LOOM: a Privacy-Preserving Linguistic Observatory of Online Misinformation

Jeremie Clos*
University of Nottingham
United Kingdom
jeremie.clos@nottingham.ac.uk

Emma McClaughlin*
University of Nottingham
United Kingdom
emma.mcclaughlin@nottingham.ac.uk

Pepita Barnard
University of Nottingham
United Kingdom
pepita.barnard@nottingham.ac.uk

Tino Tom
University of Nottingham
United Kingdom
tinotom7@outlook.com

Sudarshan Yajaman
University of Nottingham
United Kingdom
pyajamansudarshan04@gmail.com

ABSTRACT

Online misinformation is an ever-growing challenge that can have a negative impact on individuals, societies, and democracies. We report on LOOM, a project that aims to build and validate a browser-based tool to detect and respond to misinformation in a trustworthy and privacy-preserving manner to protect end users and build public resilience to untrustworthy content. LOOM uses natural language processing techniques to detect and flag linguistic misinformation markers for end users in real time, whilst preserving the end-to-end encryption that protects the privacy and security of their online browsing and communication activities. We applied a citizen science framework to test the tool as an intervention to build user resilience to false information content and assess their trust in the tool. Feedback from participants indicates that the tool has the potential to improve user awareness of subtle language cues associated with misinformation and to help them critically evaluate the information they encounter. Overall, our experiment indicates a demand for tools to combat misinformation, but also highlights the challenges in creating a tool that is both effective and user-friendly.

CCS CONCEPTS

• **Information systems** → **Web mining**; *Crowdsourcing*; • **Security and privacy** → *Usability in security and privacy*.

ACM Reference Format:
Jeremie Clos, Emma McClaughlin, Pepita Barnard, Tino Tom, and Sudarshan Yajaman. 2024. LOOM: a Privacy-Preserving Linguistic Observatory of Online Misinformation. In *Second International Symposium on Trustworthy Autonomous Systems (TAS '24), September 16–18, 2024, Austin, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3686038.3686062

*Both authors contributed equally to this research.

TAS '24, September 16–18, 2024, Austin, TX, USA

ACM ISBN 979-8-4007-0989-0/24/09.
https://doi.org/10.1145/3686038.3686062

1 INTRODUCTION

Misinformation undermines the value of factual information on a large scale and can misguide individuals on a range of pressing issues [41]. For example, misinformation has been found to be influential in politics (e.g., the 2016 US Presidential Election [19] and Brexit [3, 19]); human health (e.g., COVID-19 vaccinations [13, 18, 22, 34]); and science (e.g., climate change [24, 40, 41]). The influence of misinformation has the potential to result in societal conflict, damage to human health and limiting response to climate breakdown, all with severe consequences. As a result of these potential risks, several organisations have voiced the need to take measures to address this issue, including The Royal Society [39], UNICEF [17], and Full Fact [12].

False information may be characterised as misinformation, disinformation, satire, propaganda, rumour, hoaxes, fake news, or conspiracy theories. These labels are assigned to false information on the basis of intention or motivation behind its production. For example, 'disinformation' is generally considered false information that has been created and/or intentionally shared to mislead others (e.g., [15, 38, 44]), the motivations for which are usually to cause harm or to obtain some kind of benefit [11, 38]. On the other hand, 'misinformation' is most often described as false information, inadvertently shared without the intention to cause harm [15, 44], whilst satire is false information provided in the form of news intended to amuse [1] and signals to its audience that it is not serious and has humorous intent [36]. Because conflicting definitions make distinguishing between different types of false information difficult and, since our focus is on reception rather than intention, we use the term 'misinformation' as shorthand for 'bad information' in line with Full Fact, an independent fact checking and campaigning organisation [12]. This approach is also necessary given that the relative levels of exposure to each type of false information online is currently unknown.

To understand the nature of the misinformation that individuals are exposed to online, it is necessary to examine the language they encounter in their day-to-day browsing. Traditionally, measuring misinformation is carried out by analysing the platform, for example crawling public posts on social media. This is becoming more challenging due to the wide adoption of end-to-end encryption and ethical guidelines for data processing. Since encryption provides security benefits, which are increasingly relied upon, interventions must be 'privacy-preserving by design'. Therefore, we

need to perform measurements and interventions at either end of the information lifecycle, either by preventing the spread of misinformation, or by protecting its consumer from it. This paper reports on a project that aims to demonstrate the feasibility of a browser-based tool to perform misinformation detection and intervention in a trustworthy and privacy-preserving manner to protect end users and build public resilience to untrustworthy content, as recommended by the Royal Society [39].

In Section 2 we will review the state of the literature in privacy-preserving data collection and analysis. Sections 3 and 4 will describe the tool that was built and how it was evaluated. In Sections 5 and 5.5 we discuss our results and the responsible research aspect of our work, before concluding with a set of recommendations in Section 6.

2 PRIVACY-PRESERVING ANALYTICS

The proliferation of personal data and the increasing concerns about privacy have led to extensive research on personal data management systems and privacy-preserving technologies. Privacy-preserving technologies allow for the processing of personal data in a way that simultaneously maximises its usefulness, while minimising the risks of invading the privacy of the individuals who generated it [31].

Standard approaches to privacy-preserving analytics are trusted execution environments, homomorphic encryption, secure multi-party computation, differential privacy, and personal data stores. On the other hand, Meurisch and Mühlhäuser [27] categorise privacy-preserving approaches to machine learning and analytics in four broad types:

- Data-modifying approaches
- Data-encrypting approaches
- Data-minimising approaches
- Data-confining approaches

Several approaches have been proposed to address the challenges of data ownership, access control, and privacy protection, combining those four approaches in a sensible way.

