Cluster Computing
https://doi.org/10.1007/s10586-024-04637-7

Connecting the dots between stance and fake news detection with blockchain, proof of reputation, and the Hoeffding bound

Ilhem Salah[1] · Khaled Jouini[1] · Cyril-Alexandre Pachon[2] · Ouajdi Korbaa[1]

Received: 12 February 2024 / Revised: 11 April 2024 / Accepted: 13 June 2024

Abstract
Combating fake news is a crucial endeavor, yet the complexity of the task requires multifaceted approaches that transcend singular technological solutions. Traditional fact-checking, often centralized and human-dependent, faces scalability and bias challenges. This paper introduces a novel blockchain-based framework that leverages the wisdom of the crowd for an authority-free, scalable, automated and reputation-driven fact-checking. Within this framework, stance detection acts as an automated means of opinion retrieval, while the Proof of Reputation consensus mechanism fosters an environment where reputable contributors have greater influence in shaping news credibility. Concurrently, the Hoeffding bound is used to allow the system to adapt to evolving contexts. In contrast to Machine Learning—based approaches, our framework limits the need for periodic retraining to update a model's frozen knowledge of the world. The experimental study conducted on real-world data demonstrates that the proposed framework offers a promising and efficient solution to combat the spread of fake news.

Keywords Stance & Fake news detection · Blockchain technology · Proof of reputation · Hoeffding bound · Decentralized fact-checking

1 Introduction

In the era of ubiquitous Internet and social media platforms, where information of diverse types is readily available and where any point of view can find an audience, access to information is no longer an issue. The key challenges, instead, lie in veracity, credibility, and authenticity [1, 2]. Assessing the veracity of a news story is however a very complex and cumbersome task [3, 4], not only from a machine learning and Natural Language Processing (NLP) perspective, but also sometimes for the most experienced journalists and trained experts [3, 5]. That is why the scientific community often approaches the task from a variety of angles, moving beyond singular technological solutions and breaking down the process into independent sub-tasks. One practical step towards automated fact-checking involves estimating the opinion or point of view (*i.e., stance*) of different news sources regarding the same topic or claim [3]. This work explores the integration of *stance detection* as a key component of a decentralized, fully automated, end-to-end fact-checking pipeline built upon blockchain technology. Our ultimate objective is to derive a *credibility score* for claims based on the stances taken by various sources, with the stances being *weighted* by the respective sources' *reputation*.

Conventional fact-checking systems, often reliant on manual and centralized validation processes (i.e. central authority), are inherently constrained by scalability issues, potential biases, and censorship risks. Similarly, while valuable, conventional machine learning-based approaches often require periodic retraining to update their static and

✉ Khaled Jouini
j.khaled@gmail.com

Ilhem Salah
ilhemsalah53@gmail.com

Cyril-Alexandre Pachon
cyril.pachon@ecole-hexagone.com

Ouajdi Korbaa
ouajdi.korbaa@centraliens-lille.org

[1] MARS Research Lab LR17ES05, ISITCom, University of Sousse, 4011 Hammam-Sousse, Sousse, Tunisia

[2] École Hexagone, Parvis du Colonel Arnaud Beltrame, 78000 Versailles, France

Published online: 27 June 2024

potentially outdated knowledge of the world (as for example Large Language Models do). This necessity arises from the dynamic nature and continuous emergence of news events, posing challenges in scenarios where timely detection of fake news is crucial to prevent its spread. Additionally, conventional ML models heavily rely on manually annotated training data, which can be challenging to obtain in both quantity and quality, especially for specialized topics and live news events. Conventional ML-based approaches also often assume a centralized validation process and hence share the same limitations of conventional fact-checking regarding scalability issues and/or potential bias.

In recent years, blockchain technology, with its scalability and authority-free nature [6], is being harnessed to combat the spread of fake news [7]. Existing blockchain-based approaches have primarily focused on tracking the provenance and modification history of news articles and associated images [8–11] or on assigning veracity labels to them [12–14]. Blockchain-based veracity labeling approaches typically incorporate voting mechanisms and/or machine learning models to assess the credibility of news. While these efforts have yielded valuable insights and advancements, to the best of our knowledge, no prior research has employed blockchain technology to bridge the gap between stance detection and fake news identification, nor provided a statistical guarantee to control the risk that the label associated to a news changes over time when additional evidence (*e.g.*, votes or stances) becomes available.

In order to address the limitations of conventional fact-checking and machine learning-based approaches, this work introduces a novel decentralized framework built upon blockchain technology. By decentralizing the fact-checking process and harnessing the wisdom of the crowd, we aim to mitigate biases, improve scalability, and enhance the robustness of fake news detection. Within this framework, the credibility score of a news item is derived from the stances of participants engaging with it. To improve scalability and reduce user input, stance detection is used as an automated means to infer the opinions of participants towards a claim. A salient feature of our approach is the nuanced weighting (*i.e.*, reputation) assigned to each participant. This weight is determined according to their past positive interactions (*i.e.* supporting/sharing true news or refuting false news) and negative interactions (*i.e.* supporting/sharing false news or refuting true news). To safeguard the system against potential malicious actors and ensure trustworthy news assessment, our framework adapts the "Proof of Reputation" consensus mechanism to fake news detection by only allowing contributors with positive reputation to participate in the consensus for assessing news veracity. Stances typically arrive at irregular time intervals and waiting for all stances can delay fake news detection. To enable timely detection, our framework uses the Hoeffding bound as a statistical guarantee for labeling news items even with incomplete information.

