CritiSense: Critical Digital Literacy and Resilience Against Misinformation

Firoj Alam[1], **Fatema Ahmad**[1], **Ali Ezzat Shahroor**[1], **Mohamed Bayan Kmainasi**[1]
Elisa Sartori[2], **Giovanni Da San Martino**[2], **Abul Hasnat**[3], **Raian Ali**[4]
[1]Qatar Computing Research Institute, Qatar, [2]University of Padova, Italy
[3]APAVI.AI, France, [4]Hamad Bin Khalifa University, Qatar
fialam@hbku.edu.qa

https://critisense-web.digitqr.net/

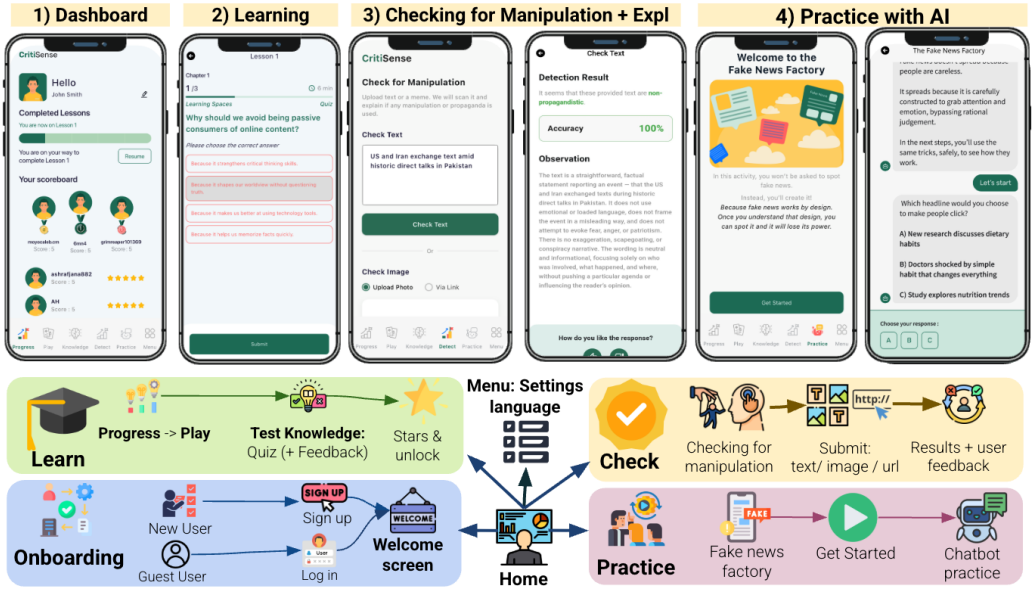

Figure 1: Overview of CRITISENSE app and its functionalities.

Abstract

Misinformation on social media undermines informed decision-making and public trust. Prebunking offers a proactive complement by helping users recognize manipulation tactics *before* they encounter them in the wild. We present CRITISENSE, a mobile media-literacy app that builds these skills through short, interactive challenges with instant feedback. It is the *first* multilingual (supporting nine languages) and modular platform, designed for rapid updates across topics and domains. We report a usability study with 93 users: 83.9% expressed overall satisfaction and 90.1% rated the app as easy to use. Qualitative feedback indicates that CRITISENSE helps improve digital literacy skills. Overall, it provides a multilingual prebunking platform and a testbed for measuring the impact of microlearning on misinformation resilience. Over 6 months, we have reached 500+ active users. It is freely available on the Apple App Store and Google Play Store.

1 Introduction

The rapid diffusion of online misinformation threatens public health, democratic governance, and social cohesion. Since false claims often spread faster than corrections, post-hoc fact-checking can arrive too late and may suffer from fatigue and limited behavioral impact. Reflecting the scale of this challenge, the World Economic Forum's Global Risks Report 2026 ranks mis- and disinformation among the world's top near-term risks (second on the two-year outlook) (World Economic Forum, 2026). To address these challenges, social media platforms rely on automatic detection, fact-checking pipelines, and warning interfaces to curb misleading content. While these measures provide essential first-line protection, they are largely reactive and claim-specific, can degrade under temporal and cross-lingual/domain shift, and do not by themselves build durable user competence (Berger et al., 2025; Stepanova and Ross, 2023).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 481–492

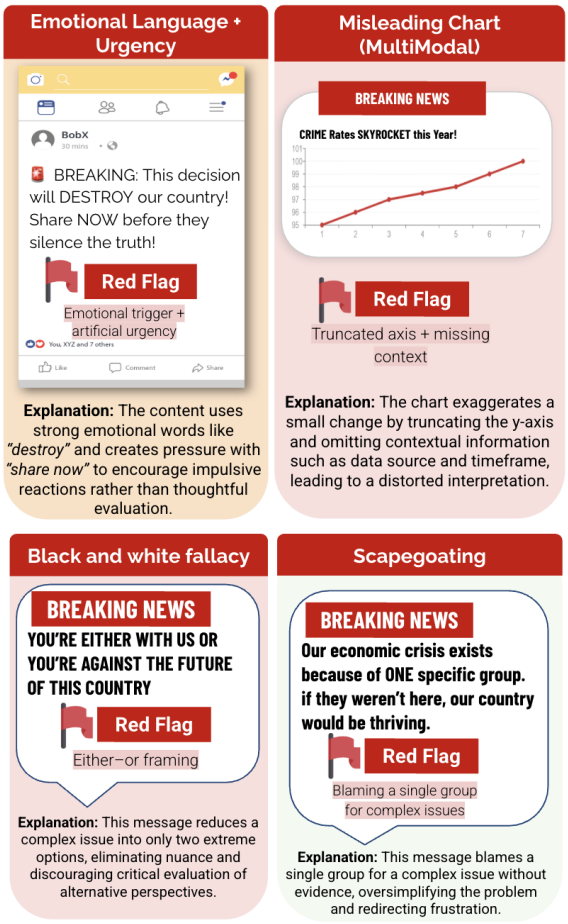

Figure 2: Examples of fictional social media posts, demonstrating different manipulation techniques.