Personal Data Stores and Vaults. Early works like Personal Data Vaults (PDVs) [29] introduced the concept of user-centric data stores where individuals retain ownership and control over their data. PDVs allow for granular access control through various mechanisms like trace-auditing. Similarly, openPDS [9, 10] is a personal metadata management framework that enables individuals to collect, store, and share their metadata with fine-grained access control. Solid [37] also follows this paradigm, providing a decentralised platform where users' data is stored in personal online datastores (pods) that are independent of applications. Open Mustard Seed (OMS) [14] further expands this concept by incorporating a trust framework that enables secure storage and processing of personal data within a legally constituted structure.

Privacy-Preserving Application Platforms. In addition to data stores, researchers explored the development of privacy-preserving platforms. Π–Box [20] is one such platform that prevents apps from misusing user information by using a sandbox that spans the user's device and the cloud, along with specialised storage and communication channels. It also incorporates differential privacy techniques for continual observation. PrivAI [26] presents a decentralised platform that focuses on privacy-by-design principles for AI-based services. It achieves this by dividing AI algorithms into cloud-based model training, local personalisation, and community-based sharing of model updates, while protecting providers' intellectual property through trusted execution environments.

Data Management and Privacy Frameworks. Several frameworks have been proposed to address the broader challenges of personal data management and privacy. Databox [7, 28] is a collection of physical and cloud-hosted software components that allow users to manage, log, and audit access to their data. WeTrace [8] is a privacy-preserving application designed for data collection during health crises like the COVID-19 pandemic, utilising Bluetooth Low Energy and asymmetric cryptography to ensure privacy.

Data Donation Workflow and Tools. Recent work has introduced a novel workflow for academic researchers to partner with individuals willing to donate their digital trace data for research purposes [5]. This workflow involves local processing of the data on participants' devices to extract only the relevant information for research after obtaining informed consent. To facilitate this workflow, a software tool called Port [6] has been developed, allowing researchers to configure the local processing step and collect specific digital traces needed for their research questions.

Other research efforts have focused on specific aspects of personal data management and privacy. The SafeAnswers mechanism in OpenPDS introduces a privacy-preserving method to share metadata by calculating answers to questions instead of directly sharing the data [9, 10]. P3 [35] is a privacy-preserving photo encoding algorithm that protects photos from unauthorised access and algorithmic recognition.

3 OVERVIEW OF LOOM

LOOM is an application which allows people to submit their observations of misinformation they encounter online. To support user recognition of misinformation, the browser-based plug-in uses natural language processing techniques to detect and flag relevant linguistic markers for end users in real time, whilst preserving the end-to-end encryption that protects the privacy and security of their online browsing and communication activities. The tool allows users to submit their browsing data/findings to researchers if (i) a relevant research study exists and (ii) they so wish. An underpinning principle of the tool is that no user is obliged to submit data. Instead, users can choose to install it for their personal use only.

In order to ensure that any data collection is privacy-preserving by design, we adhere to a set of rules, detailed in Table 1, which allow us to increase the trustworthiness of the tool.

3.1 Linguistic features

Linguistic features of false information vary according to the type of false information under examination and the approach taken. Researchers have identified discursive [2], lexico-grammatical [33, 42], morphological [42], pragmatic (e.g., through implicature [4]), semantic [33, 36], syntactic [30, 33], stylistic (e.g., more upper case

P1	Participants are aware of the purpose of the experiment.
P2	Participants are aware of the parameters (web sites, words, time scale) of the data collection.
P3	The features of interest are described in an intelligible way for the participants.
P4	Participants are aware of their right to anonymity.
P5	Participants can consult their data before it is shared with the researchers.
P6	Participants can decide to exclude selected results from the data that is shared with the researchers.
P7	Participants can decide to withdraw completely from a study at any time.
P8	If participants omit to remove personally identifiable information, the researchers should remove it before long-term storage of the data.

Table 1: Key design principles

words [33]), and prosodic (e.g., faster speech rate in satire [21]) features of the language used to communicate false information. We have selected five key lexico-grammatical features to flag to users of our tool. Practical and theoretical considerations have contributed to this selection; these are: (a) they are known to be relevant to a range of false information types and (b) they are feasible to implement according to the constraints of the underpinning packages. We outline each of the five features below.

3.1.1 Salient Use of Pronouns. Pronouns have been found to be indicators of a range of false information. Relative to truthful news, satirical fake news has been found to contain more prominent use of the first person singular pronoun ('I'), whereas, second person pronouns ('you', 'your') are more prominent in propaganda fake news relative to truthful news [36]. Second person pronouns are indicative of 'direct engagement'. Conversely, others have found false information to contain fewer pronouns (relative to truthful information). For example, Newman et al. identified fewer self-references marked by personal pronouns 'I', 'me' and 'my' in 'deceptive communications' (i.e., content containing lies about personal opinions) [30] and Memon and Carley found that informed discourse about COVID-19 contained significantly more pronouns compared with misinformed antivaxx discourses [25]. Given these inconsistencies, our aim is to support the evaluation of pronouns as a marker of misinformation in the real-world language that people encounter online.

