Predicting Misinformation Super-Spreaders

Silvia Giordano[1]

[1]Material for TSM_DataAnaCla 2025, extracted from:
Predicting Misinformation Super-Spreaders: a Time-Aware Social H-Index Approach, E. Verdolotti, L. Luceri and S. Giordano.

Contributing authors: silvia.giordano@supsi.ch;

Abstract

The spread of misinformation on social networks poses a significant challenge to online communities and society at large. While some users are more adept at spreading misinformation than others, their influence can be disproportionate, amplifying false narratives and causing widespread harm. This activity aims to build an approach to mitigate the impact of misinformation by enabling proactive interventions targeting potential threats through the identification and ranking of Super-Spreaders, the node archetype that consistently occupies the top positions in the misinformation spread hierarchy. You also have to study the critical role of time in prediction accuracy and propose methods to reduce the required data by shortening the observation window before making predictions.

Keywords: social media, misinformation, prediction, prevention, user behavior, content moderation, network analysis, early warning systems

1 Introduction

Social media platforms have revolutionized the way information spreads, allowing billions of users to create, share, and consume content at an unprecedented scale. While this interconnected environment fosters open communication and engagement, it also provides fertile ground for the rapid dissemination of misinformation. False or misleading content can shape public perception, influence decision-making, and erode trust in institutions [1]. Despite efforts to combat this issue, current moderation strategies primarily rely on reactive approaches, intervening only after misinformation has already gained traction. This delay in response limits the ability to mitigate harm before false narratives spread widely [2].

Given these challenges, a crucial research question arises: **Can we estimate the potential of an account to disseminate misinformation based on specific behavioral traits?** Addressing this question could pave the way for proactive strategies that help identify high-risk users before misinformation circulates extensively, and recommend them to platforms providing early detection and improving moderation efficiency.

To explore this, we focus on the specific behavioral archetype of **Super-spreaders**. Super-spreaders are highly influential users whose original content consistently goes viral. Their defining traits include high virality of created content and high engagement from other users. For this reason they are particularly influential in misinformation dynamics [3].

For effective mitigation of misinformation, you will have to develop and evaluate several user ranking systems.

2 Related Work

The proliferation of misinformation on digital platforms has sparked extensive research across both social science and computational disciplines. This problem is multifaceted: it involves not only detecting and mitigating false content but also understanding the mechanisms by which it spreads and identifying the users who most effectively amplify it. Broadly, the literature can be organized along two analytical dimensions: content-centered and user-centered approaches. These perspectives are not mutually exclusive, but they frame the problem in distinct ways.

Content-centered approaches focus on the features of the content itself. Key questions include: *Is this content misinformation?*, *What makes it viral?*, and *What modalities contribute to its spread?* Solutions in this space range from traditional text-based methods, such as keyword matching, sentiment analysis, and tf-idf encoding, to deep learning models including LSTMs, attention mechanisms, and transformers that can capture linguistic and semantic patterns [4–6]. In multimodal contexts, visual and textual features are often combined to model the complex signals contributing to virality [7]. While creator identity may be included as a feature, the primary unit of analysis remains the content.

User-centered approaches shift the focus to user behavior, attributes and network positions [8]. This line of work investigates how individual actions and structural roles contribute to the dissemination of misinformation. Temporal activity patterns [9], engagement metrics [10], and network centrality measures [11] are common predictors. Additionally, studies have examined ideological orientation, toxicity levels, and psychological factors such as cognitive reflection or conspiratorial thinking [12]. A growing body of work has sought to integrate these features into predictive systems capable of identifying high-risk users and assessing their potential impact on information flow.

This work will build on a user-centered perspective, particularly drawing inspiration from De Verna et al. [3], who proposed a metric for estimating users' misinformation spreading power, forecasting their impact, and characterizing their political alignment. In organizing the literature relevant to this work, we can highlight three recurring objectives:

- **Identification:** Detecting users who are already disseminating misinformation, often through heuristic scoring, supervised classification, or graph-based analysis.
- **Prediction:** Forecasting which users are likely to become active misinformation amplifiers. Models include temporal point processes, dynamic GNNs, and hybrid behavioral models.
- **Characterization:** Profiling users by ideology, sentiment, and linguistic markers to understand their role in misinformation ecosystems and susceptibility to false content.

