Truth through Democracy

Please do not distribute this work without consent of its authors

Robin Frasch*
Till Grutschus*
robin.frasch@tum.de
till.grutschus@tum.de
Technical University of Munich
Munich, Bavaria, Germany

ABSTRACT

While "Fake News" is not a new concept, recent advances in the field of Generative Artificial Intelligence make it possible to easily support fictitious claims with sources that appear believable. Even without the use of AI, simple manipulations of the geolocation or the timestamp in the case of an image, for example, hold the potential to shape public belief based on false information. In this work, we present a blockchain-based platform which journalists and news outlets, among others, can use to democratically verify content before publishing it. A verified piece of media can thus be supplemented with a token which links it to the corresponding verdict and the person claiming responsibility for it. As the authority to label new content as either real or fake is distributed among multiple parties, this approach reaches an unparalleled level of trustworthiness. Through a wide application of the proposed system we aim to strengthen fact-based reporting and protect the public from false information. The experimental code for this implementation will be available on GitHub shortly: Project blockedfake

KEYWORDS

blockchain, smart contracts, news verification, truth verification, democracy, accountability

1 INTRODUCTION

Our society, on an interpersonal as well as geopolitical level, is built on trust. But sometimes this trust is challenged or lost, which is when we need a way to prove the factual accuracy of a statement. For decades, an image or video recording was the primary method to prove that a certain event has indeed taken place and was not purely fiction. Consequently, ways to counter this method of truth verification focused on using, for example, doubles to fake the appearance of a politician at a certain event. That way a picture taken of this scene was real, while the scene itself was not. With the development of Generative Artificial Intelligence (Gen AI)[1], the creation of intentionally misleading content has become significantly easier. From AI generated images to elaborate deep fake videos, the quality of the producible content will only grow and so will the danger it poses to the informedness of society.[2][4] An example of this happening was the release of the AI generated "Trump getting arrested" photos in March of 2023. Despite the obvious mistakes in the images, the superficial air of realness was sufficient cause an outrage online, before the pictures could be discredited. [2]

The quality and quantity of this fake content will over time lead to even renowned news outlets publishing it as reliable information, either by accident or by choice, because it can no longer be expected that a single institution is able to detect a fake. This will cause the public to grow increasingly distrustful of its news providers, opening them up to manipulation.

With these advancements in the field of Gen AI it is reasonable to assume, that a race between AI detectors and generators will ensue.[6] However, due to the technical complexity and the cost of training competent AI models, the lead in this race will most likely be taken by large corporations or governments. While the developed system can be reliable and the organisation behind it well-intentioned, it requires the trust of the general public to be effective. To maintain this trust would be extremely difficult, not to mention that critics would be right to be sceptical of a single institution dictating what is accepted as the truth.

In this work we present a truth verification process that relies on democracy and accountability and can thus serve as a viable alternative to centralized verification mechanisms. This approach can be applied to all digital media. The goal is to verify solely the *realness* of a medium, not the factual correctness of arbitrary statements. We aim to achieve this by providing a tool that retains the power of media items to prove statements.

2 REQUIREMENTS

For such a truth verification system to be feasible it has to be one thing: Trusted. All other requirements serve the purpose of creating a system which is reliable and trustworthy by nature and does not require well-intentioned actions by all participants.

2.1 Distributed

To ensure that people across the political and ideological spectrum can rely on such a system, it must not be required to trust a single institution to believe the outcome of the verification process. For that reason, the entire architecture needs to be distributed on a technical as well as personal level. Distributing the ability to determine truth among multiple participants while maintaining the accountability of each individual for his or her actions increases the level of trust the public can reasonably place in the system.

2.2 Immutable

While all centralized computer systems run the risk of being hacked or otherwise compromised, our proposal requires an architecture which reduces the chances of that happening to close to zero. The democratic processes which form the core of the verification process

*Both authors contributed equally to this research.
[1]For more information about Gen AI refer to this article.
[2]For more information about the fake arrest photos refer to this article.

need to be immutable to maintain the integrity of the final decisions and the accountability of all actors.

2.3 Transparent

All steps which run the verification process need to be open source and deterministic. There can be no doubt which actions lead to which reactions, eliminating the possibility for human error completely. Furthermore, every piece of content needs to be traceable to the participant who entered it into the system.

2.4 Fair

To prevent accusations of favouritism, all contributing participants need to have equal opportunities at the time of launch of the system. However, in the interest of true fairness it is also necessary that truthful actions are rewarded and malicious ones have consequences. On an interpersonal level we are able to build trust by consistently being truthful and we can loose this trust by spreading false information. A similar mechanism in the system is necessary to systematize reputation and trust.