3.1.2 Comparatives and Superlatives. The presence of comparatives (for example, 'better', 'worse') and superlatives (for example, 'best', 'worst') also differs according to the type of information under examination. Several features of disinformation identified in prior work, including 'context inappropriate content', 'inaccuracy', 'situational dependence', and texts 'lacking conclusions or containing controversial conclusions', are characterised by the presence of superlatives, making them a key indicator of disinformation. Similarly, superlatives such as the word 'most' have been found to be more prominent in fake news propaganda relative to truthful news [36]. Satire also contains prominent adverb use (adverbs can often – but not always – be classed as comparatives and superlatives, e.g., faster, fastest) but hoax stories contain fewer comparatives and superlatives (ibid.). As comparatives are often used to provide concrete figures, they were found to appear more prominently in truthful news, when they appeared alongside words denoting money and numbers [32, 36].

3.1.3 Polarity. In sentiment analysis, polarity (i.e., the generally positive or negative nature of text segments, including words, sentences, and whole texts) has also been highlighted as a useful indicator of false information. Verma et al. identified polarity as one of the 20 most significant features of fake news using their WELFake model, a machine learning classification for fake news detection [42]. They considered polarity to be a psycholinguistic feature which encodes 'emotions, behaviours, persona, and mindset' (ibid., p.4). Compared with factual information, misinformation has been found to contain more negative words [33]; relative to truthful news, hoax fake news has been found to contain greater use of negation (e.g., 'nothing', 'not') [36] ; and deceptive communications are said to contain more negative emotion words (e.g., 'hate', 'worthless', 'sad') [30]. Double negatives are characteristic of texts containing 'rumour', and negatives are characteristic of both 'rumour' and 'inaccuracy'.

3.1.4 Subjectivity. Finally, subjective language is a known characteristic feature of false information. For example, Verma et al. considered strong subjectivity in their WELFake model [42], and relative to truthful news, both strong and weak subjectives were found to be more prominent in propaganda fake news and hoax fake news respectively by Rashkin et al.. Furthermore, strongly subjective language is present in 'context inappropriate content', another characteristic of disinformation. A number of lexico-grammatical items contribute to subjectivity. In fake news detection, the number and rate of adjectives and adverbs have been highlighted by Verma et al., who note that the "[s]ubjectivity of fake news articles is larger than for real news articles". Weak modal verbs are a feature of 'situational dependence' and 'rumour' which characterise disinformation, and Rashkin et al. also identified a greater prominence of modal adverbs (e.g., 'inevitably'), alongside action adverbs (e.g., 'foolishly'), and manner adverbs (e.g., 'deliberately') in satirical fake news. A high quantity of modifiers (adjectives and adverbs) is expected in disinformation texts offering a 'lack of supporting evidence' for claims. Moreover, hedging (e.g., use of the word 'claims') could be considered a marker of subjectivity and as such is more prominent in hoax fake news [36]. Hedging is also a feature of 'rumour' and 'lack of supporting evidence' characterises disinformation. All of these features have the capacity to increase and decrease the epistemic and deontic modality of a text (i.e., communicate how likely or possible something is respectively).

The tool we have developed is capable of flagging the presence of five features to users: (1) salient use of first-person singular pronouns (i.e., 'I') and second person pronouns (i.e., 'you', 'your'); (2) underuse of comparatives (e.g., 'better', 'greener'); (3) prominence of superlatives (e.g., 'most', 'greatest', 'best'); (4) polarity (e.g., negation words 'nothing', 'not'); (5) prominent use of subjectivity markers (e.g., 'brilliant', 'amazing', 'very clever'). Detection of

markers of polarity and subjectivity have been integrated into the tool using TextBlob a Python library for processing textual data, which provides sentiment analysis including polarity and subjectivity [23]. TextBlob is integrated into the tool via spaCy, a library for advanced natural language processing [16] in Python v.3.11.2. First person singular pronouns are identified using text matching and comparatives and superlatives are identified using POS-tagged word lists, which are pre-loaded in spaCy.

4 EVALUATION

4.1 Experimental design

Our experimental procedure involved a mixed-method approach, merging quantitative and qualitative methods, to assess the effectiveness and user experience of our browser extension. This procedure was designed to collect statistical data from the questionnaire and in-depth insights from the subsequent focus group discussions.

A total of 12 participants were recruited through various channels including email blasts, online forums, and social media posts. Attrition and difficulty in recruitment significantly reduced our sample size, affecting the statistical power of our analysis. However, our final group of participants was diverse in terms of age, gender, and technology proficiency.

The experiment was conducted in a controlled environment, with each participant assigned a pre-configured computer to use. These computers were equipped with the browser extension pre-installed and activated, set to detect misinformation in real-time. Each participant used the extension for a time period of 30 minutes, with our team overseeing the process to troubleshoot any technical issues and ensure the experiment ran smoothly.

After their session with the browser extension, participants were given a questionnaire. The questionnaire was structured to collect quantitative data about the user experience. It focused on usability, perceived effectiveness in detecting misinformation, and overall satisfaction. It incorporated both Likert scale questions for rating specific aspects and open-ended questions to gather broader feedback.

Following completion of the questionnaire, the participants were invited to a focus group discussion, moderated by a facilitator. The goal of this discussion was to dive deeper into the users' experiences, understand their views on the pros and cons of the extension, and collect recommendations for improvement. The focus group discussion was recorded and later transcribed for qualitative analysis.

The design of the experiment aimed to understand the user experience components of the tool. We believe that these insights will provide valuable information on the perceptions and experiences of end-users, which are critical for the future development of the extension.

4.2 Survey results

In this section we discuss the results of the survey. The complete set of questions can be found in Appendix A.1.