The remainder of this paper is organized as follows. Section 2 briefly reviews the main existing blockchain-based fake news detection approaches. Section 3 presents our reputation-driven stance-based fake news detection framework. Section 4 presents an experimental study validating our approach. Finally, Sect. 5 concludes the paper.

2 Related work

Early efforts to leverage blockchain technology for fake news detection primarily focused on tracing information provenance and incentivizing the creation of reliable content. Noteworthy systems in this category include Block-Proof [10], AnsaCheck [8], and the NY Times provenance project [9]. Inspired by several studies that have demonstrated the effectiveness of collective wisdom in surpassing the biases and limitations of individual experts [13, 15–17], more recent approaches have shifted towards the integration of crowdsourcing, voting, and machine learning models [18]. Our research aligns with the latter category of approaches. In this sequel, we mainly focus on two recent approaches, [12, 14], that effectively combine crowdsourcing and blockchain technology to combat fake news.

The study [12] proposes a decentralized social network using blockchain technology to manage news and an entropy-based incentive mechanism to encourage users to contribute to the system. The system uses a group of trusted human appraisers to evaluate the authenticity of news. The appraisers are chosen through an entropy-based algorithm that selects appraisers whose votes are likely to be more informative and less predictable. If a majority of appraisers vote that a news is fake, the news is flagged as such and its visibility is reduced. Appraisers are rewarded with tokens for flagging suspicious content, providing feedback on the accuracy of news articles, and participating in the consensus process. Token rewards are contingent upon the entropy of their actions. Users have the ability to rate appraisers based on the accuracy and helpfulness of their feedback. Appraiser reliability is assessed using various factors, including the percentage of correctly identified fake or real news articles, variance in votes for different news items, and reputation measured through feedback from other users. The primary concern of the approach of [12] is the reliance on manual voting by pre-selected appraisers, a method conflicting with natural interactions in social networks, and raising concerns regarding potential bias and scalability limitations. Furthermore, the effectiveness of the methodology depends on expert validators,

a form of central authority contradicting blockchain's decentralized spirit.

The approach proposed in [14] introduces a decentralized environment where users post and vote on news stories, utilizing the benefits of collective intelligence for information validation. The proposed approach incorporates Lyapunov exponents and Shannon entropy as key metrics to ascertain the equilibrium of votes on a given story and to decide when to end the voting procedure. Lyapunov exponents are employed to provide a quantitative measure of the degree of chaos in a system while Shannon entropy is used to measure the disorder of votes. Upon reaching equilibrium, the status (true or false) of a news story is determined through the utilization of machine learning classifiers. The system employs a two-stage learning architecture, comprising an Action Classifier and a Story Classifier. The Action Classifier, implemented as a convolutional neural network, is designed to discern malicious behavior, with a particular focus on identifying the involvement of bots. The Story Classifier utilizes Long Short-Term Memory (LSTM) networks to predict the label of a story based on the collective behavior of users who have voted on it. While presenting several interesting ideas, the approach of [14] exhibits two notable shortcomings. Firstly, the reliance on black-box machine learning models might pose challenges for users in understanding how decisions are made. This lack of transparency could raise concerns about the system's fairness and impartiality, crucial for gaining and maintaining user trust critical in tasks involving user-generated content. Secondly, the approach lacks a statistical guarantee when determining the veracity of a claim. This is particularly problematic since in the proposed approach the labels associated with news are immutable. Consequently, if credible votes later confirm a news story to be true after initially being labeled false, there is no recourse to alter its status. This rigidity coupled with the lack of statistical guarantee can be a limitation when dealing with evolving contexts.

It is worth noting that, to the best of our knowledge, our framework stands as the sole blockchain-based approach that avoids manual voting, provides a statistical guarantee when determining news veracity, and features dynamic and evolving reputation and credibility scores.

3 News credibility scoring with stances and reputations: a blockchain-based framework

3.1 Framework overview

In this paper, we present a novel framework that leverages stance detection and blockchain technology to establish an autonomous and decentralized system allowing to assign credibility scores to shared news and reputation scores to participants. To address the current limitations of blockchain platforms, we assume that news texts and machine learning models are stored off-chain and only participants' stances regarding news are stored on-chain. Our framework can be seen as a generalization of voting systems, wherein votes are inferred from stances rather than being explicitly provided by users. As highlighted by various studies, including [19], sharing and commenting news are the prevalent forms of news engagement among social media users and are more participatory and intuitive than voting and other paralinguistic digital affordances, which are more passive and impersonal actions.

The main steps of the news veracity assessment process are outlined in the flowchart of Fig. 1. As depicted in Fig. 1, participants interact with the system by commenting or sharing news items (Step 1). The system, in turn, detects the participant's stance regarding the news and ensures that the stance is either "Support" or "Deny" (Step 2). If not, the process concludes; otherwise, the stance is logged in