Despite significant advances, critical gaps remain. Definitions and detection criteria for emerging roles—such as super spreader are still underdeveloped.

3 Data

The provided dataset contains only posts and re-posts. Other forms of interaction, such as replies or likes, are discarded, as they do not necessarily indicate endorsement, agreement, or support. This data has also going through two main steps:

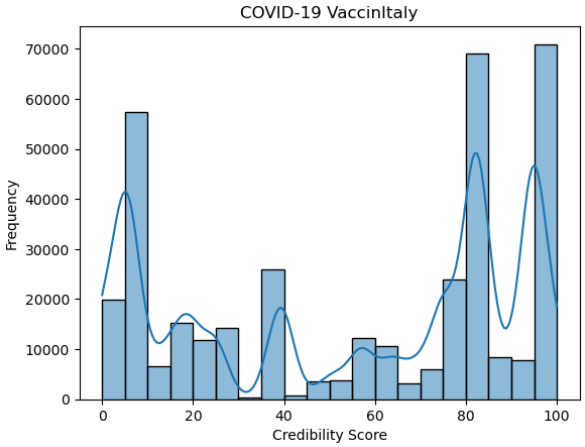
- **Labeling:** To label tweets, we relied on a common approach in the state of the art [3], consisting of labeling URLs embedded in users' posts with credibility scores provided by fact-checking agencies. We use scores from NewsGuard[1], which assigns a trustworthiness rating between 0 and 100 to news domains. Following NewsGuard guidelines, we categorized posts or re-posts linking to domains with a score of 39 or lower as *low-credibility*. Tweets without URLs or linking to unknown domains were excluded from the dataset. In cases where tweets contained multiple URLs, each was treated as a separate entry, and the average domain score was assigned to the tweet.
- **Preprocessing:** As the goal is to measure the spread, only retweet records are retained. Since retweets include metadata referencing the original tweet—such as its ID, author ID, and NewsGuard credibility score—this approach preserves all the necessary information. In this setup, a user's content is visible to the model only if it has been retweeted at least once. Additionally, platform-specific content moderation mechanisms might already be limiting the visibility of certain posts, making their inclusion potentially misleading when modeling user behavior. As a non-trivial detail, self-reposts are excluded from the analysis: they do not contribute to content diffusion and introduce noise into the re-posting pattern, as they do not reflect engagement from other users.

The data are extracted from a dataset calles **VaccinItaly.** This dataset contains COVID-19-related tweets from Italian-speaking accounts, collected as described in [13]. It focuses on discussions surrounding vaccines and public health during the pandemic. After filtering for re-posts with valid credibility scores, the final dataset includes:

- Total re-posts: 371,586
- Distinct original posts referenced: 54,916

[1] www.newsguardtech.com

- Unique users: 51,962
- Time span: December 20, 2020 – October 22, 2021 (306 days)



4 Super spreaders focused ranking

When the goal is to rank users based on how much misinformation they could spread in the future, a natural approach is to measure how much misinformation they have already spread. These users exert significant influence on social media by generating engagement from others. Engagement can be quantified through various interactions: number of likes, re-posts, comments, etc. In particular, we focus on the number of re-posts, as they directly indicate a consensual leaning toward content dissemination, whereas likes, reactions, or comments may carry ambiguous sentiment (e.g., sarcasm) and introduce noise.

The following methods are used to score users and then rank them based on their spreading capabilities:

- **Methods**
  - **H-index:** [3] This method treats posts as publications and re-posts as citations. The H-index, h, for a user is defined as the largest number of posts h that have received at least h re-posts each. Formally:

$$H(u) = \max\{h \mid u \text{ has } h \text{ posts} \in \mathcal{P}_u \text{ with re-post count} \geq h\}$$

    This method ensures that a user's influence is not dominated by a single viral post but instead reflects a consistent ability to generate engagement.

