A Blockchain-based and AI-Endorsed Mechanism to Support Social Networks on Fake News Containment

Valdemar Vicente Graciano Neto[1], Jacson Rodrigues Barbosa[1], Eliomar Araújo de Lima[1], Sérgio Teixeira de Carvalho[1], Samuel Venzi[2]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Alameda Palmeiras, Quadra D, Câmpus Samambaia – 74.690-900 – Goiânia – GO – Brasil

[2]GoLedger – Brasília – DF

{valdemarneto, jacson_rodrigues,eliomar.lima,sergiocarvalho}@ufg.br,

samuel.venzi@goledger.com.br

Abstract. *Online Social Networks (OSNs) have promoted, yet unintentionally, critical consequences of fake news dissemination. However, the mainstream OSNs are centralized, while Secure Social Networks (SSNs) are not as popular as the centralized ones. To bridge this gap, this paper proposes a solution using blockchain and artificial intelligence to enhance OSN security by introducing a mechanism for content verification, fact-checking, and rewarded participation. Preliminary proof-of-concept results demonstrate the feasibility of the approach to face misinformation.*

1. Introduction

Online Social Networks (OSNs) are platforms where massive content dissemination happens. These platforms are the target of mass manipulation by the dissemination of fake news since they have high adherence of populations of several countries, including Brazil [17]. Mainstream OSNs, such as Instagram, X, and others, have included mechanisms for detecting and containing of fake news. Instagram has labeled news detected as fake with an explicit alert about it, including the explanation on why that content is fake[1]; X, on the other hand, has allowed the users to vote about the legitimacy of content[2]. However, doubts can be raised about the interests of the owners of those private companies. Those platforms are considered centralized and monopolized since these companies' owners integrally regulate their operations. Dozens of secure OSNs also exist (such as Steemit[3], Indorse[4], Sapien[5], and SocialX[6] [11]), based on blockchains and cryptocurrencies, enabling voting, curation and rewarding of influencers and curators. But the most popular among them (OSN Kin[7]) had only 10 million users in 2020 [7], contrasted with 3.049

[1] https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram

[2] https://www.nytimes.com/2024/01/25/us/politics/elon-musk-election-misinformation-x-twitter.html

[3] https://steemit.com/

[4] https://indorse.io/

[5] https://www.sapien.network/

[6] https://socialx.network/

[7] https://kinsocial.app/

billion monthly active users on Facebook, 2 billion monthly active users on WhatsApp and Instagram, and a billion adults over the age of 18 each month on TikTok in 2024 [4].

Combined with this is the rapid proliferation of fake news on OSN, which, with the massive use of OSN, has caused significant personal, social, and economic damage [12, 15, 3]. Although there are fact-checking websites (a process conducted by experts, e.g., journalists), they need to be more adequate to address the large amount of misinformation that spreads on OSN, necessitating a framework of artificial intelligence tools to support the assessment of the content of posts on OSN [3, 10, 13, 16].

The coverage of secure OSNs is low and the most popular non-secure OSNs are proprietary, with only a limited part of fake news pipeline processing available. Hence, a research question arises from these gaps: *How can mainstream social networks be made secure?*

To answer that question, we introduce a blockchain-based and artificial intelligence (AI)-powered mechanism to allow any OSN to become secure. The mechanism is structured in terms of services that the OSN can invoke. From the blockchain side, the mechanism can (i) persist the content in an immutable way, (ii) provide a voting mechanism for fact-checkers on the legitimacy of content, and (iii) reward fact-checkers with tokens. From the AI perspective (i) the mechanism has a crawling agent that can check for public repositories and fact-verification agencies, and (ii) the human fact-checking can be complemented with a semi-automatic fact-checking based on machine learning (ML) and natural language processing (NLP), delivering the explainability of the score assigned to that content (likelihood of inauthenticity degree). A proof-of-concept is under development as a Research and Development (R&D) project between the Federal University of Goiás and the Brazilian National Telecommunications Agency (ANATEL). Results reveal that the current version of the tool is capable of supporting blockchain-based decentralized processing of content likely to be fake, using AI mechanisms with explainability, automatically delivering a score of likely fakeness, and supporting immutability, persistence, security, and human fact-checking services.

