Journal Pre-proof

Ensuring News Integrity against Online Information Disorder through Text Watermarking and Blockchain

Flavio Bertini, Alessandro Benetton and Danilo Montesi

PII: S2096-7209(25)00141-1
DOI: https://doi.org/10.1016/j.bcra.2025.100414
Reference: BCRA 100414

To appear in: *Blockchain: Research and Applications*

Received date: 1 February 2025
Revised date: 18 September 2025
Accepted date: 15 October 2025

Please cite this article as: F. Bertini, A. Benetton and D. Montesi, Ensuring News Integrity against Online Information Disorder through Text Watermarking and Blockchain, *Blockchain: Research and Applications*, 100414, doi: https://doi.org/10.1016/j.bcra.2025.100414.

Highlights

- A platform for ensuring the integrity of textual news, combining text watermarking with blockchain technology to track provenance.
- Fine-grained watermarking to enable the identification of sources, even from small text excerpts.
- Blockchain-based storage of verification information in a decentralised manner, guaranteeing the authenticity of textual news.
- A browser extension for real-time verification of textual news, flagging any detected manipulations.

Ensuring News Integrity against Online Information Disorder through Text Watermarking and Blockchain

Flavio Bertini [a,*], Alessandro Benetton[b], Danilo Montesi [b]

[a] *Department of Mathematical, Physical and Computer Sciences, University of Parma, Parco Area delle Scienze 53/A, Parma, 43124, Italy*
[b] *Department of Computer Science and Engineering, University of Bologna, Mura Anteo Zamboni 7, Bologna, 40126, Italy*

Abstract

The rapid proliferation of online information disorder poses a significant societal challenge. This phenomenon has been further exacerbated by the pervasive influence of social media, affecting a broad range of domains. Addressing the spread of information disorder through manual approaches (*e.g.*, human fact-checking) is impractical due to the vast volume of textual content daily generated, often facilitated by disruptive generative technologies. Similarly, the implementation of automated tools presents considerable obstacles, primarily due to the inherent ambiguity and complexity of natural language. This paper introduces the CERVANTES platform: an innovative application of blockchain technology integrated with text watermarking techniques. It is conceived to support news producers and enhance awareness among readers engaging with content across various social media platforms. CERVANTES is a multiplatform and language-independent solution that addresses online

*Corresponding author
Email addresses: flavio.bertini@unipr.it (Flavio Bertini), alessandro.benetton@studio.unibo.it (Alessandro Benetton), danilo.montesi@unibo.it (Danilo Montesi)

information disorder from an innovative perspective. It allows news producers to automatically embed a unique watermark within the content they create, ensuring the integrity and authenticity of the text, and safeguarding it against manipulation and misattribution. Furthermore, the association between the watermark and the corresponding news item is securely recorded on the blockchain, mitigating the risk of manipulation that might result from centralised management. We conduct an extended evaluation on twelve different social media platforms with a cohort of twenty users, observing ease of use and a high degree of satisfaction.

Keywords: Text watermarking, Text news sealing, Text provenance checking, Blockchain application, Online information disorder

1. Introduction

Over the past two decades, the proliferation of user-generated content on websites and social media platforms has significantly enhanced information exchange across ethnic, political, and geographical boundaries. This development has fostered the emergence of communities built around shared interests, perspectives, and narratives. However, in this landscape of disintermediation, it has become increasingly evident that information disorders frequently undermine human rights and the foundational principles of democracy [1]. A report by the Council of Europe identified the proliferation of information disorders, marked by the emergence of three distinct forms of misleading or false content: mis-, dis-, and mal-information [2]. These phenomena have had a significant impact across a wide range of domains, including democratic elections, public health communication, climate change

discourse, and geopolitical conflicts. Digital platforms and social media notably amplify this trend. Such false, inaccurate, or misleading content has the potential to manipulate public opinion, erode trust in institutions, and fuel scepticism about critical issues such as vaccination [3, 4].

Online information disorder represents a multifaceted research challenge, as its accurate classification necessitates addressing both malicious activities and entirely lawful behaviours. Tandoc Jr. et al. have identified a range of online behaviours that can be categorised within the framework of information disorders [5]. Information disorder manifests in various forms, including *news satire*, characterised by mock news with an explicitly humorous intent; *news parody*, which employs non-factual information for comedic purposes but, unlike news satire, fails to clearly indicate its non-journalistic nature; *news fabrication* and *photo manipulation*, where articles and images devoid of factual accuracy are intentionally altered to construct false narratives; and *propaganda*, referring to news content created by political actors to influence public opinion. Importantly, not all of these forms necessarily constitute malicious behaviour.

The global expansion of online information disorder across websites and social media platforms underscores the urgent need for effective countermeasures. This issue is particularly complex, as it spans various forms of online content, with text-based content posing distinct challenges due to its nuanced nature [6]. The literature identifies various methodologies for addressing information disorder, broadly classified into language-based, topic-agnostic, machine learning, knowledge-based, and hybrid approaches [7]. Despite their potential, automated methods exhibit significant limitations. In

particular, the inherent ambiguity of natural language poses substantial challenges to the development and application of content-based approaches, often rendering them impractical. This issue has been further exacerbated by the widespread adoption of Large Language Models (LLMs), which are capable of generating human-like text often embedded with unverified or inaccurate information [8]. Conversely, manual fact-checking approaches are equally unfeasible due to the sheer volume of text disseminated daily, which makes human intervention at scale impractical.

Moreover, one of the most common activities among online users is sharing excerpts from original news, a practice in which both automatic and manual methods fail, as this copy-and-paste behaviour removes the full context and obscures the source information. Reader deception becomes even more evident when text content is presented as legitimate news and is accompanied by seemingly authoritative markers of the user profile's integrity, such as the blue checkmark on X (formerly Twitter). Although this verification badge signifies that an account has undergone a verification process, it does not guarantee the accuracy or authenticity of the content disseminated by that account.

This paper addresses the challenge of online information disorder in textual news by introducing the CERVANTES platform designed to secure text content and trace its source provenance. The proposed platform integrates text watermarking techniques with blockchain technology to ensure the integrity of text news, even when partially re-shared, and to enable the identification of its source. Specifically, the watermarking is embedded in a fine-grained manner, enabling even small excerpts of news content to be copy-and-

pasted and shared while maintaining traceability. The embedded watermark, based on a previous text watermarking technique [9], is generated from both the data and metadata of the original textual content, seamlessly integrated into the text and securely stored on a blockchain for verification purposes. In particular, the blockchain records critical information in a decentralised and distributed manner, providing a consensus-driven framework for authenticating textual news and enabling content traceability across diverse social media platforms. A browser extension further enhances usability by allowing readers to verify in real-time whether the content they are accessing corresponds to the original one. Even when only a fragment of the original text is displayed, the extension retrieves supplementary information from the distributed ledger. Additionally, any detected manipulations are immediately flagged and highlighted to ensure transparency for the reader.

The CERVANTES platform is designed to empower readers to make more informed decisions about the content they encounter, thereby reducing reliance on third-party fact-checking organisations. It is designed to be both multiplatform and language-independent, capable of scaling to meet the needs of diverse users. It operates independently of the social media platforms utilised, supports the watermarking of texts written in the Latin alphabet irrespective of the language, and accommodates high-volume textual content production without requiring human intervention. These features enable news organizations of all sizes, as well as independent journalists, to effectively track and disseminate text content across multiple platforms. Moreover, the platform's capacity to manage the provenance, trustworthiness, and verification of online news lies in its ability to strike a balance

between protecting freedom of expression and maintaining information quality. The proposed CERVANTES platform overcomes the content-based limitations of traditional automated methods and addresses the deficiencies of manual approaches, which are often influenced by censorship dynamics and are impractical or prone to inaccuracies. To assess the effectiveness of the CERVANTES platform, we performed an extensive evaluation across twelve social media platforms. Additionally, a preliminary A/B testing experiment was conducted involving twenty participants, who were asked to evaluate the veracity of both original and manipulated news items on various topics, with and without the assistance of the CERVANTES platform. Despite the small size of the testing cohort, the experiment enabled the collection of qualitative insights into users' responses to the introduction of a new interaction element.

The paper is structured as follows. Section 2 offers a comprehensive review of the literature relevant to this study. Section 3 presents a brief overview of text watermarking attacks, to clarify the rationale behind their integration with blockchain technology. Section 4 provides a detailed overview of the CERVANTES architecture. Section 5 discusses the implementation choices and presents the evaluation results. Section 6 discusses the limitations of our work. Finally, Section 7 concludes the manuscript with key insights and outlines potential directions for future research.