In practice, detection-centric mitigation is most effective when complemented with user-facing training, for three technical reasons: ***(i)*** **Temporal drift:** misinformation is non-stationary, narratives and multimodal presentation styles evolve quickly, producing temporal distribution shift that can challenge deployed classifiers; temporally out-of-domain evaluation shows that performance can drop even for strong models (Stepanova and Ross, 2023). ***(ii)*** **Data and coverage:** robust detection benefits from large, high-quality annotations, yet such resources are uneven across languages and regions, and cross-lingual transfer remains less reliable for lower-resource settings (Ozcelik et al., 2023). ***(iii)*** **Interface design:** detection outputs must be translated into interventions (e.g., labels, banners, downranking), where careful UI choices are crucial to maximize critical evaluation and avoid unintended effects, such as "implied truth" for untagged items (Pennycook et al., 2020).

These limitations motivate complementary, user-centered approaches that improve *critical evaluation* rather than relying on perfect detection coverage. Digital media literacy interventions can improve users' ability to distinguish mainstream from false content, with effects that persist beyond immediate exposure (Guess et al., 2020). *Prebunking* (psychological inoculation) extends media literacy by targeting manipulation *techniques* (e.g., emotional (loaded) language, black white fallacy (Da San Martino et al., 2019), as shown in Figure 2). Scalable inoculation interventions have been shown to increase resistance to misinformation tactics on social media (Roozenbeek et al., 2022). Meta-analyses further support improvements in credibility judgment across studies and settings (Lu et al., 2023; Traberg et al., 2022).

Yet most media literacy and inoculation evaluations remain short, one-off web experiments, with limited attention to *(i)* sustained engagement in everyday contexts, *(ii)* multilinguality, and *(iii)* durability and behavioral outcomes beyond immediate post-intervention assessments (Lu et al., 2023; Traberg et al., 2022). CRITISENSE is motivated by these gaps. It treats automatic detection as a safety net, but prioritizes building *transferable* user competence through an iterative, mobile-first learning experience, along with an assessment method designed to measure not only immediate gains but also real-world impact.

Our contributions are as follows:

- We introduce CRITISENSE, a **mobile-first micro-lessons** app that trains users to recognize misinformation and manipulation tactics in realistic, everyday scenarios.
- To our knowledge, this is the ***first*** **multilingual** app of its kind, launched in Arabic and English and extended to Bangla, French, Hindi, Italian, Filipino, Nepali, and Urdu, offering full lesson content, quizzes, and feedback.
- We describe the complete user workflow, from learning and quizzes to simulated practice.
- We manually develop the learning content, covering core digital literacy concepts, propaganda techniques, and fake-news patterns.
- We integrate factuality, propaganda, and hate detection functionalities to provide automatic signals for textual and visual manipulation, supporting in-app feedback and learning.
- We report a formative usability evaluation measuring *(i)* ease of use, *(ii)* visual design, and *(iii)* perceived content impact.

Does CritiSense provide a usable and satisfying experience? A formative usability evaluation with 93 first-time users demonstrates strong overall us-

ability (mean construct scores of 3.99–4.20 on a 5-point Likert scale), with 90.1% of users agreeing that the app is easy to use and 83.9% reporting satisfaction with their experience.

2 CritiSense App

2.1 App Design

CritiSense is a mobile-first media-literacy app designed to help users build and strengthen skills for evaluating online information. The design of the app is grounded in *active learning*. Users read short lessons covering a wide variety of topics related to critical digital literacy, fake news and propaganda, answer short quizzes, receive immediate feedback, and practice applying the same reasoning to new examples. The app is also informed by *cognitive inoculation*, training recognition of common manipulation techniques (e.g., loaded language, name calling) to increase users' resilience before they encounter them in the wild (Roozenbeek et al., 2020).

In Figure 1, we provide an overview of CritiSense, organized around four core app components, and illustrate the end-to-end user journey from onboarding to practice. The app features a streamlined onboarding process: new users register and receive introductory guidance, while returning users authenticate directly. All pathways then converge on a dashboard/progress view, which branches into three functional modules: learning and knowledge assessment, interactive content verification, and simulated practice.

Dashboard/Progress. This module displays the user profile, including the number of completed lessons, a score summary, and a leaderboard of top users with their ratings.

Learning/Play. This module consists of lessons divided into chapters, organized by topic or manipulation technique. For example, the *Critical Digital Literacy* lesson is divided into: *(i)* the digital space and us, *(ii)* developing critical digital literacy, and *(iii)* critical digital literacy skills. Each chapter presents key definitions and concepts, followed by quizzes for reinforcement. All contents of the app are developed manually.