The paper is structured as follows: Section 2 provides a background; Section 3 introduces the mechanism, Section 4 briefly discusses the proof-of-concept, Section 5 discusses related work and Section 6 concludes the paper with final remarks.

2. Foundations on Secure Social Networks and Fake News Processing

There are several Secure Social Networks (SSNs)[7]. They share common characteristics and are essentially based on blockchain technology [9]. Decentralization is one of the guiding principles of this movement. Decentralization, in that context, consists of a disruptive paradigm, which breaks the technological absolutism of *Big Techs* and the *mainstream* OSN, such as Facebook and X. Centralized structures present multiple drawbacks, including the risk that data can be managed, sold, or stolen without the data owner's active control. This is particularly concerning in light of global legislation on user data protection. A notable example is the scandal involving Facebook, the Cambridge Analytica scandal[8], in March 2018. About 87 million Facebook users used an application that

[8] https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram.

collected profiles of users and friends. The data was handed over to Cambridge Analytica, which analyzed it for political purposes. This is one example of privacy breaches, but it is not the only problem. Another problem with today's social platforms is censorship. Facebook, for example, was banned in some countries, such as China and Iran [9].

Blockchain natively supports decentralization. The blockchain trilemma establishes that Blockchain-based systems should meet three properties: *security, decentralization, and scalability* [2]. Once a blockchain is deployed on peers, the content is replicated and distributed so that there is no single owner of the infrastructure and no single failure point. These technologies have been broadly adopted as resources to fight fake news dissemination [18, 17]. The *fake news* verification process is segmented into five stages [12]: (i) monitoring of OSNs, (ii) extraction of content to be evaluated, (iii) content classification, (iv) interpretation of results and (v) containment of dissemination. Steps (ii) and (iii) can also be *curation*. First, content dissemination in the OSN is observed to detect news to be analyzed. Given a candidate news item to be analyzed, the related media are extracted and the news is classified as likely true or false. The aim is also to interpret the automatic process results using interpretability tools. Subsequently, containment (or coercive) actions are carried out based on the result obtained. Content classification is often known as *fact-checking*, and can be semi-automatic or human-performed. The semi-automatic approach is often conducted using AI mechanisms [13, 8], while the human-performed approach is often carried out by human fact-checkers from fact-checking agencies [5].

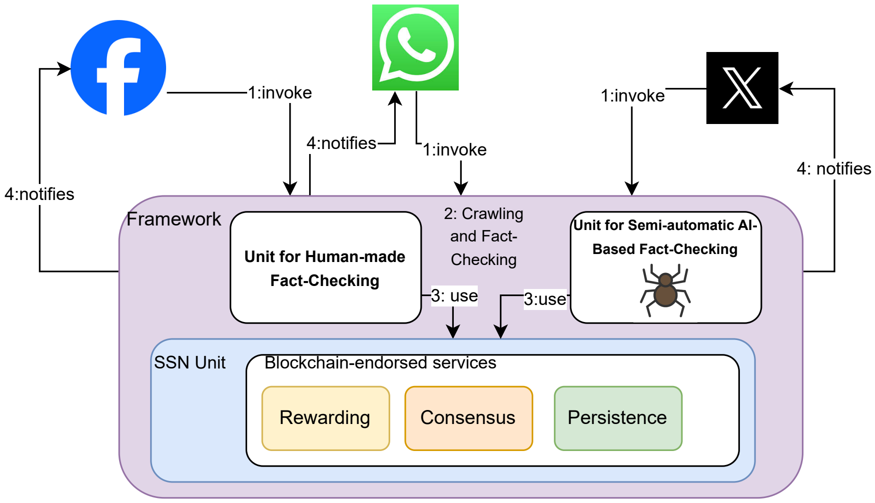

1:invoke
4:notifies
1:invoke
1:invoke
4: notifies
4:notifies
Framework
Unit for Human-made Fact-Checking
2: Crawling and Fact-Checking
Unit for Semi-automatic AI-Based Fact-Checking
3: use
3:use
SSN Unit
Blockchain-endorsed services
Rewarding
Consensus
Persistence

Figure 1. A conceptual illustration of the framework modules and workflow.

3. A Mechanism for Content Curation and Fake News Containment

Figure 1 brings a conceptual illustration of the mechanism and its workflow to support OSNs about fake news processing.