2. Related Work

In this section, we review relevant literature spanning multiple domains that converge in the scope of our research. Specifically, we examine studies on

online information disorder, text watermarking techniques, and blockchain technologies, highlighting their potential to verify the provenance of textual content and ensure its integrity.

In recent years, the phenomenon of information disorder has witnessed a significant escalation. While the internet has enabled the amplification of diverse voices and democratised access to information, it has also introduced a series of technological vulnerabilities that exacerbate the spread of information disorder. Notably, a study conducted by MIT demonstrated that false information spreads more rapidly, extensively, and deeply compared to truthful content, with falsehoods being 70% more likely to be retweeted than accurate information [10]. Wardle et al. [2] proposed a tripartite classification of information disorder, distinguishing between misinformation, disinformation, and malinformation. However, their framework does not encompass other forms of expression, such as satire and parody, which introduce additional complexities in efforts to address the dissemination of fake news [5]. Various methodologies address information disorder, including language-based, topic-agnostic, machine learning, knowledge-based, and hybrid approaches [7]. Language-based methods detect textual inconsistencies. Topic-agnostic strategies rely on meta-information, such as advertisement density or sensationalist language. Machine learning uses annotated datasets to identify fake news patterns. Knowledge-based methods combine machine learning with knowledge engineering but struggle to keep pace with fabricated content. Hybrid approaches merge human analysis and machine learning, focusing on text, audience engagement, and source credibility. The multi-dimensional nature of this issue, compounded by the inherent ambiguity of natural language

and the sheer volume of information shared online daily, renders both manual and automated approaches to combating information disorder frequently impractical and often imprecise [7]. Many existing EU projects require extensive end-user involvement to evaluate individual posts and assess their trustworthiness [11], incorporate overly complex content evaluation procedures [12], or lack the capability to address fine-grained textual content [13]. More recently, a collaborative effort by Microsoft Research, BBC, CBC/Radio-Canada, The New York Times, and Truepic introduced a platform aimed at curbing the dissemination of fraudulent images and videos online [14].

The CERVANTES platform offers a fully automated solution for tracking and verifying textual content. By integrating text watermarking techniques with blockchain technology, the platform ensures both the integrity and traceability of information, while simultaneously fostering greater awareness among online readers regarding content authenticity.

Watermarking embeds a mark in digital content to prove ownership, authorship, and verify content [15]. Text watermarking, due to its low embedding capacity and limited syntactic and semantic variations, presents significant challenges. These have led to the development of various approaches, broadly categorised as follows.

Zero-watermarking - In these methods, no direct watermark is embedded in the text itself; instead, defining characteristics of the text are stored on a third-party server [16]. A key limitation of zero-watermarking lies in its centralised nature and the associated lack of transparency.

Image-based approaches - These methods necessitate the conversion of the text into an image to apply the watermark [17, 18]. This approach alters

the original nature of the document and it is often impractical in contexts such as online social media, where image-based methods are incongruent and significantly constrain sharing practices.

Syntactic methods - These techniques operate on the syntactic structure of natural language text by modifying the syntactic tree of a sentence to embed a watermark. Examples of such syntactic operations include clefting, passivization, or activization [19, 20, 21]. A notable limitation of these methods is their language dependency and low embedding capacity. Additionally, the assumption that different syntactic forms convey identical meanings is not always accurate, further constraining their applicability.

Semantic methods - These approaches leverage the semantic similarity between words by replacing them with their synonyms [22]. Other techniques operate at the sentence level, capitalizing on the implicit presuppositions embedded in each sentence [23, 24]. However, even when combined with syntactic approaches to enhance the embedding capacity [25], semantic methods share the same limitations as syntactic methods, particularly their low embedding capacity and heavy dependence on language.

Structural methods - These methods do not modify the textual content directly; instead, they alter its structure, such as by inserting empty lines [26] or utilising different Unicode whitespace characters [27, 28]. In recent work, the authors employed an alphabet comprising five visually indistinguishable Unicode whitespace characters to encode and embed a secret message within a cover text [29]. Because the watermark is embedded within the underlying representation of the text, these techniques offer the notable advantage of preserving the original content without the need for an external database.

CERVANTES integrates a structural text watermarking technique for sealing text, ensuring the preservation of content and length [9].

Blockchain technology enables the storage of transactional data in an open, decentralised ledger, ensuring both verifiability and immutability. To enhance efficiency, transactional data (*i.e.*, tokens) are organized into time-stamped blocks, each linked to its predecessor [30]. The decision to add a new block is made collectively by peers using a consensus protocol. Once added, a block becomes immutable and cannot be removed. Transaction validity can be publicly verified by any client connecting to a peer that holds the full blockchain. Originally designed for decentralized digital currencies, this technology has since found applications in contexts where ensuring the authenticity and verification of data is critical, without the reliance on a centralized, trusted authority [31, 32].

Several pioneering works have proposed blockchain-based solutions to address the issue of online misinformation [33, 34, 35]. In [36], Arquam et al. introduced a blockchain-based framework for verifying information propagation. A blockchain-based model incorporating a secure voting system is proposed in [37]. News reviewers provide feedback on news items, and a probabilistic mathematical model is then used to estimate the truthfulness of each item based on the feedback received. Alexandrescu et al. propose a decentralised blockchain-based architecture for news retrieval and aggregation, separating crawling and scraping phases and ensuring information reliability via a majority-based mechanism [38]. A conceptual framework for fake news detection based on machine learning and blockchain is proposed in [39], where both experts and users rate content to assess its authenticity. To counter the

spread of fake news on social media, [40] proposes a multilevel model based on blockchain and deep learning techniques, designed to detect and prevent the propagation of rumours. More recently, several approaches combining watermarking and blockchain have been developed specifically for multimedia content to combat fake news. In [41], the authors combine basic blockchain and watermarking techniques to trace the origin of fake news and limit its spread on social media. A theoretical framework combining blockchain and watermarking is proposed in [42]. The authors claim that the framework ensures the integrity of posted content and enables accountability of the post's owner or user. In [43], the authors present a blockchain-based system for tracking photographs, which accommodates image transformations while ensuring provenance and integrity, and demonstrates advantages over traditional watermarking approaches. A recent study proposes a semi-fragile, blind watermarking system within a blockchain-based framework, designed to detect fake images on social media, limit their spread, and support traceability, manipulation tracking, and blacklisting [44]. The authors in [45] also propose a blockchain-enabled watermarking technique to address deepfake-related challenges, integrating content tracking, decentralised identity, and cryptographic verification.

These solutions are primarily aimed at audio and video content, and the reliability depends on social network parameters, whereas our approach does not require any network structure to validate textual content. Moreover, in contrast to these systems, our solution leverages blockchain technology to track textual content directly at the paragraph level, without relying on content evaluation methods. These features enhance the versatility of CERVANTES,

making it readily applicable across multiple platforms and websites and facilitating its integration into various online information production systems.

3. Background on Text Watermarking Attacks

As this paper brings together expertise from distinct disciplines, we deemed it necessary to include a section outlining some fundamental concepts related to text watermarking attacks. This should later help clarify the rationale behind the integration of blockchain technology within the proposed platform.

Focusing on structural text watermarking techniques - considered more flexible and promising, as discussed in the Section 2 - several attacks can be identified that aim to remove or compromise the watermark embedded in the text. Based on a careful analysis of the literature [46, 47, 48], we identified the following types of attack:

- *Deletion* - this type of attack consists in the removal of segments of text to compromise the integrity of the watermark. This strategy can be employed to reshape the message, aligning it with a specific narrative or obscuring particular information, while preserving the overall coherence of the text.

- *Insertion* - this attack is complementary to the previous one, as it involves adding new words or characters to alter its meaning. A simple example would be the insertion of the word "not" before a statement, thereby reversing its intended message.

- *Replacement* - this attack involves replacing certain words or characters in the text with different ones. It may be viewed as a combination of

deletion and insertion at the same location. A typical example is the modification of a pronoun to alter its gender expression.

- *Copy&Paste* - this type of attack involves copying and reusing portions of text, and is particularly insidious as it often relies on extracting only small segments from a watermarked source. Such attacks can be employed, for instance, to misrepresent the statements of a political opponent by selectively quoting only those fragments that support the attacker's narrative.

- *Retyping* - this attack occurs when a malicious user manually rewrites the text in a different file or platform. As all structural methods are inherently vulnerable to such attacks, this case is excluded and instead informs the design of the three evaluation states of the CERVANTES browser extension (see Section 4.2 for details).

- *Reformatting* - this attack involves altering the formatting characteristics of a text. However, within the context of our work and the intended application of the CERVANTES platform, such attacks can be excluded without loss of generality. This is because social platforms typically do not allow stylistic modifications that would affect the watermarking method employed, unlike several earlier approaches (for further details, see [49]).