4.2.1 *Participants' experience of online misinformation.*

I encounter misinformation online. Participants expressed awareness of encountering misinformation online, primarily citing social media as the source. Even those not active on social media recognised potential biases and subjectivity on other websites they frequent (news, music, blogs, etc.). It was noted that personal choices and preferences, such as news sources and friend groups, can shape exposure to misinformation. Some individuals indicated trust in the accounts they follow on platforms like Facebook and Instagram but acknowledged scepticism toward the information they consume.

I am confident I can detect misinformation online. The majority of participants expressed confidence in their ability to detect misinformation online, primarily citing factors like hyperbolic language and the reputation of certain websites. However, some individuals acknowledged difficulties, particularly with content that presents statistics or appears to come from authoritative sources. In addition, participants noted the increasing sophistication of misinformation, especially on social media, where short, attention-grabbing posts with questionable grammar are becoming more prevalent. Visual content such as images and videos was identified as particularly challenging to assess, as they can be easily manipulated or taken out of context.

I often check if information is true or false before sharing it online. The participants varied in the frequency with which they verified the information before sharing it online. Half reported frequently checking information, a quarter occasionally checked, and another quarter rarely or never checked. Reasons for checking information included concerns about fake posts on social media and a desire to verify facts before sharing. Those who rarely or never checked cited reasons such as not posting on social media or only sharing their own content. One participant mentioned sharing unverified information to spark conversations.

I have knowingly shared misinformation. The majority of participants reported never having knowingly shared misinformation. However, a significant proportion admitted to having done so at least occasionally, and most of these individuals indicated that it was rare. Reasons for sharing misinformation included humour, pushing an agenda, and sharing unverified information to convey the general idea of a story.

I have unintentionally shared misinformation in the past. Participants were divided on whether they had unintentionally shared misinformation in the past. The majority admitted to having done so rarely, often due to not fact-checking information before sharing it. Some indicated that they may have unintentionally shared misinformation, but were unsure. The remaining participants reported never having unintentionally shared misinformation, either because they rarely share information or because they rely on what they consider to be trusted sources.

I am concerned about the negative impacts of misinformation. Participants overwhelmingly expressed concern about the negative impacts of misinformation, with 80% strongly agreeing and 20% agreeing. They cited various reasons for their concern, including:

- Intentional manipulation and division: Some participants believed that misinformation is deliberately spread to create conflict and for financial profit.

- Spread of harmful narratives: Misinformation was seen as a tool to promote harmful ideologies such as racism, sexism, and transphobia.
- Real-world consequences: Participants pointed to the impact of misinformation on political discourse and public health, citing examples such as the COVID-19 pandemic and violence against marginalised groups.
- Herd mentality: The amplifying effect of social media was highlighted as a concern, as misinformation easily spreads and warps views.
- In general, the participants showed a strong awareness of the potential dangers of misinformation and its far-reaching consequences for individuals and society.

4.2.2 *Participants' experience of the tool.*

Did the tool do what you expected it to do? The feedback regarding whether the tool met user expectations was mixed. Four users confirmed that it performed generally as described, providing detailed information. However, six users expressed uncertainty, indicating they didn't know what to expect or desired additional features like an overall trustworthiness score or clearer explanations of the scores. Two users explicitly stated that the tool did not meet their expectations, citing issues with its discreteness and the lack of prominent misinformation alerts. One user also mentioned initially misunderstanding the tool's functionality but gaining clarity after reading the data analysis.

In general, the responses suggest that while the tool provides useful information, there is room for improvement in terms of user experience, clarity of results, and prominence of misinformation alerts.

The tool was easy to use. The majority of participants (11 out of 12) found the tool easy to use, with 4 strongly agreeing and 7 agreeing. They appreciated quick analysis, easy clicks, and the expansion of selections to show subjectivity. However, some suggestions are as follows:

- Clearly defined variables: One participant desired clearer definitions of the variables used in the analysis.
- Explanation of scores: Multiple participants expressed confusion about the meaning of the percentage and decimal scores, suggesting the use of percentages or colour gradients to convey the extent of (potential) misinformation more intuitively.
- Clarification of terminology: Participants wanted to know what the scores were out of and what the terms meant to better understand the results.
- Simplified presentation: One participant found the tool complex due to the scales used to present results, suggesting the use of nominal values or colour coding for easier interpretation.

Overall, feedback indicates that the tool's usability is generally positive, but clarifying terminology, explaining scores, and potentially simplifying the presentation of results would improve user experience and understanding.

Would you recommend this tool to others? The majority of participants (7 out of 12) expressed a willingness to recommend the tool to others, but often with caveats or conditions. These conditions included improvements in clarity, accuracy, and user-friendliness. Some specifically mentioned recommending it to family and friends who are prone to misinformation, but only after the tool is refined. The remaining participants were unsure about recommending the tool, citing concerns about its current form, accuracy, and lack of layman's terms. They emphasised the need for further development, particularly in terms of user experience and providing more specific information about misinformation.

Overall, the responses indicate a positive sentiment towards the tool's potential, but they also highlight areas for improvement before it can be confidently recommended to a wider audience.

The information provided by the tool was easy to understand. The feedback on the intelligibility of the information provided by the tool is divided. While a few participants found it easy to understand, the majority found it somewhat difficult, primarily due to the lack of clarity regarding the scores, terminology, and reasoning behind the identified misinformation markers.

Positive feedback:

- One participant strongly agreed that the information was easy to understand.
- Five participants agreed, but some mentioned initial difficulties and suggested improvements in the presentation of scores and colour schemes.

Negative feedback:

- Six participants disagreed, expressing confusion about the meaning of scores, the absence of definitions and explanations, and the lack of information on why certain markers were selected.
- Participants specifically requested more detailed explanations of the scores, the reasoning behind the identified markers, and clearer definitions of terms.
- A participant with dyslexia suggested using a more colourful display to improve comprehension.

