Decentralized social media app to combat misinformation

CS 426 - Introduction to Blockchains

Ayush Mallick
CS22BT008
220010008@iitdh.ac.in

Ameya Langer
CS22BT064
220030003@iitdh.ac.in

Pratyush Kaurav
MC22BT015
220120015@iitdh.ac.in

Link to GitHub Repository

Indian Institute of Technology (IIT) Dharwad

Contents

Abstract

Misinformation on digital platforms poses a critical threat to society. From false cures during pandemics to political conspiracy theories, the virality of misinformation can result in widespread panic, mistrust, and real-world harm. While digital platforms have accelerated the spread of falsehoods, emerging technologies like blockchain offer potential countermeasures. This project presents a decentralized social media platform built on Ethereum blockchain that implements verification of data provenance, online identity and reputation management, and incentive mechanisms to mitigate the spread of misinformation. This report details our approach, architecture, and implementation to prove the system's capability in combating misinformation.

1 Introduction

Misinformation poses a serious threat to the digital information ecosystem. With billions of users online, false claims—often spread intentionally for political, ideological, or financial motives—can quickly go viral, reaching a wide audience before fact-checkers can react. Traditional social media platforms have struggled to address this issue due to their centralized nature, opaque content moderation practices, and susceptibility to internal and external biases.

This project proposes a blockchain-based decentralized platform to counter misinformation by ensuring content transparency, source verification, and collective accountability. Through cryptographic techniques and smart contracts, we aim to foster an environment where truth is encouraged, bad actors are penalized, and content integrity is preserved.

2 Literature Review

Several academic and industry-led efforts have explored blockchain's potential for combating misinformation. For example, a Harvard Business Review article highlights how blockchain-based registries can provide immutable records of published media, including text, images, and metadata such as timestamps, authorship, and geolocation, enabling transparent verification of content provenance. Similarly, IBM's Blockchain Transparent Supply (BT Supply) initiative has demonstrated the viability of tracking digital assets and verifying authenticity across complex distribution networks, underscoring the technology's applicability beyond physical goods to digital media.

In the research domain, projects such as Factom and Civil have implemented decentralized timestamping and content validation services, but typically focus on a single dimension—either provenance tracking or reputation management. Factom uses a hash-based anchoring mechanism to timestamp documents on Bitcoin's blockchain, whereas Civil created a token-curated registry for news outlets, relying on token-based voting to curate credible publishers.

3 Problem Statement

Digital platforms have made the dissemination of misinformation alarmingly efficient. False claims—ranging from health remedies like ”boiled garlic water curing COVID-19” to conspiracy theories such as ”GPS chips in currency notes”—spread rapidly and have real-world consequences. This type of misinformation, which is deliberately crafted to mislead for political or financial gain, undermines public trust and safety.

While technology has accelerated the problem, blockchain presents a promising countermeasure. This project aims to design and implement a decentralized system that addresses the multifaceted nature of digital disinformation. Specifically, the platform tackles:

- **Verification of Data Provenance:** Ensuring the origin and authenticity of digital content.
- **Online Identity and Reputation Management:** Verifying the identity of content creators and tracking their credibility.
- **Incentivization of Quality Content:** Using smart contracts to reward accurate information and penalize misinformation.

By integrating these blockchain-enabled solutions, the project seeks to mitigate the spread of misinformation on digital platforms, particularly social media, through transparency, accountability, and community-driven validation.

4 Project Goals

- Design and develop a decentralized social media platform that allows users to create, view, and interact with content in a transparent and tamper-proof environment.
- Use IPFS for decentralized media storage and store unique content hashes on the Ethereum blockchain to track data provenance and prevent duplication or forgery.
- Introduce dynamic reputation scores based on users' content reporting and voting accuracy, helping to distinguish trustworthy users from potential misinformation spreaders.
- Create a verifier network comprising of high-reputation users to participate in a consensus-based validation mechanism to vote on reported content, promoting a decentralized decision-making process for flagging misinformation.
- Use Solidity smart contracts to automate incentivization of good behavior and penalization of misinformation. Rewards are granted for accurate reporting and voting, while penalties apply for false reports or incorrect validation votes.
- Design the architecture to uphold blockchain's core principles by eliminating central points of control, maintaining on-chain transparency, and using scalable off-chain storage (IPFS) for efficient content management.