The mechanism supports the whole lifecycle of fake news processing. Monitoring the OSN is indirectly achieved, given that once the OSN itself diagnoses a suspect content, it submits that content for analysis by invoking the framework's services (Step 1 in Figure 1). The extraction of content is conducted by the framework itself during preprocessing steps not illustrated by the figure; an alternative solution is that the OSNs themselves invoke the framework services using textual descriptive content of the news or image for evaluation[9]. After the content is available, the curation and examination processes are started[10]. The human fact-checking process (in the Unit to Support Examination from Human Fact-Checking) is triggered in parallel with the Semi-automatic fact-checking process, boosted by sentiment analysis and AI mechanisms. During this step, a crawling activity (Step 2 in the Figure 1) takes place to check whether other fact-checking agencies and portals have already labeled the same content as fake or real. If so, this interrupts the semi-automatic and human fact-checking process, labeling the content as true or false, linking with the source, and persisting in the blockchain, also returning the content to the invoking OSN. If the content is not found by the crawler, then the semi-automatic fact-checking proceeds. Since the semi-automatic process is often faster than human fact-checking (which demands human analysis and consensus between the fact-checkers from the pool involved in the analysis of the content), the labeling is performed by the Unit for Semi-automatic fact-checking verification. Then, the labeled content is delivered to the human fact-checkers for the final classification. This is needed because the AI mechanism is not 100% accurate, and a misclassification could reduce the tool's credibility. The semi-automatic unit also delivers together an explanation for the results (an inherent part of the fake news processing workflow). After the consensus and decision, the result is delivered back to the requiring OSN (Step 3 in Figure 1) and also persisted in the blockchain to enrich and feed the machine learning mechanisms of the framework. Once the process is concluded, the OSN can take action for containment of dissemination, which is the last step of the fake news pipeline. Note that rewarding, which is a typical resource in SSN, is also supported by the framework, which can be returned as cryptocurrency or tokens, particularly to the fact-checkers who act as content curators.

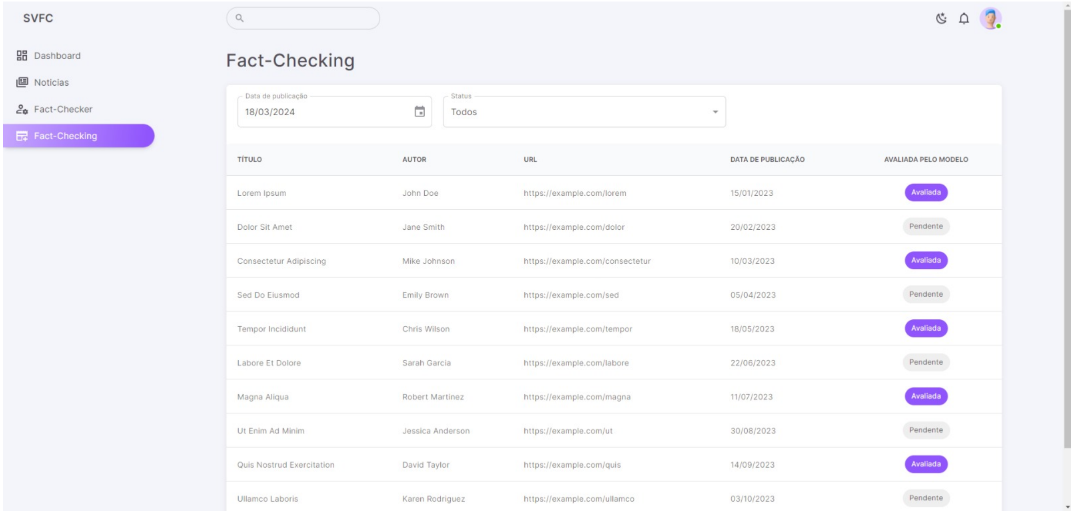
An important advancement here is that, once content being propagated in a particular OSN (such as Facebook) is labeled as false, the framework can also contribute to containment by notifying other OSNs about the likely fakeness of that particular content, contributing to a cross-OSN dissemination of labeled content and containment mechanism.

4. Proof-of-Concept

A Proof-of-Concept of this tool is being developed in the context of a Research and Development (R&D) project. The project is a partnership established between ANATEL and UFG, also supported by the regulatory entities to combat fake news dissemination, particularly in 2024, the year of municipal elections in Brazil. Figure 2 illustrates the user interface for human fact-checkers. The preliminary results reveal that the tool is capable of supporting blockchain-based decentralized processing of likely fake content, using AI

[9] We are not dealing with video, audio and the DeepFake phenomenon yet unless their textual transcription is extracted and submitted to the framework analysis.