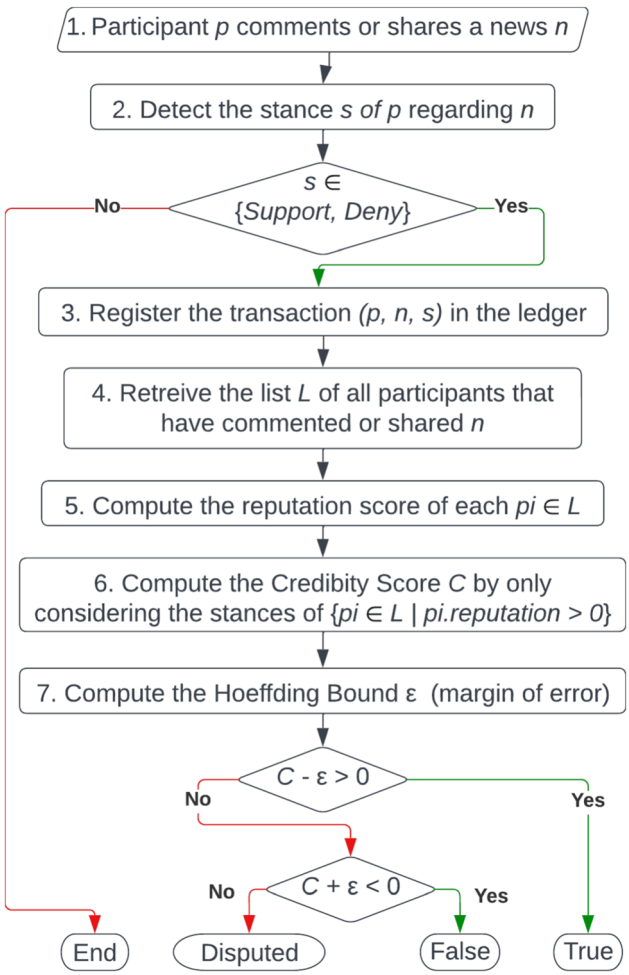


Fig. 1 Comment handling process

the distributed ledger ("Neutral" and "Unrelated" stances are not registered) (Step 3). Serving as a platform for tracking stance history, the distributed ledger facilitates the determination of both positive contributions and negative contributions made by each participant. The distributed ledger ensures that the history of contributions cannot be tampered with, thereby enhancing the overall integrity of the system. The smart contract associated with the distributed ledger enables the determination of participants' reputation and news credibility scores on the fly.

The credibility score of a news item depends on the stances and reputation scores of participants who commented or shared it. Reputation scores draw inspiration from conventional cryptocurrency balances, dynamically increasing with positive contributions and decreasing for negative actions. These "balances" reflect the overall behavior of participants within the system. The reputation score of a participant is computed by scanning their entire stance history (Step 5). This involves summing up positive contributions (similar to incoming transactions in a standard blockchain), deducting negative contributions (akin to outgoing transactions), and subsequently normalizing the result. The reputation score reflects therefore the trustworthiness of participants based on their interactions within the system.

Reputation acts as a trust indicator, guiding the consensus on news veracity assessments (Step 6). The consensus mechanism adopted in our framework bears similarities with Proof of Reputation (PoR) and Proof of Stake (PoS), akin to how these consensus mechanisms verify transactions in conventional blockchain. As shown in Fig. 1, only participants with positive reputations are allowed to contribute to the consensus, *i.e.*, only the stances of participants with positive reputations are taken into account when determining the credibility score of a news. This is analogous to how stakeholders in Proof of Stake systems exert more influence in blockchain consensus. Analogous to reputation scores, news credibility scores are dynamically recalculated as new stances become available. The credibility score C of a news item n is computed as the sum of the stances of participants who have a positive reputation score, where each stance is weighted according to the corresponding participant's reputation score.

After calculating a credibility score, the next step involves determining the status of the news story (that is, whether the news is "True," "False" or in a "Disputed" state) (Step 7). Accurate status determination is crucial for maintaining result consistency, as associating an incorrect label with news directly impacts the reputation scores of participants, potentially leading to adverse consequences for the entire system. While status determination is straightforward when all stances regarding a news item are available, in real-world scenarios and dynamic contexts, stances often arrive at irregular time intervals, and decisions need to be made dynamically before all stances become available. Using fixed credibility thresholds for status determination introduces various drawbacks, including over-generalization, limited adaptability to evolving data, and potential biases. In response to these challenges, our approach uses the Hoeffding Bound [20] to derive an adaptive margin of error ϵ with a statistical guarantee. As shown in the flowchart of Fig. 1, our framework evaluates if the credibility score, considering ϵ, aligns with labeling the news as "True," "False" or "Disputed."

In the remainder of this section, we first provide a detailed explanation of how the reputation and credibility scores are computed. Subsequently, we elaborate on how the Hoeffding Bound is used for determining news status.

3.2 On-demand assessment of participant reputation and news credibility

Participants' understanding of the reputation and credibility scores is crucial for ensuring transparency and encouraging active participation. Consequently, we have opted for highly straightforward formulas that make the concepts of reputation and credibility intuitive and easily understandable. By presenting these simplified formulas, participants can gain a clear understanding of how their actions directly influence their scores and contribute to the system's assessment of news article credibility.

3.2.1 Reputation score

As mentioned earlier, in our work we do not explicitly store participant reputation and news credibility scores. Instead, we employ an on-demand, real-time reputation balance approach, akin to traditional blockchain systems. Similar to a standard blockchain's ledger that maintains records of all transactions, we keep a record for each participant interaction within the network. Each interaction record includes the participant's address, news identifier, and the participant's stance regarding a news. The reputation score of a participant is determined based on the history of their stances regarding news items identified as either true or false. Since reputation is linked to past behavior and contributions, participants have a vested interest in maintaining a positive reputation. Additionally, the reputation score makes it difficult for malicious actors to manipulate the network by creating multiple identities (*i.e., Sybil attacks*), as their influence is constrained by cumulative reputation scores.