Learning/Test Knowledge. This module provides a flexible, non-linear self-assessment experience, allowing users to attempt targeted question banks and localized quizzes to gauge their understanding and earn rewards. The quizzes use multiple-choice questions to assess both conceptual knowledge and

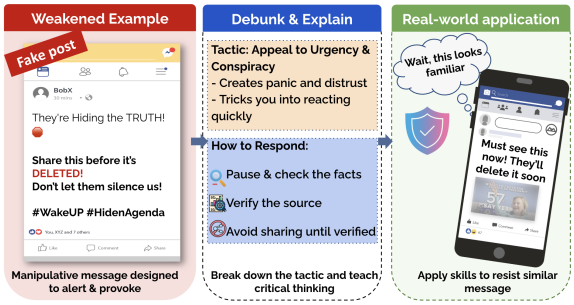

Figure 3: A pictorial example illustrating inoculation theory: expose a weakened misleading message, explain the tactic, then test transfer to a novel post.

applied reasoning beyond individual lessons.

Detect. This module serves as a practical verification tool where users can submit multimodal content through text input, image uploads, or direct URLs for real-time analysis. In addition to providing automatic predictions, it generates explanations that highlight linguistic cues, emotional framing, persuasive patterns, and other indicators of misleading information, helping users better understand why a piece of content may be unreliable.

Practice. This module operationalizes inoculation theory (Banas and Rains, 2010) through an interactive simulation titled *"The Fake News Factory"* (Figure 3). In this environment, users engage with a conversational system to practice identifying and understanding disinformation tactics through guided interaction with simplified, fictional examples and immediate feedback. By exposing users to weakened forms of misleading arguments paired with explanations and refutations, the module aims to build cognitive resistance to future persuasion attempts and strengthen users' ability to recognize manipulation in the wild. To mitigate dual-use risks, the Practice module is implemented as a fully scripted, branching dialogue rather than open-ended generation. All prompts, options, and feedback are pre-authored and curated for educational purposes. Users select from predefined choices illustrating manipulation tactics (e.g., emotional language, vague evidence, manufactured urgency), each followed by an explanation of *why* the tactic is persuasive. The system does not generate novel disinformation or target real entities, ensuring that users learn to recognize manipulation patterns without producing deployable persuasive content.

Design goals. CritiSense is guided by several design goals. **First**, it adopts a *technique-first* approach, training users to recognize recurring propaganda and misinformation strategies rather than fo-

cusing only on individual claims. This helps users apply what they learn across different topics and contexts. *Second*, it emphasizes practice in realistic formats: examples mirror the kinds of posts and media users encounter in everyday online settings. *Third*, it provides immediate feedback with explanations, going beyond binary correctness to highlight the reasoning behind each decision. *Fourth*, it is designed to be accessible and scalable through short microlearning lessons that require minimal setup and support learning beyond classroom contexts. *Finally*, it is multilingual, expanding from its initial launch to nine supported languages: Arabic, English, Bangla, French, Hindi, Italian, Filipino, Nepali, and Urdu.

2.2 Functionalities

CRITISENSE provides: *(i)* interactive microlessons with short quizzes, *(ii)* technique-first prebunking modules that teach common manipulation tactics through examples, *(iii)* immediate, explanation-rich feedback, and *(iv)* prompts for repeatable verification habits (e.g., check sources, separate claims from evidence, and cross-check across outlets). The app also offers progress tracking and topic-level assessments for continued reinforcement. Finally, an AI-assisted analysis tool helps users examine external text and images for potential manipulative cues. In Table 1, we summarize the current functionality coverage across languages.

Module	EN	AR	BN	FR	HI	IT	TL	NE	UR
Lessons	✓	✓	✓	✓	✓	✓	✓	✓	✓
Quizzes	✓	✓	✓	✓	✓	✓	✓	✓	✓
Detect (text)	✓	✓	⧖	⧖	⧖	⧖	⧖	⧖	⧖
Detect (image)	✓	✓	⧖	⧖	⧖	⧖	⧖	⧖	⧖
Practice	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Language coverage and status of the modules. EN=English, AR=Arabic, BN=Bangla, FR=French, HI=Hindi, IT=Italian, TL=Filipino (Tagalog), NE=Nepali, UR=Urdu. ✓ = available; ⧖ = in progress at the time of writing this paper.

2.3 Models

CRITISENSE currently supports *(i)* text-based fact-checking and propaganda detection, and *(ii)* image-based propaganda detection for Arabic and hateful meme detection for English. These models are deployed through an in-house API platform.[1]

[1] https://apihub.tanbih.org/docs

Task/Dataset	Modality	Train	Dev	Test
Factuality (En)	Text	467,830	67,389	134,594
Factuality (Ar)	Text	28,221	4,507	7,847
Propaganda (En)	Text	4,472	621	922
Propaganda (Ar)	Text	18,453	1,318	1,326
ArMeme (Ar)	Image	3,604	522	1,021
Hateful meme (En)	Image	8,500	540	2,000

Table 2: Dataset statistics for each task and modality, showing the number of instances in the train/dev/test splits used in our experiments.

At present, automated detection is available for English and Arabic, while instructional components, including lessons, quizzes, and the Practice module, are accessible across all nine supported languages (Table 1). Extending detection to additional languages requires curated factuality and propaganda resources, which remain uneven across languages. We aim to add these functionalities for other languages as part of our ongoing work.

2.3.1 Datasets

Factuality (text). We train the factuality model on a curated collection of publicly available datasets in Arabic and English. For Arabic, we include AraFacts2 (Sheikh Ali et al., 2021), ANS-Claim (Khouja, 2020), CT22Claim (Nakov et al., 2022), NewsCredibilityDataset (Samdani et al., 2023), and COVID19Factuality (Alam et al., 2021). For English, we include PolitiFact (Da San Martino et al., 2023), CT22T3_Factuality (Alam et al., 2021), CheckThat!-style misinformation datasets (Köhler et al., 2022; Shahi et al., 2021), and AVeriTeC (Schlichtkrull et al., 2023).