In general, the feedback from the participants provided constructive and practical points for improvement relating to the clarity and intelligibility of the information provided by the tool. This involves improving the clarity of definitions and explanations, using more intuitive visualisations, and offering additional context for the identified misinformation markers.

The tool is useful for detecting language often found in misinformation. The majority of participants (11 out of 12) found the tool useful to detect the language often found in false information, with 3 strongly agreeing and 8 agreeing. They appreciated the identification of markers like polarity, first-person pronouns, and superlatives, which helped them become more aware of the language used in false information in their day-to day online browsing.

However, there were some concerns raised:

- False positives: Some participants noted that the tool occasionally flagged content that was not actually misinformation.
- Need for context: One participant suggested that the tool should consider the broader context of the text when analysing language.

- Limited effectiveness: One participant felt that the tool was too subtle and failed to pick up on some grammatical issues, particularly in strongly opinionated pieces.

Overall, feedback indicates that the tool is considered valuable for detecting language patterns associated with misinformation, but there is room for improvement in terms of reducing false positives, considering context, and enhancing its effectiveness in identifying misinformation in various types of content. Adding information to explain that flagged language only indicates potential misinformation (and does not guarantee its presence) would help users understand the need for further investigation of context and content.

I trust this tool to detect language often found in misinformation. The majority of participants (10 out of 11) expressed trust in the tool's ability to detect language often found in misinformation, with 1 strongly agreeing and 9 agreeing. They acknowledged its effectiveness in identifying certain language patterns, but also raised concerns about its limitations in handling short-form content and the intentional use of lower language levels by some media outlets.

One participant disagreed, stating that while the tool has potential, its current accuracy and functionality need improvement.

Overall, the feedback suggests a generally positive sentiment towards the tool's ability to detect misleading language, but it also highlights the need for further refinement to address its limitations and improve accuracy in diverse content formats.

The tool picked up on things I may not have noticed. The majority of participants (10 out of 11) indicated that the tool picked up on things they might not have noticed, with 2 stating "a great deal," 4 stating "a moderate amount" and 4 stating "occasionally." This suggests that the tool was effective in highlighting subtle language patterns and potential biases that might otherwise have been overlooked.

Specific examples of elements identified by the tool that participants might not have noticed themselves include:

- Comparatives and superlatives
- First-person pronouns in formal articles
- Polarity
- Grouping of certain words and phrases

However, one participant indicated that the tool rarely picked up on things they would not have noticed. This could arise in users with individual differences in reading habits, but this particular participant did report prior knowledge and interest in misinformation tactics. Overall, the feedback suggests that the tool has real potential to enhance user awareness of subtle language cues associated with false information and to help them critically evaluate the content they encounter.

I would use this tool in future. The majority of participants (11 out of 11) indicated a willingness to use the tool in the future, but with varying frequencies.

- Yes, frequently (4): These participants found the tool useful and expressed a desire to incorporate it into their regular browsing habits. However, some noted that the tool could be improved by simplifying the language and terminology used in the analysis, making it more accessible to a wider audience.
- Yes, occasionally (2): These participants saw the tool as having potential value, but anticipated that its usefulness might diminish over time as they became more adept at recognising misinformation patterns themselves.
- Yes, rarely (5): These participants expressed interest in using the tool, but primarily for situations where they are already suspicious of a website's content. Some also suggested that the tool would be more valuable with further development and refinement.

Overall, the feedback suggests general interest in using the tool, but with varying levels of enthusiasm and frequency of use. While some participants see it as a valuable tool for regular use, others envision it as a resource to be consulted in specific situations or after further development.

4.3 Interview analysis

Participants across four focus groups shared their perspectives on the misinformation detection tool, highlighting both potential benefits and areas for improvement.

4.3.1 Perceived impact of misinformation. Participants expressed concern about the negative impacts of misinformation, ranging from individual harm (e.g. self-harm, poor health choices) to societal consequences (e.g., political polarisation, violence). They emphasised the need for tools to help individuals navigate the information landscape and make informed decisions about the truthfulness of the content they encounter.

Accuracy and trustworthiness. While participants appreciated the tool's ability to identify potentially misleading language, many expressed uncertainty about its accuracy and underlying methodology. They desired clearer explanations of how the tool works, what the scores represent, and how to interpret the results. Participants also highlighted the risk of false positives, which could erode trust in the tool if the potential for this to occur is not adequately communicated.

Usability and accessibility. Opinions on the tool's ease of use were mixed. Some found it intuitive and easy to navigate, while others found the terminology confusing and the results difficult to interpret. Several participants (including one with with dyslexia) suggested using colour to improve readability. Additionally, concerns were raised about the tool's accessibility for visually impaired users, though none of our participants identified as such.

Desired features and improvements. Participants suggested several enhancements to improve the tool's functionality and user experience:

- Clearer explanations: Provide definitions of terms, explain how scores are calculated, and offer context for why certain language is flagged.
- Simplified presentation: Use plain language, consistent scales, and visual aids (e.g., colour gradients) to make results easier to understand.

- Enhanced features: Offer an overall score for articles, summaries of key findings, and the ability to compare articles across sources.
- Accessibility: Ensure compatibility with screen readers and consider alternative formats for visually impaired users.
- Transparency: Include a disclaimer explaining the tool's limitations and its focus on language analysis rather than definitive truth determination.

Target audience and susceptibility. Participants identified various groups as being particularly susceptible to misinformation, including older adults, those unfamiliar with the internet, and individuals from marginalised or isolated communities. They emphasised the need for the tool to be accessible and relevant to diverse audiences.