Section 5.1 will illustrate how these attacks were adapted to reflect realistic user behaviour, thereby emulating practical usage scenarios of the CERVANTES platform.

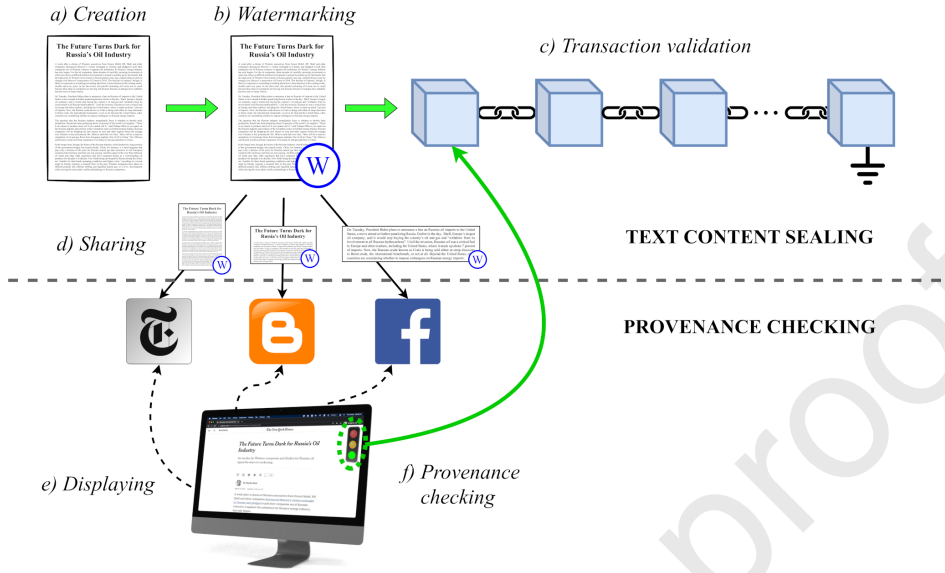


Figure 1: Overview of the architecture of the CERVANTES platform, illustrating its key components and functionalities.

4. The CERVANTES Architecture

This section outlines the architecture of the proposed platform, highlighting its two key components. First, we examine the sealing of text news through text watermarking, as shown above the dotted line in Figure 1. Second, we provide a detailed explanation of the verification process, depicted below the dotted line in Figure 1.

4.1. Text Content Sealing

The content sealing process through structural text watermarking employs homoglyph-based character substitution. The technique has already been evaluated in a previous publication [9]. In the present work, it is subject only to minor modifications, with negligible impact on its performance and robustness, in order to meet the requirements of the proposed platform.

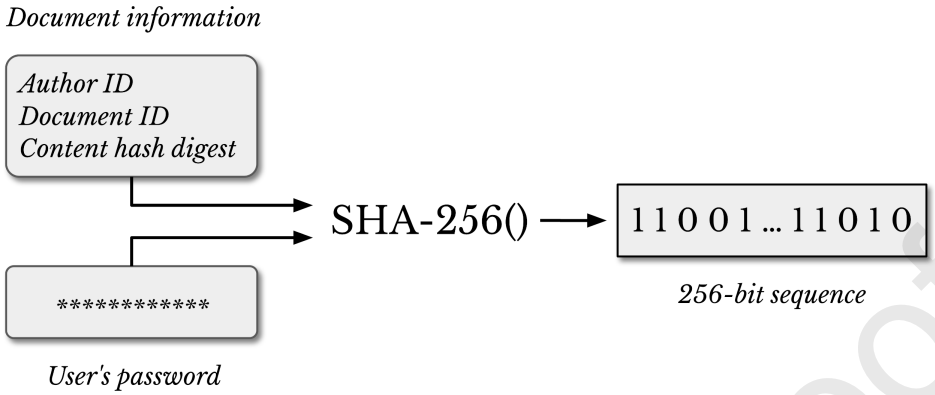


Figure 2: Generation of the watermark: given the document information and a secret user password, the hash function generates a bit sequence representing the watermark to be embedded.

Homoglyphs[1] are Unicode characters that visually resemble each other but differ in their underlying representations. This visual similarity is leveraged to embed a bit sequence watermark, which is generated using a hash function that combines the document's information with the user's password, thereby enhancing both security and resilience. Figure 2 shows the use of the SHA-256 hash function; however, the function is parametric, and a more secure variant can be selected. This would simply require a greater number of characters to embed the resulting, longer bit sequence.

Continuous and fine-grained watermarking, applied at the level of a few dozen words, is accomplished by concatenating the watermark throughout the document. This approach enables the validation of any sufficiently long copied portion that contains a complete watermark sequence. This requirement does not pose a limitation, even when considering character restrictions imposed

[1] `https://www.unicode.org/reports/tr36`

by certain social media platforms, as the watermarking technique operates effectively with between 46 and 101 characters [49]. To maintain content integrity, the CERVANTES platform stores hashes for each sub-portion, which are explicitly identified to contain the full payload (*i.e.*, the watermark). An invisible separator character is introduced before and after each sub-portion during the watermark embedding process to ensure consistent separation between the sealing and validation phases. The embedded payload to generate the watermark comprises the *author ID*, *document ID*, and a *seal* (*i.e.*, the hash digest), thereby safeguarding the integrity of the text and confirming that it has not been altered.

The lists of characters and whitespace symbols, along with their respective homoglyphs - which ensure that the final watermarked text remains visually indistinguishable across all evaluated social media platforms [49] - are shown in Table 1 and Table 2, respectively (further details are provided in Section 5.3). The embedding process will employ the homoglyphs listed in Tables 1 and 2 in order to enhance embedding capacity. Furthermore, Table 3 illustrates the frequency distribution of the original symbols and letters across six Latin-based languages (*i.e.*, English, Spanish, French, Portuguese, German, and Italian). For the English language, the frequencies were derived from the *New York Times* corpus, comprising approximately 14 million words [50]. For other languages, frequency calculations were based on the *Full-text Corpus Data* website for Spanish and Portuguese[2], the *88milSMS* corpus for French [51], the *DeReWo* corpus for German [52], and the *Il Post* corpus

[2] `https://www.corpusdata.org`

Table 1: The set of Latin letters and punctuation homoglyphs utilised during the watermarking process.

Character	Homoglyph	Original Code	Homoglyph Code
-	‐	U+002D	U+2010
C	С	U+0043	U+216D
D	Ⅾ	U+0044	U+216E
L	Ⅼ	U+004C	U+216C
M	Ⅿ	U+004D	U+216F
V	Ⅴ	U+0056	U+2164
X	Ⅹ	U+0058	U+2169
c	ⅽ	U+0063	U+217D
d	ⅾ	U+0064	U+217E
i	ⅰ	U+0069	U+2170
j	ϳ	U+006A	U+0458
l	ⅼ	U+006C	U+217C
v	ⅴ	U+0076	U+2174
x	ⅹ	U+0078	U+2179

for Italian [53]. In addition to the homoglyphs listed in Table 1, the invisible character U+200B, commonly referred to as the *Zero Width Space*, is employed as a separator during the multiple stages of watermark embedding.

As shown in Figure 3, the embedding algorithm operates through the following sequence of steps:

1. the invisible character U+200B (*i.e.*, *Zero Width Space*) character is inserted to delineate the start of the embedding process (highlighted in red in Figure 3);

Table 2: Name and Unicode code of the eight whitespace characters, including the encodable bit sequences.

Whitespace	**Unicode Code**	**Sequence Encodable**
Space	U+0020	000
En quad	U+2000	001
Three-per-em space	U+2004	010
Four-per-em space	U+2005	011
Punctuation space	U+2008	100
Thin space	U+2009	101
Narrow no-break space	U+202f	110
Medium mathematical space	U+205f	111

2. the text is scanned sequentially until a character or whitespace listed in Table 1 and Table 2 is identified (all highlighted in green in Figure 3);
3. if the subsequent bit in the watermark sequence is '1', the identified character is substituted with its equivalent homoglyph; otherwise, the original character is retained. If the identified character in the previous step is a whitespace, this step remains unchanged. However, the availability of eight distinct homoglyphs for whitespace enables the consumption of three bits of the watermark per character (as illustrated in Table 2 third column, and in Figure 3);
4. steps *2* and *3* are iterated until the entire sequence of watermark bits has been embedded. Upon reaching the final bit of the watermark, another *Zero Width Space* character is introduced, and the watermark bit sequence is reset to its initial state to restart a new round of embedding.

Table 3: Frequency distribution of original symbols and letters across major Latin-based languages.