Propaganda (text). For text-based propaganda detection, we use PropXplain (Hasanain et al., 2025), which contains labeled instances in both Arabic and English.

Propaganda (image). For image-based Arabic propaganda detection, we use ArMeme (Alam et al., 2024) (propagandistic vs. non-propagandistic), a meme-centric dataset.

Hateful memes (image). For meme English hate detection, we use the Facebook Hateful Memes dataset (Kiela et al., 2020) (hate vs. non-hate).

Table 2 reports the data used for each task, grouped by modality and language, along with the train/ dev/ test split sizes. Across tasks, development splits are used for model selection and tuning, while test splits are held out for final evaluation.

2.3.2 Model Training and Evaluation

Model Training. We prioritize models that are low-cost to deploy in resource-constrained settings, including CPU-based servers. Accordingly, we use lightweight, widely adopted backbones. BERT-base for English (Devlin et al., 2019), AraBERT for Arabic (Antoun et al., 2020), and ViT-B/16 for image-based classification models (Dosovitskiy et al., 2020). These choices provide a strong accuracy *vs.* efficiency trade-off and allow seamless integration into our in-house API platform. Note that all trained models are binary classifiers.

Results. Table 3 reports performance across tasks. Overall, the text models perform well on factuality and propaganda detection. Arabic factuality is highest (Micro-F1=0.868), suggesting that the training data is well represented. English factuality is lower (Micro-F1=0.726), which likely reflects broader topic diversity and more varied claim formulations. Propaganda detection performs similarly across languages, with Micro-F1 scores of 0.772 for English and 0.762 for Arabic. However, the task remains challenging as misleading information often relies on subtle rhetorical framing, implicit meanings, and contextual interpretation.

In contrast, image-based hateful/propagandistic meme detection remains more challenging. ViT-B/16 achieves moderate Macro-F1 on ArMeme (0.554) and lower performance on Hateful Memes (0.507). This gap is expected, as memes often rely on implicit cultural context, sarcasm, and fine-grained image–text interactions that lightweight vision-only models struggle to capture. While stronger multimodal models exist (e.g., Shahroor et al., 2026), we adopt lightweight models to meet deployment and CPU-serving constraints.

Error analysis. We conducted an error analysis of the multimodal model outputs and found that most errors occur in memes requiring multimodal grounding, especially those with implicit references or tightly coupled text-image semantics. To address these issues, we are exploring two directions: *(i)* integrating stronger multilingual vision-language models for more accurate inference, and *(ii)* expanding the training data with additional annotated meme datasets. We will retain lightweight models as a fallback to support efficient deployment in resource-constrained environments.

Task/Dataset	Metric	Model	Performance
Factuality (En)	Mi-F1	BERT-Base	0.726
Factuality (Ar)	Mi-F1	AraBERT	0.868
Propaganda (En)	Mi-F1	BERT-Base	0.772
Propaganda (Ar)	Mi-F1	AraBERT	0.762
ArMeme (Ar)	Ma-F1	ViT-B/16	0.554
Hateful meme (En)	Ma-F1	ViT-B/16	0.507

Table 3: Model performance on each task and dataset, reporting the evaluation metric (Micro-F1 for text; Macro-F1 for image) and the backbone used.

3 Evaluation of CritiSense

We evaluate CritiSense through a structured usability study covering both interface usability and perceived learning/behavioral outcomes. The questionnaire measures five usability constructs, along with single-item ratings for overall satisfaction, ease of completing the lesson, quiz flow, intention to continue using the app, and likelihood of recommending it. We also collect brief contextual information and open-ended feedback to identify concrete design priorities. The study was conducted for both Arabic and English languages.

Participants. We recruited 93 participants and administered the study through SurveyMonkey. Participants were primarily aged 18–24 (68.8%). English was the most common language preference (62.4%), followed by bilingual English/Arabic use (20.4%) and Arabic (17.2%). The educational background of the participants was predominantly bachelor's degree level (80.6%), and their fields of study were skewed toward STEM disciplines. Participation was voluntary, however, participants received a $14 voucher as modest compensation.

Instrument. The instrument combines quantitative ratings with qualitative feedback (Section B). It includes 17 five-point Likert items (strongly disagree-strongly agree), a five-point overall satisfaction item, a seven-point ease-of-completion rating for the lesson/quiz flow, and a Net Promoter Score (0-10).[2] We also measured participants' intention to continue and collected two multi-select responses on *(i)* activities completed and *(ii)* area of expertise or educational background. In addition, we included three open-ended questions on *(i)* points of confusion, *(ii)* highest-priority fixes, and *(iii)* most-liked features.

Usability constructs and scoring. We group the 17 Likert items into five theoretically motivated constructs (Appendix Table 5): **Usability/UX** (ease

[2] *How likely are you to recommend [product/app] to a friend or colleague?* (Reichheld et al., 2003)

Construct	Mean (SD)	α	% Positive
Usability / UX	4.12 (0.85)	0.653	77.4%
Visual Design	4.16 (0.70)	0.774	72.0%
Navigation	4.20 (0.84)	0.837	73.1%
Content Effectiveness	4.09 (0.62)	0.746	65.6%
Behavioral Impact	3.99 (0.72)	0.840	63.4%
Overall (17 items)	—	**0.921**	—

Table 4: Construct-level summary statistics, including mean (**M**), standard deviation (**SD**), internal consistency (α), and the percentage of positive responses. Overall ratings for each questionnaire item in the construct were out of 5.

of use and perceived usefulness), **Visual Design** (aesthetics and clarity of visual elements), **Navigation** (ease of moving through the interface and finding content), **Content Effectiveness** (quality of examples/quizzes and cross-language consistency), and **Behavioral Impact** (perceived changes in critical evaluation behaviors). For each respondent, we compute each construct score as the mean of its constituent items, supporting both item-level analysis and construct-level comparisons.

