An Aid to Mitigate Online Misinformation with Self-Sovereign Identity and Artificial Intelligence

Calogero Turco
Department of Computer Science
University of Pisa
Pisa, Italy
calogero.turco@phd.unipi.it

Andrea De Salve
ISASI
National Research Council (CNR)
Lecce, Italy
andrea.desalve@isasi.cnr.it

Damiano Di Francesco Maesa
Department of Computer Science
University of Pisa
Pisa, Italy
damiano.difrancesco@unipi.it

Paolo Mori
IIT
National Research Council (CNR)
Pisa, Italy
paolo.mori@iit.cnr.it

Laura Ricci
Department of Computer Science
University of Pisa
Pisa, Italy
laura.ricci@unipi.it

Accepted manuscript. This author-created version incorporates the peer-review revisions accepted for publication. It is not the IEEE-formatted Version of Record. The final published article is: C. Turco, A. De Salve, D. Di Francesco Maesa, P. Mori, and L. Ricci, "An Aid to Mitigate Online Misinformation with Self-Sovereign Identity and Artificial Intelligence," in *2025 7th International Conference on Blockchain Computing and Applications (BCCA)*, pp. 169–177, 2025. DOI: 10.1109/BCCA66705.2025.11229690.

***Abstract*—The issue of misinformation in Online Social Networks (OSNs) has been a persistent challenge traditionally addressed through automated artificial intelligence tools and human fact-checkers. However, existing solutions do not assess the credibility of a post based on the author's competence in the subject matter. This work presents an approach to fill this gap by leveraging emerging Self-Sovereign Identity (SSI) technology to generate verifiable competence certificates for content authors. Our solution combats misinformation in OSNs by offering a new perspective on content credibility. SSI is a paradigm of digital identity that enables individuals to present cryptographically provable certificates, known as Verifiable Credentials (VCs), which attest to their attributes, such as skills and qualifications. We propose a system in which authors attach VCs to their posts, providing readers with verifiable proof of the authors' competence regarding their posts' topics and thus enhancing trust in the authors. To match skills attested in the VCs with post content, we employed a system equipped with a Named Entity Recognition (NER) model to extract keywords from posts and map them to an ontology of skills and competencies. Experimental results on a synthetic dataset generated using GPT-4o demonstrate that our approach effectively reduces the spread of misinformation in a simulated OSN environment. We strongly believe that integrating verifiable author expertise attestations enhances the credibility of posts.**

***Index Terms*—Self-Sovereign Identity, Misinformation, Artificial Intelligence, Trust, Online Social Networks, Fact-Checking**

I. Introduction

More than ever during the last few years, the Internet of Contents has fulfilled its promise of being a worldwide platform for the free exchange of information. This phenomenon has been possible due to the decentralized rise of user-created content. Online Social Networks (OSNs) have hence emerged as the most popular online application [1]. As people spend more and more of their time online on OSNs, their impact on everyday people's lives and decisions grows. More and more OSN users rely on them not only to interact with their friends and families but also to work and to get informed on important topics ranging from health care to science and finance.

The diversity of opinions and freedom of expression fostered by decentralized content creation empower users to access information once difficult — or even impossible — to acquire, driving personal and societal progress. [2]. However, as different OSN platforms compete for ever-fleeting attention and users compete for visibility, the presented content must elicit stronger reactions. This systemic issue, coupled with the ever-rising influence of OSNs, has led to a proliferation of fake news and disinformation campaigns [2]. The increased ease of content generation worsens the situation thanks to the advances of generative Artificial Intelligence (AI) [3]. Never more than today, misbehaving entities can flood the network with highly curated malicious content. Current proposed solutions to counter online misinformation cannot keep up with malicious content volumes. Moreover, approaches with curators and human fact-checkers may appear to users as censorship. For instance, fact-checkers have recently been removed from all Meta OSNs[1].

Fake news is not a recent phenomenon [4]. We can say it existed since the first humans gathered in small cave-dwelling groups. As such, albeit on a smaller scale, fake news has always plagued social interactions. We notice that

[1]https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

communities have often mitigated this issue by following an 'authority-based' model, i.e., social constructs where information is deemed more trustworthy, the creator's authority on the subject is higher. Historically, individuals would undergo years of dedicated studies, after which they were awarded recognition certificates—like university degrees. These certificates were a formal attestation to their competence on a subject, allowing them to teach and guide novices. As experts in their fields, they would pass on their knowledge from a position of recognized authority. In this paper, we propose to take inspiration from this millennia-old 'authority' based model to aid OSN users in assessing social content trustworthiness by indicating the author's authority on the subjects of the content produced. However, we aim to foster the freedom of expression enabled by modern technologies, while allowing the experts to lend credibility to their words in alignment with ethical principles. This harmoniously couples with the Self-Sovereign Identity (SSI) paradigm. The choice remains with content consumers on what importance to give to authors' competence scores, including ignoring them altogether, similar to how users retain control on what SSI issuers to trust. This paper contributes by proposing, implementing, and testing an approach enabling content creators to attach to their posts a set of Verifiable Credentials (VCs) attesting their expertise on given topics automatically extracted from the content using AI techniques. Our proposal uses VCs issued by widely recognized institutions, such as universities. This enables content consumers to assess the reliability of each post by evaluating the credibility of the issuing institution and ensuring that the credential is genuinely relevant to the post's context. VCs are linked to decentralized identifiers (DIDs), enabling authors to remain pseudo-anonymous, being identified through their DID rather than by their full name on the OSN, ensuring their safety in disclosing controversial opinions. Furthermore, we propose to refine our approach by introducing third-party services called Credibility Assessment and Verification Systems (CAVS) which automatically assesses the relevance of the submitted credentials about the posted content. In particular, CAVS utilizes an AI-driven Natural Language Processing model that automatically processes the post's text and identifies the key topics. Afterward, CAVS filters the relevant authors' presented credentials based on the key topics to assess the relevance of the authors' expertise. CAVS ensures that only credentials pertinent to the issues touched by a post should be considered, thereby enhancing the author's credibility. For example, if we possess a certification in financial analysis, we can use it as valid proof of authoritativeness for a post discussing investment strategies, making our post more trustworthy. Furthermore, this approach ensures that receiving a Nobel Prize in chemistry cannot be spent as proof of authority for a post on modern poetry. Each CAVS service runs the model over an author's submitted post and generates a relevance assessment embedded in a signed VC. The author cannot alter the assessment result, preserving the reliability of the assessment presented on the OSN. A CAVS automatically verifies a post's link to credentials but leaves the evaluation of their authoritativeness to the consumer's judgment. Moreover, we advocate for this solution to be adopted in the context of professional OSNs where discussions involve specific technical knowledge, such as in scientific and financial domains, as it naturally aligns with an expert authority-based model. Thus, we target specialized communities on professional and technical topics (such as Reddit communities r/finance, r/wallstreetbets, r/health).