In general, the participants recognised the potential value of the misinformation detection tool, but emphasised the need for further development to enhance its accuracy, transparency, and user-friendliness. By addressing these concerns and incorporating user feedback, the tool can evolve into a valuable resource for individuals navigating the complex information landscape.

5 DISCUSSION

5.1 Summary of key findings

While participants expressed concern about the negative impacts of misinformation and a desire for tools to combat it, their experiences with our browser extension were varied.

The survey results highlighted a general awareness of participants' encounters with online misinformation, primarily through social media. Although most of the participants were confident in their ability to detect misinformation, many acknowledged challenges, especially with visual content and increasingly sophisticated tactics. The tool was generally considered easy to use and potentially useful for detecting misleading language. A significant proportion of users were willing to recommend the tool, but many emphasised the need for improvements in clarity, accuracy, and user-friendliness.

Focus group discussions provided further information on user perceptions. Participants further emphasised the need for accuracy, transparency, and improved usability. They wanted clearer explanations of how the tool works, what the scores represent, and how to interpret the results. They also highlighted the risk of false positives and suggested several features that could enhance the tool's functionality and user experience.

Overall, the evaluation indicates a demand for tools to combat misinformation, but it also highlights the challenges in creating a tool that is both effective and user-friendly for a range of users with different needs and abilities. The mixed feedback on the current tool suggests a need for further development to address the issues of accuracy, transparency, and usability. Adding the functionality for users to personalise their experience of using the tool, for example by proving options to show/hide lay explanations or choose between numerical or colour-coded interpretations of the flagged results would provide users the flexibility they desire.

5.2 Interpretation of results

The mixed results from the survey and focus groups suggest that the current tool, while promising, is not yet fully satisfying user needs. The high level of concern about misinformation, coupled with the overall willingness of participants to use and recommend the tool, indicates a strong demand for such a solution.

The desire for clearer explanations and more intuitive presentation of results suggests that the tool's current interface may be overwhelming or confusing to some users. Furthermore, concerns about false positives underscore the importance of accuracy and transparency in building user trust. The mixed feedback on intelligibility and usefulness indicates that the tool's value is not yet fully realised for all users.

Overall, these results suggest that the tool has potential but needs further refinement to fully meet user needs and expectations.

5.3 Implications for design

The evaluation findings have several implications for the design and development of misinformation detection tools, including the further development of LOOM.

- **Prioritise transparency and clarity:** Ensure clear explanations of the tool's methodology, how scores are calculated, and what the results mean. Consider using plain language, visual aids, and consistent scales to make the information more accessible to a wider audience.
- **Address accuracy concerns:** Investigate the causes of false positives and explore ways to improve the tool's accuracy. This could involve refining the algorithms, incorporating additional contextual information, or providing users with more control over the detection process.
- **Enhance user experience:** Simplify the interface, provide more intuitive visualisations, and consider incorporating user feedback into future iterations. Address accessibility concerns by ensuring compatibility with screen readers and exploring alternative formats for visually impaired users.
- **Expand functionality:** Consider adding features such as an overall score for articles, summaries of key findings, and comparisons across sources. This could enhance the tool's usefulness and provide users with a more comprehensive understanding of the information they encounter.
- **Tailor to diverse audiences:** Recognise that different user groups may have varying levels of understanding and familiarity with misinformation. Tailor the tool's interface, language, and features to accommodate diverse audiences, including older adults, those unfamiliar with the internet, and individuals from marginalised communities.

5.4 Limitations and future directions

This evaluation was limited by the small sample size and the specific demographics of the participants. Although efforts were made to recruit a diverse group, the final sample may not be fully representative of the general population. Additionally, LOOM being used through a browser extension, its usefulness is constrained by the browsing habits of the users, and a range of misinformation cannot be captured (e.g. messaging apps).

Following refinement of LOOM, future research should aim to explore the effectiveness of the tool with a larger and more diverse sample, including individuals from different age groups, educational backgrounds, and cultural contexts. It would also be valuable to compare the tool's performance with other misinformation detection approaches and to assess its long-term impact on user behaviour and beliefs. The authors will incorporate the suggestions of the participants, such as clearer explanations, simplified presentation, and improved features. The tool could also be adapted to address the specific challenges of detecting misinformation in different content formats and language patterns.

5.5 Responsible research and innovation

This project anticipated responsible research and innovation (RRI) challenges and integrated them into the research design. We reflected on challenges surrounding 'digital information' and 'critical thinking skills' to better support people who would benefit most from using the tool. As such, we engaged in dialogue with end users to understand the following perspectives and experiences: the concerns people have about misinformation; the problems they have encountered with misinformation in relation to digital information literacy; and who is most often harmed, influenced, and targeted by misinformation. Upon completion of user testing, the tool has been made open source and freely available. Future work will include the production of a typology of misinformation markers (including markers not yet integrated into the flagging tool) as a tool for digital literacy training. Using 'logic-based' examples to 'inoculate' people against misinformation as part of our training materials [43], this research ultimately aimed to raise awareness of the different forms of misinformation and its known harms. In this way, we hoped that participants might learn to critically evaluate information for potential misinformation markers in both online and offline contexts. The principles used to develop the tool, outlined in Table 1, aim to be compatible with modern regulations in Internet privacy, such as the General Data Protection Regulation of the European Union and its counterpart in the United Kingdom, allowing users to have complete control of their data. This important step in the design of our application serves as a way to increase trust in the system by allowing users to remain in control of the way they interact with it.