Algorithm 1 Reputation score calculation

```
function getReputation(particiantHash)
    for each news in participants[participantHash].commentedAndSharedNews do
        status ← newsMap[newsID].status
        stance ← participants[participantHash].stances[newsID]
        if (status × stance > 0) then
            positiveContributions++
        else if (status ≠ 0) then
            negativeContributions + +
        end if
    end for
    totalContributions ← positiveContributions + negativeContributions
    if (totalContributions) > 0) then
        return (positiveContributions − negativeContributions)/(totalContirbutions)
    end if
    return 0
end function
```

Algorithm 1 outlines the calculation process for the reputation score. The stance of a participant regarding a news can either be 1 for "Support" or -1 for "Deny", while news status can either be 1 for "True", − 1 for "False" or 0 for "Disputed". Consequently, the resulting reputation score ranges from − 1 to 1, with the magnitude indicating the participant's perceived trustworthiness.

Given the impact of stances on both reputation and credibility scores, we advocate presenting the stance classification results to participants. This approach not only enhances transparency and engagement, but also provides a foundation for reinforcement learning and empowers participants to adjust classification errors.

3.2.2 Credibility score

The credibility score assesses the reliability of a news based on participants' interactions and consensus. Similar to reputation scores, credibility scores are not explicitly stored but calculated on the fly, by only considering the stances of reputable participants (*i.e*. participants with a positive reputation) and their respective reputation scores. This participatory reputation-based approach enhances the system's ability to identify and combat fake news by leveraging the wisdom of the crowd and giving greater weight to trusted participants. The credibility score of a news, denoted C, is calculated using the following formula:

$$C = \frac{\sum_{i=1}^{N}(\text{Stance}_i \times \text{Reputation}_i)}{N}, \tag{1}$$

where N is the total number of participants with positive reputations engaging with the news, $\text{Stance}_i \in \{-1,1\}$ and $\text{Reputation}_i \in [-1,1]$ are respectively the stance and the reputation score of participant i.

The above formula yields a value $C \in [-1,1]$. This normalized score serves as an indicator of the overall trustworthiness of the news article. A positive score suggests a general consensus in favor of the news, while a negative score indicates a general consensus against the news. The magnitude of the score reflects the strength of this consensus.

3.3 Hoeffding bound—based news credibility assessment

The Hoeffding Bound [20] states that, with probability $1-\delta$, the true average of a random variable of range R will not differ from the estimated average after N independent observations by more than :

$$\epsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2N}} \tag{2}$$

The Hoeffding bound is useful because it is assumption-free and holds true regardless of the distribution generating the values, and depends only on the range of values R, the number of observations N and the desired confidence $1-\delta$. With R and δ fixed, the only variable left to change the Hoeffding bound ϵ is the number of observations N. As N increases, ϵ will decrease, in accordance with the estimated average getting ever closer to its true value. From a practical point of view, the confidence level $\delta \in [0,1]$ allows to fine-tune the "acceptable" margin of error ϵ.

For the purpose of deciding on the status of a news, the random variable being estimated is $p = \text{Stance} \times \text{Reputation}$, the stance of a participant regarding a news, weighted by its reputation score. The average of p corresponds to $C = \frac{\sum_{i=1}^{N}(\text{Stance}_i \times \text{Reputation}_i)}{N}$, the credibility score of the news. Since the *Stance* can either be 1 or -1 and *Reputation* ranges from -1 to 1, C ranges from -1 to 1, meaning that $R = 2$.

When $C - \epsilon > 0$, this implies that C is confidently positive, and the news can be considered "True": *i.e*., we can assert with a confidence level of $1-\delta$ that C will

always remain positive, and hence that the news will remain "True" when additional stances become available. Similarly, when $C + \epsilon < 0$, this implies that C is confidently negative, and we can assert, with a confidence level of $1 - \delta$, that the news is "False". As shown in the flowchart of Fig. 1, when the conditions making a news either "True" or "False" are not met, the news is considered to be in a "Disputed" state.

Example 1 Let's consider the following stances regarding a claim and suppose that $\delta = 0.25$.

i	Stance$_i$	Reputation$_i$
1	1 (Support)	0.1
2	− 1 (Deny)	0.9
3	− 1 (Deny)	0.9

According to Eqs. 1 and 2, the credibility score and the margin of error are resp. −0.57 and 0.96. Since $|C| < \epsilon$, we cannot conclude that the news is False and it remains in a "Disputed" state. The relatively large value of the margin of error ϵ is essentially due to the low number of participants.

Example 2 Let's suppose that 2 more participants commented the news as follows.

i	Stance$_i$	Reputation$_i$
1	1 (Support)	0.1
2	− 1 (Deny)	0.9
3	− 1 (Deny)	0.9
4	− 1 (Deny)	0.9
5	− 1 (Deny)	0.9

In this example $C = -0.7$, and $\epsilon = 0.69$. Since $C + \epsilon < 0$, we conclude, with a confidence level of 75% (corresponding to $1 - \delta$), that C will always remain negative as new stances become available and consider the news as "False."

4 Experimental study

4.1 Tools and dataset

One of the major difficulties we faced in our experimental study was the lack of datasets covering all aspects of stance-based fake news detection. Existing fake news detection datasets do not yet support a fully automated end-to-end setup, and often treat stance detection, source credibility, and news classification as separate tasks. This explains why the majority of existing blockchain-based fake news detection approaches primarily concentrate on attack scenarios rather than the system's efficiency in identifying fake news.