The paper is organized as follows: the background section, Section II, introduces our solution's basic SSI concepts. Section III presents some related works. Section IV illustrates some use cases, while Section V details our proposal and discusses its ethical considerations. Section VI presents the implementation of our system, while Section VII shows the experimental results. Finally, Section VIII draws some conclusions and presents future works.

II. Background

This section presents the fundamental components of the SSI paradigm and the involved actors. SSI is a paradigm shift in the concept of digital identity, which empowers individuals to own and manage their digital identity fully [5]. In SSI, all actors are identified by a Decentralized Identifier (DID) [6], a type of Uniform Resource Identifier that resolves to a DID Document describing an identity and its attributes and containing a public key for verifying a signature made with an associated private key. DID Documents do not expose any Personally Identifiable Information (PII) about the subject to whom they refer; this allows holders to be identifiable while being pseudo-anonymous. Blockchains, such as Ethereum, are pivotal in enabling SSI and DIDs, providing a verifiable data registry (VDR) that is distributed and replicated storage for storing DID Documents. For example, the Ethereum blockchain is used as a VDR by the *did:eth* method: a DID-compliant implementation that interacts with the blockchain in order to publish and retrieve public keys of the DIDs.

In addition to identification using DIDs, the SSI paradigm allows the generation of digital certificates that can be cryptographically validated. It allows the sharing of information embedded in them under the control of the individual to whom they refer. These digital certificates are called VC by the W3C Data Model specification [7]. The main actors in the SSI scenario are the Holder, the Issuer, which signs a VC with its private key, and the Verifier, which verifies the VC signature with the Issuer's public key. VCs are presented from a holder to a verifier in the form of a Verifiable Presentation (VP), in which the Holder aggregates multiple VCs and signs the presentation with its DID private key, demonstrating in this way a proof of possession of them. VCs are securely stored in a wallet under the Holder's control, and any stakeholder can act as a verifier, checking the validity of VCs and the integrity of data in them by cryptographically verifying the Issuer's signature on VCs. The workflow for issuing and verifying a VC is illustrated in Figure 1. In this scenario, the holder Alice receives two VCs—one from the University of Pisa and another from a Language Certification Center—and presents them in a VP to Tech Corp, the verifier.

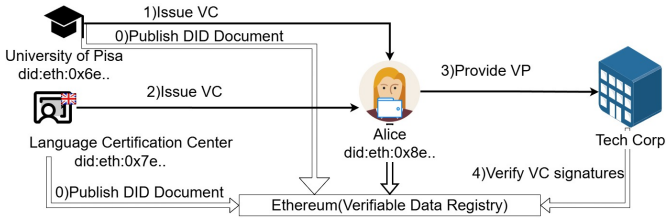

Fig. 1. Workflow of VCs issuing, storing and verification

A. ESCO ontology

The ESCO ontology is a project of the European Commission (EC) that was created to classify skills, competencies, and occupations in a standard dictionary for the European Union. The ontology provides each skill with a primary and optional secondary label, description, and URI. We define the contents of VCs as including skills from this ontology or descriptions mappable to them.

III. Related Works

There exist many definitions of fake news. Broadly defined as false news or narrowly as intentionally false news published by a news outlet [4]. Some definitions include its propagation through media, such as "a news article or message shared via media that carry false information, regardless of intent" [8]. Some researchers found a direct correlation between fake news and the stock market for a company. A conclusion reached by researchers is that fake news has a detrimental effect (albeit small) on a company's valuation, potentially leading to a decrease in its stock value [9]. To correctly identify what we are mitigating, we will define fake news as: *Fake news refers to any statement published online across various platforms, including OSN, that is disseminated by sources that may or may not have malicious intent and is propagated by users who do not critically assess the veracity of the information and the authoritativeness of the source.*

Misinformation mitigations can be, according to Rod et al. Framework [10], based on regulations, education, politics, psychology, and technology. We merely focus on technological solutions. The proposed fake news solutions in the literature are categorized into three groups: Content-based identification, Feedback-based methods, and Intervention-based solutions as described in [8].

- Content-based identification: analyzing textual characteristics (Linguistic-based Cue Set, N-gram Approach) with text features or patterns in the text. Generally done with AI models.
- Feedback-based identification: examining propagation patterns, as fake news often spreads faster than true news on OSNs [11]. These approaches generally use AI models on propagation trees or temporal pattern analysis on posts and their authors.
- Intervention-based strategies: these involve countering fake news posts with a set of true news posts. Alternatively, flagging fake news by a community of humans adopting a crowd-sourcing approach [12].

The main issue with solutions that exploit post characteristics is that by beeing based on AI, they need purposely made models with trained high-quality datasets and labels for good accuracy. In particular, early detection is difficult when there are few data features. Also, models should be trained to deal with different contexts, which requires data in each context. In AI-based models or models that require human intervention, AI services or curators flag posts as true or false, but those strategies are implemented after a post has been shared. As a result, some readers may have already formed an opinion about the news content and missed the newly applied label. These challenges emphasize the need for preemptive solutions that inform readers at the moment of content exposure.

Some proposals leverage blockchain for fact-checkers content verification [13]. TruthSeeker [14] has explored SSI-based identity verification, primarily focusing on authenticating content creators by linking their real-world identities and reputations to their online accounts. While it optionally allows users to upload certificates of expertise, these certificates are not directly used to evaluate the reliability of their posts before dissemination. Our contribution builds upon these foundations by integrating AI-driven content analysis with cryptographically verified author credentials (SSI/VCs) to assess not only the author's trustworthiness but also the credibility of the content itself before it is published. This preemptive verification mechanism aims to mitigate misinformation at the exposure point, addressing key limitations of existing post hoc verification strategies. This novel approach has the potential to significantly reshape misinformation prevention efforts by providing a proactive content-based trustworthiness assurance framework.