6 CONCLUSION

Detection of five language features is adequate for assessing the feasibility of flagging misinformation markers to individual users in a privacy-preserving way, including user trust in such a tool. However, we recognise that additional features will be required to establish greater overall reliability in misinformation detection. Verma et al. model draws on 20 features from four categories including writing patterns, psycholinguistic features, readability, and quantity to achieve a 96.73% accuracy in categorising real and fake news [42]. It is also clear that the absence or lower-than-expected quantity of linguistic features may be used in conjunction with prominent features to flag potential misinformation to users. Future development of the tool will need to account for linguistic variation across and within discourses (i.e., scientific texts vs. news discourse, tabloid vs. broadsheet news). For example, passive structures, longer sentence lengths and more obscure lexis (less frequent terms) are more prevalent in scientific discourse, whilst shorter sentence length, and common, high-frequency words are more prevalent in tabloid news. These differences will influence the relative word counts and thus thresholds for flagging potential misinformation. Moreover, the language characterising the range of false information types (misinformation, disinformation, rumour, hoax, satire, etc.) will also have a bearing on flagging thresholds.

Ultimately, the tool will allow researchers to analyse the consumption of false information without infringing on the privacy of the participants. At the same time, it will offer users privacy enhanced data processing for the detection of false information markers, and the opportunity to submit findings as participants in follow-on privacy-preserving research. In the future, user-submitted data will support the regular testing and updating of the tool to maintain accuracy as trends in misinformation emerge. This is particularly important since the subject matter is known to have a bearing on the accuracy of established detection models [30]. By providing a way for automated systems to analyse data without undoing progress in end-to-end encryption, the tool contributes to safeguarding personal freedoms, which is essential for protecting the privacy and security of people's online communications, and one of the grand challenges outlined by the Trustworthy Autonomous Systems hub.

ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/V00784X/1] UKRI Trustworthy Autonomous Systems Hub and Responsible AI UK [grant number EP/Y009800/1]. All data created during this research are openly available from the University of Nottingham data repository at https://doi.org/10.17639/nott.7451.

REFERENCES

[1] Diaa Salama Abdelminaam, Fatma Helmy Ismail, Mohamed Taha, Ahmed Taha, Essam H Houssein, and Ayman Nabil. 2021. CoAID-DEEP: an optimized intelligent framework for automated detecting COVID-19 misleading information on Twitter. *Ieee Access* 9 (2021), 27840–27867.

[2] Aseel Addawood, Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Linguistic cues to deception: Identifying political trolls on social media. In *Proceedings of the international AAAI conference on web and social media*, Vol. 13. 15–25.

[3] Eiríkur Bergmann. 2020. Populism and the politics of misinformation. *Safundi* 21, 3 (2020), 251–265.

[4] Marcella Bertuccelli Papi. 2018. Satire as a genre. *Pragmatics & Cognition* 25, 3 (2018), 459–482.

[5] Laura Boeschoten, Jef Ausloos, Judith Moeller, Theo Araujo, and Daniel L Oberski. 2020. Digital trace data collection through data donation. *arXiv preprint arXiv:2011.09851* (2020).

[6] Laura Boeschoten, Niek C de Schipper, Adriënne M Mendrik, Emiel van der Veen, Bella Struminskaya, Heleen Janssen, and Theo Araujo. 2023. Port: A software tool for digital data donation. *Journal of Open Source Software* 8, 90 (2023), 5596.

[7] Andy Crabtree, Tom Lodge, James Colley, Chris Greenhalgh, Richard Mortier, and Hamed Haddadi. 2016. Enabling the new economic actor: data protection, the digital economy, and the Databox. *Personal and Ubiquitous Computing* 20 (2016), 947–957.

[8] Alessandro De Carli, M Franco, Andreas Gassmann, Christian Killer, Bruno Rodrigues, E Scheid, David Schönbächler, and Burkhard Stiller. 2020. WeTrace–a privacy-preserving mobile COVID-19 tracing approach and application. *arXiv preprint arXiv:2004.08812* (2020).

[9] Yves-Alexandre De Montjoye, Erez Shmueli, Samuel S Wang, and Alex Sandy Pentland. 2014. openpds: Protecting the privacy of metadata through safeanswers. *PloS one* 9, 7 (2014), e98790.

[10] Yves-Alexandre de Montjoye, Samuel S Wang, Alex Pentland, Dinh Tien Tuan Anh, Anwitaman Datta, et al. 2012. On the Trusted Use of Large-Scale Personal Data. *IEEE Data Eng. Bull.* 35, 4 (2012), 5–8.