5 System Overview

Our platform differentiates itself by integrating four key components into a unified and decentralized system for combating misinformation:

1. **Verifiable Content Provenance:** Every piece of content—image or video—is uploaded by the user and stored on IPFS using the Pinata platform. The resulting IPFS hash is recorded on the Ethereum blockchain via the `PostContract.sol` smart contract, ensuring immutability and traceability. During posting, users must declare whether their content is original or sourced.

2. **Community Voting Mechanism(Consensus):** Users can report content they suspect to be misinformation. Once a report crosses a predefined threshold, the content enters a verification phase. Here, a group of randomly selected high-reputation users (verifiers) cast votes to determine whether the flagged content is indeed misinformation. The decision is based on a majority threshold and executed on-chain.

3. **Reputation Management:** Each user is assigned a dynamic reputation score upon registration. This score evolves based on the user's actions—reporting accuracy, validation performance, and general engagement. Accurate reports and correct votes increase the reputation, while false reports or incorrect validations reduce it. Users with high enough scores gain validator privileges in the community consensus process.

4. **Smart Contract Incentives:** Incentivization is governed by the `Misinformation.sol` contract. Users are rewarded with reputation boosts for correctly identifying misinformation or contributing to accurate validation. Conversely, penalties are applied for incorrect reports and votes, disincentivizing malicious behavior and low-quality contributions.

This unified system—comprising verifiable provenance, decentralized consensus, adaptive reputation, and incentive structures—provides a holistic approach to misinformation detection and mitigation.

The post submission workflow functions as follows:

- **If the content hash is already present:** It is accepted as a sourced post without the ”original content” tag.

- **If not found and marked as original:** The post is accepted and tagged as original content.

- **If not found and marked as sourced:** The post is submitted but flagged for review due to its unverifiable origin, initiating a community validation cycle.

Flagged posts undergo verification by a reputation-based validator pool. Depending on the outcome of the vote, posts are either retained, marked as misinformation, or moved to a disputed state. Throughout, transparency, decentralization, and community trust form the backbone of the system.

6 Technology Stack

- **Blockchain Platform:** Ethereum
- **Deployment Network:** Sepolia Testnet
- **Smart Contracts:** Solidity
- **Web3 Integration:** Ether.js
- **Frontend:** React.js, TailwindCSS
- **Backend:** Node.js
- **Storage:** IPFS via Pinata Cloud

7 System Architecture and Design

The architecture of our decentralized social media platform is designed to achieve transparent, scalable, and community-driven validation of content authenticity. This section elaborates on how different components interact to mitigate misinformation, grounded in blockchain principles and cryptographic verification.

7.1 User Interaction and Content Submission

Users begin their interaction through the application's login interface, which leads them to a personalized dashboard. From here, they can post content (images, videos, text) and engage with other users' posts. Each post submission includes metadata where users must indicate whether the content is original or sourced.

7.2 Content Hashing and Storage via IPFS

Upon submission, the platform processes the content by generating a cryptographic hash and uploading the actual content to IPFS using the Pinata Cloud API. The generated hash, which uniquely identifies the content, is recorded on the blockchain through our `PostContract.sol`. This guarantees immutability and prevents tampering.

7.3 Initial Validation through Reputation System

The user who submits content is evaluated by the on-chain Reputation System. If the user has sufficient reputation (a threshold based on historical behavior, voting patterns, and reporting accuracy), their post is immediately marked as **verified**. Otherwise, the post is tagged as **neutral**, making it visible but not validated.

7.4 Flagging and Reporting Suspicious Content

Once published, all content is publicly accessible. Other users can interact with it and flag it if deemed suspicious. When the number of unique reports on a post exceeds a predefined threshold, the post is escalated for formal verification via the consensus system.

7.5 Consensus Mechanism via Verifier Network

Posts flagged for review enter a consensus phase governed by the `Misinformation.sol` contract. A random subset of users from the validator network—comprised of high-reputation participants—is selected to vote on the post's authenticity.

Each validator is presented with the flagged post and supporting context. If the majority votes that the post is real and not misleading, the post is marked as **verified**. If the majority deems it false or misleading, it is labeled as **misinformation** and optionally removed or flagged with a visible warning.