– **Page-Rank:** This method consist in counting importance (ranking) of the accounts that have re-posted an account across all its low-credibility posts:

$$\text{Page-Rank}(u) = \sum_{v \in \mathcal{V}_u} \frac{\text{Page-Rank}(v)}{L(v)}$$

where $\mathcal{V}_u$ is the set of users who re-posted low-credibility posts created by user u, Page-Rank(v) is the page-rank of user v, and $L(v)$ is the set containing all the re-posts by v of low-credibility posts created by user u.

– **Influential:** This method consist in counting the total number of re-posts generated by an account across all its low-credibility posts:

$$\text{Influential}(u) = \sum_{p \in \mathcal{P}_u} \text{Re-posts}_p$$

where $\mathcal{P}_u$ is the set of low-credibility posts created by user u, and Re-posts$_p$ is the number of re-posts for post p. Posts created by user u that do not receive any re-shares contribute zero to this method, reinforcing the idea that only re-shared posts should be considered in a ranking system. This decision was made during data pre-processing, ensuring that all experiments use datasets structured in this format.

These methods are computed entirely within an observed period, meaning that depending on the size of the period and the activity frequency of each user, they may yield different rankings for users with different activity patterns. Notably, the H-index is an interesting approach to mitigate this issue by being more robust to elapsing time since users must continue their activity and originate more data to get higher H-index scores.

However, since user activity varies over time, it is smart to partition the training set into non-overlapping, contiguous time intervals (or time slots) and compute scores for each slot. An exponential moving average (EMA) is then applied to smooth the scores over time, producing a single aggregated value that balances historical and recent influence. The EMA helps capture long-term trends while smoothing out short-term fluctuations, and the H-index calculated for each time slot:

$$\begin{cases} \hat{f}(u)_t = \alpha \cdot \hat{f}(u)_{t-1} + (1-\alpha) \cdot f(u)_t \\ \hat{f}(u)_0 = f(u)_0 \end{cases}$$

where $f(u)_t$ is the original metric at time slot t spanning δ days and α controls the trade-off between long-term and short-term influence. Running through several testing an α value of 0.5 and δ of 14gg are suggested to achieve general good performances.

5 Network Dismantling

To ensure a fair comparative analysis with previous work [3], we employed a metric based on re-share network dismantling. Consider a re-share network constructed from the dataset described in the preceding section. This network is modeled as a weighted, directed graph where nodes represent users, and directed edges represent re-share activities. The weight of an edge corresponds to the total number of distinct pieces of content re-shared. For example, if there exists an edge from A to B with a weight of 3, it indicates that B re-shared three distinct pieces of content originally created by A.

It is important to note that, on certain platforms (e.g., X, formerly Twitter), users cannot re-share the same content multiple times, while on others (e.g., Facebook), they can. Therefore, the re-share count must be computed per unique piece of content. For instance, $A \xrightarrow{3} B$ signifies that B re-shared exactly three unique pieces of content created by A. If B re-shared any of these contents multiple times, such additional shares would not be counted.

The re-share network only captures direct connections between content creators and re-sharers, even when the re-share is not inspired by a direct follow. For example, if a user C re-shared a piece of content originally created by A but discovered it through another user B, the edge is still established directly between A and C. This limitation reflects the constraints of current data availability, as social media platforms typically do not provide information about the specific source from which a user saw the content (e.g., "I re-shared this content because I saw it from this user"). As such, reconstructing full re-share cascades is not currently reliable [14].

By applying this method and considering only content with a credibility label below a predefined threshold (e.g., 39 for the NewsGuard platform), it is derived a **low-credibility re-share network**. This network represents the dissemination of misinformation among users over a specific time period.

Removing a node from this network eliminates all its incoming and outgoing edges, effectively preventing the associated misinformation re-shares. The impact of removing a specific user is quantified by summing the weights of all incident edges (both incoming and outgoing) and normalizing this value as a fraction of the total weight of the network.