Analysis. We compute descriptive statistics, including the mean (**M**), standard deviation (**SD**), and median, for all individual items and aggregated construct scores. We assess internal consistency for each construct using Cronbach's alpha (α. Finally, we examine relationships among constructs using Pearson correlations (r)

4 Findings

Usability and satisfaction. Overall usability and satisfaction was strong (see Figure 4 and 5 in Appendix). All five constructs exceeded the agreement threshold (M>3.0) as reported in Table 4, with Navigation rated highest (M=4.20, SD=0.84) followed by Visual Design (M=4.16, SD=0.70). At the item level, *easy to use* received the highest rating (M=4.43, SD=0.88), with 90.1% of respondents selecting agree or strongly agree. Self-reported satisfaction was similarly high: 83.9% of users indicated they were satisfied or very satisfied (M=4.14, SD=0.87).

Learning design and engagement. Participants responded positively to the pedagogical structure, particularly the quiz-based reinforcement. The item *Quizzes reinforce learning* scored M=4.30 (SD=0.79), and 90.3% of participants completed at least one quiz during their session. At the construct level, *content effectiveness* was high (M=4.09, SD=0.62), suggesting that lessons communicated the targeted critical-thinking concepts.

Behavioral impact. The *behavioral impact construct* scored M=3.99 (SD=0.72), with 63.4% of users providing positive ratings. Notably, Arabic-language users reported the highest behavioral impact (M=4.36), which is consistent with the hypothesis that the app might be valuable in contexts where localized and language specific critical-thinking resources are less available.

Instrument reliability. It was high overall. Cronbach's α ranged from 0.653 for Usability/UX (2 items) to 0.840 for behavioral impact (4 items), with an overall scale reliability of α=0.921. The lower value for the two-item construct is expected since α is sensitive to the number of items, while the full scale shows excellent reliability.

Qualitative feedback and improvement targets. Open-ended questions had high response rates (86–100%), providing actionable feedback. Participants frequently highlighted *ease of use* and the quiz-based learning flow. Responses also identified some issues that can help us to improve the app.

Interpretation and implications. Overall, results indicate that CRITISENSE provides a usable and engaging first-release learning experience, with clear priorities for iteration. Navigation was rated highest (M=4.20), suggesting that the chapter→lesson→quiz flow is easy to follow. Quiz-based reinforcement was also well received (M=4.30), and 90.3% of participants completed at least one quiz.

Correlations are consistent with a "UX→content→impact" pattern. UX ratings correlate with content effectiveness ($r = 0.58$), and content effectiveness correlates with behavioral impact ($r = 0.63$). While not causal, these associations suggest that improving UX may yield downstream gains in perceived learning outcomes. Finally, NPS reflects an early-release profile (NPS$= -8.6$): 31.2% of users are promoters (mode=10/10), with detractors accounting for 39.8% of responses.

Pre/post app-use learning gains. To complement the usability findings, we conducted a preliminary pre/post app-use study through Prolific using misinformation and propaganda recognition quizzes. Among n=52 participants, accuracy in identifying misleading information increased from 70.2% to 77.4%, an absolute gain of +7.2 points and a relative improvement of +10.3%. The largest gains were observed for propaganda technique iden-

tification. Each participant received £8 as compensation. A full longitudinal study with a control group, and behavioral measures is ongoing and will be reported in future work.

5 Related Work

Detection-based and platform interventions. There has been substantial amount of work addressing misinformation through automatic detection, fact-checking pipelines, and platform UI interventions (e.g., labels and warnings) (Hasanain et al., 2024b,a; Alam et al., 2022, 2021; Zhou and Zafarani, 2020; Shaar et al., 2022; Abouzied et al., 2025). While these systems provide important first-line safeguards, their effectiveness can degrade under temporal shift (Stepanova and Ross, 2023), vary across languages in cross-lingual transfer (Ozcelik et al., 2023), and yield limited average changes in real-world beliefs or consumption (Aslett et al., 2022). Moreover, selective labeling can increase perceived accuracy of untagged content (the "implied truth" effect) (Pennycook et al., 2020).

Prebunking and inoculation. Psychological inoculation builds resistance by teaching manipulation techniques rather than correcting individual claims. Games such as *Bad News* and *Harmony Square* improve recognition of propaganda tactics and reduce perceived credibility of misleading content (Roozenbeek and van der Linden, 2019, 2020), with similar effects reported in topic-focused and multilingual variants (Basol et al., 2021). Scalable prebunking videos can also improve technique recognition on social media (Roozenbeek et al., 2022). However, many interventions are delivered as one-off web experiences and are evaluated primarily with immediate post-intervention assessments.

Media literacy tools. Mobile-first media literacy tools remain comparatively less common. *Cranky Uncle* uses humor and fallacy training to build resilience against climate misinformation (Cook et al., 2023), while feed simulations such as *Fakey* provide ecologically realistic practice and report improved source discernment in longitudinal deployments (Micallef et al., 2021). Overall, prior tools tend to emphasize either short-session tactic inoculation or practice-based news literacy, but rarely combine both within a multilingual, mobile-first experience designed for sustained everyday use beyond classroom settings.