7.6 Incentive Layer and Reputation Adjustment

The incentive system is enforced through smart contracts. Users who accurately report misinformation or vote correctly in consensus receive token-based rewards. Conversely, those who make inaccurate reports or malicious votes are penalized with reputation loss or token deductions. This dynamic encourages honest participation and strengthens the integrity of the validation process.

7.7 System Flow Summary

The overall post-processing flow, as depicted in the architecture diagram, is summarized below:

1. User logs in and posts content via the dashboard.
2. The content is uploaded to IPFS via Pinata, and a hash is generated.
3. The reputation system checks the user's credibility.
    - If validated: the post is immediately verified.
    - If not: the post is marked neutral.
4. If the post receives reports beyond a threshold, it enters the consensus system.
5. Validators vote on the post's authenticity.
    - If deemed real: the post is verified.
    - If false: it is marked as misinformation.
6. Based on voting accuracy, users are rewarded or penalized through smart contracts.

This modular, decentralized design ensures that the system remains censorship-resistant, transparent, and adaptable to the evolving nature of online misinformation.

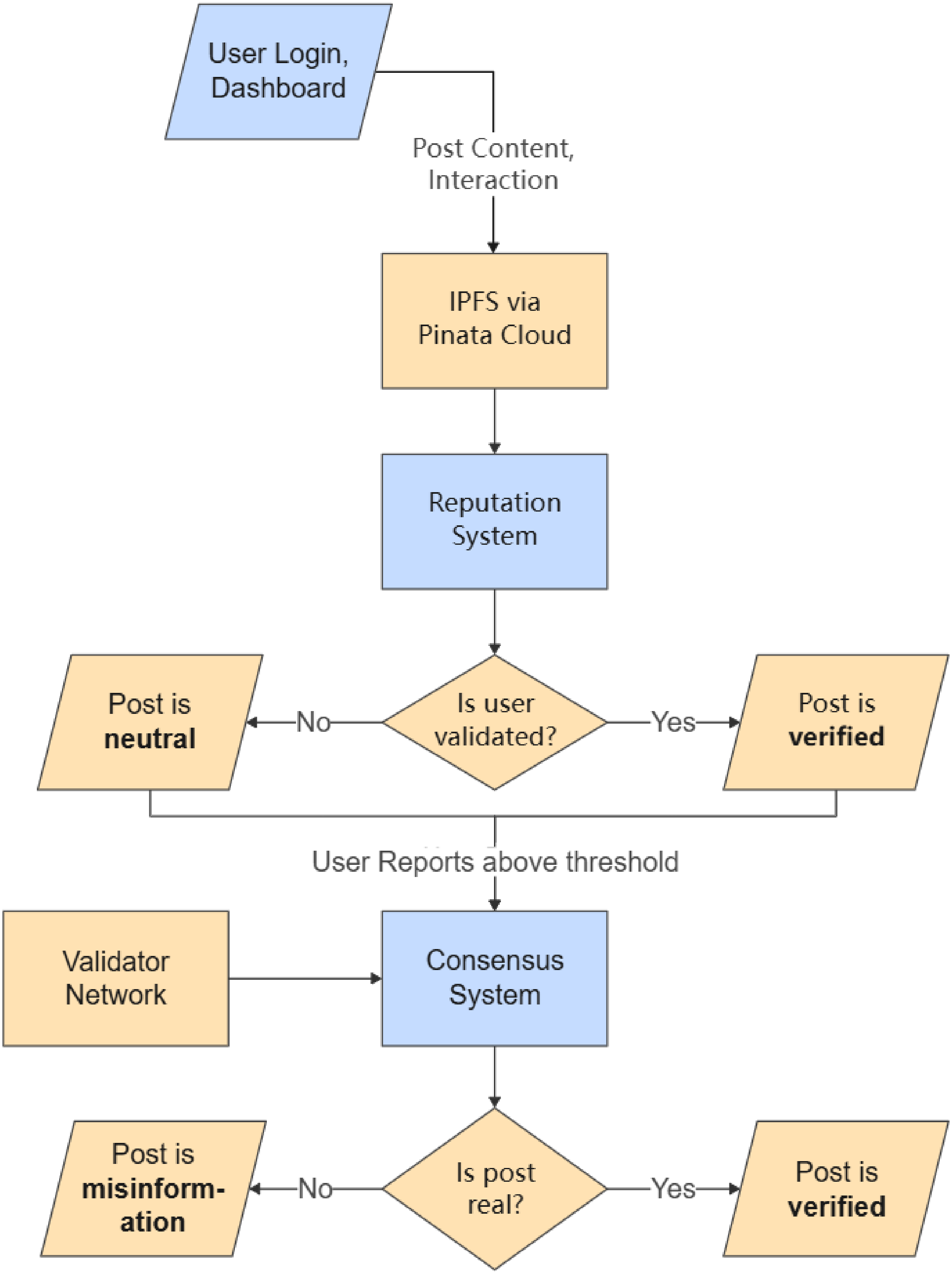

Figure 1: Flowchart of a new post

7.8 Post Provenance Logic

When a user submits content, they must indicate whether it is original (i.e., created by them) or sourced (i.e., obtained from elsewhere). The system computes a cryptographic hash of the content and checks it against the existing records stored on the blockchain:

- **If the hash already exists:** The system recognizes that the content has been previously posted. It is accepted as a sourced post without the original content flag and is not subject to further verification.