6 Evaluation

To evaluate your method you need to apply network dismantling.

6.1 Network dismantling

Network dismantling refers to the procedure of iteratively removing users from the misinformation re-share network based on a given ranking and measuring the fraction of misinformation prevented at each step. While this method does not account for potential cascading effects (i.e., whether removing one user influences the behavior of others), it acts as a proxy to estimate the impact of targeted user moderation, quantifying the share of misinformation that would no longer circulate if certain users were deplatformed.

6.2 Evaluation Metrics

To quantitatively assess the performance of the proposed ranking models, you must use two primary metrics: **Quality@K** (elaborated from [3]). and **nDCG@K** (normalized Discounted Cumulative Gain) [15].

Quality@K estimates the proportion of misinformation that would be removed from the platform by entirely moderating the activity of the top-K ranked users. For example, Quality@5 represents the share of low-credibility content that would be eliminated by suppressing the behavior of the top 5 users in the ranking. This metric provides an intuitive measure of ranking effectiveness in practical moderation scenarios.[2]

nDCG@K is a standard metric in Information Retrieval used to evaluate the quality of ranked outputs. It accounts for both the relevance of items and their positions in the ranking, assigning higher scores when relevant users appear earlier in the list. In our context, each user's relevance is determined by the volume of misinformation they disseminate during the evaluation period. The truncated version, nDCG@K, ensures comparability with Quality@K by focusing only on the top-K users in both predicted and ground-truth rankings.

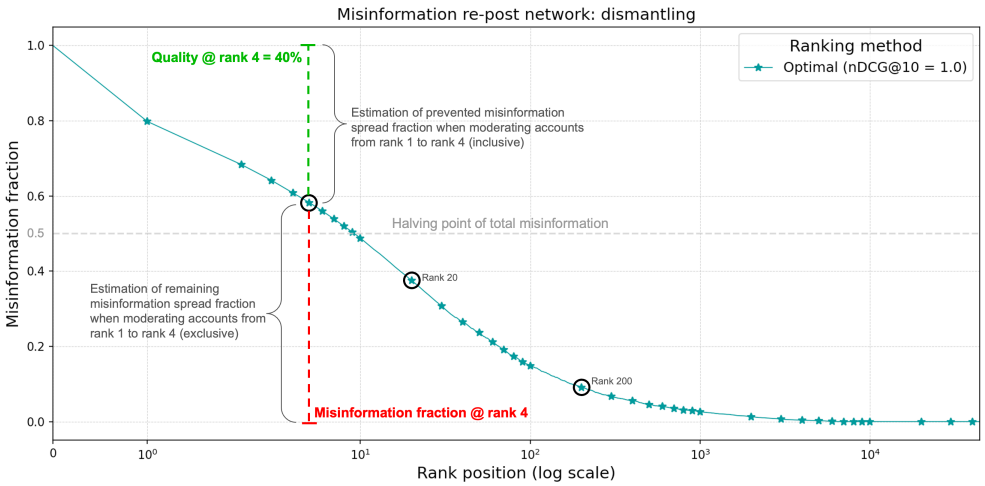


Fig. 1: Illustration of the network dismantling process. The figure highlights how the removal of highly ranked users, according to different scoring functions, affects the total volume of misinformation. It also conveys the concentration of misinformation among a small subset of users, as seen in the ground truth curve.

References

[1] Erisen, C., Erisen, E.: Populist attitudes and misinformation

[2] This is a conservative estimate of real-world moderation impact, as it assumes complete removal of content from the selected users.

challenging trust: The case of turkey. International Journal of Public Opinion Research **37**(1), 056 (2025) https://doi.org/10.1093/ijpor/edae056 https://academic.oup.com/ijpor/article-pdf/37/1/edae056/62374318/edae056.pdf

[2] Truong, B.T., Kim, S., Nogara, G., Verdolotti, E., Sahneh, E.S., Saurwein, F., Just, N., Luceri, L., Giordano, S., Menczer, F.: Delayed takedown of illegal content on social media makes moderation ineffective. arXiv. https://doi.org/10.48550/arXiv.2502.08841 . http://arxiv.org/abs/2502.08841 Accessed 2025-04-07