IV. Application Scenario

Information verification is an important requirement for several data platforms. However, there are certain platforms where the reliability and credibility of information and the author of that information become even more critical because they directly impact the platform's effectiveness and user trust. These platforms include services designed to help people build and maintain professional relationships, find job opportunities, and share knowledge on specific topics. Authors may seek to demonstrate their expertise to enhance the credibility of their content, while readers may avoid scams by discerning truth from falsehood to judge the author's content. We identified three types of platforms where information verification is vital: community-driven platforms, user-contributed content sites, and collaborative content platforms. Community-driven platforms exploit the skills of users to create specialized clusters or groups, such as *MeetUp* (https://meetup.com), which allows users with specific expertise to organize events or groups based on common interests as *SpiceWorks* (https://community.spiceworks.com), an online community of experts in ICT. User-contributed content sites help users find jobs, connect with other professionals, and participate in discussion groups, such as *BeBee* (https://bebee.com) and *XING* (https://xing.com). Finally, collaborative content platforms, such as *LinkedIn* (https://linkedin.com)

and HackerNews(https://news.ycombinator.com/), which are OSNs for professionals that allow the sharing of professional and learning content based on competencies and specific. Also, *Reddit* (https://reddit.com) communities (e.g., *r/Health*, *r/LegalAdvice*, *r/PersonalFinance*, *r/Fitness*, and *r/TechSupport*), in which authors participate to discussion under a about a topic of interest under a pseudonym. Verifying the expertise and skills of authors helps users make informed decisions across all these platforms. One case could be that Mark, eager to invest, finds a GameStop analysis by Roaring Kitty (Keith Gill) on r/WallStreetBets. Initially dismissed as a gambler, Gill was later proven right — though his Chartered Financial Analyst (CFA) credentials were not visible in his posts [2]. Our approach could have let Gill prove his expertise anonymously, helping users like Mark trust him sooner. Before describing our approach, we address trust frameworks in SSI and the ethical guidelines for AI inclusion in the proposed platforms. Note that our system does not check factual accuracy but merely assesses the author's expertise. However, an expert author generally will refrain from posting misleading and bad-quality content.

V. Assessing Social Content Trustability

In today's OSNs, the rampant sharing of opinions on any subject has led to pervasive misinformation propagated by uninformed or malicious authors on a wide range of topics. It is a stark reality that OSN users might place their trust in a post from an unfamiliar author simply because it is engaging and detailed, potentially leading to the spread of misinformation. On the other hand, if an expert acts on the OSN under a pseudonym, their posts could be ignored by OSN users because the latter do not know the author and, consequently, they do not know the author's expertise in the topic of the post [15]. Our proposed solution, leveraging SSI and Machine Learning (ML) technologies, helps assess the author's credibility about a specific subject. This approach enables readers to perceive the author's content as more trustworthy and less likely to constitute misinformation. Our proposed approach allows experts who wish to add credibility to their statements on OSNs to present a set of VCs issued by well-established institutions to a competence assessment system to obtain a competence assessment that can be paired with their post. Authors retrieve VCs with attributes specifying skills relevant to the statement for this process. As authors using this approach, we can show our authoritativeness for the statement we want to publish, thereby positively influencing our audience's trust in us. In particular, our approach is based on three fundamental concepts: *i)authors' attributes representing their skills and competencies encapsulated within certificates*, *ii) verifiability* of the credential, *iii) relevancy* of the attributes in the credentials concerning the topic of the posted statements. *Authors' attributes certification* is implemented through the SSI technology. More in detail, each author exploits the VCs they hold as a means to certify on the OSN that they have expertise in some topics. The *verifiability of the credentials* is addressed by SSI technology as well.

[2]https://www.cbsnews.com/news/gamestop-stock-share-price-roaring-kitty/

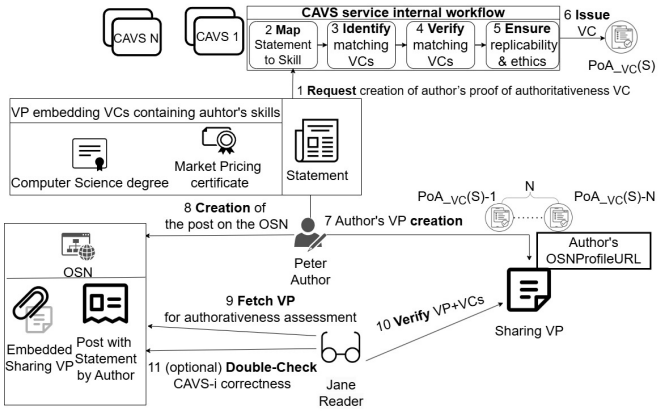

Fig. 2. Overview of the approach

As a matter of fact, the VCs proving authors' expertise are cryptographically verifiable since their Issuers sign them, and their DID-associated public key can be retrieved from a VDR according to the SSI standards. In our approach, the VDR is a blockchain, providing a distributed, resilient, and tamper-proof solution for storing DID data. The *relevancy* of the attributes in the VCs presented by the author concerning the content of the statement is assessed by exploiting ML models to extract the relevant keywords from the statement and to match them with the attributes of the VCs describing the author's expertise. We identify three key actors in our scenario:

- **Issuer** I**:** is the entity responsible for producing VCs attesting to a person's skills. This entity could be an institution, an organization, or an authority capable of certifying an individual's capabilities. For instance, a university or a course provider.
- **Author** A**:** is the subject who posts their statements on the OSN.
- **Reader** R**:** The Reader is the final consumer of the content. They want to have the ability to verify its trustworthiness by checking the VCs associated with the Author's A statement.

In the next section, we describe in detail the main phases of our approach.

