Distributed Ledger and Text Watermarking for Fine-Grain Provenance Checking of Textual Content

Flavio Bertini
Department of Mathematical, Physical and Computer Sciences
University of Parma
Parma, Italy
flavio.bertini@unipr.it

Alessandro Benetton
Department of Computer Science and Engineering
University of Bologna
Bologna, Italy
alessandro.benetton@studio.unibo.it

Danilo Montesi
Department of Computer Science and Engineering
University of Bologna
Bologna, Italy
danilo.montesi@unibo.it

Abstract

Information disorder has become a major societal challenge, impacting public discourse and democracy. This phenomenon has been exacerbated by the spread of social media platforms, affecting various areas, ranging from national elections to public health. Addressing fake news through a manual approach (*e.g.*, human fact-checking) is unfeasible due to the rapid production of textual content. At the same time, applying automatic tools is equally challenging, primarily due to the ambiguity of natural language. In this paper, we addressed online information disorder from a different perspective by proposing a platform that supports trustworthy and reputable news producers and enhances awareness among readers across various social media. Specifically, the proposed platform enables news producers to automatically embed a unique watermark in the text they create, ensuring that the news cannot be manipulated or misattributed. The watermarking is embedded in a fine-grained way, allowing even small extracts of the news to be shared while preserving traceability. Additionally, the association between the watermark and the news item is recorded in a distributed ledger, preventing further manipulation that could arise from centralised management. The aim is to enable readers to make more informed decisions about the content they encounter, even when engaging with excerpts of the original document, minimising reliance on external fact-checking organisations.

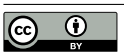
CCS Concepts

• **Human-centered computing** → **Social content sharing**; • **Applied computing** → *Document management*.

Keywords

Distributed ledger, Text watermarking, Text news sealing, Text provenance checking, Online misinformation

ACM Reference Format:
Flavio Bertini, Alessandro Benetton, and Danilo Montesi. 2025. Distributed Ledger and Text Watermarking for Fine-Grain Provenance Checking of Textual Content. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3701716.3717536



WWW Companion '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3717536

1 Introduction

Over the past fifteen years, the widespread availability of user-generated content on websites and online social media has promoted information sharing across ethnic, political, and geographical boundaries. This phenomenon has facilitated the formation of communities around shared interests, worldviews, and narratives. However, within this context of disintermediation, it has become increasingly apparent that information disorder has often undermined human rights and the core principles of democracy [22]. The proliferation of information disorder refers to the widespread increase of misleading or false information, including misinformation, disinformation, and malinformation, particularly amplified by digital platforms and social media. Specifically, false, inaccurate, or misleading information can mislead and manipulate individuals, erode trust in institutions and democratic processes, disrupt elections, and foster scepticism regarding critical issues such as vaccination and climate change [35], [39].

Information disorder constitutes a complex research challenge, primarily because accurately classifying it requires addressing both malicious behaviours and entirely lawful activities. Tandoc Jr et al. discussed several online behaviours that can be categorised as information disorders [31]. In particular, information disorder can manifest in several forms, such as *news satire*, where mock news has a transparent humorous motivation; *news parody*, which uses non-factual information for humour and, unlike news satire, does not clearly signal its non-journalistic nature; *news fabrication* and *photo manipulation*, where articles and photos with no factual basis are manipulated to construct a false narrative; and *propaganda*, referring to news stories produced by political entities to shape public perceptions. Not all of these forms represent malicious behaviour.

The phenomenon of information disorder on websites and online social media is expanding globally, making it imperative to tackle this issue. The complexity of the problem, which impacts diverse types of online content, is particularly pronounced with text content [10]. The literature presents various methodologies for addressing information disorder, categorised as language-based, topic-agnostic, machine learning, knowledge-based, and hybrid approaches [42]. Language-based approaches leverage linguistic and syntactic structures to detect inconsistencies. Topic-agnostic approaches utilise meta-information, such as a high volume of advertisements or the presence of sensational phrases, to identify potential misinformation. Machine learning approaches rely on annotated datasets to recognise patterns indicative of fake news. Knowledge-based approaches combine machine learning with knowledge engineering,

though they often struggle to keep pace with the rapid production of fake news. Hybrid approaches integrate human analysis with machine learning, drawing on fundamental attributes such as the article's text, the responses it generates, and the sources cited to support it. All these automatic methods present several shortcomings. In particular, the inherent ambiguity of natural language complicates the implementation of content-based solutions, rendering them challenging, if not impossible, to apply. Recently, the phenomenon has been exacerbated by the massive use of Large Language Models (LLMs), which can generate human-like text infused with unverified facts [7]. On the other hand, any manual approach is unfeasible: the vast volume of text shared daily makes the implementation of human-based fact-checking solutions impractical.