Character	English	Spanish	French	Portuguese	German	Italian
-	4.34%	0.11%	0.18%	0.79%	$<0.00\%$	0.04%
C	0.35%	0.23%	0.26%	0.34%	0.06%	0.24%
D	0.20%	0.14%	0.22%	0.21%	0.20%	0.12%
L	0.16%	0.20%	0.16%	0.20%	0.24%	0.25%
M	0.40%	0.16%	0.33%	0.28%	0.35%	0.19%
V	0.05%	0.05%	0.05%	0.08%	0.25%	0.06%
X	0.01%	0.01%	0.02%	0.01%	$<0.00\%$	0.01%
c	3.01%	4.35%	2.95%	3.44%	2.79%	4.13%
d	3.63%	4.86%	2.95%	5.22%	2.12%	3.80%
i	6.94%	6.32%	7.55%	6.39%	6.78%	11.33%
j	0.10%	0.38%	1.28%	0.28%	0.09%	0.01%
l	3.92%	5.21%	4.20%	2.75%	4.64%	6.11%
v	1.00%	0.95%	1.77%	1.28%	0.64%	1.44%
x	0.19%	0.18%	0.36%	0.24%	0.10%	0.03%

The four steps are iterated continuously until the entire document has been processed. The extraction procedure mirrors these steps: it begins by locating the *Zero Width Space* character, subsequently identifies homoglyph-compatible characters listed in Table 1 and Table 2, and reconstructs the binary watermark sequence based on the presence or absence of homoglyph. It is worth noting that the use of the invisible character U+200B as a watermark delimiter can be further secured. Currently, malicious users familiar with the watermarking scheme and the role of this special character could

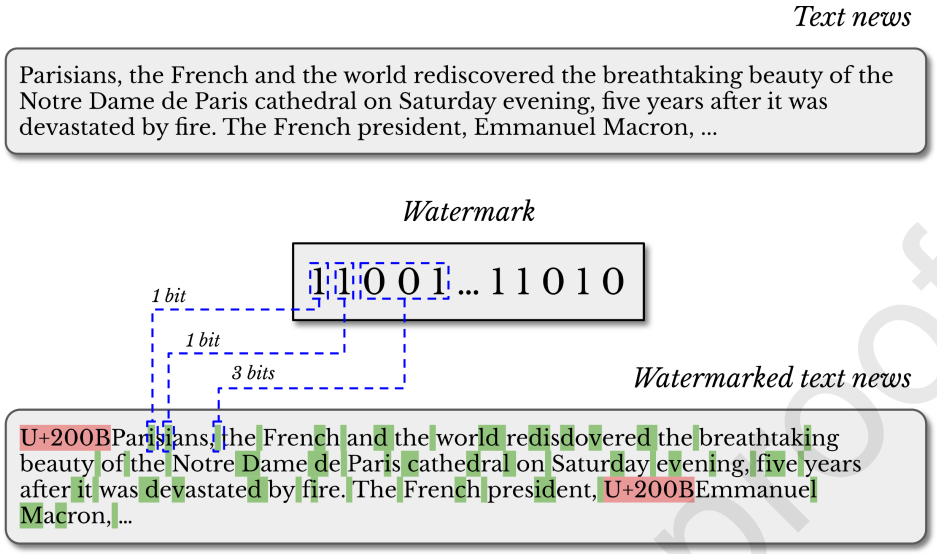


Figure 3: Embedding watermark bits by replacing confusable symbols. Only those characters with a corresponding duplicate in Tables 1 and 2 (highlighted in green) are used to embed the watermark bits according to the bit sequence. One bit can be embedded using Latin letters, and three bits using space characters. The U+200B character is used to delimit successive rounds of embedding. For the sake of simplicity, a shortened bit sequence is shown.

easily remove it, potentially resulting in a false negative (*i.e.*, the news item is reliable but cannot be verified). This aspect could be strengthened by obfuscating the generation of the character or sequence of characters used as delimiters, for example by incorporating the Innamark method proposed in [29].

The proposed watermarking approach enables content creators to secure their textual content, as described in phases *a* and *b* of Figure 1, and offers two key advantages. Firstly, by repeatedly applying the watermark and distributing uniquely signed versions across multiple social media platforms

(as illustrated in phase *d* of Figure 1), authors can enhance traceability and pinpoint the origins of specific quotes. Secondly, the system enables the immediate detection of alterations, even in brief excerpts of the original document, while leveraging the pointer to the blockchain to provide readers with supplementary metadata - such as the author's identity, timestamps, and the original source URL - thereby enhancing transparency and fostering trust in the authenticity of the content.

4.2. Provenance Checking

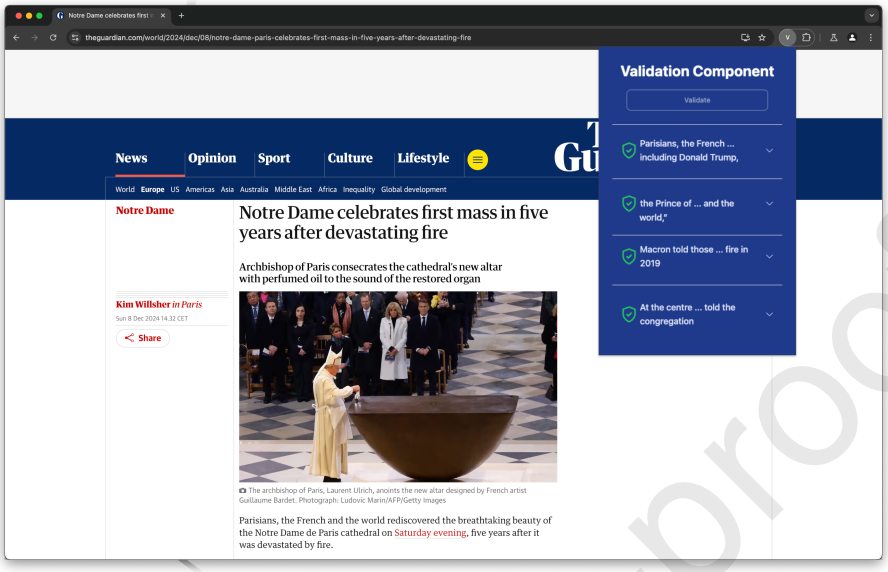
Relying exclusively on embedded information for text validation presents several limitations. First, it is relatively straightforward to create a document containing a valid watermark without utilizing the designated sealing tool. Additionally, embedding extensive metadata directly into the text via watermarking is inefficient, as minimizing the payload size is essential to reduce the required embedding length. Larger payloads necessitate more text for successful embedding, which becomes impractical when it is necessary to validate excerpts of the original document. To address this, the CERVANTES platform employs a minimal payload design, ensuring efficient and scalable verification. Furthermore, the platform incorporates a provenance-checking protocol that tracks watermarks via a blockchain, enabling the validation of text fragments without requiring duplication of the entire document on the ledger.

The CERVANTES platform integrates blockchain technology as a second key component (depicted as phase *c* in Figure 1). This integration ensures a consensus-driven approach for recording supplementary metadata and provides a decentralized framework for verifying the integrity of text content

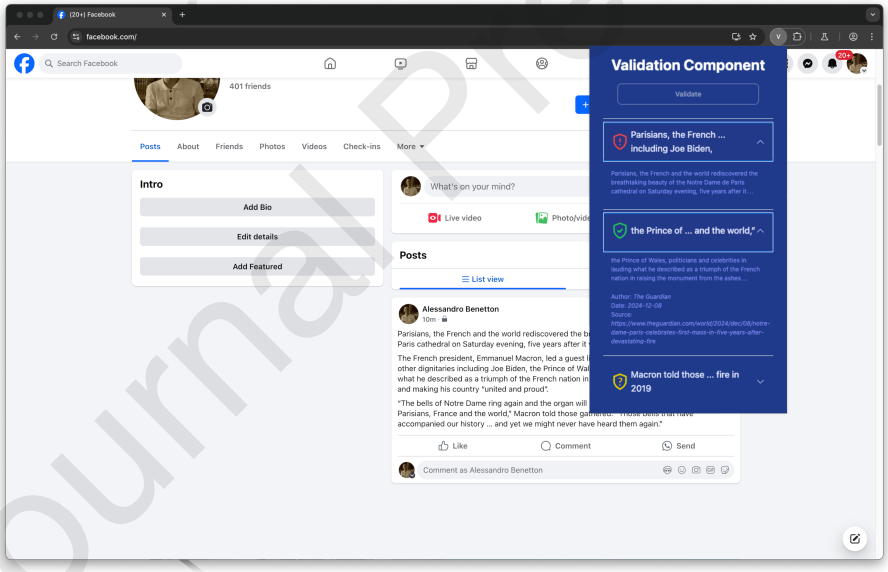
(illustrated as phase *f* in Figure 1). The distributed ledger functions as the authoritative source of truth within the system and is designed to meet three critical criteria: reliability, ensuring that all recorded data is accurate and trustworthy; independence, guaranteeing that no single entity has control over the stored information; and immutability, ensuring that once data is entered into the ledger, it remains permanently unalterable.