A. Creation of author's proof of authoritativeness VC

For each statement S posted on the OSN, the author A requests the creation of a *proof of authoritativeness* to services defined by our approach: Credibility Assessment and Verification System (CAVS). Such evidence of authoritativeness is modeled as a VC issued by a CAVS that attests to the author's qualifications for the content of such a statement. In the following, we refer to this VC as Proof of Authoritativeness VC on the statement, in short, $PoA_{VC}(S)$. A before publishing S selects a CAVS instance (or even more) to request a $PoA_{VC}(S)$. Indeed, distinct CAVS services could be run by distinct entities using different algorithms. Indeed, distinct CAVS services could be run by distinct entities using different algorithms. Each CAVS is implemented as an off-chain microservice whose DID and associated public

key are anchored on-chain for verification purposes. For instance, a CAVS may be operated by a university, a nonprofit organization, a media watchdog NGO or any organization or individual willing to provide the service. Distinct readers may trust distinct CAVS services, e.g., because they are operated by entities they know and perceive as trustworthy. As illustrated in figure 2, the steps taken to produce $PoA_{VC}(S)$ are as follows:

1) A requests the creation of a proof of authoritativeness related to the statement S. To this aim, A forwards to a CAVS instance S, their DID DID_A, and a Verifiable Presentation VP_A. Such a VP consists of a set of VCs concerning A, i.e., $VP_A = (vc_1, .., vc_n)$, where vc_i contains a set of attributes describing the skills of A, i.e., $SK_i = (s_1^i, .., s_m^i)$. The ESCO ontology described in Section II-A can be used to take the skills labels. The ESCO ontology skills have been incorporated by Certificate Issuers as attributes into the VCs when issued.
2) CAVS extracts a number of keywords $(k_1, .., k_n)$ from S and maps them to skills $SK' = (s'_1, .., s'_n)$ from the ESCO ontology. A CAVS can use one or more ML models m_x for this aim. For instance, in our CAVS reference implementation, we used the KeyBert tool, based on the BERT model, for keyword extraction and the Nesta Skill extractor, using spaCy's NER model, for mapping keywords to skills. Those models are detailed in Section VI. In this paper, the attributes in the VCs submitted by the author are always expressed by exploiting the chosen ontology. We leave the case where such attributes are not expressed exploiting the chosen ontology format as a future work, and we plan to investigate the same approach used for the keyword extracted from the statement.
3) CAVS identifies the subset $VCS_{\text{match}} \subseteq VP_A$ of VCs where each $vc_i \in VCS_{\text{match}}$ embeds at least one skill matching the ones extracted from S. Formally, the set VCS_{match} is defined as follows: $VCS_{\text{match}} = \{vc_i \mid vc_i \in VP_A \wedge SK_i \cap (k_1, \ldots, k_n) \neq \emptyset\}$, and the matching skills SK_{match} are defined as: $\{SK_i \cap (k_1, \ldots, k_n) \mid vc_i \in VP_A \wedge SK_i = (s_1^i, \ldots, s_m^i)\}$.
4) The VCs in $\mathcal{VCS}_{\text{match}}$ are verified according to the W3C Data Model and, as part of the verification, CAVS checks that the holder specified in each VC corresponds to the DID that submitted them, i.e., DID_A. If verification fails for vc_i, such a credential is removed from the set $\mathcal{VCS}_{\text{match}}$.
5) CAVS generates SK_{mismatch}=SK'-SK_{match} to highlight the unmatched skills from author-provided VCs, the hash of the statement $h(S)$, and the hashes of the AI model used to extract keywords $h(BERT)$ and the model to map them to skills $h(NER)$. In particular, the hash of the AI models includes the models' architectures and trained weights. Those hashes are fundamental for reproducing results for any R, guaranteeing that models with the same h() should output the same result for S. Based on the skills in SK, CAVS addresses ethical concerns we will express in section V-E by embedding disclaimers in the VC, such as D_i = "This credential confirms expertise but does not indicate correctness".
6) CAVS issues $PoA_{VC}(S)$ consisting of the following elements: SK_{match}, SK_{mismatch}, h(BERT), h(NER), $h(S)$, a set of disclaimers D_i. and for replicability's sake, the name of the hashing algorithm.

Optionally, suppose the author detects inconsistencies with SK extracted that differ from the expected skills or suspects the CAVS attempted to censor them during the $PoA_{VC}(S)$ creation. In that case, they may take action to lower the CAVS instance's reputation within the trust framework. At the end of these steps, A obtain a $PoA_{VC}(S)$ from a CAVS certifying their competencies concerning the specific statement S. To be compliant with the SSI principles, the $PoA_{VC}(S)$ issued by a CAVS does not include any information about the original VCs of A and does not reveal any attributes regarding the specific titles and qualifications held by A. Consequently, following the SSI philosophy, a verifier will consider only the $PoA_{VC}(S)$s produced by the CAVS services they trust. The trust concept is discussed in section V-D. Trusting a CAVS involves accepting the inherent risk that the $PoA_{VC}(S)$ generation process may be manipulated — for example, by misrepresenting the AI models used or bypassing AI altogether to select skills manually.

B. Creation of the post on the OSN

After A receives the $PoA_{VC}(S)$ created by a CAVS service for the statement S, they create a VP to be posted on the OSN alongside the statement. Given that different CAVS services may be operated by various entities under different configurations, A could request the creation of a $PoA_{VC}(S)$ from multiple CAVS services. Hence, A embeds a set of $PoA_{VC}(S)$s in their VPs, as described in Section V-A and shown in step 7 of Figure 2. In addition to the $PoA_{VC}(S)$s, A includes in their VPs their ***OSNProfileURL*** representing where the statements will be published. This prevents replay attacks where someone else may reuse the VP for their post. Once A's VP is ready, they publish the statement S and the corresponding VP as a post on their OSN profile feed (step 8 of Figure 2).