A prominent activity among online users is the re-sharing of content. When it comes to extensive or subscription-based text content, this re-sharing often manifests as copying and pasting selected portions of the text. This behaviour enables users to disseminate specific segments of the text content without sharing the entire document, which is particularly common with lengthy articles or content behind paywalls. Consequently, the re-sharing of content frequently results in the obfuscation of the source of the text, undermining reader trust. Furthermore, verifying the source of text content presented as legitimate news can be challenging or even impossible, especially if it includes seemingly authoritative brand integrity markers, such as the blue checkmark on X (formerly Twitter), which the company grants after an account verification process but does not actually guarantee the content disseminated by that account.

In this paper, we address the issue of online information disorder in text news by proposing a platform designed to seal text content and trace its source provenance. The platform combines text watermarking techniques with distributed ledger technologies to ensure that the news has not been manipulated, even when re-shared partially, and to enable the identification of the text content's source. Specifically, text watermarking enables continuous and fine-grained marking of lengthy news articles, facilitating watermark identification even after portions of the text are copied and pasted. The embedded watermark is created using the data and metadata of the original text content, seamlessly integrated into the text and recorded on a distributed ledger for verification purposes. Additionally, the distributed ledger component stores essential information in a decentralised and distributed manner, ensuring a consensus-based approach to authenticating textual news content and enabling the traceability of this content across various social media platforms. In particular, a browser extension enables the reader to check in real-time whether the content they are reading conforms to the original. Even if the content represents only a portion of the original, the extension can provide additional information by retrieving it from the distributed ledger, while any manipulations are promptly brought to the reader's attention.

The proposed platform is designed to be cross-platform and language-independent, capable of operating at scale. Specifically, it functions independently of the social media platform used, allows watermarking texts that use the Latin alphabet regardless of the language, and supports high-volume textual content production without human intervention. The Latin alphabet is widely used by many Western democracies, which are often the primary target of online disinformation [11, 26]; therefore, it does not constitute a significant limitation. Britannica highlights that the Latin alphabet is the most widely used writing system, utilised by nearly 70% of the global population[1]. Currently, the Latin alphabet serves as the primary writing system for a wide variety of languages, including English, Spanish, French, Portuguese, German, and Italian, spoken by approximately 1.5 billion, 595 million, 321 million, 300 million, 200 million, and 81 million people worldwide, respectively. It influences a significant proportion of online users.

Moreover, unlike most approaches in the literature, which primarily focus on assessing the semantic content of news, our platform addresses disinformation from a novel perspective, offering a straightforward and user-friendly solution specifically designed for online readers, who are often inattentive or hesitant to adopt complex systems. Existing methodologies encounter significant practical limitations, including the absence of real-time applicability, reliance on extensive language-specific datasets, and dependence on platforms that lack incentives to mitigate misinformation, as their engagement-driven models may, instead, amplify its spread [19]. Also, integrating a distributed ledger directly mitigates challenges associated with centralisation.

These functionalities enable both small and large news organisations, as well as freelancers, to track and share text content across diverse platforms effectively. Furthermore, the platform's robustness in managing online news provenance, trustworthiness, and verification lies in its ability to balance freedom of expression and information quality. The proposed platform overcomes the content-based limitations of conventional automatic approaches and mitigates the shortcomings of impractical and inaccurate manual methods, which censorship dynamics may influence. The platform can later be supplemented with truth and consensus-based mechanisms, such as the solution proposed by Arquam et al. in [3], to overcome any centralised assessment of the reputation of news creators, which is beyond the scope of this paper.

The remainder of the manuscript is organised as follows. In Section 2, we review existing research closely related to our study. The proposed platform is detailed in Section 3, while implementation details are discussed in Section 4. Finally, we conclude with remarks and suggestions for future research in Section 5.

2 Related Work

In this section, we describe literature from different domains, at the intersection of which our work lies. Specifically, we delve into online disinformation, text watermarking, and distributed ledger topics to verify the provenance of textual content.

In recent years, the prevalence of information disorder has escalated significantly. While the internet has facilitated a plurality of voices and democratised access to information, it has simultaneously introduced new technological vulnerabilities. A study conducted by MIT revealed that fake news propagates more rapidly, deeply, and broadly than truthful information, with falsehoods being 70% more likely to be retweeted [36]. Wardle et al. [40] identified three categories of information disorder: misinformation, disinformation, and malinformation. However, they did not

[1] https://www.britannica.com/list/the-worlds-5-most-commonly-used-writing-systems

account for other forms of expression, such as satire and parody, which can complicate efforts to combat fake news [31]. Given the multi-dimensional nature of the problem, the ambiguity inherent in natural language, and the vast volume of information shared online daily, both manual and automated approaches to combating disinformation are often impractical and imprecise [42]. Existing EU projects often demand excessive end-user involvement to analyse a specific post and assess its trustworthiness [34], involve complex content evaluation phases [8], or fail to address fine-grained textual content [41]. Recently, a collaborative initiative involving Microsoft Research, BBC, CBC/Radio-Canada, The New York Times, and Truepic has proposed a platform designed to mitigate the spread of fraudulent images and videos on the internet [30]. Building on these initiatives, we propose a fully automated tracking and verification solution that integrates text watermarking techniques with distributed ledger technology to ensure the integrity and provenance of information. This solution aims to enhance awareness among online readers.