- **If the hash does not exist and is marked original:** The post is considered the first known instance of the content. It is stored with an `isOriginal = true` flag to indicate that this is the original submission.
- **If the hash does not exist and is marked sourced:** The system cannot verify the source and flags the post for review, as it may represent potential misinformation or unauthorized use of unverified content.

This logic ensures that content provenance is established and consistently enforced across the platform.

7.9 Effect of Consensus on Reputation System

When a post is reported and enters the consensus phase, validators are selected to assess its legitimacy. Their responses are aggregated, and if a strong consensus is achieved (i.e., a majority agreement above a defined threshold), the platform adjusts reputations accordingly:

- **Correct voters and reporters:** Users whose verdict aligns with the consensus outcome receive an increase in their reputation score. This incentivizes truthful behavior and active participation in moderation.
- **Incorrect judgments:** Users whose decisions consistently diverge from the consensus suffer a decrease in reputation. This mechanism discourages malicious or careless voting and helps maintain the credibility of the verifier network.

The reputation system is entirely on-chain, using smart contracts to maintain transparency and resistance to manipulation.

7.10 Post Interactions

In addition to content creation and moderation, users can engage with posts through a range of interactions:

- Users can like or dislike posts, contributing to informal sentiment feedback.
- Each post displays metadata such as timestamps, the original poster's blockchain address, and the outcome of any consensus validation it underwent.
- Posts also show whether they were flagged as original content to highlight pioneering contributions.