CRITISENSE. CRITISENSE bridges these strands by combining tactic-level prebunking with practical verification skills (e.g., distinguishing opinion from evidence, fallacy spotting, and verification habits) through short, interactive exercises with explanation-rich feedback. Unlike primarily browser-based prebunking tools, it is mobile-first, multilingual, and designed for repeated microlearning.

More concretely, CritiSense differs from prior prebunking and media-literacy tools along three axes. *First*, browser-based inoculation games such as *Bad News* (Roozenbeek and Van Der Linden, 2019) and *Harmony Square* (Roozenbeek and van der Linden, 2020) are typically delivered as one-off sessions, whereas CritiSense supports continuous, repeated engagement through microlearning. *Second*, tools such as *Cranky Uncle* (Cook et al., 2023) and *Fakey* (Micallef et al., 2021) are often limited to specific domains or primarily English-language settings, while CRITISENSE provides consistent lesson, quiz, and Practice functionality across nine languages. *Third*, whereas prior systems commonly emphasize either tactic-level inoculation or practice-based literacy, CRITISENSE integrates both with explanation-rich feedback and an AI-assisted detection module within a single deployed platform.

6 Conclusions

We presented CRITISENSE, a mobile-first media-literacy app that delivers technique-focused prebunking and practical verification skills through short, interactive exercises with explanation-rich feedback. CRITISENSE is, to our knowledge, the first effort of this kind released in multiple languages, starting with Arabic and English and extending to additional seven languages. A formative usability study with 93 users shows strong user experience, including high usability (Navigation M=4.20; ease-of-use M=4.43), high satisfaction (83.9% positive), excellent internal consistency of the evaluation instrument (Cronbach's $\alpha = 0.92$) $+10.3\%$ relative) gain in quiz accuracy after app use. Qualitative feedback further validates the chapter→quiz learning loop and highlights functionality improvements. Overall, these findings support CRITISENSE as a scalable multilingual prebunking platform and motivate future work on long-term learning, behavioral impact, and multimodal analysis.

Limitations. CRITISENSE is designed to strengthen awareness and skills for recognizing fake news, mis/disinformation, and propaganda. Our current evaluation is a pilot mixed-method study focused on usability, engagement, and short-term learning signals; it does not yet measure long-term retention or real-world behavioral change (e.g., sharing on users' own platforms). The app presently covers a curated set of techniques and item types, and the content requires ongoing updates and localization.

To support scalability, the platform adopts a modular design: lessons, quizzes, and Practice scenarios are stored as structured content, allowing rapid updates without requiring app-store re-releases. We are also piloting an LLM-assisted authoring workflow, where generated lesson drafts are curated and validated by expert reviewers, enabling faster adaptation to emerging narratives while maintaining content quality.

While the design aims to generalize across topics, we do not claim coverage across all domains or user groups, and larger longitudinal and cross-cultural studies are needed. Finally, CRITISENSE is intended to complement platform detection and fact-checking systems not to replace them.

Ethics and broader impact. CRITISENSE operates in a sensitive domain where design choices can inadvertently amplify harmful narratives. To mitigate this, the app teaches manipulation patterns (e.g., emotional framing, scapegoating) using a technique-centered approach and avoids presenting harmful misinformation. Feedback emphasizes actionable verification steps (e.g., source checks, evidence tracing) rather than restating false claims. We also acknowledge dual-use risks: explanations of propaganda strategies could be misused to craft persuasive misinformation. CRITISENSE reduces this risk by prioritizing recognition and critical questioning, not step-by-step guidance for manipulation, and by curating examples for educational intent. Regarding privacy, CRITISENSE does not collect sensitive user information. Overall, we expect positive impact. Improved media literacy can support informed participation and reduce susceptibility to manipulation, particularly in multilingual settings. We plan longitudinal studies to assess durability and monitor potential unintended effects (e.g., overconfidence or blanket skepticism).

Acknowledgments

The work was supported by NPRP grant 14C-0916-210015 from the Qatar National Research Fund, part of the Qatar Research Development and Innovation Council (QRDI). The findings reported herein are solely the responsibility of the authors.

References

Azza Abouzied, Firoj Alam, Raian Ali, and Paolo Papotti. 2025. Combating misinformation in the arab world: Challenges and opportunities. *Communications of the ACM*, 68(10):48–53.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Firoj Alam, Abul Hasnat, Fatema Ahmad, Md. Arid Hasan, and Maram Hasanain. 2024. ArMeme: Propagandistic content in Arabic memes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21071–21090, Miami, Florida, USA. Association for Computational Linguistics.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Kevin Aslett, Andrew M Guess, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2022. News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science advances*, 8(18):eabl3844.

John A Banas and Stephen A Rains. 2010. A meta-analysis of research on inoculation theory. *Communication monographs*, 77(3):281–311.

Melisa Basol, Jon Roozenbeek, Manon Berriche, Fatih Uenal, William P McClanahan, and Sander van der Linden. 2021. Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against covid-19 misinformation. *Big Data & Society*, 8(1):20539517211013868.

Lara Marie Berger, Anna Kerkhof, Felix Mindl, and Johannes Münster. 2025. Debunking "fake news" on social media: Immediate and short-term effects of fact-checking and media literacy interventions. *Journal of Public Economics*, 245:105345.

John Cook, Ullrich KH Ecker, Melanie Trecek-King, Gunnar Schade, Karen Jeffers-Tracy, Jasper Fessmann, Sojung Claire Kim, David Kinkead, Margaret Orr, Emily Vraga, and 1 others. 2023. The cranky uncle game—combining humor and gamification to build student resilience against climate misinformation. *Environmental Education Research*, 29(4):607–623.