[3] DeVerna, M.R., Aiyappa, R., Pacheco, D., Bryden, J., Menczer, F.: Identifying and characterizing superspreaders of low-credibility content on twitter **19**(5), 0302201 https://doi.org/10.1371/journal.pone.0302201 . Accessed 2025-03-25

[4] Shaeri, P., Katanforoush, A.: A Semi-supervised Fake News Detection using Sentiment Encoding and LSTM with Self-Attention. arXiv. https://doi.org/10.48550/arXiv.2407.19332 . http://arxiv.org/abs/2407.19332 Accessed 2025-04-09

[5] Guo, Q., Kang, Z., Tian, L., Chen, Z.: TieFake: Title-Text Similarity and Emotion-Aware Fake News Detection. arXiv. https://doi.org/10.48550/arXiv.2304.09421 . http://arxiv.org/abs/2304.09421 Accessed 2025-04-09

[6] Karim, A.A.J., Asad, K.H.M., Azam, A.: Strengthening Fake News Detection: Leveraging SVM and Sophisticated Text Vectorization Techniques. Defying BERT? arXiv. https://doi.org/10.48550/arXiv.2411.12703 . http://arxiv.org/abs/2411.12703 Accessed 2025-04-09

[7] Abdali, S., shaham, S., Krishnamachari, B.: Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities. arXiv. https://doi.org/10.48550/arXiv.2203.13883 . http://arxiv.org/abs/2203.13883 Accessed 2025-04-09

[8] Lü, L., Chen, D., Ren, X.-L., Zhang, Q.-M., Zhang, Y.-C., Zhou, T.: Vital nodes identification in complex networks **650**, 1–63 https://doi.org/10.1016/j.physrep.2016.06.007 1607.01134 [physics]. Accessed 2025-04-02

[9] Stockinger, E., Gallotti, R., Hausladen, C.I.: The connection between the spread of misinformation, time of day, and individual user activity patterns. arXiv. https://doi.org/10.48550/arXiv.2307.11575 . http://arxiv.org/abs/2307.11575 Accessed 2025-04-09

[10] Avram, M., Micallef, N., Patil, S., Menczer, F.: Exposure to social engagement metrics increases vulnerability to misinformation https://doi.org/10.37016/mr-2020-033 2005.04682 [cs]. Accessed 2025-04-09

[11] Zhou, F., Lü, L., Liu, J., Mariani, M.S.: Beyond network centrality: Individual-level behavioral traits for predicting information superspreaders in social media. https://doi.org/10.1093/nsr/nwae073 . http://arxiv.org/abs/2112.03546

Accessed 2025-04-09

[12] Karami, M., Nazer, T.H., Liu, H.: Profiling fake news spreaders on social media through psychological and motivational factors. In: Proceedings of the 32st ACM Conference on Hypertext and Social Media, pp. 225–230. https://doi.org/10.1145/3465336.3475097 . http://arxiv.org/abs/2108.10942 Accessed 2025-04-06

[13] Pierri, F., Tocchetti, A., Corti, L., Giovanni, M.D., Pavanetto, S., Brambilla, M., Ceri, S.: VaccinItaly: monitoring Italian conversations around vaccines on Twitter and Facebook (2021). https://arxiv.org/abs/2101.03757

[14] DeVerna, M.R., Pierri, F., Aiyappa, R., Pacheco, D., Bryden, J., Menczer, F.: Information diffusion assumptions can distort our understanding of social network dynamics. arXiv. https://doi.org/10.48550/arXiv.2410.21554 . http://arxiv.org/abs/2410.21554 Accessed 2025-03-25

[15] Wang, Y., Li, Y., Chen, W., Wang, L., Li, D., He, W., T.-Y, C., Liu: A theoretical analysis of ndcg ranking measures. (2013). https://api.semanticscholar.org/CorpusID:7050659