3 CURRENT LANDSCAPE

The need to differentiate between true and false information is evident. For that reason there are multiple different products being developed to approach this issue from various angles. In the following, we show existing approaches in this area and analyze the problem that is being solved and potential deficiencies with respect to our stated requirements.

3.1 Harmful Content Detection

With over 80 years worth of video footage being uploaded to the internet every day[7], human content moderators are not able to screen all of it for adherence to the code of conduct of the respective platform. For that reason, AI tools like Unitary [7] are being used to automatically detect and flag harmful content. During time periods in which false information is actively being spread, some social media platforms have even implemented a method which detects specific content in posts and is able to supplement them with a link to official information concerning the topic.
These systems have the purpose to remove hate speech and other harmful content and are designed and maintained by a single company. Thus, they lack the transparency and decentralization to be trusted.

3.2 Community-based fact checking

The Fact Foundation[5] aims to build a decentralized autonomous organization (DAO)[3] and a community of *News Registrars* and *News Validators*, i.e. fact-checkers, and provide a blockchain-based platform for their collaboration, to combat fake news. Contribution is incentivised by a cryptocurrency, the *FACT token*. Because the organization is blockchain-based, decision processes are decentralized, immutable and transparent by nature. Therefore, the platform has the potential of generating trust. However, it runs separate from traditional news outlets and needs to be the point source for all published media before providing a viable alternative and generating trust. By providing a tool to a diverse group of news outlets across the political spectrum which they can use validate specific media items, we aim to overcome this barrier.
Another community-based approach are the community notes, offered by X(Twitter).[8] To be able to leave community notes with additional information on posts, users have to register as so called *contributors*. These contributors can then propose a note, which will be displayed to all other users if enough other contributors rate it as helpful. Eligible to become contributors are all users who satisfy the following criteria: 1) No recent notice of violations of X's Rules. 2) Joined X at least 6 months ago. 3) Has a verified phone number registered. While these criteria improve the reliability of the system and is a good first step, there remains the issue of accountability. The contributors have generally no stake in the system and have no incentive to judge a community note objectively, instead of based on whether it supports their preferred narrative.

3.3 Gen AI Detection

A supporting approach to verifying realness of media items, is the automated detection of AI generated media. Companies like Reality Defender [3] develop algorithms to classify content as real or fake. These algorithms provide an invaluable contribution towards solving the problem outlined in this paper. However, this solution has the same limitations concerning decentralization and transparency as Unitary. We posit that a single entity should not hold the monopoly on content classification. Nonetheless, the algorithms may be a powerful tool employed by actors that are part of the platform, we propose.

4 IMPLEMENTATION ON THE BLOCKCHAIN

The aforementioned requirements make the blockchain the ideal ecosystem to develop such a platform, since blockchain-powered applications are naturally decentralized, transparent and have immutable states.

4.1 Blockchain 101

In principle, the blockchain is a way to store and interact with data in a distributed infrastructure that does not require a central control mechanism. [4] Every piece of information is stored simultaneously on every device, so called nodes, which are part of the chain. The latest state of the stored data is called the world state. New information can be processed by and added to the blockchain in regular time intervals, in which a state transition takes place. During this state transition all participating nodes need to agree via a consensus algorithm on the next state, after which it will be added to the chain. "Added to the chain" means in this case, that the world state is updated on all nodes, but the information about the previous state is saved in the form of a hash. Should now some information on one of the nodes be changed, the resulting hash would differ from all other nodes and the fraud could be easily detected.[1]

[3]For more information about DAOs refer to this article.

[4]For more information about Blockchain refer to this article.

4.2 Benefits of the Blockchain

In the following, we discuss how the blockchain-based application naturally fulfills the stated requirements.

Distribution. The fact that the blockchain operates as a decentralized entity with a built-in consensus algorithm ensures, that the entire architecture is decentralized. This can be reinforced by relying on bare metal hosting in multiple locations instead of being dependent on cloud providers or a small number of server farms. It could be beneficial, that select participants contribute complete nodes to further increase diversity of providers.

Immutability. By linking consecutive blocks of the chain with a unique hash, a manipulation of the on-chain data can be considered almost impossible. Depending on the consensus algorithm practiced by the chain, a majority stake either in computational power or native currency would be necessary to enforce a manipulation which are both infeasible options.