Giovanni Da San Martino, Firoj Alam, Maram Hasanain, Rabindra Nath Nandi, Dilshod Azizov, and Preslav Nakov. 2023. Overview of the CLEF-2023 CheckThat! lab task 3 on political bias of news articles and news media. In *Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum*, CLEF '2023, pages 250–259, Thessaloniki, Greece.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv preprint*, arXiv:2010.11929.

Andrew M Guess, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences*, 117(27):15536–15545.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024a. Can GPT-4 identify propaganda? annotation and detection of propaganda spans in news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, LREC-COLING '24, pages 2724–2744, Torino, Italia. ELRA and ICCL.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024b. Large language models for propaganda span annotation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14522–14532, Miami, Florida, USA. Association for Computational Linguistics.

Maram Hasanain, Md Arid Hasan, Mohamed Bayan Kmainasi, Elisa Sartori, Ali Ezzat Shahroor, Giovanni Da San Martino, and Firoj Alam. 2025. PropXplain: Can LLMs enable explainable propaganda detection? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23855–23863, Suzhou, China. Association for Computational Linguistics.

Jude Khouja. 2020. Stance prediction and claim verification: An Arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 8–17, Online. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Juliane Köhler, Gautam Kishore Shahi, Julia Maria Struß, Michael Wiegand, Melanie Siegel, and Thomas Mandl. 2022. Overview of the CLEF-2022 CheckThat! lab task 3 on fake news detection. In *Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum*, CLEF '2022, Bologna, Italy.

Chang Lu, Bo Hu, Qiang Li, Chao Bi, and Xing-Da Ju. 2023. Psychological inoculation for credibility assessment, sharing intention, and discernment of misinformation: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 25:e49255.

Nicholas Micallef, Mihai Avram, Filippo Menczer, and Sameer Patil. 2021. Fakey: A game intervention to improve news literacy on social media. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Preslav Nakov, Alberto Barrón-Cedeño, Firoj Alam, Giovanni Da San Martino, Tamer Elsayed, Maram Hasanain, Reem Suwaileh, and Wajdi Zaghouani. 2022. Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets. In *Working Notes of CLEF 2022 (CEUR Workshop Proceedings)*.

Oguzhan Ozcelik, Arda Sarp Yenicesu, Onur Yildirim, Dilruba Sultan Haliloglu, Erdem Ege Eroglu, and Fazli Can. 2023. Cross-lingual transfer learning for misinformation detection: Investigating performance across multiple languages. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 549–558, Vienna, Austria. NOVA CLUNL, Portugal.

Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management science*, 66(11):4944–4957.

Frederick F Reichheld and 1 others. 2003. The one number you need to grow. *Harvard business review*, 81(12):46–55.

Jon Roozenbeek and Sander Van Der Linden. 2019. The fake news game: actively inoculating against the risk of misinformation. *Journal of risk research*, 22(5):570–580.

Jon Roozenbeek and Sander van der Linden. 2019. Fake news game confers psychological resistance against online misinformation. *Humanities and Social Sciences Communications*, 5(65):1–10.

Jon Roozenbeek and Sander van der Linden. 2020. Breaking harmony square: A game that "inoculates" against political misinformation. *Harvard Kennedy School (HKS) Misinformation Review*, 1(8).

Jon Roozenbeek, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. 2022. Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, 8(34):eabo6254.

Jon Roozenbeek, Sander van der Linden, and Thomas Nygren. 2020. Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School (HKS) Misinformation Review*, 1(2).

D. Samdani, M. Taileb, and N. Almani. 2023. Arabic news credibility on twitter using sentiment analysis and ensemble learning. *Zenodo*.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shaden Shaar, Nikola Georgiev, Firoj Alam, Giovanni Da San Martino, Aisha Mohamed, and Preslav Nakov. 2022. Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2069–2080, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021. Overview of the clef-2021 checkthat! lab task 3 on fake news detection. *Working Notes of CLEF*.

Ali Ezzat Shahroor, Mohamed Bayan Kmainasi, Abul Hasnat, Dimitar Dimitrov, Giovanni Da San Martino, Preslav Nakov, and Firoj Alam. 2026. MemeLens: Multilingual multitask vlms for memes. *ArXiv preprint*, arXiv:2601.12539.

Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2021. AraFacts: The first large Arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nataliya Stepanova and Björn Ross. 2023. Temporal generalizability in multimodal misinformation detection. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 76–88, Singapore. Association for Computational Linguistics.

Cecilie S Traberg, Jon Roozenbeek, and Sander Van Der Linden. 2022. Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science*, 700(1):136–151.

World Economic Forum. 2026. The global risks report 2026. Technical report, World Economic Forum. Published: 14 January 2026.

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

A Usability constructs

Table 5 summarizes the five usability constructs used in our evaluation and the number of Likert items mapped to each construct. These constructs span usability/UX, visual design, navigation, content effectiveness, and behavioral impact, and are used to compute construct-level scores by averaging their constituent items.

B Usability Study Questionnaire

The following questionnaire was administered to $N = 93$ participants via SurveyMonkey after interacting with the CritiSense application. All Likert items used a 5-point scale: *Strongly Disagree* (1) – *Strongly Agree* (5), unless stated otherwise.

Construct	#	Description
Usability / UX	2	General usability: capabilities meet needs; ease of use
Visual Design	4	Aesthetics and clarity: design appeal; readability; icon clarity; animation clarity
Navigation	3	Ease of navigation; finding lessons; speed and responsiveness
Content Effectiveness	4	Misinformation identification; example relevance; quiz reinforcement; EN/AR consistency
Behavioral Impact	4	Confidence evaluating information; questioning reliability; noticing manipulation tactics; helping others

Table 5: Usability constructs and item counts used in the evaluation.