Despite the fact that the RumourEval dataset [21] does not associate comments with the corresponding participants, and, hence, does not allow us to track participants' good and bad contributions, we used it in our experiments because it is one of the very few datasets that provides social media conversations annotated for both stance and veracity. The dataset contains 3342 conversations from Twitter covering a variety of topics, including politics, natural disasters, and celebrity gossip. RumourEval contains two subtasks, "Subtask A—Stance Detection" and "Subtask B—Veracity Detection". The goal of Subtask A is to classify the stance of a social media post towards a rumored claim (SDQC: Support, Deny, Query, Comment). The goal of subtask B is to classify the veracity of a rumored claim (True, False, Undetermined). Figure 2 provides an example retrieved from the dataset. Due to class imbalance, the RumourEval 2019 benchmark [21] adopts the macro-averaged F1 metric to evaluate competing approaches. The relatively low F1-scores achieved by most competing approaches (Table 4) underscore the non-trivial nature of the task. The benchmark creators [21] offer baseline models for both subtasks. For Subtask A, they provide a Keras implementation of BranchLSTM [22], the winning system of RumourEval 2017 Subtask A. Utilizing the conversation's structure, BranchLSTM divides it into linear branches and employs LSTM layers to process tweet sequences, generating a stance label at each time step. For subtask B, [21] modified BranchLSTM to produce a single output for each branch. The veracity prediction for the thread is then determined by majority voting over per-branch outcomes [21]. In this study, we configured our framework using BranchLSTM [22] as the stance detector, implemented the core logic using a Solidity 0.9.0 [23] smart contract, and conducted testing using Ganache [24].

For the purpose of tracking participants' actions, we simulated 42 participants,[1] generating approximately the same number of stances. Randomly assigning comments to participants could lead to borderline low magnitudes reputations (*i.e.*, close to zero). Such random assignment fails to accurately mirror real-world scenarios where participants exhibit varying degrees of reliability and where we encounter a mix of reputable participants, malicious actors, and those falling in between. Accordingly, to introduce diversity and capture the heterogeneity present in real-world social networks, we categorized participants into

[1] With 41 being the highest number of comments on one tweet in the dataset.

Fig. 2 Example extracted from the RumourEval dataset [21]

u1: These are not timid colours; soldiers back guarding Tomb of Unknown Soldier after today's shooting #StandforCanada PICTURE **[support]**
- **u2:** @u1 Apparently a hoax. Best to take Tweet down. [deny]
- **u3:** @u1 This photo was taken this morning, before the shooting. [deny]
- **u4:** @u1 I dont believe there are soldiers guarding this area right now. [deny]
  - **u5:** @u4 wondered as well. Ive reached out to someone who would know just to confirm that. Hopefully get response soon. [comment]
    - **u4:** @u5 ok, thanks. [comment]

three reputation score ranges: lower than -0.5, higher than 0.5, and between -0.5 and 0.5. These categories represent, respectively, malicious participants, reputable participants, and borderline participants. We also explored varying fractions within each category.

4.2 Results and discussion

4.2.1 Results

The primary objective of our experimental study is to assess the overall effectiveness of our reputation-driven, stance-based fake news detection framework and to identify the conditions necessary for its success. To this end, we conducted our experiments with four specific objectives: (*i*) illustrating and quantifying the impact of the efficiency of the stance detector on the system performance; (*ii*) examining the influence of varying fractions of malicious and reputable participants; (*iii*) exploring the system sensitivity to variations in δ, the confidence level associated with the Hoeffding bound; and (*iv*) analyzing the impact of the length of participants' historical trajectories.

To illustrate the impact of the stance detector on overall performance, we compare the results obtained using the BranchLSTM stance detector [22] (Table 1) to those that would be achieved when all stances are correctly determined (Table 2). The latter scenario corresponds to a voting system where participants explicitly provide their stances on news. It's worth mentioning that such a scenario is covered by our framework, as participants have the ability to correct the detected stances. Tables 1 and 2 show that improving the performance of the stance detector module has the potential to yield an average improvement of approximately 12.2‘% in fake news detection.

To evaluate the influence of different participant distributions on the framework's performance, we examined four scenarios represented by distributions $Dist_1$ to $Dist_4$: $Dist_1$: [0.05 *Reputable*, 0.9 *Borderline*, 0.05 *Malicious*], $Dist_2$ [0.1 *Reputable*, 0.8 *Borderline*, 0.1 *Malicious*], $Dist_3$: [0.15 *Reputable*, 0.7 *Borderline*, 0.15 *Malicious*], and $Dist_4$: [0.2 *Reputable*, 0.6 *Borderline*, 0.2 *Malicious*]. As expected and reflected in Tables 1 and 2, the framework effectiveness is not impacted by the fraction of malicious participants. However the effectiveness is highly influenced by the fraction of reputable participants, with poor performance observed at very low fractions (*e.g.* $Dist_1$). As shown in Tables 1 and 2, a 5% increase in the fraction of reputable participants results in an average 8.75% improvement in the F1-score obtained by the system.