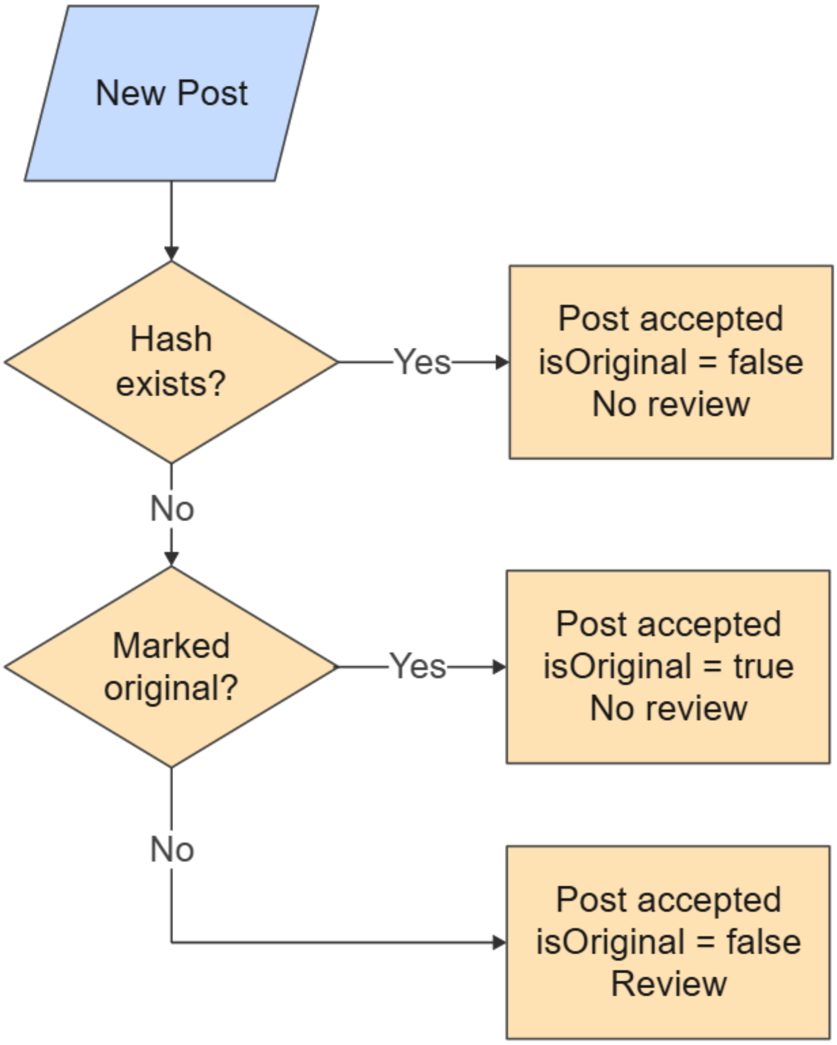

Figure 2: Logic for Post Provenance

8 Smart Contracts Overview

Our system employs two custom smart contracts deployed on the Ethereum Sepolia testnet to manage the core functionalities of the platform.

PostContract.sol

This contract handles the core social media functionality and content verification logic. It is responsible for:

- **User registration and login:** While authentication is handled off-chain for simplicity and scalability, user identity mappings and reputation tokens are managed on-chain.
- **Post creation:** Users submit posts by uploading media content to IPFS (via Pinata), and submitting a reference hash to the blockchain through this contract.
- **Hash lookup for original/sourced logic:** When content is posted, the contract checks whether the submitted hash already exists. If it does, the post is accepted as a sourced repost; if it is new, the post is tagged as original if indicated by the user.
- **Likes and dislikes:** This contract records user interactions such as likes or dislikes, allowing for engagement metrics and feedback mechanisms.

Misinformation.sol

This contract governs the reputation system and community validation workflow. It ensures that moderation is decentralized, transparent, and fair. Responsibilities include:

- **Report and flag mechanism:** Users can report posts they believe to be misinformation. When the number of reports exceeds a threshold, the post is flagged for consensus-based validation.
- **Voting and consensus:** High-reputation users (validators) are selected at random to participate in consensus voting. Their responses are aggregated to determine whether a post should be verified, marked disputed, or removed.
- **Reputation tracking:** Each user's reputation is stored and updated on-chain. Correct voting or reporting increases a user's reputation, while malicious or inaccurate contributions reduce it. This incentivizes accurate participation and helps maintain validator quality.

9 Conclusion

This project demonstrates how decentralized technologies can be applied to combat digital misinformation. By leveraging blockchain for data provenance, community consensus for content validation, and smart contracts for reputation and incentive management, we present a censorship-resistant and transparent alternative to conventional social media moderation. Our design shifts the responsibility of content validation from centralized authorities to a distributed community, enhancing both accountability and resilience. The integration of cryptographic proofs and reputation-based trust ensures that the system scales sustainably while upholding content integrity.

10 Future Work

While the current implementation provides a functional prototype, there are several directions for future enhancement:

- **Integration with Decentralized Identity (DID) protocols:** To ensure stronger user authenticity and reduce the risk of Sybil attacks, future versions can incorporate DID standards such as ERC-725 or Ceramic Network.
- **Expanding the reputation model using machine learning:** Machine learning techniques can be used to dynamically analyze behavioral patterns, detect collusion or voting anomalies, and enhance reputation score calibration.
- **NFT-based tagging of verified content:** Posts that pass consensus and are deemed original could be minted as NFTs, ensuring immutable ownership and provenance while enabling broader interoperability across decentralized platforms.

11 References

1. `https://hbr.org/2021/07/how-blockchain-can-help-combat-disinformation`
2. Ethereum Documentation: `https://ethereum.org`
3. Pinata and IPFS: `https://www.pinata.cloud`