Section 1: Consent & App Usage

Q1. Which device did you use for CritiSense?
[Android phone / iPhone / Tablet / Computer]

Q2. Which language did you use in the app?
[English / Arabic / Both English & Arabic]

Q3. Is this your first time using CritiSense?
[Yes / No]

Q4. Which activities did you complete? *(Select all that apply)*
[Browsed the home screen / Completed a quiz / Switched app language / Explored multiple sections / Other]

Q5. What feature do you see in the app after reading a chapter?
[Multiple choice: questions related to the chapter / another lesson / another chapter]

Section 2: Usability & UX

Q6. CritiSense's technical capabilities meet my needs.

Q7. CritiSense is easy to use.

Q8. Overall, how easy or difficult was it to complete a lesson and quiz?
[1 = Extremely difficult 7 = Extremely easy]

Section 3: Visual Design

Q9. The visual design of the app is appealing.

Q10. The text is easy to read (size, contrast, spacing).

Q11. Icons and interface elements are clear and understandable.

Q12. Animations and transitions add clarity to the experience.

Section 4: Navigation

Q13. It is easy to navigate through the app.

Q14. I can find lessons and quizzes easily.

Q15. The app feels fast and responsive.

Section 5: Content Effectiveness

Q16. The lessons helped me better identify misinformation and manipulation techniques.

Q17. The examples used in the lessons feel relevant to real situations.

Q18. The quizzes reinforce what I learned.

Q19. The English and Arabic content feel consistent in quality and coverage.

Section 6: Behavioral Impact

Q20. After using CritiSense, I feel more confident evaluating information I see online.

Q21. I am more likely to question the reliability of social media posts after using CritiSense.

Q22. CritiSense helped me notice common tricks used to manipulate people online.

Q23. After using CritiSense, I am more likely to help friends or family spot misleading or manipulative online content.

Section 7: Overall Evaluation

Q24. Overall, how satisfied are you with CritiSense?
[1 = Very Dissatisfied 5 = Very Satisfied]

Q25. I would like to continue using CritiSense in the future.

Q26. How likely are you to recommend CritiSense to a friend or colleague?
[0 = Not at all likely 10 = Extremely likely]

Section 8: Open-Ended Feedback

Q27. What part of the app (if any) felt confusing or difficult?

Q28. If we fix one thing first, what should it be?

Q29. What did you like most about using CritiSense?

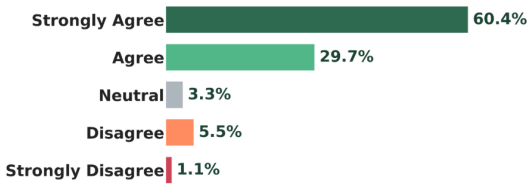

Figure 5: User perception of system usability ($N = 93$). A significant majority (90.1%) expressed agreement, with 60.4% indicating they "Strongly Agree."

Section 9: Demographics

Q30. What is your highest level of education? *[High school or equivalent / Some college / Bachelor's (pursuing/completed) / Master's (pursuing/completed) / PhD or Doctorate / Other]*

Q31. What is your field of study or expertise? *(Select all that apply)* *[Education / Media & Communication / Technology & IT / Business & Management / Engineering / Health & Medicine / Social Sciences / Prefer not to say / Other]*

Q32. What is your age group? *[Under 18 / 18–24 / 25–34 / 35–44 / 45–54 / 55+]*

C Usability and Satisfaction

Figures 4 and 5 summarize participants' overall satisfaction with CRITISENSE and their perceived *ease of use*. The satisfaction distribution (Figure 4) is strongly skewed toward positive responses: 36.6% of participants reported being *Very Satisfied* and 47.3% *Satisfied*, yielding a combined positive sentiment of 83.9%. Neutral responses accounted for 11.8%, while negative responses were marginal at 4.4% combined (2.2% *Dissatisfied* and 2.2% *Very Dissatisfied*). This pattern indicates broad platform acceptance among first-time users across both Arabic and English participants.

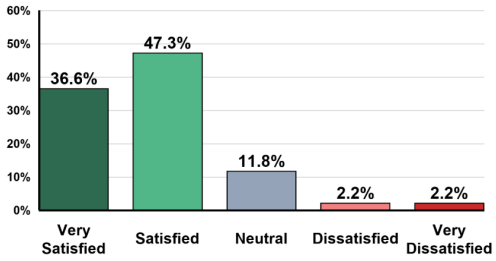

Figure 4: User satisfaction survey results (N=93). The distribution indicates high platform acceptance, with 83.9% of participants reporting positive sentiment (36.6% Very Satisfied and 47.3% Satisfied). Neutral responses accounted for 11.8%, while combined negative sentiment (Dissatisfied and Very Dissatisfied) remained minimal at 4.4%.

Perceived *ease of use* (Figure 5) follows a similar but even stronger positive skew. A combined 90.1% of participants agreed the app was *easy to use*, with 60.4% selecting *Strongly Agree* and 29.7% *Agree*. Neutral responses were minimal at 3.3%, and disagreement was limited to 5.5% *Disagree* and 1.1% *Strongly Disagree*. The clustering of responses at the *Strongly Agree* end suggests that the chapter→lesson→quiz workflow and onboarding flow lower the entry barrier for new users. Taken together, these distributions corroborate the construct-level results in Table 4, where Usability/UX, Visual Design, and Navigation all exceeded a mean score of 4.1 out of 5.