[10] The term 'examination' will be used as the translation for *perícia* in Portuguese. We did not find a straightforward satisfactory translation, since the most common translations are 'expertise' (to denote the higher skills of someone in a field) or 'investigation' (to denote the act of exploiting some subject). We understand that *examination* is an inherent part of curation, which is a broader concept.



Fact-Checking

Data de publicação: 18/03/2024

Status: Todos

TÍTULO	AUTOR	URL	DATA DE PUBLICAÇÃO	AVALIADA PELO MODELO
Lorem Ipsum	John Doe	https://example.com/lorem	15/01/2023	Avaliada
Dolor Sit Amet	Jane Smith	https://example.com/dolor	20/02/2023	Pendente
Consectetur Adipiscing	Mike Johnson	https://example.com/consectetur	10/03/2023	Avaliada
Sed Do Eiusmod	Emily Brown	https://example.com/sed	05/04/2023	Pendente
Tempor Incididunt	Chris Wilson	https://example.com/tempor	18/05/2023	Avaliada
Labore Et Dolore	Sarah Garcia	https://example.com/labore	22/06/2023	Pendente
Magna Aliqua	Robert Martinez	https://example.com/magna	11/07/2023	Avaliada
Ut Enim Ad Minim	Jessica Anderson	https://example.com/ut	30/08/2023	Pendente
Quis Nostrud Exercitation	David Taylor	https://example.com/quis	14/09/2023	Avaliada
Ullamco Laboris	Karen Rodriguez	https://example.com/ullamco	03/10/2023	Pendente

Figure 2. A screenshot showing the interface for human fact-checking.

mechanisms with explainability, automatically delivering a score of likely fakeness and supporting immutability, persistence, security and human fact-checking services. The tool supports typical steps of fake news processing, including (i) extraction of content to be evaluated, (ii) human and semi-automatic content curation, and (iii) interpretation of results. Monitoring and containment of dissemination can be indirectly supported once the OSN uses the service and allows notification of results.

5. Related Work

As formerly stated, various SSNs exist [7, 9] and conventional OSNs are also implementing fake news detection mechanisms. Then, the focus herein is on the proposition of mechanisms for supporting existing OSNs to use fake news fighting services.

Dhall et al. (2021) [6] proposes a Blockchain-based Framework that preserves the integrity of the posted content as well as ensures accountability of the author of the post. Apart from many other studies similar to ours, the authors do not propose a new SSN, they provide an infrastructure that can be used by existing OSNs instead. However, they deal only with labeling the content and tracking the origin.

Arquam et al. (2021) [1] establishes a system to check the information authenticity of content circulating in OSNs. Their model can detect misinformation (fake news or rumors), as well as the source of information propagating nodes by applying blockchain technology. However, they still use an intrusive model, accessing the OSN to obtain content and data. We invert that paradigm by offering a non-intrusive service to be used by existing OSNs.

Salim et al. (2021) [14] works closer to our proposal. The authors rely on social media (SM) 3.0, which integrates SM platforms, such as Facebook and Twitter, with the Internet of Things (IoT). Their work (i) uses blockchain to support security and decentralization, and (ii) adopts machine learning (ML). However, their framework involves IoT devices, which is not our focus, and they do not support fake news lifecycle.

To the best of our knowledge, we are not aware of any more recent studies focused on deploying a blockchain-based decentralized infrastructure to enhance existing OSNs with capabilities for fake news detection, containment, and cross-social network services.

6. Final Remarks

To answer the research question *how can mainstream social networks be made secure?*, the main contribution of this paper is introducing a blockchain-based and artificial intelligence (AI) reinforced mechanism to combat fake news in Online Social Networks (OSNs). The mechanism is non-intrusive, and allows OSNs to invoke its services, delegating part of the fake news processing lifecycle to a decentralized solution. A prototype is under developed in a Research and Development (R&D) project conducted in a partnership between ANATEL and UFG. Preliminary results show that the tool supports typical steps of fake news processing, including (i) extraction of content to be evaluated, (ii) human and semi-automatic content curation, and (iii) interpretation of results. Monitoring and containment of dissemination can be indirectly supported once the OSNs use the service and allow notification of results. Future work includes a rigorous evaluation of the tool and a pilot study during the municipal elections of 2024.