[11] Don Fallis. 2015. What is disinformation? *Library trends* 63, 3 (2015), 401–426.
[12] FullFact. 2023. *Who We Are.* FullFact. https://fullfact.org/about/
[13] Renee Garett and Sean D Young. 2021. Online misinformation and vaccine hesitancy. *Translational behavioral medicine* 11, 12 (2021), 2194–2199.
[14] Thomas Hardjono, Patrick Deegan, and John Henry Clippinger. 2014. Social use cases for the ID3 open mustard seed platform. *IEEE Technology and Society Magazine* 33, 3 (2014), 48–54.
[15] Peter Hernon. 1995. Disinformation and misinformation through the internet: Findings of an exploratory study. *Government information quarterly* 12, 2 (1995), 133–139.
[16] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. GitHub.
[17] Philip N Howard, Lisa-Maria Neudert, Nayana Prakash, and Steven Vosloo. 2021. Digital misinformation/disinformation and children. *UNICEF. Retrieved on February* 20 (2021), 2021.
[18] Brian Hughes, Cynthia Miller-Idriss, Rachael Piltch-Loeb, Beth Goldberg, Kesa White, Meili Criezis, and Elena Savoia. 2021. Development of a codebook of online anti-vaccination rhetoric to manage COVID-19 vaccine misinformation. *International journal of environmental research and public health* 18, 14 (2021), 7556.
[19] Ali Khan, Kathryn Brohman, and Shamel Addas. 2022. The anatomy of 'fake news': Studying false messages as digital objects. *Journal of Information Technology* 37, 2 (2022), 122–143.
[20] Sangmin Lee, Edmund L Wong, Deepak Goel, Mike Dahlin, and Vitaly Shmatikov. 2013. {πBox}: A Platform for {Privacy-Preserving} Apps. In *10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*. 501–514.
[21] Saskia Leymann, Tomas O Lentz, and Christian Burgers. 2022. Prosodic markers of satirical imitation. *Humor* 35, 4 (2022), 509–529.
[22] Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour* 5, 3 (2021), 337–348.
[23] Steven Loria et al. 2018. textblob Documentation. *Release 0.15* 2, 8 (2018).
[24] Rakoen Maertens, Frederik Anseel, and Sander van der Linden. 2020. Combatting climate change misinformation: Evidence for longevity of inoculation and consensus messaging effects. *Journal of Environmental Psychology* 70 (2020), 101455.
[25] Shahan Ali Memon and Kathleen M Carley. 2020. Characterizing covid-19 misinformation communities using a novel twitter dataset. *arXiv preprint arXiv:2008.00791* (2020).
[26] Christian Meurisch, Bekir Bayrak, and Max Mühlhäuser. 2020. Privacy-preserving AI services through data decentralization. In *Proceedings of The Web Conference 2020*. 190–200.
[27] Christian Meurisch and Max Mühlhäuser. 2021. Data protection in AI services: A survey. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–38.
[28] Richard Mortier, Jianxin Zhao, Jon Crowcroft, Liang Wang, Qi Li, Hamed Haddadi, Yousef Amar, Andy Crabtree, James Colley, Tom Lodge, et al. 2016. Personal data management with the databox: What's inside the box?. In *Proceedings of the 2016 ACM Workshop on Cloud-Assisted Networking*. 49–54.
[29] Min Mun, Shuai Hao, Nilesh Mishra, Katie Shilton, Jeff Burke, Deborah Estrin, Mark Hansen, and Ramesh Govindan. 2010. Personal data vaults: a locus of control for personal data streams. In *Proceedings of the 6th International COnference*. 1–12.
[30] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin* 29, 5 (2003), 665–675.
[31] Alison Noble, Guy Cohen, Jon Crowcroft, Adrià Gascón, Marion Oswald, and Angela Sasse. 2019. *Protecting Privacy in Pracice: the current use, development and limits of Privacy Enhancing Technologies in data analysis.* Technical Report. The Royal Society.
[32] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557* (2011).
[33] Wei Peng, Sue Lim, and Jingbo Meng. 2023. Persuasive strategies in online health misinformation: a systematic review. *Information, Communication & Society* 26, 11 (2023), 2131–2148.
[34] Francesco Pierri, Brea L Perry, Matthew R DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. 2022. Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Scientific reports* 12, 1 (2022), 5966.
[35] Moo-Ryong Ra, Ramesh Govindan, and Antonio Ortega. 2013. P3: Toward {Privacy-Preserving} Photo Sharing. In *10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*. 515–528.
[36] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2931–2937.
[37] Andrei Vlad Sambra, Essam Mansour, Sandro Hawke, Maged Zereba, Nicola Greco, Abdurrahman Ghanem, Dmitri Zagidulin, Ashraf Aboulnaga, and Tim Berners-Lee. 2016. Solid: a platform for decentralized social applications based on linked data. *MIT CSAIL & Qatar Computing Research Institute, Tech. Rep.* (2016).
[38] Karishma Sharma, Emilio Ferrara, and Yan Liu. 2022. Characterizing online engagement with disinformation and conspiracies in the 2020 US presidential election. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. AAAI, 908–919.
[39] The Royal Society. 2022. The online information environment. *The Royal Society* (2022). https://royalsociety.org/topics-policy/projects/online-information-environment
[40] 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.
[41] Sander Van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. 2017. Inoculating the public against misinformation about climate change. *Global challenges* 1, 2 (2017), 1600008.
[42] Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. 2021. WELFake: word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems* 8, 4 (2021), 881–893.
[43] Emily K Vraga, Sojung Claire Kim, John Cook, and Leticia Bode. 2020. Testing the effectiveness of correction placement and type on Instagram. *The International Journal of Press/Politics* 25, 4 (2020), 632–652.
[44] Claire Wardle. 2018. The need for smarter definitions and practical, timely empirical research on information disorder. *Digital journalism* 6, 8 (2018), 951–963.

A APPENDIX

A.1 Survey questions

(1) I encounter misinformation online
(2) I am confident I can detect misinformation online.
(3) I often check if information is true or false before sharing it online.
(4) I have knowingly shared misinformation
(5) I have unintentionally shared misinformation in the past
(6) I am concerned about the negative impacts of misinformation.
(7) Did the tool do what you expected it to do?
(8) The tool was easy to use.
(9) Would you recommend this tool to others?
(10) The information provided by the tool was easy to understand.
(11) The tool is useful for detecting language often found in misinformation.
(12) I trust this tool to detect language often found in misinformation.
(13) The tool picked up on things I may not have noticed.
(14) I would use this tool in future.