As mentioned earlier, the Hoeffding bound is known for its conservative nature. This characteristic is confirmed in our experimental study, where relatively low values of the confidence level $(1-\delta)$ result in good performance. This is attributed to the fact that when δ is low the Hoeffding bound refrains from making decisions regarding the status of news until it reaches a high level of confidence. Consequently, a significant portion of news remains categorized as "Disputed" rather than being identified as "True" or "False". As demonstrated in Tables 1 and 2, regarding this specific dataset, the system's performance stabilizes or exhibits marginal improvement starting from a $\delta = 0.4$.

The continuous cycle of interaction, stance classification, reputation and credibility scores updates strengthens the overall effectiveness of our framework over time. However, a common challenge for crowdsourcing and user-generated data systems is the cold start problem, arising when there are few participants and stances during the initial stages. To evaluate the impact of the length of participants' interaction history on system performance, we considered scenarios where only 50% and 75% subsets (randomly selected) of the training dataset are used to derive the initial reputation scores. As shown in Table 3, the system's performance improves significantly as it matures, with participants experiencing less fluctuations in their reputation scores.

4.2.2 Discussion

Our framework demonstrates promising results on the RumourEval 2019 benchmark. However, as is often the case, benchmarks may not fully capture the complexities of the real world. This underscores the importance of considering the factors discussed below.

The findings from our experimental study have shed light on the key factors influencing the efficiency of our framework. In particular, results revealed that, while participants with low reputations have negligible influence on the system's performance, the overall effectiveness hinges critically on the proportion of reputable participants. This reliance can be attributed to two factors: the Proof of

Table 1 Macro-averaged F1 achieved by our framework using BranchLSTM [22] as Stance Detector

δ	$Dist_1 : [0.05, 0.9, 0.05]$ (%)	$Dist_2 : [0.1, 0.8, 0.1]$ (%)	$Dist_3 : [0.15, 0.7, 0.15]$ (%)	$Dist_4 : [0.2, 0.6, 0.2]$ (%)
0.2	15.63	27.75	30.59	34.07
0.3	19.74	31.47	37.88	38.55
0.4	24.37	34.59	40.77	49.83
0.5	28.72	37.57	42.58	50.22
0.6	28.72	37.90	49.29	50.22

Table 2 Macro-averaged F1 achieved by our framework when all stances are correctly identified (*i.e.* similar to voting)

δ	$Dist_1 : [0.05, 0.9, 0.05]$ (%)	$Dist_2 : [0.1, 0.8, 0.1]$ (%)	$Dist_3 : [0.15, 0.7, 0.15]$ (%)	$Dist_4 : [0.2, 0.6, 0.2]$ (%)
0.2	21.64	30.40	33.28	41.49
0.3	30.50	37.15	39.90	51.97
0.4	39.20	41.03	49.29	78.20
0.5	40	49.29	54.83	78.20
0.6	40	49.29	67.04	78.20

Table 3 Macro-averaged F1 score as a function of the fraction of the training dataset used to derive initial reputations ($\delta = 0.5$)

Fraction	$Dist_1 : [0.05, 0.9, 0.05]$ (%)	$Dist_2 : [0.1, 0.8, 0.1]$ (%)	$Dist_3 : [0.15, 0.7, 0.15]$ (%)	$Dist_4 : [0.2, 0.6, 0.2]$ (%)
Proposed approach with BranchLSTM stance detector				
0.5	9.97	13.65	18.29	22.36
0.75	17.52	22.31	26.24	31.74
Proposed approach with voting				
0.5	18.70	24.96	31.03	38.48
0.75	27.14	36.22	43.72	54.62

Reputation mechanism, which filters out stances from low-reputation participants, and the Hoeffding bound, which requires a minimum level of corroborating evidence (support) before reaching a verdict on news veracity. This highlights a potential weakness in scenarios with sparse reliable contributors or during the initial stages, as the system may struggle in reaching a consensus, leaving a significant portion of news in "Disputed" status.

Results also showed to what extent the efficiency of the stance detector, represented by the BranchLSTM model, influences the system's overall performance (with an observed average potential improvement of approximately 12.2% in fake news detection). This underscores the need for further exploration of more sophisticated stance detectors, especially as stance detection in real-world scenarios may face challenges, such as sarcasm and evolving language patterns.

Regarding the Hoeffding bound, its conservative nature introduces a trade-off between confidence levels and precision. As show in our experimental study, lowering the confidence level prompts the Hoeffding bound to make decisions, resulting in more decisive judgments and improved F1-scores (as fewer news items are classified as "Disputed"). However, this also increases the risk of misclassifications, especially in situations with limited data or high uncertainty. While this adaptability allows the framework to be fine-tuned based on specific application requirements, it also prompts considerations regarding the balance between decisiveness and precision.

As shown in Table 3, the effectiveness of our framework is impacted by the length of participants' historical stance data. To mitigate this cold start problem, two "airdrop" strategies (using reputation as a token) deserve further exploration: (*i*) Uniform seeding: similar to some existing cryptocurrency blockchains, a (low) fixed reputation score is uniformly distributed to new participants; and (*ii*) Weighted seeding: high initial reputation scores are assigned to trusted entities such as fact-checking

Table 4 Macro-averaged F1 achieved by the baseline model and the best-performing models

Approach	Score (%)
Baseline Majority Class [21]	22.41
BranchLSTM [22]	49.29
eventAI [25]	57.65
Fine-tuned Longformer [26]	58.68
Shared Multi-channel Interactions (MTL-SMI) [27]	68.5
Proposed approach with BranchLSTM stance detector ($\delta = 0.5$)	up to 50.22
Proposed approach with voting ($\delta = 0.5$)	up to 78.20

organizations, governmental press agencies, and academic institutions. These "seeded" participants would serve as anchors for the reputation system during the initial phase, providing a foundation for credibility assessment until new participants establish their reputation through interaction.