References

[1] Arquam, M., Singh, A., and Sharma, R. (2021). A blockchain-based secured and trusted framework for information propagation on online social networks. *Social Network Analysis and Mining*, 11(1):49.

[2] Buterin, V. (2014). Ethereum: A next-generation smart contract and decentralized application platform. *Bitcoin Magazine*, 20.

[3] Caravanti de Souza, M., Silva Gôlo, M. P., Mário Guedes Jorge, A., Carvalho Freire de Amorim, E., Nuno Taborda Campos, R., Marcondes Marcacini, R., and Oliveira Rezende, S. (2024). Keywords attention for fake news detection using few positive labels. *Information Sciences*, 663:120300.

[4] DateReportal (2024). Global social media statistics. Available at: `https://datareportal.com/social-media-users#:~:text=Detailed%20analysis%20by%20the%20team,of%20the%20total%20global%20population.`

[5] de Lima et al., E. A. (2024). Projeto Web 3.0 - Avaliação de Impacto da Web 3.0: Descentralizada, Imersiva, Semântica, Centrada no Usuário e Conectada com o Mundo Ciberfísico; Relatório Técnico - Fake News – Etapa 4 – Relatório 2 – PoC dApp. Technical Report 02-2024, Universidade Federal de Goiás. In Portuguese – Restricted Access.

[6] Dhall, S., Dwivedi, A. D., Pal, S. K., and Srivastava, G. (2021). Blockchain-based framework for reducing fake or vicious news spread on social media/messaging platforms. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(1).

[7] Freni, P., Ferro, E., and Ceci, G. (2020). Fixing social media with the blockchain. In *Proc. of the 6th EAI ICSOTSG*, pages 175–180.

[8] Gomes, J., Graciano Neto, V. V., Barbosa, J., and de Lima, E. A. (2023). A Rapid Tertiary Review at the Fake News Domain. In *XI ERI-GO*, pages 1–10, Goiânia, Brazil. SBC.

[9] Guidi, B. (2020). When blockchain meets online social networks. *Pervasive and Mobile Computing*, 62:101131.

[10] Gôlo, M. P. S., de Souza, M. C., Rossi, R. G., Rezende, S. O., Nogueira, B. M., and Marcacini, R. M. (2023). One-class learning for fake news detection through multimodal variational autoencoders. *Engineering Applications of Artificial Intelligence*, 122:106088.

[11] Li, C. and Palanisamy, B. (2019). Incentivized blockchain-based social media platforms: A case study of steemit. In *Proc. of the 10th WebSci*, page 145–154, New York, NY, USA. ACM.

[12] Morais, J. I. d., Abonizio, H. Q., Tavares, G. M., da Fonseca, A. A., and Jr, S. B. (2020). A multi-label classification system to distinguish among fake, satirical, objective and legitimate news in brazilian portuguese. *iSys - Brazilian Journal of Information Systems*, 13(4):126–149.

[13] Reis, J. and Benevenuto, F. (2022). Detecção automática de desinformação em diferentes cenários: Eleições nos estados unidos e no brasil. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 1–12, Porto Alegre, RS, Brasil. SBC.

[14] Salim, S., Turnbull, B., and Moustafa, N. (2021). A blockchain-enabled explainable federated learning for securing internet-of-things-based social media 3.0 networks. *IEEE Transactions on Computational Social Systems*, pages 1–17.

[15] Santana, C., Claro, D. B., and Souza, M. (2022). Fake news detection in tweets: Challenges and adaptations imposed by the covid-19. *iSys - Brazilian Journal of Information Systems*, 15(1):11:1–11:26.

[16] Testoni, G., Souza, M., Freire, P. M., and Goldschimidt, R. (2021). Um método linguístico que combina polaridade, emoção e aspectos gramaticais para detecção de fake news em inglês. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 151–162, Porto Alegre, RS, Brasil. SBC.

[17] Torky, M., Nabil, E., and Said, W. (2019). Proof of credibility: A blockchain approach for detecting and blocking fake news in social networks. *IJACSA*, 10(12).

[18] Zhou, X. and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.