Transparency. When deployed on a blockchain, a verification system is realised as a Smart Contract. Such a Smart Contract is a piece of code which is publicly accessible and immutable once it has been deployed. This enables everyone to analyze the verification process themselves and assures them of the integrity of the system. Furthermore, due to the immutability after deployment, significant changes in the process can not be implemented without setting up a new contract to which all participants would need to be transitioned.

Fairness. Given the aforementioned properties of Smart Contracts and the blockchain, the complete verification process is public which also means that all stakeholders in the system can and should serve as an oversight committee. This reduces the possibility of bias or partisanship. The active contribution of major participants to the infrastructure via bare metal nodes improves the technical and social reliability of the system further while also preventing the perception of a operator-user imbalance. The process of verifying a medium is specifically designed to be fair but not equal to all as the voice of an honest actor must hold a higher value than the voice of a dishonest one. This is achieved through a reputation system, which will be presented in the following section.

5 DESIGN OF A TRUTH VERIFICATION SMART CONTRACT

5.1 Problem Statement

The goal is the creation of a system that enables trusted interaction between untrusting parties, allowing participants to rely on the results without having to trust the good intention of a single institution, be it from industry, academia or politics.

5.2 Design and Architecture

The following section will present the proposed system and its use depending on the role a participant takes. Based on a hybrid blockchain[5], the main features of our solution are the traceability of authorship of a medium and the verifiability of a medium based on a democratic process which incentivizes truthful behavior. By using a hybrid blockchain we are able to regulate the access rights of participants in the system and create two groups of members: The *watchers* and the *actors*. Actors are for example news outlets or journalists and they are entitled to entering media into the system as well as participating in the voting process. Watchers are all other members of society and they are able to see all verified media as well as media pending verification. This distinction is made to prevent an unregulated flood of media entries which would overload the voting capacity of other members and break the system. By using a hybrid blockchain, actors are the only members who can set up nodes with validation rights, so that only those participants with actual stake in the system contribute to the security of the infrastructure. They can not remain anonymous which prevents other anonymous parties from hijacking the system by skewing the outcome of certain votes.

5.2.1 *Traceability.*

From a watchers' perspective, we propose that each medium which is displayed on a website can have a unique token which links it immutably to its creator. With this token a watcher can verify with one click, if this medium has someone claiming accountability or if the source is unknown. This is achieved by an actor uploading the medium to our platform, which mints it as a Non-fungible token (NFT)[6]. This token is then saved on the blockchain where the actor has an account as well, which is used to track the NFTs they are the creator of. The actual image file is saved on the Interplanetary File System (IPFS)[7], which provides a distributed solution to store larger files and keeps the memory consumption on the blockchain low.

5.2.2 *Verifiability.*

The process of verifying a medium relies on *Reputation*, a digital currency used to quantize the honesty of participants within the system. It can be earned, lost and used to increase the weight of ones own vote. This approach creates a purposefully unequal but fair ecosystem, in which the voice of a consistently truthful and reliable actor can be worth more than that of a still unknown or known to be malicious actor. Such a mechanism is necessary, to avoid the need for manual censorship which could always be seen as partisan. That way the consequences for malicious behaviour would be a deterministic part of the process, independent from personal bias or favouritism.

Step by step, the verification process starts with an actor minting a medium as NFT and claiming it to be either *real* or *fake*. The creator of this NFT has the choice to immediately call for a vote or leave the claim unverified for the time being. If they choose to call for a vote immediately, all other actors in the system get the ability to vote on whether they support the initial claim or oppose it. For this, each actor can stake a variable amount of their own reputation to increase the weight of their vote, which is then locked in a deposit until the vote closes. To ensure a fair voting process, the vote has to remain open for a set amount of time and can not close until a minimal percentage of all actors have entered their vote. Once the vote is closed, there are two possible outcomes for the creator and the voters. Option one, their claim or vote were in line with

[5]For more information on hybrid blockchains, refer to this article.

[6]For more information on NFTs, refer to this article.

[7]The Interplanetary File System is a distributed storage platform. More information can be found onthis site.

the final outcome. In this case they earn their staked reputation back, with interest based on the originally staked amount. Option two, the final result disagrees with their initial claim or vote. Then some of their staked reputation is burned. While a vote that has been closed, can not be reopened again, it is possible to initiate a new vote with a differently worded claim based on the same or a similar medium.