An integral feature of the platform is its ability to automatically enable real-time content validation for readers. This functionality is facilitated through a browser extension that actively scans the content of the displayed web page for embedded watermarks. Upon detecting a watermark, the extension initiates a blockchain-based verification process to authenticate the content. The results of this process are then presented to the user in the form of a reliability indicator and any additional metadata retrieved (refer to phases *e* and *f* in Figure 1). Notably, the validation mechanism remains effective regardless of the length of the text excerpt being evaluated. Furthermore, any detected discrepancies between the watermark and the blockchain record are flagged, promptly alerting the reader to potential issues with the content's integrity.

Figure 4 illustrates the result of the validation process for the displayed text content, as observed both on the original news website (*i.e.*, The Guardian, Figure 4a) and on the social media platform where the news was shared (*i.e.*, Facebook, Figure 4b). Upon completion of the validation, the extension provides the user with a detailed report, including supplementary information alongside the identified watermarked content. Specifically, each paragraph is assigned one of three status labels: *valid*, *unknown*, or *invalid*. A *valid*



(a)



(b)

Figure 4: The browser extension for provenance verification: enhancing reader awareness on websites (*i.e.*, The Guardian) *(a)* and social media platforms (*i.e.*, Facebook) *(b)*.

status indicates that the paragraph has been copied from a known source without modification, as the extracted watermark matches an entry in the blockchain (the green shields in Figure 4b). The *unknown* status signifies an incomplete or corrupted watermark, which may occur when the copied portion fails to capture the full watermark or when part of the seal is missing (the yellow shield in Figure 4b). In contrast, an *invalid* status denotes that the watermark is absent or does not correspond to any entry in the blockchain, suggesting that the original seal has been deliberately altered or replaced (the red shield in Figure 4b).

4.3. Content Sharing and Verification Protocol

This section provides an overview of the processes involved in content sealing, dissemination, and verification, building upon the foundational components outlined in the preceding sections.

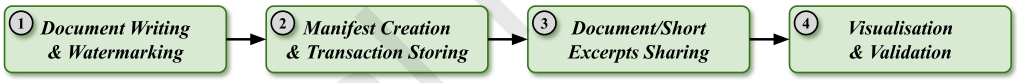


Figure 5: The procedural flow of text content sharing and verification.

The protocol for content sharing and verification consists of four primary steps, as illustrated in Figure 5:

1. The initial stage, referred to as *document writing & watermarking*, involves the creation of textual content by the news producer. During this process, the system automatically applies a unique watermark to the content through the functionalities provided by the *text content sealing* module.

2. In the subsequent stage, referred to as *manifest creation & transaction storing*, the manifest associated with the newly created document is constructed and securely stored on a blockchain. The corresponding address is then registered on a distributed ledger. This manifest encapsulates key metadata, including details about the author, the document, and the hash digest of the original text, thereby enabling a robust framework for validation and provenance verification of all the excerpts of the original document.
3. In *document/short excerpts sharing* stage, the text - either in its entirety or as selected portions - can be distributed across websites and social media platforms.
4. In the final stage, referred to as the *visualisation & validation* phase, any sufficiently lengthy segment of the original document that encompasses a complete watermark sequence can undergo validation through the use of the *provenance checking* components.

The subsequent section will discuss the various implementation decisions involved in the development of the platform prototype, as well as its evaluation across twelve social media platforms with a cohort of twenty users.

5. Prototyping and Evaluation

The following section outlines the rationale behind key technical choices to ensure the replicability of the platform. We begin by presenting results on the robustness of the text watermarking technique against the attacks outlined in Section 3, in order to complement the prior evaluation in [9]. Additionally, we conducted an evaluation of CERVANTES on twelve of the

most widely utilised social media platforms. Lastly, we report the results of the user assessment.

5.1. Text Watermarking Robustness

Section 3 outlined several types of attacks; however, an important aspect to consider is the proportion of text that must be affected to simulate a realistic and credible scenario. In [54], the authors suggest that altering 10% of a text constitutes a significant attack size for any watermarking technique. In the case of *copy & paste* attack, this percentage should be interpreted in reverse: as the proportion of copied text increases, watermark reconstruction becomes easier. In our tests, we consider portions of text copied from the original up to a minimum of 5% of its length. A set of 1,000 articles from the New York Times Corpus was used to carry out the tests, and the results are presented in Table 4.

Table 4: Success rate of watermark reconstruction following each specific attack involving 10% of the original text (*i.e.*, 5% in the case of the *copy&paste* attack).

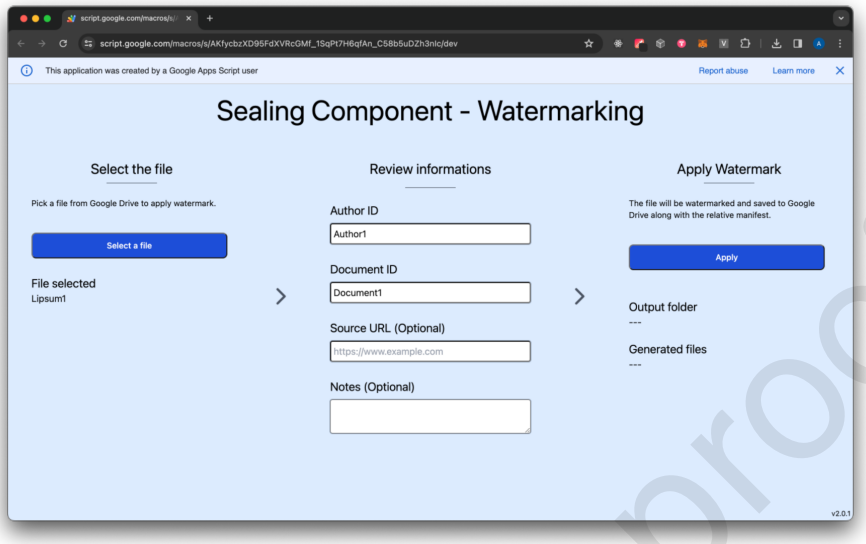
Deletion	**Insertion**	**Replacement**	**Copy&Paste**
98.20%	96.40%	98.30%	99.92%

The retyping and reformatting attacks have been excluded from the evaluation. The former renders structural watermarking methods ineffective by definition, while the latter is not applicable in the context of social media platforms, which do not support formatting changes. The results obtained indicate that the watermarking technique integrated into the CERVANTES platform demonstrates strong robustness, largely attributable to its ability to watermark the original text repeatedly.

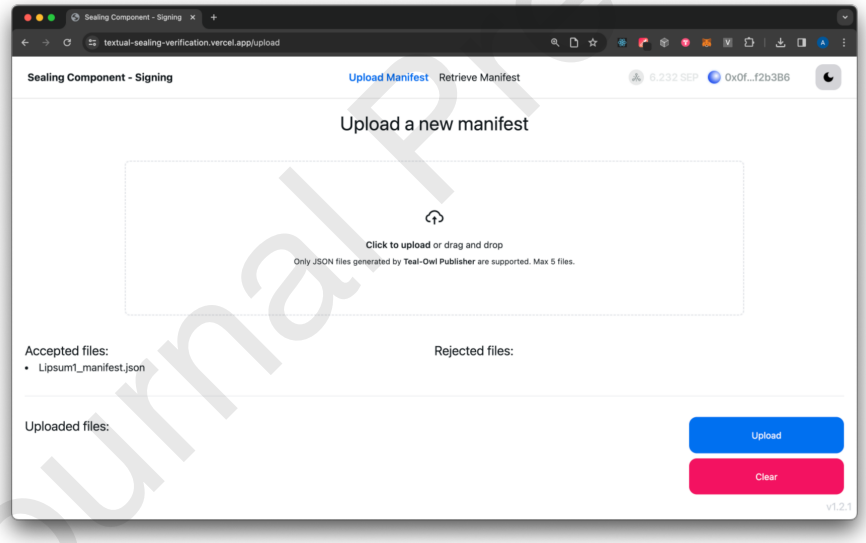
5.2. Platform Prototyping

CERVANTES platform comprises two primary components: *watermarking & sealing* and *storing & validation*. The watermarking process embeds a unique identifier into the text document with minimal computational overhead, generating a specific file, referred to as the manifest. The manifest contains critical metadata designed for secure storage on the blockchain, facilitating subsequent validation and provenance verification. The Google Apps Script environment was selected as the scripting language for implementing the watermarking process within Google Docs. As illustrated in Figure 6a, the script seamlessly and automatically embeds a watermark into the text document and generates a JSON file, *i.e.*, the manifest. The manifest includes essential metadata such as the *author ID*, *document ID*, *sealing timestamp*, *source URL*, and a *list of hash digests* corresponding to each subsection of the text. These hash digests are crucial for verifying the provenance of each of the short excerpts from the original document.