Watermarking is the practice of embedding a mark in digital content to provide proof of ownership, authorship attribution, and verification [9]. Among the various digital content watermarking techniques, text watermarking poses the greatest challenge due to its low embedding bandwidth and the limited number of alternative syntactic and semantic permutations available [5]. The text watermarking approaches can be categorised as follows.

Zero-watermarking - In these approaches, no actual mark is embedded in the text content; instead, characterizing features of the text are stored on a third-party authority server [16]. The primary drawbacks of zero-watermarking include the centralization of the service and a lack of transparency.

Image-based approaches - These methods require that the text be first scanned as a screenshot, after which a watermark is applied to the corresponding image [17], [13]. This approach modifies the nature of the original document and is often impractical for scenarios such as online social media, where image-based methods would be unnatural and significantly limit sharing activities.

Syntactic methods - These techniques work on the syntax of natural language text by altering the syntactic tree of a sentence to embed a watermark. Examples include syntactic operations such as clefting, passivization, or activization [4], [18], [20]. The low embedding capacity of these methods is a significant limitation. Moreover, the assumption that different syntactic forms have identical meanings is not always valid, further restricting their use.

Semantic methods - These approaches exploit the similarity in meaning between different words, replacing words with their synonyms [33]. Other techniques operate at the sentence level, leveraging the implicit presuppositions of each sentence [37], [38]. However, even when combined with syntactic approaches to increase the payload [32], semantic methods share the limitations of syntactic methods and are highly dependent on the language.

Structural methods - These methods do not alter the text content but instead modify its structure, for example, by filling an empty line [25] or using different Unicode whitespace characters [1], [21]. Since the watermark is embedded in the underlying representation of the text, these methods have the significant advantage of preserving the original content without relying on an external database. The proposed platform incorporates a structural text watermarking approach for sealing text, which preserves content and length based on our previous findings [5].

Blockchain technology enables storing transactional data on an open, decentralised ledger in a verifiable and immutable manner. For efficiency reasons, transactional data (*i.e.*, tokens) are stored in time-stamped blocks and linked to a previous block [23]. The decision to append a new block of transactions to the already formed chain of blocks is collectively taken by the peers running the system, upon running a distributed consensus protocol. Once appended, a block cannot be deleted from the blockchain and checking the validity of a transaction can be publicly done by any client that connects to any peer that stores the entire blockchain. Initially developed for implementing decentralised digital currencies, this technology has found applications in scenarios in which it is relevant to guarantee authenticity and verification of any kind of data without relying on a centralised, trusted party [6], [14]. Several solutions based on blockchain have been proposed to combat online fake news [24], [27], and [28]. Unlike other proposed solutions, we use a distributed ledger to track text directly down to the paragraph level, without relying on content evaluation methodologies. In [3], Arquam et al. proposed an interesting solution that employs a blockchain-based framework to verify information propagation. Their framework's reliability system is based on social network parameters, whereas in our case, we do not require a network structure to validate the textual content. This makes our approach applicable to websites and more easily integrable with various online information production solutions.

3 The Proposed Platform

This section presents the proposed platform, detailing its two primary components. Firstly, we discuss the sealing of text news through text watermarking, as depicted above the dotted line in Figure 1. Secondly, we elaborate on the verification process, illustrated below the dotted line in Figure 1.

3.1 Text Content Sealing

The structural text watermarking method for content sealing relies on homoglyph-based character substitution. Homoglyphs[2] are Unicode characters that appear similar but have different internal representations. This similarity can be exploited to embed a watermark generated using a hash function to enhance robustness. Continuous and fine-grained watermarking, at the level of a few dozen words, is achieved by concatenating the watermark throughout the document, enabling the validation of any copy-pasted sub-portion long enough to hold a complete watermark. This is not a limitation, even when considering the character limits imposed by some social media platforms since the watermarking method requires between 46 and 101 characters to work properly [29]. To ensure integrity, the proposed platform stores hashes for all sub-portions, clearly defined to contain the entire payload (*i.e.*, the watermark). Introducing an invisible separator character before and after each sub-portion during watermark embedding ensures consistency in sub-portion separation between the sealing and validation phases. The embedded payload includes the *author ID*, *document ID*, and