While achieving comparable performance to machine learning approaches (Table 4), our framework stands out due to its fully decentralized and scalable nature, coupled with its dynamic reputation-based news credibility assessment. This dynamic approach offers a more nuanced evaluation of news veracity compared to static ML models and fosters an environment where participant influence aligns with their established reliability. These features are achieved without relying on a central authority or requiring frequent model retraining to update a frozen knowledge of the world. This combination positions our framework as a promising solution in the complex landscape of fake news detection.

Finally, it is worth to noticing that the proposed framework and conventional ML-based approaches are not mutually exclusive. The reputation and credibility scores derived within our framework can serve as valuable inputs to ML models. These scores provide additional contextual information that has the potential to enhance the performance of ML algorithms in detecting fake news.

5 Conclusion

This paper introduces a novel framework that integrates stance detection and blockchain technology for a decentralized, scalable and collaborative fact-checking. The proposed reputation-driven system operates autonomously, eliminating the need for human intervention or reliance on a central authority. Within this framework, stance detection acts as an automated means of opinion retrieval regarding a claim, facilitating evidence gathering. The Proof of Reputation mechanism fosters an environment where reputable participants have a greater influence in shaping credibility judgments, ensuring that the evaluation of news benefits from the collective wisdom of high-reputation contributors and minimizing the potential impact of malicious actors. The incorporation of the Hoeffding bound provides statistical guarantees for news label assignment, addressing challenges posed by evolving data and ensuring robustness in dynamic environments. While allowing comparable results with approaches based on conventional (batch) machine learning models, the proposed framework stands out by its authority-free nature and by limiting the need for periodic retraining to update a model's static knowledge of the world.

Our envisioned future directions encompass several key aspects. Firstly, we aim to dynamically adjust the confidence level δ in the Hoeffding Bound, to ensure that the system's certainty in credibility assessments adapts to evolving contexts and dynamic news interactions. Additionally, we plan to introduce fine-grained categories beyond the binary true/false classification, such as "half-true" and "doubtful", etc. for a more nuanced analysis, particularly valuable for borderline cases. Furthermore, we envision leveraging external knowledge sources, including fact-checking databases, to enrich the system's evidence base and context. Finally, we intend to investigate adversarial attacks and test the system's robustness against potential manipulation attempts, such as coordinated bot campaigns or fake accounts.

Author contributions I.S and K.J drafted the main manuscript, I.S.,K.J., and C.A.P carried out the experimental study. O.K., K.J., and C.A.P. supervised and ensured the validation of each milestone. O.K. reviewed and proofread the manuscript.

Funding The authors have not disclosed any funding.