The second component consists of a web application designed to upload the generated JSON file (*i.e.*, the manifest) to the InterPlanetary File System (IPFS) and record its corresponding address on a smart contract. This process is facilitated through the user-friendly interface shown in Figure 6b. Future developments aim to streamline the workflow by automating both the watermarking and upload operations through an integrated interface. The IPFS functions as a decentralized and distributed storage network, leveraging a peer-to-peer architecture to improve content availability and fault tolerance. By generating addresses derived from the content's hash, IPFS ensures document immutability while providing an efficient mechanism for identify-



(a)



(b)

Figure 6: The Google Apps Script interface facilitates the watermarking of text documents within Google Docs *(a)* and enables the subsequent upload of the generated manifest *(b)*.

ing duplicates. The smart contract, deployed on Ethereum-based *testnets* for testing purposes, securely stores the IPFS address, facilitating the identification and retrieval of the associated manifest. A Chromium-based browser extension has been developed to validate text documents accessed by users. When activated, the extension analyses the webpage, detects paragraphs containing watermarked text, and cross-references them with data stored on IPFS and entries recorded in the smart contract, ensuring authenticity and provenance.

5.3. Platform Evaluation Across Social Media

The prototype of CERVANTES was tested in a controlled environment using basic HTML pages to conduct a preliminary evaluation. Subsequently, we assessed the platform's performance in real-world web environments through the following protocol: *i)* posting a text containing all the special characters listed in Tables 1 and 2 on a social media platform to verify their compatibility, and *ii)* posting multiple encrypted contents to examine the effectiveness of the watermark extraction and validation process.

The proposed platform was evaluated across the following twelve social media platforms: Bluesky, Discord, Facebook, LinkedIn, Mastodon, Quora, Reddit, Telegram web, Whatsapp web, Wordpress, X (formerly Twitter), and Youtube.

All the evaluated platforms demonstrated full compatibility with the special characters outlined in Table 1, enabling the successful extraction and validation of the watermarked text. Minor adjustments were required for certain platforms; for instance, Reddit required the implementation of dedicated code to remove predictable formatting characters (*i.e.*, additional lines and

spaces at the start and end of HTML section containing text) that would otherwise interfere with the watermark extraction process. Similarly, on Facebook, the watermark could only be fully detected if the entire text was displayed (*i.e.*, the "show more" is clicked).

Whitespace support varies significantly across platforms. Discord, LinkedIn, WordPress, and X do not support the "En quad" whitespace (code U+2000). In contrast, Facebook, Quora, and Telegram Web do not support any of the whitespace characters listed in Table 2. For the first group of social media, the remaining whitespace still allows for the embedding of two bits of the watermark. While this limitation increases the length of the text required for watermark embedding, it does not significantly impact overall performance and the verification process.

It is worth noting the following characteristics of Quora and Mastodon. Quora imposes a 250-character limit on question text, although no such restriction applies to answers. Mastodon, on the other hand, features a multi-server architecture in which each server is highly customizable, which may require tailored solutions, as was the case with Reddit.

5.4. Users Evaluation

To preliminarily assess the effectiveness of the CERVANTES platform in enhancing user awareness, we conducted an A/B test involving 20 participants. The cohort was structured to ensure a balanced representation in terms of age (ranging from 18 to 65 years), education level (including both secondary school and university graduates), and gender. It is worth noting that the sample size is limited: achieving statistical significance in theoretical terms would require several thousand participants. Nevertheless, the cohort

enabled us to gather qualitative insights into users' responses to the introduction of a new interaction element (*i.e.*, the browser plug-in) within the content evaluation process.

Participants were then divided into two groups, with only one granted access to CERVANTES. Both groups were asked to evaluate ten news articles covering five different topics (*i.e.*, politics, economy, culture, science, and sports). For each topic, one article was slightly altered to reverse its original meaning. Participants were instructed to assess the reliability of the entire article and, if deemed unreliable, identify the specific paragraphs that had been modified. The CERVANTES group had the option to use the platform at their discretion, ensuring a realistic scenario without imposing any constraints.

To assess and compare the performance of the two groups, we employed a set of well-established metrics commonly used for classifier evaluation, namely sensitivity, specificity, F1-score, and accuracy. The results are presented in Table 5.

Table 5: A/B test results for the two participant groups.

Group	**Sensitivity**	**Specificity**	**F1-score**	**Accuracy**
Without CERVANTES	0.778	0.311	0.631	0.544
Using CERVANTES	0.818	0.655	0.756	0.736

In our context, sensitivity represents the ability to correctly identify reliable news, while specificity denotes the ability to detect manipulated articles. These metrics assess the participants' ability to distinguish reliable news and

evaluate the effectiveness of the CERVANTES platform in supporting this process. In particular, the participants who had access to CERVANTES chose to use it in 68.2% of cases. The improvement of the group that autonomously chose to use CERVANTES is clearly reflected in the summary metrics, with F1-score and accuracy increasing by 19.81% and 35.29%, respectively. These results highlight the platform's effectiveness in enhancing users' ability to assess news reliability. Also, a significant improvement was observed in the ability to identify altered paragraphs, with the group that did not have access to CERVANTES achieving only 20.0%, compared to 61.8% for the group that used the platform.

6. Discussion

In this section, we critically analyse the crucial aspects of the CERVANTES platform, assessing the potential limitations and outlining possible strategies for enhancement.

The text watermarking technique is designed for languages that use the Latin script, which may limit its applicability in non-Western contexts. However, according to Britannica[3], the Latin alphabet is the most widely adopted writing system globally, used by nearly 70% of the world's population. Furthermore, many Western democracies, which are frequent targets of online information disorder campaigns [55, 56], predominantly employ the Latin alphabet. Therefore, this constraint does not represent a significant limitation. Currently, the Latin alphabet serves as the primary writing system for

[3]"The World's 5 Most Commonly Used Writing Systems", `https://www.britannica.com/list/the-worlds-5-most-commonly-used-writing-systems`.

numerous widely spoken languages, including English, Spanish, French, Portuguese, German, and Italian, with approximately 1.5 billion, 595 million, 321 million, 300 million, 200 million, and 81 million speakers worldwide, respectively. As such, it encompasses a substantial proportion of global online users.

It is worth noting that the proposed watermarking method, which relies on the use of specific Unicode characters with well-defined visual properties, is font-dependent and therefore susceptible to reformatting attacks. It constitutes only a partial limitation of the CERVANTES platform. Firstly, the platform is designed to interface with twelve social media services, offering broad coverage in terms of user reach. The list of homoglyphs presented in Tables 1 and 2 has been verified to be supported by all twelve platforms. It is important to emphasise that, on these platforms, users are not given the option to select or modify the font used for publishing content. Furthermore, in the event of a reformatting attack involving a change of font, the watermark is not removed; rather, some characters may appear slightly different. As a result, the provenance verification process would remain feasible. Only the complete removal of the watermark would compromise the verification process, a situation that, by design, would prompt the browser extension to report the content as unverifiable, even if the content itself remains unchanged.

Another critical aspect concerns the time required to register the manifest, the costs associated with maintaining the blockchain, and scalability. According to the protocol, news content can only be circulated after establishing the necessary groundwork for validation (*i.e.*, manifest storage).

While this process may introduce a slight delay, it represents a reasonable trade-off, as CERVANTES is not intended for real-time applications, such as stock trading, where immediate execution is critical. Regarding maintenance costs, these primarily consist of a one-time expense for uploading a manifest to the smart contract and recurring costs associated with maintaining storage on IPFS. The initial cost is determined by the gas fees required to execute an operation on an Ethereum smart contract. For ongoing storage, several platforms provide "pinning" services to ensure the persistence of IPFS documents, typically at an average rate of $0.1 per GB. For instance, assuming an average JSON manifest containing 20 paragraphs with a size of 2KB, the monthly storage cost would approximate $0.1 for 20,000 manifests. Moreover, the modular architecture of the CERVANTES platform enables the substitution of this component with alternative blockchain technologies that may offer enhanced performance or cost-efficiency. Finally, with the increasing volume of news content sealing and provenance checking requests, the CERVANTES platform - with minimal adjustments - could also address the scalability challenge. Specifically, in the event of a substantial influx of data (*i.e.*, new texts to be sealed) and transactions (*i.e.*, provenance verification requests), the platform could adopt modular and complementary strategies to maintain efficiency and cost-effectiveness. These might include the use of multiple blockchains to distribute certification requests, as well as a multi-tiered verification process designed to optimise provenance verification requests, which are computationally less intensive but expected to be more numerous. For example, the sealing phase could be managed through different blockchain technologies offering diverse service characteristics, while the

browser extension may interface with multiple blockchains operating at varying levels of granularity, such as national systems, journalist associations, or individual newspapers, in a coordinated manner. This architecture would enhance both performance and operational efficiency. Therefore, high volumes of content and frequent transactions do not pose a scalability constraint.