[2]https://www.unicode.org/reports/tr36

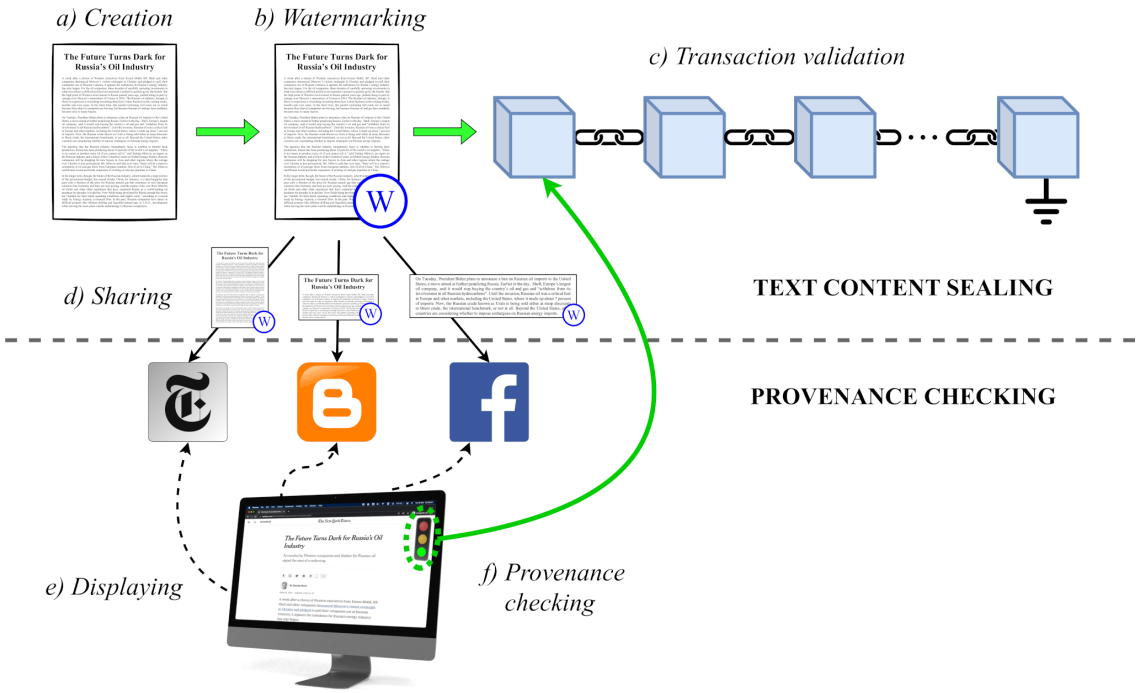


Figure 1: The overview of the proposed platform.

Table 1: The list of the used homoglyphs.

Character		Code	
Original	Homoglyph	Original	Homoglyph
-	-	U+002D	U+2010
C	C	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	j	U+006A	U+0458
l	l	U+006C	U+217C
v	v	U+0076	U+2174
x	x	U+0078	U+2179

Table 2: Frequency of the original symbols and letters used in the most common Latin languages. Languages are identified by the ISO 639-1:2002 codes.

Character	EN	ES	FR	PT	DE	IT
-	4.34%	*n.a.*	*n.a.*	*n.a.*	*n.a.*	*n.a.*
C	0.35%	3.87%	3.15%	3.35%	3.16%	4.30%
D	0.20%	4.67%	3.55%	4.21%	4.98%	3.39%
L	0.16%	5.24%	5.68%	3.00%	3.60%	5.70%
M	0.40%	3.08%	3.23%	5.07%	2.55%	2.87%
V	0.05%	1.05%	1.83%	1.72%	0.84%	1.75%
X	0.01%	0.14%	0.42%	0.28%	0.05%	0.00%
c	3.01%	3.87%	3.15%	3.35%	3.16%	4.30%
d	3.63%	4.67%	3.55%	4.21%	4.98%	3.39%
i	6.94%	5.28%	6.94%	5.49%	8.02%	10.18%
j	0.10%	0.52%	0.71%	0.30%	0.24%	0.00%
l	3.92%	5.24%	5.68%	3.00%	3.60%	5.70%
v	1.00%	1.05%	1.83%	1.72%	0.84%	1.75%
x	0.19%	0.14%	0.42%	0.28%	0.05%	0.00%

a *seal* (*i.e.*, hash digest) to ensure the integrity of the text content and confirm that it has not been tampered with.

Table 1 lists the used characters and their corresponding homoglyphs, which are compatible with most social media platforms [29]. Meanwhile, Table 2 reports the frequency of the original symbols and letters used in six Latin-based languages (*i.e.*, English, Spanish, French, Portuguese, German, and Italian), identified in the table by their ISO 639-1:2002 codes. The frequencies in English were computed using the *New York Times* corpus, which contains approximately 14 million words [15], while the frequencies for other languages were calculated using the Stefan Trost service[3], which employs texts with a good mix of different literary genres. In addition to the homoglyphs shown in Table 1, the invisible character

[3]https://www.sttmedia.com/characterfrequencies

U+200B, known as *Zero Width Space*, is used as a separator for the various rounds of embedding the watermark.
In practice, the embedding algorithm follows these steps: *1)* a separator character is inserted; *2)* text characters are scanned until a character from Table 1 is encountered; *3)* if the next watermark bit is 1, the character is replaced with its homoglyph counterpart; otherwise, the original character remains unchanged; *4)* steps *2* and *3* are repeated until all watermark bits are processed. Once the last bit of the watermark is consumed, another separator character is inserted, resetting the watermark bit sequence to its initial position to begin a new embedding round until the end of the document. The extraction process follows the same steps: initially searching for separator characters, then identifying characters listed in Table 1 to determine the corresponding binary sequence of the watermark.