C. Authoritativeness assessment by Reader

The OSN user R who reads a post published by A on the OSN could be interested in assessing the authoritativeness of A to decide whether to trust the post content or not. To this aim, at first, R should verify the VP posted by A, thus acting as a verifier of the SSI model (steps 9 and 10 of figure 2). In particular, R verifies the cryptographic proof of the VP (i.e., the signature) using the public key associated with the DID specified in the credential subject field of the VP. The key is retrieved by resolving the DID on its corresponding blockchain, as specified by its DID method. In the same way, R verifies the validity of the $PoA_{VC}(S)$s embedded in the VP by checking the related CAVS services' signatures. To complete the VP verification,

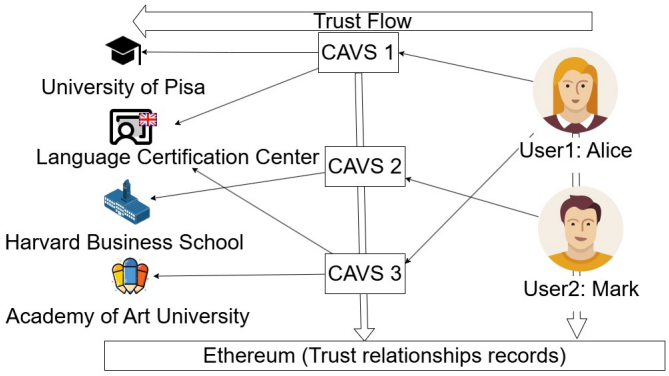

Fig. 3. Graph of trust flow

R checks that the credential subject of each $PoA_{VC}(S)$ is the same as the VP. Finally, R checks that the hash of the statement matches the hashes stored in the $PoA_{VC}(S)$s. This confirms that such $PoA_{VC}(S)$s have been created for the specific statement. To verify that the VP is for the specific post, R checks that $OSNProfileURL$ from the VP corresponds to the one where it finds the posts. Once the verification procedure has been successfully terminated, to decide whether to trust the posted statement, R considers only the $PoA_{VC}(S)$s produced by the CAVS service they trust. In the optional step 8 of Figure 2, R may execute the AI tools on the statement S to independently derive the set of extracted skills, denoted as $SK_{\text{verification}}$. If $SK_{\text{verification}} \neq SK$, this discrepancy may indicate that the CAVS service has acted dishonestly or incorrectly. Furthermore, R can compute the hashes of the locally executed models, $h(BERT)$ and $h(NER)$, and compare them with the corresponding hashes embedded in $PoA_{VC}(S)$. If a mismatch is detected, the CAVS instance may have employed a different model than claimed. Depending on the underlying trust framework, R may take appropriate actions to reduce the reputation of the CAVS within it.

D. Trust Framework

No consolidated approaches to handling trust in SSI-based systems like ours have been established. Multiple proposals have featured centralized, decentralized, and distributed models with trust approaches that can be reputation-based, policy-based, and evidence-based [16]. A solution we base our approach on represents a Web Of Trust over OSN with trust scores for issuers according to different expertise on topics [17]. We model trust over the OSN as a directed graph $G = (V, E)$,as in figure 3 where:

- The set of nodes V consists of issuers I, CAVS instances C, and users U, i.e., $V = I \cup C \cup U$. C and U are all registered on the same OSN.
- The set of directed edges E represents trust relationships.
- $E_C = \{(c, i) \mid c \in C, i \in I\}$, where (c, i) denotes that CAVS c recognizes issuer i.
- $E_U = \{(u, c) \mid u \in U, c \in C\}$, where (u, c) denotes that user u trusts CAVS c. The trust relationships are the following on the OSN and are recorded over an Ethereum Smart Contract as notarization.
- Trust propagates transitively: if $(u, c) \in E_U$ and $(c, i) \in E_C$, then u implicitly trusts issuer i through CAVS c.

The resulting trust relation T between a user u and an issuer i is defined as: $T(u, i) = \exists c \in C$ such that $(u, c) \in E_U$ and $(c, i) \in E_C$. The trust relationships are public and are mapped over the OSN "follow" relationships; each CAVS is an entity on the OSN, and the reputation of a CAVS depends on the number of u trusting it. As for misbehaving, CAVS may be removed from the following relationship, and their reputation may be reduced. We do not enforce a minimum system-wide threshold for judging a CAVS as trustworthy. Each user is free to define what constitutes an acceptable reputation level depending on personal and contextual criteria.

E. AI Ethical guidelines

We address ethical concerns to ensure our proposal aligns with United Nations [18] and European Union AI principles and the EU 2024 AI Act [19]. Our approach uses LLM and NLP models to extract skills from statements, focusing solely on these aspects of the statement to avoid harming or discriminating against authors. While the models lack direct explainability (XAI), authors can see how changes in their input affect outputs. According to the EU AI Act, our CAVS system is not a Social Scoring system and does not restrict authors' access to online platforms. To comply with Article 50, a CAVS instance will inform the author of the AI models used. In a production environment, a mechanism to complain about CAVS decisions to the manager of one instance must be allowed. We do not intend for OSN platforms to treat authors presenting posts with VCs preferentially. The VCs serve as a mark of trustworthiness for the readers. We avoid harmful manipulation as each CAVS assessment is accompanied by disclaimers clarifying that it reflects a skill-match evaluation rather than an absolute ground truth. Human oversight remains essential — for instance, in the financial domain or other sensitive areas such as healthcare or legal advice. Our system does not fall under the high-risk AI categories defined in Article 6 of the EU AI Act; rather, it is designed to support human decision-making, not replace it. The proposed approach is not intended to assess or influence political opinions but to provide context based on factual information and scientific evidence. Our approach wants to be a proof of concept for research purposes. Anyone intending to implement this system in a real environment should work with ethics experts to ensure it is compliant with regulations.

VI. CAVS AND SSI WORKFLOW IMPLEMENTATION

In order to conduct an experimental evaluation of the proposed approach, an implementation of the CAVS service has been developed, combining Veramo[3] for SSI implementation and AI tools. All the actors involved in the proposed system, including the CAVS services, are identified using DIDs belonging to the *did:eth* method, which implements the

[3]https://veramo.io/

DID standard that exploits Ethereum addresses as DIDs. The Ethereum blockchain is used as a data registry to support the resolution and management of DIDs, allowing the retrieval of the public keys during the VC/VP verification process. Veramo allows the holder to store their VC securely in a digital wallet. The `SHA-256` hashing algorithm is used to create the hashes of the workflow by CAVS instances. The proposed implementation uses the ESCO ontology for Skills and Competencies, introduced in Section II-A, in order to automate the extraction of keywords that best describe the statement; the provided implementation exploits Nesta's Skills Extractor (OJD DAPS Skills): a library that uses a pre-trained model to perform Named Entity Recognition (NER) by identifying potential skills in text. Keywords are extracted from a statement before being mapped to skills by a BERT model [20] (all-MiniLM-L6-v2 model) used by the KeyBert tool [20]: the tool exploits a model that processes text to extract contextual embeddings (vector space representations) and then uses cosine similarity between embeddings of the whole document and candidate keywords embeddings.