Although this lies outside the scope of the present work, as it concerns a purely blockchain-related security issue, we note that the platform is not entirely immune to poisoning attacks on the blockchain. In such attacks, a malicious actor introduces manipulated data with the aim of compromising the integrity of the ledger. In its current configuration, the platform would produce false positives in the presence of such attacks. Nevertheless, the adoption of multiple blockchains with cross-verification mechanisms, combined with a decentralised consensus-based system for verifying the identity and trustworthiness of authors prior to publication, may prove effective in mitigating these risks.

CERVANTES addresses the challenge of online information disorder from an orthogonal perspective, offering a user-friendly solution tailored to the behaviour of online readers who often skim headlines, exhibit a limited inclination to delve deeper into topics or face difficulties in distinguishing fake news [57]. The platform aims to maintain independence from social media platforms, which have limited incentives to curb the spread of partially true information that enhances user engagement [58]. To this end, while the selection of platforms analysed may initially appear restrictive, it encompasses 333 widely used platforms, including Facebook, X, Reddit, and Threads, which collectively account for approximately 2.9 billion, 611 million, 267.5

million, and 275 million monthly active users, respectively. Furthermore, given the orthogonal nature of our approach to online information disorder, a quantitative and qualitative comparison with machine learning methods focused on direct content evaluation is scarcely feasible. Nonetheless, we do not consider the CERVANTES platform a definitive solution to the problem, but rather an additional tool that can be made available to users. We believe that future developments may well integrate multiple approaches, both provenance verification and content evaluation, for instance, by triggering a semantic content analysis only when the provenance verification platform is unable to resolve the pointer to the blockchain. This would have the dual advantage of reducing the computational workload and relying on a set of previously certified reference news items for any given topic.

Finally, we briefly clarify the role of LLMs in online information disorder and the potential contribution of CERVANTES in this context. Previous research has explored the idea of embedding watermarks into textual content generated by service providers, such as LLMs. Major industry players are actively seeking solutions along these lines[4,5]. The main reason is that text generation services based on LLMs are widely used by malicious actors to produce information disorder and create information pollution on a global scale [59]. CERVANTES would enable the use of such generative tools by

[4] "Google Is Paying Publishers to Test an Unreleased Gen AI Platform" https://www.adweek.com/media/google-paying-publishers-unreleased-gen-ai.

[5] "Meta has created a way to watermark AI-generated speech" https://www.technologyreview.com/2024/06/18/1094009/meta-has-created-a-way-to-watermark-ai-generated-speech.

legitimate news producers while enhancing reader awareness in situations where content verification is not feasible.

In the future, the platform may be enhanced with truth and consensus-based mechanisms to mitigate the reliance on centralised evaluations of the reputation of news creators, an aspect that lies beyond the scope of this paper. An alternative approach could involve collaboration with established organizations, such as professional journalist associations or international fact-checking networks. This proposed platform aligns with the perspective outlined by historian Y. N. Harari, who emphasizes that the fundamental distinction between dictatorships and democracies lies in their approach to information management [60]. While dictatorships prioritize control and restriction, democracies focus on the sharing, consensus-driven evaluation, and dissemination of accurate information.

7. Conclusion

The widespread dissemination of online information disorder has emerged as a critical societal challenge, exerting a profound influence on public discourse. This phenomenon increasingly undermines societal values and democratic processes, fostering opinion polarisation on critical issues and reshaping perceptions of facts, truths, and beliefs without substantiated foundations. While notable progress has been made in the development of automated systems for detecting fake news, research on the human factors driving individuals to believe and share such information disorder remains in its infancy.

Instead of tackling the inherently complex task of reliably detecting online information disorder based solely on content, this study introduces the

CERVANTES platform, designed to assist end users in identifying trustworthy news sources. The platform employs an innovative integration of text watermarking techniques with blockchain technology to achieve fine-grained marking of textual news. This approach ensures that any sufficiently long excerpt of the original news, shared across web pages or social media platforms and containing a complete watermark sequence, can be authenticated for its provenance.

The primary objective of the CERVANTES platform is to empower online readers to form informed opinions about textual news independently, without reliance on third-party fact-checking organisations. It effectively secures even short excerpts of text, addressing the prevalent trend of users engaging only with news headlines. Its robustness in managing the provenance, trustworthiness, and verification of online news lies in its capacity to strike a balance between freedom of expression and the assurance of information quality. As future work, CERVANTES can be further enhanced with mechanisms grounded in truth and consensus, mitigating potential biases associated with centralised assessments of the reputation of news creators.

CRediT authorship contribution statement

Flavio Bertini: Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing. **Alessandro Benetton**: Conceptualization, Software, Investigation. **Danilo Montesi**: Conceptualization, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work is partially supported by the INCA project, funded under the EU Horizon Europe Programme, grant agreement n° 101061653.

References

[1] L. Monsees, Information disorder, fake news and the future of democracy, Globalizations 20 (1) (2023) 153–168.

[2] C. Wardle, H. Derakhshan, Information disorder: Toward an interdisciplinary framework for research and policymaking, Vol. 27, Council of Europe Strasbourg, F-67075 Strasbourg Cedex, 2017.

[3] K. M. d. Treen, H. T. Williams, S. J. O'Neill, Online misinformation about climate change, Wiley Interdisciplinary Reviews: Climate Change 11 (5) (2020) e665.

[4] Y. Wang, M. McKee, A. Torbica, D. Stuckler, Systematic literature review on the spread of health-related misinformation on social media, Social science & medicine 240 (2019) 112552.

[5] E. C. Tandoc Jr, Z. W. Lim, R. Ling, Defining "fake news" a typology of scholarly definitions, Digital journalism 6 (2) (2018) 137–153.

[6] D. De Beer, M. Matthee, Approaches to identify fake news: a systematic literature review, Integrated science in digital age 2020 136 (2021) 13–22.

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

[8] C. Chen, K. Shu, Can llm-generated misinformation be detected? (2024). arXiv:2309.13788.

[9] S. Branchetti, Testing attacks on structural text watermarking techniques, Master's thesis, Department of Computer Science and Engineering, University of Bologna (2023).

[10] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, science 359 (6380) (2018) 1146–1151.

[11] L. Toumanidis, R. Heartfield, P. Kasnesis, G. Loukas, C. Patrikakis, A prototype framework for assessing information provenance in decentralised social media: The eunomia concept, in: International Conference on e-Democracy, Springer, Springer, Cham, 2019, pp. 196–208.

[12] M. Choraś, M. Pawlicki, R. Kozik, K. Demestichas, P. Kosmides, M. Gupta, Socialtruth project approach to online disinformation (fake news) detection and mitigation, in: Proceedings of the 14th International Conference on Availability, Reliability and Security, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–10.

[13] B. Yousuf, M. A. Qureshi, B. Spillane, G. Munnelly, O. Carroll, M. Runswick, K. Park, E. Culloty, O. Conlan, J. Suiter, Provenance:

An intermediary-free solution for digital content verification (2021). arXiv:2111.08791.

[14] E. Strickland, This election year, look for content credentials: Media organizations combat deepfakes and disinformation with digital manifests, IEEE Spectrum 61 (01) (2024) 24–27.

[15] I. Cox, M. Miller, J. Bloom, J. Fridrich, T. Kalker, Digital watermarking and steganography, Morgan kaufmann, Burlington, Massachusetts, United States, 2007.

[16] S. Kaur, G. Babbar, A zero-watermarking algorithm on multiple occurrences of letters for text tampering detection, International Journal on Computer Science and Engineering 5 (5) (2013) 294.

[17] Y.-W. Kim, I.-S. Oh, Watermarking text document images using edge direction histograms, Pattern Recognition Letters 25 (11) (2004) 1243–1251.

[18] D. Huang, H. Yan, Interword distance changes represented by sine waves for watermarking text images, IEEE Transactions on Circuits and Systems for Video Technology 11 (12) (2001) 1237–1245.

[19] M. J. Atallah, V. Raskin, M. Crogan, C. Hempelmann, F. Kerschbaum, D. Mohamed, S. Naik, Natural language watermarking: Design, analysis, and a proof-of-concept implementation, in: Information Hiding: 4th International Workshop, IH 2001 Pittsburgh, PA, USA, April 25–27, 2001 Proceedings 4, Springer, Springer, Berlin, Heidelberg, Berlin, Heidelberg, 2001, pp. 185–200.