The described watermarking method allows text content creators to seal their news (phases *a* and *b* in Figure 1). The advantages are twofold: firstly, authors can improve traceability by sealing the text multiple times and uploading each signed version to different social media platforms (phase *d* in Figure 1), enabling them to track the origins of specific quotes; secondly, it provides an immediate means to detect citation tampering and grants readers access to additional information (*e.g.*, author, timestamps and source URL), thereby enhancing transparency and reliability.

3.2 Provenance Checking

Relying solely on embedded information to validate text content is not optimal for several reasons. Firstly, it is relatively easy to produce a document with a legitimate watermark without using the sealing tool. Moreover, embedding all additional information within text content using watermarking is impractical, as the payload ideally needs to be minimised to reduce the number of characters required for embedding. The longer the payload, the more text is required to embed it. Since the platform must be able to verify sub-portions of the text content, we designed a minimal payload to ensure efficient verification. Furthermore, the proposed provenance checking protocol allows tracking watermarks in a distributed ledger, enabling the validation of portions of the original text without duplicating the entire document on the ledger.

Therefore, the platform includes a distributed ledger as the second component (phase *c* in Figure 1), ensuring a consensus-based approach for storing additional information and a decentralised mechanism for verifying text content (phase f in Figure 1). The distributed ledger serves as the single point of truth for the platform and must satisfy the following requirements: reliability, ensuring all stored information is valid and dependable; independence, ensuring no single organisation controls the stored information; and immutability, ensuring that once information is stored, it remains permanent.

This secondary component of the platform provides readers with a tool to automatically validate the content they encounter in real-time. It includes a browser extension that, upon activation, scans the displayed web page for watermarked content. If a watermark is detected, the plugin initiates the validation process using the distributed ledger and provides the user with a reliability indicator and any additional retrieved information (phases *e* and f in Figure 1). This validation process operates regardless of the length of the original text excerpt being read by the user. This process also highlights eventual inconsistencies between the watermark and the storage, alerting the reader that something might be wrong with part of the content.
Figure 2 shows the outcome of the validation process of the visualised text content. Once the validation is complete, the extension presents the results to the user, offering additional information alongside the identified watermarked content. In particular, each paragraph receives a status that can be *valid*, *unknown*, or *invalid* and is coloured using green, yellow, and red, respectively. The *valid* status indicates that the paragraph has been copied from a known source without alteration, meaning the extracted watermark matches a valid entry in the distributed ledger. The *unknown* status indicates an incomplete or invalid watermark, which may occur if the copied text portion does not contain the entire watermark or if part of the seal has been lost. Lastly, the *invalid* status implies that the extracted watermark does not match any entry in the distributed ledger, suggesting an active attempt to replace the content's seal.

3.3 Content Sharing & Verification Protocol

In this section, we briefly describe the sealing, sharing and verification steps based on the two basic components described in the previous sections.

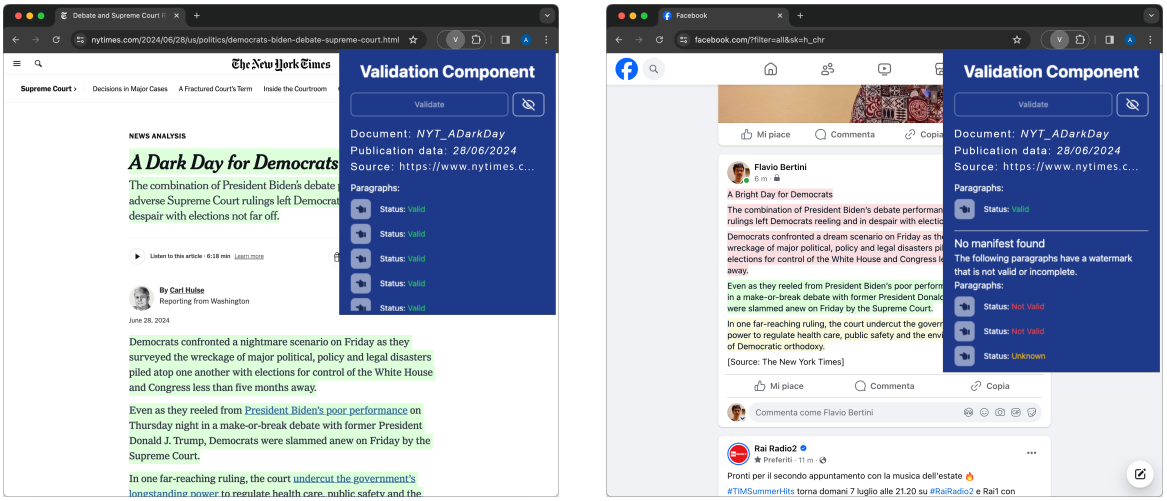
Figure 3 illustrates the four main steps of the text content sharing and verification protocol. In the first stage, *document writing & watermarking*, the news creator types the text content, which is automatically watermarked using the *text content sealing* components. Next, in the *manifest creation & transaction storing* stage, the manifest of the newly created document is packaged, stored in a distributed storage system, and its address is recorded on a distributed ledger. The manifest contains essential information about the author, the document, and each sub-portion of the original text, facilitating validation. The text can then be shared in whole or in part on web pages and social media platforms, which constitutes the *document/short excerpts sharing* stage. Finally, during the *visualisation & validation* stage, any sub-portion of the original document that is sufficiently long to contain a complete watermark sequence can be validated using the *provenance checking* components.