We deployed a Docker container where we installed and configured the two AI-based services and Veramo agents. The CAVS service is also implemented as a Docker container. It exploits the Veramo agent for SSI operations and communicates with the service deployed in the other container via REST API. We implemented a web app to simulate the actors in the workflow: Issuers, Authors, and Readers as a verifier. This architecture eases the customization of AI-based services to exploit different AI models, as they can be swapped with others. Since CAVS could adopt different implementations of AI-based services, as previously stated, $PoA_{VC}(S)$ also includes a hash of the AI model used for such services. The services and web app implementation are on GitHub [4].

VII. Experimental Results

To validate our approach, we evaluate how the proposed system's use could affect the spread of fake news. To evaluate the spread of fake news under identical network conditions, we devised two simulations of an OSN platform where users publish posts and read those from accounts they follow. In the simulated platforms, users may follow each other with a random probability, and there are two types of users: Base Users and Checkers. The former will create posts that can be either true or false, while the latter are consumers of information who read posts and decide whether or not to reshare them. A simulation called Ground Truth (**GT**) is designed to model the ideal scenario where checkers of information know to accurately determine whether a post is true or false. In contrast, the other simulation, called Skill-based (**SK**) simulation, is meant to test the proposed CAVS system. In **GT** simulation checkers **share the post only if the post contains true information**. The Skill-Based simulation, instead, assumes that Checkers do not know the ground truth, and they decide whether to reshare posts by assessing how much the author's skills match the post content using the proposed solution. Each simulation implements a Susceptible-Infectious-Recovered (SIR) model already used in the literature for fake news propagation [21]. The whole experiment is run as a Montecarlo simulation, allowing us to account for different probabilities of state changes in the SIR model. Inside the Montecarlo simulation, as done in previous works [22], simulations with a set of probabilities are run for a certain number of epochs. In each epoch, the users decide what action to perform. The behavior of the Base Users in both simulations is modeled by the SIR model, where a Base User may be:

- in the Susceptible (S) state: they may create a fake post with probability $P_{newPost} * P_{fake}$. They can share a fake post generated from the users they follow with probability $(1 - P_{newPost} * P_{sharedFake})$ where the probability $P_{sharedFake}$ depends on the number of posts carrying fake news in the aggregated personal feeds over all the post in the feeds. They go to the infected state (I) if they post/reshare a fake post. Otherwise, they post/reshare true posts and remain in the (S) state.
- in the Infected (I) state: they post/reshare only fake posts. They pass to the recovered (R) state with probability $P_{Recovered}$.
- in the Recovered (R) state: they only share/reshare true posts. They return to state (S) with $P_{Susceptible}$.

Checkers in the **GT** Simulation only share/reshare true news and stay in the (R) state. Meanwhile, Checkers in the Skill-Based simulation start from the (S) state and can go to the (I) when sharing fake news. This can happen even if the skill match is above their set threshold.

A. Dataset Generation for the Testbed

To run our simulation, we created a synthetic dataset with the GPT-4o LLM because it is a resource-effective way to obtain labeled data [23], considering that no other dataset publicly accessible provides us with the author's background alongside a set of posts, with true/false labels. As such, generating a synthetic dataset is the most resource-effective way to generate experiment data. The data generated are users' profiles, including a biography containing their skills defined according to the ESCO ontology and posts with fake and true news labels. We processed the data with KeyBert and Nesta Skill Extractor to obtain the final synthetic dataset, which embeds authors' labeled statements (articles) that users can post, and for each statement, a count of the users's skills matching the ones associated with the statement itself. Each user has a pool of five statements, where **three** have the label **"true news"** and **two** have the label **"false news"**. Users may post statements from this pool. A given statement may be posted more than once. The dataset is available at Kaggle[5]. The code to create the synthetic dataset, including the prompt and the GPT-4o configuration parameters $(temperature, top\_p)$, are available on a GitHub repository [6]

[4]https://github.com/caltr98/CAVS

[5]https://doi.org/10.34740/KAGGLE/DSV/9080191

[6]https://github.com/caltr98/CAVS/blob/master/ProfilesAndNewsGenerator/gen.py

B. Simulations statistics

To conduct our tests, we set an assumed probability of a user creating a new post in each epoch as $P_{\text{newPost}} = 0.3$, while the likelihood of resharing a post from their feed is $1 - P_{\text{newPost}}$. We maintain this probability to align with the 80-20 rule, which states that 20% of users account for 80% of a community's content production [24]—reflecting the expectation that users engage more frequently in resharing than creating new posts. The parameters P_{fake}, $P_{\text{Recovered}}$, and $P_{\text{Susceptible}}$ are assigned values within the range $[0.1, 0.9]$, drawn from a uniform distribution. We ensure that there is consistency across experiments by setting the random number generator with the same seed for both (**GT**) and (**SK**) simulation types. We conducted 1000 MonteCarlo simulations, each lasting 100 epochs with a network of 100 users, and collected the resulting metrics for analysis. Table I shows the (**SK**) vs (**GT**) simulation and the final users count for each state: Recovered, Infected, and Susceptible on average for both simulations. As can be seen in Figure 4, increasing the number of Checkers resulted in having a baseline of users in the state Recovered in the **GT** simulation, as the Checkers are always in the Recovered state. In **SK** simulations, Checkers can spread fake news and become Infected, but the impact is minimal. Overall, **SK** Checkers avoid fake news nearly as well as **GT** Checkers. Additional figures of SIR states configurations are available at the GitHub of simulation code [7] in the plots directory.

TABLE I
AVERAGE NUMBER OF USERS IN EACH STATE (RECOVERED, INFECTIOUS, SUSCEPTIBLE) AT THE END OF 1000 MONTECARLO SIMULATIONS WITH 100 USERS, WHERE EACH ROW REPRESENTS A DIFFERENT NUMBER OF CHECKERS.