[20] K. Lambrecht, A framework for the analysis of cleft constructions, Linguistics 39 (3) (2001) 463–516.

[21] H. M. Meral, B. Sankur, A. S. Özsoy, T. Güngör, Sevinç, Emre, Natural language watermarking via morphosyntactic alterations, Computer Speech & Language 23 (1) (2009) 107–125.

[22] U. Topkara, M. Topkara, M. J. Atallah, The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions, in: Proceedings of the 8th workshop on Multimedia and security, Association for Computing Machinery, New York, NY, USA, 2006, pp. 164–174.

[23] O. Vybornova, B. Macq, A method of text watermarking using presuppositions, in: Security, Steganography, and Watermarking of Multimedia Contents IX, Vol. 6505, SPIE, SPIE, San Jose, CA, United States, 2007, pp. 613–622.

[24] O. Vybornova, B. Macq, Natural language watermarking and robust hashing based on presuppositional analysis, in: 2007 IEEE International Conference on Information Reuse and Integration, IEEE, IEEE, Las Vegas, NV, USA, 2007, pp. 177–182.

[25] M. Topkara, C. M. Taskiran, E. J. Delp III, Natural language watermarking, in: Security, Steganography, and Watermarking of Multimedia Contents VII, Vol. 5681, SPIE, SPIE, San Jose, California, United States, 2005, pp. 441–452.

[26] L. Y. Por, K. Wong, K. O. Chee, Unispach: A text-based data hiding method using unicode space characters, Journal of Systems and Software 85 (5) (2012) 1075–1082.

[27] R. A. Alotaibi, L. A. Elrefaei, Improved capacity arabic text watermarking methods based on open word space, Journal of King Saud University-Computer and Information Sciences 30 (2) (2018) 236–248.

[28] N. Mir, Copyright for web content using invisible text watermarking, Computers in Human Behavior 30 (2014) 648–653.

[29] M. Hellmeier, H. Norkowski, E.-C. Schrewe, H. Qarawlus, F. Howar, Innamark: A whitespace replacement information-hiding method, arXiv preprint arXiv:2502.12710 (2025).

[30] A. Narayanan, J. Bonneau, E. Felten, A. Miller, S. Goldfeder, Bitcoin and cryptocurrency technologies: a comprehensive introduction, Princeton University Press, Princeton, New Jersey, United States, 2016.

[31] C. Di Ciccio, G. Meroni, P. Plebani, On the adoption of blockchain for business process monitoring, Software and Systems Modeling 21 (3) (2022) 915–937.

[32] F. Donini, A. Marcelletti, A. Morichetta, A. Polini, Coordinating rest interactions in service choreographies using blockchain, Blockchain: Research and Applications (2024) 100241.

[33] S. Paul, J. I. Joy, S. Sarker, S. Ahmed, A. K. Das, et al., Fake news detection in social media using blockchain, in: 2019 7th international Con-

ference on smart computing & communications (ICSCC), IEEE, IEEE, Sarawak, Malaysia, 2019, pp. 1–5.

[34] Z. Shae, J. Tsai, Ai blockchain platform for trusting news, in: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), IEEE, IEEE, Dallas, TX, USA, 2019, pp. 1610–1619.

[35] W. Shang, M. Liu, W. Lin, M. Jia, Tracing the source of news based on blockchain, in: 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), IEEE, IEEE, Singapore, 2018, pp. 377–381.

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

[37] E. Sengupta, R. Nagpal, D. Mehrotra, G. Srivastava, Problock: a novel approach for fake news detection, Cluster Computing 24 (2021) 3779–3795.

[38] A. Alexandrescu, C. N. Butincu, Decentralized news-retrieval architecture using blockchain technology, Mathematics 11 (21) (2023) 4542.

[39] A. R. Faridi, R. Singh, F. Masood, M. Y. Salmony, Machine learning based novel framework for fake news detection and prevention using blockchain, in: 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, 2023, pp. 751–755.

[40] P. Rani, V. Jain, J. Shokeen, A. Balyan, Blockchain-based rumor detection approach for covid-19, Journal of Ambient Intelligence and Humanized Computing 15 (1) (2024) 435–449.

[41] A. D. Dwivedi, R. Singh, S. Dhall, G. Srivastava, S. K. Pal, Tracing the source of fake news using a scalable blockchain distributed network, in: 2020 IEEE 17th international conference on mobile ad hoc and sensor systems (MASS), IEEE, 2020, pp. 38–43.

[42] S. Dhall, A. D. Dwivedi, S. K. Pal, G. Srivastava, Blockchain-based framework for reducing fake or vicious news spread on social media/messaging platforms, Transactions on Asian and Low-Resource Language Information Processing 21 (1) (2021) 1–33.

[43] X. Li, L. Wei, L. Wang, Y. Ma, C. Zhang, M. Sohail, A blockchain-based privacy-preserving authentication system for ensuring multimedia content integrity, International journal of intelligent systems 37 (5) (2022) 3050–3071.

[44] T. K. Araghi, D. Megías, V. Garcia-Font, M. Kuribayashi, W. Mazurczyk, Disinformation detection and source tracking using semi-fragile watermarking and blockchain, in: Proceedings of the 2024 European Interdisciplinary Cybersecurity Conference, 2024, pp. 136–143.

[45] Q.-u.-A. Mastoi, M. F. Memon, S. Jan, A. Jamil, M. Faique, Z. Ali, A. Lakhan, T. A. Syed, Enhancing deepfake content detection through blockchain technology, Int. J. Adv. Comput. Sci. Appl. 16 (6) (2025).

[46] Z. Jiang, H. Wang, S. Han, A robust pdf watermarking scheme with versatility and compatibility, Multimedia Tools and Applications 83 (24) (2024) 64341–64367.

[47] S. N. KL, B. K. R, Text steganography: enhanced character-level embedding algorithm using font attribute with increased resilience to statistical attacks, Multimedia Tools and Applications (2024) 1–26.

[48] F. N. Al-Wesabi, F. Alrowais, H. G. Mohamed, M. Al Duhayyim, A. M. Hilal, A. Motwakel, Heuristic optimization algorithm based watermarking on content authentication and tampering detection for english text, IEEE Access 11 (2023) 86104–86111.

[49] C. Stomeo, Text watermarking e social network: uno studio sperimentale, Master's thesis, Department of Computer Science and Engineering, University of Bologna (2016).

[50] M. N. Jones, D. J. Mewhort, Case-sensitive letter and bigram frequency counts from large-scale english corpora, Behavior research methods, instruments, & computers 36 (3) (2004) 388–396.

[51] Panckhurst, Rachel and Détrie, Catherine and Lopez, Cédric and Moïse, Claudine and Roche, Mathieu and Verine, Bertrand, 88milSMS. A corpus of authentic text messages in French, Banque de corpus CoMeRe. Chanier T.(éd)-Ortolang: Nancy (2016).

[52] Kupietz, Marc and Lüngen, Harald and Kamocki, Paweł and Witt, Andreas, The German reference corpus DeReKo: New developments–new

opportunities, in: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 2018.

[53] Landro, Nicola and Gallo, Ignazio and La Grassa, Riccardo and Federici, Edoardo, Two new datasets for italian-language abstractive text summarization, Information 13 (5) (2022) 228.

[54] M. Bashardoost, M. S. Mohd Rahim, T. Saba, A. Rehman, Replacement attack: A new zero text watermarking attack, 3D Research 8 (1) (2017) 8.

[55] M. Reglitz, Fake news and democracy, J. Ethics & Soc. Phil. 22 (2022) 162.

[56] D. Gordon, Targeted systems and democracy: Russia, iran, and china's cyber threats and disinformation campaigns to weaken and undermine western democracies, Master's thesis, Utica College (2020).

[57] N. Newman, R. Fletcher, C. T. Robertson, A. R. Arguedas, R. K. Nielsen, Reuters institute digital news report 2024, Reuters Institute for the study of Journalism (2024).

[58] N. Marchal, B. Kollanyi, L.-M. Neudert, P. N. Howard, Junk news during the eu parliamentary elections: Lessons from a seven-language study of twitter and facebook, University of Oxford (2019).

[59] J. Zhou, Y. Zhang, Q. Luo, A. G. Parker, M. De Choudhury, Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–20.

[60] Y. N. Harari, Nexus: A Brief History of Information Networks from the Stone Age to AI, Random House, 1745 Broadway New York, NY 10019, 2024.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Flavio Bertini: Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing.

Alessandro Benetton: Conceptualization, Software, Investigation.

Danilo Montesi: Conceptualization, Writing – review & editing, Supervision.