The robustness of the proposed text watermarking method was demonstrated against various attacks, including partial copy-and-paste, insertion, deletion, replacement, and retyping, with up to 20% of the text being affected [5]. The results confirm the effectiveness of the method even when a substantial portion of the original text is modified. The following section will address various implementation choices in the development of the platform prototype.

4 Prototyping and Discussion

In this section, we briefly present the rationale behind some technical decisions to make the platform replicable. We also evaluated the platform using four of the most widely used social media platforms that support text sharing.

The proposed platform has two main components: *watermarking & sealing* and *storing & validation*. The first component embeds a watermark into the text document in a negligible amount of time and generates a specific file, the manifest, which contains additional



(a) (b)

Figure 2: The browser extension for provenance checking: enhancing reader awareness on websites *(a)* and social media *(b)*.

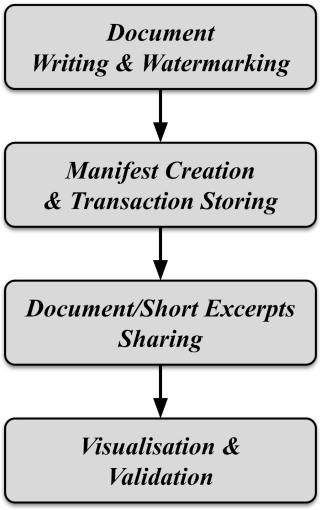


Figure 3: The flow of the text content sharing and verification protocol.

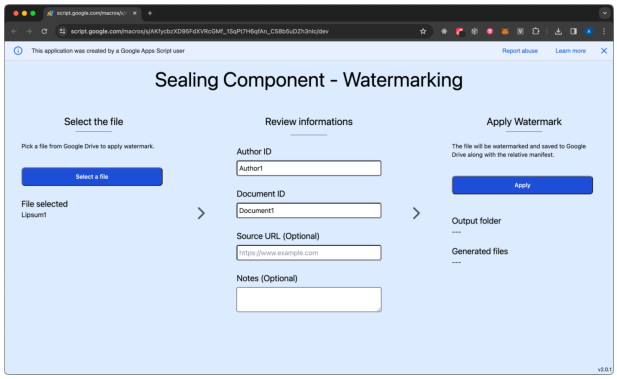
information intended for storage on the distributed ledger. We chose Google Apps Script as the scripting language for watermarking text content created using Google Docs. As shown in the interface depicted in Figure 4a, the script watermarks the text document and produces a JSON file (*i.e.*, the manifest) containing the *author ID*, *document ID*, *sealing timestamp*, *source URL*, and a *list of hash digests* for each sub-portion of the original text document. The list of hash codes will be useful for validating the provenance of short text excerpts.

The second component is a web app that stores the generated JSON file (*i.e.*, the manifest) in the InterPlanetary File System (IPFS) and registers the upload address within a smart contract. This activity can be easily performed using the intuitive interface depicted in Figure 4b. In the future, we plan to automate both the watermarking and uploading processes by integrating the interface. IPFS is a decentralised and distributed storage system that operates as a peer-to-peer network to enhance content availability and fault tolerance. IPFS generates addresses based on the content's hash, ensuring document immutability and immediate recognition of duplicates. The smart contract is implemented using *testnets*, an Ethereum-based blockchain for testing purposes, that holds the IPFS address, enabling identification and retrieval of the stored manifest. A browser extension built on the Chromium framework validates text documents viewed by the user. Upon request, the extension scans the page, identifies paragraphs with watermarked text, and verifies them against IPFS data and the smart contract.

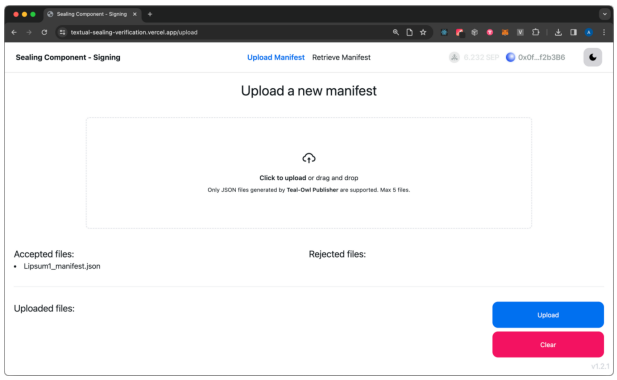
The prototype underwent testing in a controlled environment using basic HTML pages for a preliminary evaluation. However, the complexity of the actual web environment, particularly in social media contexts, necessitates further assessment. We evaluated the platform's efficacy in real-world web settings using the following protocol: *1)* posting a text containing all special characters listed in Table 1 on a social media platform to verify their support, and *2)* posting multiple encrypted contents to test the watermark extraction and validation process.