5.3 Reputation

As a core feature of the system, Reputation is implemented in the form of a cryptocurrency with the exception, that it can not be bought, sold or transferred in any way by the participants. It is distributed and collected by a separate Smart Contract which also manages the voting process. As the Reputation is used to vote, an actor who has earned more reputation can increase their influence on the outcome of future votes. The goal of this feature is to fulfill the fairness requirement by recreating the interpersonal process of building and losing trust. To repeat, actors are able to earn Reputation by stating claims which end up being verified and voting in line with the final outcome of the vote. This is deemed truthful behaviour and warrants their vote to have a higher impact. As the amount of Reputation to stake on a claim or a vote can be chosen freely by the actor, as long as it doesn't exceed their total amount of earned Reputation, they can also use it to reflect their level of confidence in their choice. This enables newer actors with a low level of reputation but high certainty about a specific medium to have a strong impact on the vote as well. The danger of loosing ones Reputation, however, functions as an effective deterrent against emotional or ill-considered decisions.

6 LIMITATIONS AND COUNTERMEASURES

The fundamental assumption of the truth verification process proposed in this work is, that the final decision achieved by a majority vote is the correct one. All additional features like the concept of earning and loosing Reputation, the flexibility to stake variable amounts of it to vote and the separation in actors and watchers have the purpose to promote knowledgeable and honest actors in the hopes of them coming to a reliable consensus. As they are free to form their decision based on various sources of information and utilize different tools to analyze the medium, the final outcome rests on a wide foundation which mitigates potential biases. However, errors are still a possibility as well as actors voting with malicious intent. The following sections will present possible attacks on the system and ways to counter them.

Pushing a narrative. The success of the proposed system would entail a high level of trust in its results by the general public. This creates an incentive for actors within the system to band together, either voluntarily or motivated by external entities, to achieve their desired outcome and further increase of their Reputation. Thus, it is imperative that the circle of participants is as pluralistic as possible, i.e. news outlets must not be denied access to the system based on their political orientation.

Public Reasoning. It could be beneficial for the actors to provide a reasoning for their vote, as that would let the public know, how the final result came to be. In the case of controversial decisions, these statements could be anonymised and presented on a website to inform the public about the reasons given to support both choices.

Duplicates. The traceability aspect of the proposed system is only a tool to keep track of the entered content and should not be viewed as a legal form of copyright. However, the same content being entered multiple times by different actors can lead to voting fatigue and reduce trust in the process. For that reason, once new content is minted as an NFT and entered into the system, a duplicate search will begin and alert both actors of the possible overlap. However, because the claim and its wording can differ, a pre-existing similar or equal content should not automatically forbid the entering of the new one.

Voting Speed. The primary goal of our truth verification process is to provide reliable actors with a tool to jointly verify content before it is being published. However, under special circumstances, like a controversial image being posted online, it can be necessary to quickly produce a provisional verdict on whether or not this information can be trusted. In those cases, the actor who enters this content into the system can decide to call for an emergency vote, in which case the top 10 to 20% of actors, Reputation wise, would be notified of the request through various means and be asked to place an immediate vote. Should they not be reachable or decline the request, the same call will go to the next in line, until a minimal threshold of participation could be reached.

7 CONCLUSION

In an ideal world there would exist an all knowing entity, trusted by everyone, which could legitimate information with absolute certainty. But as we do not live in an ideal world, we are certain that a transparent and democratic process is the only sustainable way to ensure the integrity of what we view as the truth. Our proposal is shaped by our desire to produce the best possible compromise between a centralization of power and unlimited participation of the public. All of this is done to create a user-friendly system that is fair and trustworthy and is perceived by the public as such, without requiring a detailed understanding of the technical specifications.

REFERENCES

[1] Andreas M. Antonopoulos and Gavin Wood. 2019. *Mastering Ethereum: building smart contracts and DApps* (first ed.). O'Reilly, Sebastopol, CA.
[2] Rotem Baruchin. 2023. GenAI: A Weapon of Mass Disinformation. https://cyabra.com/genai-a-weapon-of-mass-disinformation/.
[3] Reality Defender. 2023. Reality Defender. https://realitydefender.com/.
[4] Dan Lohrmann. 2023. How to Combat Misinformation in the Age of AI. https://www.govtech.com/blogs/lohrmann-on-cybersecurity/how-to-combat-misinformation-in-the-age-of-ai.
[5] Damodar Kalyan Mohit Agadi. 2023. Fact Protocol, A Web3 Verifiability Layer. https://fact.technology/
[6] The New York Times. 2023. How Easy Is It to Fool A.I.-Detection Tools? https://www.nytimes.com/interactive/2023/06/28/technology/ai-detection-midjourney-stable-diffusion-dalle.html.
[7] Unitary. 2023. Unitary. https://www.unitary.ai/.
[8] X. 2023. About Community Notes on X. https://help.twitter.com/en/using-x/community-notes.