

Table 5

Overview of the Twitter News Dataset 2022.

As illustrated in the dataset, analyzing the frequency of retweets within a specified time frame is simplified by the `retweet_count` and `created_at` attributes. This allows for an efficient investigation of trends in content virality and engagement.

4.2.1. Rule1: Rapid spread of suspicious content

To verify Rule 1, we applied a query to identify users who post the highest number of tweets within a short time frame (e.g., 30 minutes). In this case, we do not detect duplicates (as in the previous scenario) because the dataset contains only unique tweets, with no repetitions.

As a result, the table 6 presents the most active users in each 30-minute interval, sorted by the total number of tweets they published.

The data analysis reveals that some users post tweets at a very high frequency, with some exceeding 80 tweets in 30 minutes. Users such as *ZeetAli3*, *techinjektion*, and *kumamonz_masa* were among the most frequent posters in certain time frames. These results could indicate malicious behaviours like: i) High tweet frequency (large number of tweets in a short period); ii) Repetitive content (indicating bots or advertising campaigns); iii) Coordinated activity (multiple accounts tweeting similar content simultaneously).

4.2.2. Rule2: Abnormal engagement from low-profile users

To identify rule 2, we defined the number of interactions as:

$$\text{interactions} = \text{retweets} + \text{replies} + \text{likes} \quad (1)$$

Time (UTC)	Most Active User	Tweets Published
2022-08-14 15:30	ZeetAli3	86
2022-08-13 14:30	politic_patriot	84
2022-08-13 11:00	techinjektion	75
2022-08-14 15:00	ZeetAli3	75
2022-08-13 12:00	X_161927A	57
2022-08-14 01:30	kumamonz_masa	85
2022-08-14 11:00	techinjektion	75
2022-08-13 20:00	techinjektion	33
2022-08-14 01:00	BcuJgODhTCQPZjK	56
2022-08-14 14:00	techinjektion	48
2022-08-13 12:00	Bettylong2	26
2022-08-14 07:00	Mummichogblogd1	38
2022-08-13 19:30	GarridoPlagas	15
2022-08-13 17:00	techinjektion	46
2022-08-14 02:30	ATGBrokers	16
2022-08-14 14:30	politic_patriot	84

Table 6

Most Active Users and Number of Tweets Published in Each 30-Minute Interval

This rule analyzes user engagement in 30-minute intervals, focusing on users with fewer than 100 followers. It ranks the top two users per interval based on total engagement, calculated as the sum of retweets, likes, replies, and quotes. Only users with at least one engagement are included. The rule also retrieves the selected users' follower counts and basic profile information. By applying this rule to Open Search, we have identified users who, despite having fewer followers, have high engagement levels, grouped in 30-minute intervals. Results are shown in table 7.

Date & Time	Username	Followers	Total Engagement
Aug 13, 2022, 09:30	MaboTofusauce	8	117
	AnoshhaKhan	43	113
Aug 13, 2022, 10:00	Already_Taken_9	48	153
	elecjazz	48	131
Aug 13, 2022, 10:30	tomo_21148	98	106
	memezon5	22	103
Aug 13, 2022, 11:00	SinghSandhu_	52	276
	specialcash1376	8	220

Table 7

Major Examples of User Engagement by Date and Time

This rule helps to spot lesser-known users who receive a lot of engagement despite having a small following. This could mean:

- **Viral Content from Small Accounts** – A post might have struck a chord with a wide audience, getting shared well beyond the user's usual reach.
- **Manipulation or Spam** – The engagement could be artificial, driven by bots or coordinated efforts to inflate interactions.
- **Algorithmic or Organic Boost** – The platform might have pushed the content to more people, or it naturally gained traction through interactions.

Understanding these cases can help distinguish between organic growth, platform dynamics, and potential manipulation. To respond to this threat, a *blocking the access to the resources is suggested (in table 2)*. In this case, we can send an alert to security administrators and suggest blocking.