N. Checkers&Type	N. Recovered	N. Infected	N Susceptible
0	36.47 ± 15.26	36.52 ± 17.80	27.01 ± 14.13
20 GT	48.66 ± 12.39	28.53 ± 13.61	22.81 ± 11.40
20 SK	47.34 ± 12.51	30.13 ± 14.71	22.53 ± 11.35
40 GT	61.11 ± 9.44	20.85 ± 10.17	18.04 ± 9.11
40 SK	59.65 ± 10.62	22.69 ± 12.01	17.66 ± 9.07
60 GT	73.70 ± 6.52	13.86 ± 6.87	12.44 ± 6.33
60 SK	72.65 ± 7.32	15.12 ± 8.09	12.24 ± 6.44
80 GT	86.79 ± 3.62	6.73 ± 3.77	6.48 ± 3.59
80 SK	86.55 ± 4.13	7.13 ± 4.38	6.32 ± 3.55

Table II compares the number of fake and true news posts in **SK** and **GT** simulations with 0 to 80 Checkers. Increasing Checkers significantly reduces fake news posts, with **SK** simulations showing improvements close to those of **SK** simulations.

C. Performance evaluation

We provide a set of benchmark tests on the critical operations performed by the CAVS systems and other actors, such as the Readers. The evaluation focuses on the performance *of skills matching, which is the most crucial operation for the functioning of the system*, considering that it is a multi-step process with first keyword extraction and then mapping. As such, we measure the performance of the AI tools. However, we also take into account the performance metrics of Veramo when generating VC and them with different numbers of attributes. Verification requires retrieval of the DID public keys, in our case, from the Ethereum Blockchain. The tests

[7]https://github.com/caltr98/SimulationFakeNews

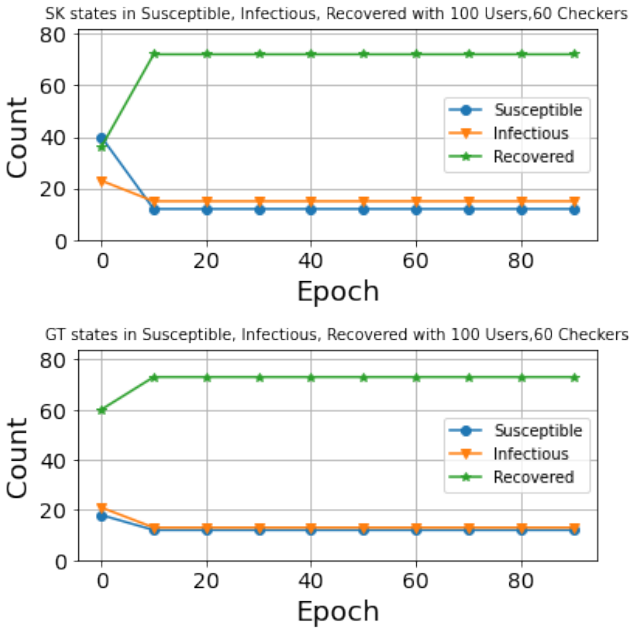

Fig. 4. SIR states over time with 100 users and 60 checkers in SK (top) and GT (bottom) for 100 epochs averaged for all Montecarlo simulations.

TABLE II
AVERAGE NUMBER OF FAKE POSTS AND FAKE RESHARED POSTS ACROSS 1000 MONTECARLO SIMULATIONS WITH 100 USERS, WHERE EACH ROW REPRESENTS A DIFFERENT NUMBER OF CHECKERS.

Checkers	**n. Fake**	**n. Fake Reshared**
0	1001.64 ± 470.58	977.86 ± 742.62
20 GT	805.74 ± 372.65	520.34 ± 366.42
20 SK	838.70 ± 396.31	669.78 ± 497.43
40 GT	604.95 ± 276.97	249.17 ± 166.83
40 SK	654.62 ± 312.93	433.35 ± 318.72
60 GT	404.57 ± 184.44	95.83 ± 63.10
60 SK	446.74 ± 210.35	231.39 ± 163.16
80 GT	201.82 ± 91.44	20.66 ± 14.28
80 SK	213.42 ± 98.41	49.56 ± 39.66

TABLE III
AVERAGE TIME (IN SECONDS) FOR MAPPING TO SKILLS DIFFERENT NUMBERS OF KEYWORDS ACROSS FIVE TRIALS.

N.Keywords (5 trials)	Time in seconds (Average ± Stdev)
5	20.9498 ± 0.6424
10	20.6783 ± 0.1821
15	21.2301 ± 0.0835

TABLE IV
AVERAGE TIME (IN MILLISECONDS) FOR DIFFERENT NUMBERS OF ATTRIBUTES ACROSS 50 TRIALS OF VC ISSUE AND VERIFICATION.

N. of Attributes	**Issuing Time (ms)**	**Verification Time (ms)**
10	11.9618 ± 6.0303	398.6447 ± 90.2456
20	13.8403 ± 6.6283	419.9820 ± 71.1129
50	29.8172 ± 13.7308	411.5480 ± 99.5790

were conducted on a Lenovo Legion 5 15ARH7H, equipped with an AMD Ryzen 7 6800H CPU with 8 cores, 16 threads, and 16 GB SO-DIMM DDR5-4800 RAM. The AI tool's performance is based on CPU execution time. KeyBert's keyword extraction times for varying text sizes are, based on 15 trials: **1)** 1.94 seconds for 630 characters (80 words) and **2)** 2.07 seconds for 1280 characters (170 words).In comparison, the Nesta skills extractor takes at least 20 seconds to map keywords to skills, as seen in Table III. We see in Table IV how SSI in the workflow adds a low overhead. The table shows the issuing as the CAVS instance crafting and signing a $PoA_{VC}(S)$. The verification is the operation done by the Reader on a $PoA_{VC}(S)$; we can see that reading from the Ethereum blockchain, the public key of the DID associated with the CAVS instance has a negligible overhead.