We evaluated the proposed platform across Facebook, X (formerly Twitter), Reddit, and Threads, which have respective monthly active user bases of 2.9 billion, 611 million, 267.5 million, and 275 million. All platforms fully support the special characters detailed in Table 1 and the validation of the watermarked text content. Notably, Reddit introduces additional lines and spaces at the start and end of HTML pages containing text, affecting hash calculation during validation. Thus, we included a special rule that detects Reddit as the platform and removes these predictable characters to ensure accurate validation results.

One of the significant considerations in utilising a distributed ledger is the associated cost. The proposed platform incurs a one-time expense for uploading a manifest to the smart contract, as



(a)



(b)

Figure 4: The Google Apps Script interface for watermarking text documents created in Google Docs *(a)* and for uploading the manifest *(b)*.

well as ongoing costs for data storage on IPFS. The one-time cost is determined by the gas fees required to execute operations on the Ethereum smart contract. The recurring cost is covered through platforms offering "pinning" services, typically at an average rate of $0.10 per GB. For instance, storing 20,000 JSON manifests, each 2KB in size, would result in an approximate monthly cost of $0.10. However, the modular architecture of the platform enables the substitution of this component with any other distributed ledger that provides improved performance or cost efficiency.

The rationale behind the choice of a distributed ledger is to eliminate a single point of failure and validation, as would occur with traditional databases. From an ethical standpoint, the platform cannot completely preclude the potential misuse of watermarking technology; however, the three-colour flag system empowers users to recognise that there is no absolute reliability in the information they are currently engaging with. In combating online misinformation, the proposed approach complements existing solutions that address the problem from an orthogonal perspective, such as identifying websites responsible for creating and disseminating misleading content [2]. Furthermore, the platform does not assume responsibility for assessing the credibility of content uploaded by users. Practically, a viable solution would be to collaborate with established organisations (*e.g.*, professional journalist associations), while also implementing a system for feedback and evaluation from other news creators. In practice, the proposed platform aligns with the direction described by historian Y. N. Harari, who highlights that the distinction between dictatorships and democracies lies in how they manage information [12]. The former is more focused on control, whereas the latter prioritise the sharing, consensus-based evaluation, and dissemination of accurate information. Also, the proposed platform can be combined with LLMs either to track text content produced by these models or to invalidate manipulated news content through their use. In the first case, fair use of the LLMs does not hinder the dissemination of validated, automatically generated content. On the other hand, any malicious manipulation of information would be rendered invalid.

5 Conclusion

The proliferation of online information disorder has become a major societal concern, significantly impacting public discourse. It increasingly affects societal values and democratic processes, polarising opinions on critical issues and redefining facts, truths, and beliefs without foundation. While advances in automated systems for detecting fake news have been made, research into the human factors that lead individuals to believe and share fake news remains in its early stages.

Rather than focusing on the challenging task of reliably detecting fake news, the proposed platform aims to help the end user identify trustworthy news. In particular, the proposed platform integrates a text watermarking technique with a distributed ledger to enable fine-grained watermarking of text documents. This allows any sufficiently long sub-portion of the original document shared on web pages and social media platforms containing a complete watermark sequence to be validated for provenance.

The proposed platform assists online readers in forming informed opinions about text news without depending on third-party fact-checking organisations. The watermarking method effectively seals short text excerpts, addressing the growing trend of users reading only news titles. The platform's strength in managing online news provenance, trustworthiness, and verification lies in its ability to balance freedom of expression with information quality.

Acknowledgments

This work is carried out under the framework of the INCA project, funded by the EU Horizon Europe Programme, grant agreement n° 101061653.

References

[1] Reem A Alotaibi and Lamiaa A Elrefaei. 2018. 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.

[2] Leandro Araujo, Joao MM Couto, Luiz Felipe Nery, Isadora C Rodrigues, Jussara M Almeida, Julio Reis, and Fabricio Benevenuto. 2024. Finding fake news websites in the wild. *arXiv e-prints* (2024), arXiv–2407.

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

[4] Mikhail J Atallah, Victor Raskin, Michael Crogan, Christian Hempelmann, Florian Kerschbaum, Dina Mohamed, and Sanket Naik. 2001. 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, 185–200.

[5] Simone Branchetti. 2023. *Testing Attacks on Structural Text Watermarking Techniques.* Master's thesis. Department of Computer Science and Engineering, University of Bologna.

[6] Bert-Jan Butijn, Damian A Tamburri, and Willem-Jan van den Heuvel. 2020. Blockchains: a systematic multivocal literature review. *ACM Computing Surveys (CSUR)* 53, 3 (2020), 1–37.

[7] Canyu Chen and Kai Shu. 2024. Can LLM-Generated Misinformation Be Detected? arXiv:2309.13788 [cs.CL] https://arxiv.org/abs/2309.13788

[8] Michał Choraś, Marek Pawlicki, Rafał Kozik, Konstantinos Demestichas, Pavlos Kosmides, and Manik Gupta. 2019. 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, 1–10.