5. Conclusions and Future Works

The spread of disinformation presents a significant challenge for society, especially within social networks. Distinguishing between accurate and misleading information has become increasingly difficult. Despite existing policy frameworks like ABCDE and DISARM, there remains a gap in operational tools for practical implementation. This paper proposed an architecture based on the DISARM framework, specifically designed to monitor and counter disinformation in real-time. By integrating detection mechanisms with security response policies, our approach offers an automated and scalable solution for identifying and mitigating disinformation campaigns. Through experiments using Twitter datasets, we demonstrated the effectiveness of our detection rules and response strategies in real-world scenarios, showing that our method can be applied to active social media environments. The results confirm the importance of an integrated solution combining social media monitoring, automated detection, and timely responses. Future work will focus on broadening the scope of evaluation to include diverse types of data and platforms, ensuring that the proposed approach remains adaptable and robust across various social networks.

Acknowledgments

This work has been partially supported by project SERICS (PE00000014) - Spoke 2 “Misinformation and Fakes” (CUP D43C22003050001) under the MUR National Recovery and Resilience Plan, which is funded by the European Union - NextGenerationEU.

Declaration on Generative AI

The authors used *ChatGPT* as a tool for grammar checking and proofreading of the manuscript. The content, ideas, and arguments presented in the paper were developed solely by the authors, who assume full responsibility for the final text.

References

- [1] J. Pamment, The EU’s Role in Fighting Disinformation: Crafting A Disinformation Framework, Carnegie Endowment for International Peace., 2020.
- [2] S. Terp, P. Breuer, Disarm: a framework for analysis of disinformation campaigns, in: 2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), 2022, pp. 1–8. doi:10.1109/CogSIMA54611.2022.9830669.
- [3] L. Coppolino, R. Nardone, A. Petruolo, L. Romano, Increasing the cybersecurity of smart grids by prosumer monitoring, *IEEE Transactions on Industrial Informatics* (2024).
- [4] G. Lax, R. Nardone, A. Russo, Enabling secure health information sharing among healthcare organizations by public blockchain, *Multimedia Tools and Applications* 83 (2024) 64795–64811.
- [5] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, *ACM SIGKDD Explorations Newsletter* 19 (2017) 22–36.
- [6] J. Kalyanam, M. Quezada, G. Lanckriet, B. Poblete, Early prediction and characterization of high-impact world events using social media., *arXiv preprint arXiv:1511.01830* (2015).
- [7] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media, *Big data* 8 (2020) 171–188.
- [8] A. Dhawan, M. Bhalla, D. Arora, R. Kaushal, P. Kumaraguru, Fakenewsindia: A benchmark dataset of fake news incidents in india, collection methodology and impact assessment in social media, *Computer Communications* 185 (2022) 130–141.

- [9] G. Santia, J. Williams, Buzzface: A news veracity dataset with facebook user commentary and egos, *Proceedings of the International AAAI Conference on Web and Social Media* 12 (2018) 531–540. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14985>. doi:10.1609/icwsm.v12i1.14985.
- [10] M. Khalil, M. Azzeh, Fake news detection models using the largest social media ground-truth dataset (TruthSeeker), *International Journal of Speech Technology* 27 (2024) 389–404.
- [11] L. Coppolino, R. Nardone, A. Petruolo, L. Romano, A. Souvent, Exploiting digital twin technology for cybersecurity monitoring in smart grids, in: *Proceedings of the 18th international conference on availability, reliability and security*, 2023, pp. 1–10.
- [12] D. Granata, M. Rak, Systematic analysis of automated threat modelling techniques: Comparison of open-source tools, *Software quality journal* 32 (2024) 125–161.
- [13] J. Kalyanam, M. Quezada, B. Poblete, G. Lanckriet, Prediction and characterization of high-activity events in social media triggered by real-world news, *PloS one* 11 (2016) e0166694.
- [14] DataGuy, G. Amoako, *twitter-news*, 2022. URL: <https://www.kaggle.com/dsv/4086173>. doi:10.34740/KAGGLE/DSV/4086173.