VIII. Conclusions

This paper proposes a novel approach to mitigate misinformation and the risk of uncritical acceptance of information in OSNs by allowing authors to provide proof of their authoritativeness on a statement published on an OSN if they wish to. Thus, we provide readers the statements with additional information that helps them decide whether or not to trust its content. While current solutions are based on human fact-checkers or AI detection of fake news, our solution integrates blockchain-based SSI technology and easily portable VCs with AI models for misinformation mitigation. The experimental results show that our solution reduces the phenomenon of misinformation. Additionally, the performance evaluation shows that the VC creation, verification and the use of the AI tools added a tolerable overhead that we plan to reduce in future works. As SSI gains mainstream adoption, a solution like ours offers a promising method for mitigating the spreading of fake news. This must be done following strict ethical considerations to preserve freedom of expression and self-sovereignty, which we support by upholding the SSI principles of individual autonomy, voluntary participation in the decentralized ecosystem, and control over one's credentials. In future work, we plan to leverage Selective Disclosure techniques to conceal potential Personally Identifiable Information (PII) within VC presented to the CAVS. Furthermore, we plan to analyze the AI tool's biases to enhance CAVS fairness and prevent misjudgment by incorporating XAI models. Additionally, we plan to develop and devise a trust framework that ensures all actors have a reliable and secure ecosystem.

Acknowledgment

This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

[1] DataReportal.com, "Digital 2024 october global statshot report," 2024, accessed: 2025-01-20. [Online]. Available: https://datareportal.com/reports/digital-2024-october-global-statshot

[2] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: a review," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 30, 2023.

[3] A. Loth, M. Kappes, and M.-O. Pahl, "Blessing or curse? a survey on the impact of generative ai on fake news," *arXiv preprint arXiv:2404.03021*, 2024.

[4] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surv.*, vol. 53, no. 5, Sep. 2020. [Online]. Available: https://doi.org/10.1145/3395046

[5] C. Mazzocca, A. Acar, S. Uluagac, R. Montanari, P. Bellavista, and M. Conti, "A survey on decentralized identifiers and verifiable credentials," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2025.

[6] D. Reed, M. Sporny, D. Longley, C. Allen, M. Sabadello, and O. Steele, *Decentralized identifiers (dids) v1. 0 - Core architecture, data model, and representations*, M. Sporny, A. Guy, M. Sabadello, and D. Reed, Eds. W3C Recommendation, 2022.

[7] M. Sporny, D. Longley, and D. Chadwick, *Verifiable Credentials Data Model v1.1*, M. Sporny, G. Noble, D. Longley, D. C. Burnett, B. Zundel, and K. D. Hartog, Eds. W3C Recommendation, 2022.

[8] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, "Combating fake news: A survey on identification and mitigation techniques," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 3, p. article no. 21, May 2019.

[9] K. Zhou, S. Scepanovic, and D. Quercia, "Characterizing fake news targeting corporations," 2024. [Online]. Available: https://arxiv.org/abs/2401.02191

[10] B. Rød, C. Pursiainen, and N. Eklund, "Combatting disinformation – how do we create resilient societies? literature review and analytical framework," *European Journal for Security Research*, March 2025. [Online]. Available: https://doi.org/10.1007/s41125-025-00105-4

[11] A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng, "Rumor cascades," in *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, vol. 8, May 2014, pp. 101–110.

[12] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, and M. Gomez-Rodriguez, "Leveraging the crowd to detect and reduce the spread of fake news and misinformation," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, ser. WSDM '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 324–332. [Online]. Available: https://doi.org/10.1145/3159652.3159734

[13] M. Choraś, M. Pawlicki, R. Kozik, K. Demestichas, P. Kosmides, and M. Gupta, "Socialtruth project approach to online disinformation (fake news) detection and mitigation," in *Proceedings of the 14th International Conference on Availability, Reliability and Security*, ser. ARES '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3339252.3341497

[14] M. Guerar and M. Migliardi, "Truthseekers chain: Leveraging invisible cappcha, ssi and blockchain to combat disinformation on social media," in *Computational Science and Its Applications – ICCSA 2022 Workshops: Malaga, Spain, July 4–7, 2022, Proceedings, Part IV*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 419–431. [Online]. Available: https://doi.org/10.1007/978-3-031-10542-5_29

[15] E. R. Velasco, "Roaring kitty returns with $180 million gamestop investment, sends stock price up by 70%," 2024, available: https://shorturl.at/Vijdf/ [Accessed: 2025-01-20].

[16] S. Y. Lim, O. B. Musa, B. A. S. Al-Rimy, and A. Almasri, *Trust Models for Blockchain-Based Self-Sovereign Identity Management: A Survey and Research Directions*. Cham: Springer International Publishing, 2022, pp. 277–302. [Online]. Available: https://doi.org/10.1007/978-3-030-93646-4_13

[17] A. De Salve, D. Di Francesco Maesa, P. Mori, L. Ricci, and A. Puccia, "A multi-layer trust framework for self sovereign identity on blockchain," *Online Social Networks and Media*, vol. 37-38, p. 100265, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468696423000241

[18] United Nations System Chief Executives Board for Coordination (CEB), "Ai ethics principles," 2022, accessed: March 12, 2025. [Online]. Available: https://unsceb.org/sites/default/files/2023-03/CEB_2022_2_Add.1%20%28AI%20ethics%20principles%29.pdf

[19] European Commission, "Approval of the content of the draft communication from the commission - commission guidelines on prohibited artificial intelligence practices established by regulation (eu) 2024/1689 (ai act)," 2024, accessed: 2025-03-12. [Online]. Available: https://ec.europa.eu/newsroom/dae/redirection/document/112366

[20] M. Grootendorst, "Keybert," https://github.com/MaartenGr/KeyBERT, 2024.

[21] A. Concepcion and C. Sy, "Modeling the spread of fake news on social networking sites using the system dynamics approach," *ASEAN Engineering Journal*, vol. 13, no. 4, pp. 69–78, Oct. 2023, available: https://journals.utm.my/aej/article/view/19251.

[22] L. Qiu, W. Jia, W. Niu, M. Zhang, and S. Liu, "Sir-im: Sir rumor spreading model with influence mechanism in social networks," *Soft Comput.*, vol. 25, no. 22, p. 13949–13958, Nov. 2021. [Online]. Available: https://doi.org/10.1007/s00500-020-04915-7

[23] V. Veselovsky, M. H. Ribeiro, A. Arora, M. Josifoski, A. Anderson, and R. West, "Generating faithful synthetic data with large language models: A case study in computational social science," *arXiv preprint arXiv:2305.15041*, 2023.

[24] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. E. Zhao, "Analyzing patterns of user content generation in online social networks," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 369–378. [Online]. Available: https://doi.org/10.1145/1557019.1557064