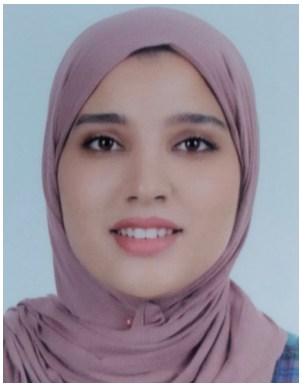
Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Salah, I., Jouini, K., Korbaa, O.: Augmentation-based ensemble learning for stance and fake news detection. In: Advances in Computational Collective Intelligence - 14th International Conference, ICCCI 2022 Proceedings. Communications in Computer and Information Science, vol. 1653, pp. 29–41. Springer, Hammamet, Tunisia (2022). https://doi.org/10.1007/978-3-031-16210-7_3
2. Salah, I., Jouini, K., Korbaa, O.: On the use of text augmentation for stance and fake news detection. J. Inf. Telecommun. **7**(3), 359–375 (2023). https://doi.org/10.1080/24751839.2023.2198820
3. Slovikovskaya, V.: Transfer learning from transformers to fake news challenge stance detection (fnc-1) task. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 1211–1218. European Language Resources Association, Marseille, France (2019). https://www.aclweb.org/anthology/2020.lrec-1.152
4. Alhassani, Z., Obaid, A.: A systemic literature overview of fake news challenge (fnc-1) dataset and its use in fake news detection schemes. J. Dis. Math. Sci. Cryptogr. **26**, 1197–1206 (2023). https://doi.org/10.47974/JDMSC-1567
5. Jemili, F., Meddeb, R., Korbaa, O.: Intrusion detection based on ensemble learning for big data classification. Cluster Comput. (2023). https://doi.org/10.1007/s10586-023-04168-7
6. Khan, A.U., Javaid, N., Khan, M.A., Ullah, I.: A blockchain scheme for authentication, data sharing and nonrepudiation to secure internet of wireless sensor things. Cluster Comput. **26**(2), 945–960 (2023). https://doi.org/10.1007/s10586-022-03722-z
7. Zarrin, J., Phang, H.W., Saheer, L.B., Zarrin, B.: Blockchain for decentralization of the internet: prospects, trends, and challenges. Cluster Comput. **24**(4), 2841–2866 (2021). https://doi.org/10.1007/s10586-021-03301-8
8. ANSAcheck. https://www.ansa.it/sito/static/ansa_check.html. Accessed 31 Oct 2023
9. New York Times News Provenance Project. https://newsprovenanceproject.com. Accessed 31 Oct 2023
10. Avelino, M., Rocha, A.A.D.A.: Blockproof: a framework for verifying authenticity and integrity of web content. Sensors (2022). https://doi.org/10.3390/s22031165
11. Wang, X., Xie, H., Ji, S., Liu, L., Huang, D.: Blockchain-based fake news traceability and verification mechanism. Heliyon **9**(7), 17084 (2023). https://doi.org/10.1016/j.heliyon.2023.e17084
12. Chen, C., Du, Y., Peter, R., Golab, W.M.: An implementation of fake news prevention by blockchain and entropy-based incentive mechanism. Soc. Netw. Anal. Min. **12**(1), 114 (2022). https://doi.org/10.1007/S13278-022-00941-5
13. Soprano, M., Roitero, K., La Barbera, D., Ceolin, D., Spina, D., Mizzaro, S., Demartini, G.: The many dimensions of truthfulness: crowdsourcing misinformation assessments on a multidimensional scale. Inf. Process. Manage. (2021). https://doi.org/10.1016/j.ipm.2021.102710
14. Yilmaz, T., Ulusoy, O.: Modeling and mitigating online misinformation: a suggested Blockchain approach (2023). https://doi.org/10.48550/arXiv.2303.10765
15. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '08, pp. 453–456. Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1357054.1357127
16. Bhuiyan, M.M., Zhang, A.X., Sehat, C.M., Mitra, T.: Investigating differences in crowdsourced news credibility assessment: raters, tasks, and expert criteria. Proc. ACM Hum.-Comput. Interact. (2020) https://doi.org/10.1145/3415164
17. Denaux, R., Merenda, F., Gómez-Pérez, J.M.: Towards crowdsourcing tasks for accurate misinformation detection. In: Joint Proceedings of Workshops AI4LEGAL2020, NLIWOD, PROFILES 2020, QuWeDa 2020 and SEMIFORM2020 Colocated with the 19th International Semantic Web Conference (ISWC 2020), Virtual Conference, November, 2020. CEUR Workshop Proceedings, vol. 2722, pp. 159–167. CEUR-WS.org. https://ceur-ws.org/Vol-2722/semiform2020-paper-2.pdf
18. Marx, J., Cheong, M.: Decentralised social media: scoping review and future research directions. In: Australasian Conference on Information Systems (2023)
19. Boot, A.B., Dijkstra, K., Zwaan, R.A.: The processing and evaluation of news content on social media is influenced by peer-user commentary. Human. Soc. Sci. Commun. **8**(1), 209 (2021). https://doi.org/10.1057/s41599-021-00889-5
20. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**, 13–30 (1963). https://doi.org/10.1080/01621459.1963.10500830
21. Gorrell, G., Bontcheva, K., Derczynski, L., Kochkina, E., Liakata, M., Zubiaga, A.: RumourEval 2019: Determining rumour veracity and support for rumours (2018). https://doi.org/10.48550/arXiv.1809.06683
22. Kochkina, E., Liakata, M., Zubiaga, A.: All-in-one: Multi-task learning for rumour verification. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3402–3413. Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018). https://doi.org/10.48550/arXiv.1806.03713
23. Solidity. https://soliditylang.org/. Accessed 31 Dec 2023
24. Ganache. https://www.trufflesuite.com/ganache. Accessed 31 Dec 2023
25. Li, Q., Zhang, Q., Si, L.: eventAI at SemEval-2019 task 7: Rumor detection on social media by exploiting content, user credibility and propagation information. In: May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M., Mohammad, S.M. (eds.) Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 855–859. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019). https://doi.org/10.18653/v1/S19-2148
26. Khandelwal, A.: Fine-tune longformer for jointly predicting rumor stance and veracity. CODS-COMAD '21, pp. 10–19. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3430984.3431007
27. Liu, Y., Yang, X., Zhang, X., Tang, Z., Chen, Z., Zheng, L.: Predicting rumor veracity on social media with cross-channel interaction of multi-task. Neural Comput. Appl. (2024). https://doi.org/10.1007/s00521-024-09519-y



Ilhem Salah is currently pursuing a Ph.D. in Computer Science at Sousse University (Tunisia) after earning her Master's degree in Distributed Computing from the same institution. Her research primarily focuses on fake news detection in social media, as well as machine learning, deep learning, distributed ledgers, and natural language processing.

Khaled Jouini received the Ph.D. degree in Computer Science from Paris-Dauphine University (France) in 2008. He was a research staff member at Telecom ParisTech (France). Since 2011, he has been with University of Sousse (Tunisia), where he is currently an Associate Professor. His research interests include data engineering, natural language processing, and large-scale data management and mining.

Cyril-Alexandre Pachon received his Ph.D. in Computer Science, Systems, and Communication from Université Joseph Fourier (France) in 2005. He is currently the Director of Studies at École Hexagone (France). Previously, he spent over a decade at SUPINFO International University (France), where he managed the Robotics Laboratory and led initiatives in robotics competitions. His research interests include robustness testing, test case generation, data science, and artificial intelligence applications.

Ouajdi Korbaa is a full-time professor at the University of Sousse (Tunisia). He received his Engineering Diploma from the Ecole Centrale de Lille (France) in 1995 and his Master's degree in Production Engineering and Computer Science from the University of Lille (France) in the same year. He obtained his Ph.D. in Production Management, Automatic Control, and Computer Science from the University of Science and Technologies of Lille (France) in 1998. Pr. Korbaa has published approximately 190 research papers on optimization, applied and computational mathematics, manufacturing engineering, and computer engineering.