[9] Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. 2007. *Digital watermarking and steganography.* Morgan kaufmann, Burlington, Massachusetts, United States.

[10] Dylan De Beer and Machdel Matthee. 2021. Approaches to identify fake news: a systematic literature review. *Integrated science in digital age 2020* 136 (2021), 13–22.

[11] Daniel Gordon. 2020. *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.

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

[13] Ding Huang and Hong Yan. 2001. 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.

[14] Huawei Huang, Wei Kong, Sicong Zhou, Zibin Zheng, and Song Guo. 2021. A survey of state-of-the-art on blockchains: Theories, modelings, and tools. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–42.

[15] Michael N Jones and Douglas JK Mewhort. 2004. Case-sensitive letter and bigram frequency counts from large-scale English corpora. *Behavior research methods, instruments, & computers* 36, 3 (2004), 388–396.

[16] Sukhpreet Kaur and Geetanjali Babbar. 2013. 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] Young-Won Kim and Il-Seok Oh. 2004. Watermarking text document images using edge direction histograms. *Pattern Recognition Letters* 25, 11 (2004), 1243–1251.

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

[19] Nahema Marchal, Bence Kollanyi, Lisa-Maria Neudert, and Philip N Howard. 2019. Junk news during the EU parliamentary elections: Lessons from a seven-language study of Twitter and Facebook. *University of Oxford* (2019).

[20] Hasan Mesut Meral, Bülent Sankur, A Sumru Özsoy, Tunga Güngör, and Emre Sevinç. 2009. Natural language watermarking via morphosyntactic alterations. *Computer Speech & Language* 23, 1 (2009), 107–125.

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

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

[23] Arvind Narayanan, Joseph Bonneau, Edward Felten, Andrew Miller, and Steven Goldfeder. 2016. *Bitcoin and cryptocurrency technologies: a comprehensive introduction.* Princeton University Press, Princeton, New Jersey, United States.

[24] Shovon Paul, Jubair Islam Joy, Shaila Sarker, Sharif Ahmed, Amit Kumar Das, et al. 2019. Fake news detection in social media using blockchain. In *2019 7th international Conference on smart computing & communications (ICSCC)*. IEEE, IEEE, Sarawak, Malaysia, 1–5.

[25] Lip Yee Por, KokSheik Wong, and Kok Onn Chee. 2012. UniSpaCh: A text-based data hiding method using Unicode space characters. *Journal of Systems and Software* 85, 5 (2012), 1075–1082.

[26] Merten Reglitz. 2022. Fake news and democracy. *J. Ethics & Soc. Phil.* 22 (2022), 162.

[27] Zonyin Shae and Jeffrey Tsai. 2019. AI blockchain platform for trusting news. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, IEEE, Dallas, TX, USA, 1610–1619.

[28] Wenqian Shang, Mengyu Liu, Weiguo Lin, and Minzheng Jia. 2018. Tracing the source of news based on blockchain. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*. IEEE, IEEE, Singapore, 377–381.

[29] Carlo Stomeo. 2016. *Text watermarking e Social Network: uno studio sperimentale.* Master's thesis. Department of Computer Science and Engineering, University of Bologna.

[30] Eliza Strickland. 2024. This Election Year, Look for Content Credentials: Media organizations combat deepfakes and disinformation with digital manifests. *IEEE Spectrum* 61, 01 (2024), 24–27.

[31] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. 2018. Defining "fake news" A typology of scholarly definitions. *Digital journalism* 6, 2 (2018), 137–153.

[32] Mercan Topkara, Cuneyt M Taskiran, and Edward J Delp III. 2005. Natural language watermarking. In *Security, Steganography, and Watermarking of Multimedia Contents VII*, Vol. 5681. SPIE, SPIE, San Jose, California, United States, 441–452.

[33] Umut Topkara, Mercan Topkara, and Mikhail J Atallah. 2006. 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, 164–174.

[34] Lazaros Toumanidis, Ryan Heartfield, Panagiotis Kasnesis, George Loukas, and Charalampos Patrikakis. 2019. A prototype framework for assessing information provenance in decentralised social media: The eunomia concept. In *International Conference on e-Democracy*. Springer, Springer, Cham, 196–208.

[35] Kathie M d'I Treen, Hywel TP Williams, and Saffron J O'Neill. 2020. Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change* 11, 5 (2020), e665.

[36] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science* 359, 6380 (2018), 1146–1151.

[37] Olga Vybornova and Benoit Macq. 2007. 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, 613–622.

[38] Olga Vybornova and Benoit Macq. 2007. 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, 177–182.

[39] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine* 240 (2019), 112552.

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

[41] Bilal Yousuf, M. Atif Qureshi, Brendan Spillane, Gary Munnelly, Oisin Carroll, Matthew Runswick, Kirsty Park, Eileen Culloty, Owen Conlan, and Jane Suiter. 2021. PROVENANCE: An Intermediary-Free Solution for Digital Content Verification. arXiv:2111.08791 [cs.CY] https://arxiv.org/abs/2111.08791

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