Supporting Young Learners in Identifying Toxic Content and Misinformation in Social Media – a Virtual Learning Companion Approach

Von der Fakultät für Informatik
der Universität Duisburg-Essen

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.)

genehmigte Dissertation

von

Farbod Aprin
aus
Teheran (Iran)

Erstgutachter / Betreuer: Prof. Dr. H. Ulrich Hoppe
Zweitgutachterin: Prof. Dr. Irene A. Chounta
Drittgutachter (extern): Prof. Dr. Rwitajit Majumdar, Kumamoto University (Japan)

Tag der mündlichen Prüfung: 19-12-2023

Abstracts

This dissertation explores the orchestration and evolution of Virtual Learning Companion (VLC), support as an add-on in both simulated and real social media environments. Furthermore, it investigates the impact of the VLC on users' judgments, critical thinking, levels of agreement, and awareness of both non-toxic and toxic content, using a blend of qualitative and quantitative statistical methodologies. Before shaping the VLC system and its potential scenarios, we created a minigame that enables users to categorize internet images based on various dimensions such as hate speech, cyber mobbing, verbal violence, and discrimination. This game not only promotes critical thinking skills but also serves as a tool for learning through conflict. The data collected from the game provide invaluable insights into students' perceptions of harmful content and contribute to fostering responsible online behavior. Moreover, the application enhances users' ability to analyze and evaluate online content. To address any potential controversies, we examined various methodologies for assessing agreement levels and selected the most appropriate one for calculating this index among raters. This chosen methodology was also applied to gauge the level of agreement among learners in the context of image manipulation within the VLC scenario.

Learning Companion Systems (LCS) enhance traditional Intelligent Tutoring Systems (ITS) by incorporating a computer companion that offers personalized guidance and interactivity. This results in an enriched learning experience. Referred to as 'learning companions,' these intelligent and autonomous robots or virtual conversational agents are capable of maintaining long-term user relationships. They find applications across various domains, including education and business, and focus on specific tasks such as teaching math and programming or assisting users in online transactions. Our VLC system is designed as a Chrome plugin and as a simulated version of Instagram with modular functionalities. This design allows for flexible application across different scenarios. We also crafted a series of statements in xAPI standard format, specifically designed to log user interactions.

We conducted empirical studies in Germany and Italy, focusing on daily social network issues such as hate speech, discrimination, racism, conspiracy theories, fake news, and especially image manipulation. The VLC browser plugin employs various chat dialogues to facilitate interactive learning experiences. It actively engages learners by posing knowledge-activating questions and providing pre-defined, insightful responses in certain cases.

In the initial scenario, a qualitative study was carried out with a limited user group. Feedback from users led to the optimization of both the VLC system and the simulated environment, influencing the design and user interaction for better qualitative outcomes. Subsequently, a second study was conducted with a larger sample, providing quantitative data that corroborated our initial findings. Here, we honed in on image manipulation that could be addressed by reverse image search engines, offering significant insights. The results of the analysis from the German school indicate that the time learners spent engaging with VLC recommendations from reverse image search and metadata correlated with improved judgment.

Comparative studies were also conducted in Italy to discern the impact of cultural variables, reinforcing the VLC's effectiveness in different educational settings for image manipulation scenario. These investigations confirmed the VLC's consistent influence on fostering critical evaluation skills, irrespective of cultural background.

Moreover, we also executed a targeted study dealing with racism through the VLC system in a simulated social media platform. This study unveiled the VLC's potential and adaptability to educating young minds about the adverse effects of racism, thereby contributing to a more inclusive society.

Keywords: Learning Companion Systems (LCS), Intelligent Tutoring System (ITS), Recommendation, Resilience, Increase Awareness, Fake news, social media, AI-based learning support systems, chatbot, xAPI. AI-based learning support systems, Recommendations, Awareness tools, Misinformation, Fake news, Image manipulation, Chatbot, xAPI, Gamification, COVID-19.

Zusammenfassung

Diese Dissertation untersucht die Orchestrierung und Entwicklung der Unterstützung durch den Virtual Learning Companion (VLC) als Ergänzung zu simulierten oder echten sozialen Medienumgebungen. Weiterhin wird der Einfluss des VLC auf die Urteilsfähigkeit, das kritische Denken, die Übereinstimmungsgrade und das Bewusstsein für sowohl nicht-toxische als auch toxische Inhalte anhand einer Kombination aus qualitativen und quantitativen statistischen Methoden untersucht. Bevor wir das VLC-System und seine potenziellen Szenarien gestaltet haben, haben wir ein Minispiel entwickelten, das es den Benutzern ermöglicht, Internetbilder anhand verschiedener Dimensionen wie Hassrede, Cyber-Mobbing, verbale Gewalt und Diskriminierung zu kategorisieren. Dieses Spiel fördert nicht nur das kritische Denken, sondern dient auch als Werkzeug für das Lernen durch Konfliktsituation. Die aus dem Spiel gesammelten Daten bieten unschätzbare Einblicke in die Wahrnehmung schädlicher Inhalte durch die Schüler und tragen dazu bei, ein verantwortungsbewusstes Online-Verhalten zu fördern. Um mögliche Kontroversen zu adressieren, haben wir verschiedene Methoden zur Bewertung von Übereinstimmungsgraden untersucht und die am besten geeignete ausgewählt, um Indizes unter den Bewertern zu berechnen. Diese ausgewählte Methode wurde ebenfalls eingesetzt, um den Grad der Übereinstimmung unter den Lernenden im Kontext der Bildmanipulation im VLC-Szenario zu messen.

Lernbegleitsysteme (LCS) verbessern traditionelle Intelligent Tutoring Systems (ITS), indem sie einen Computerbegleiter integrieren, der personalisierte Anleitung und Interaktivität bietet. Dies führt zu einer bereicherten Lernerfahrung. Als ‘Lernbegleiter’ bezeichnet, sind diese intelligenten und autonomen Roboter oder virtuellen Gesprächsagenten in der Lage, langfristige Benutzerbeziehungen aufrechtzuerhalten. Sie finden Anwendung in verschiedenen Bereichen, einschließlich Bildung und Geschäft, und konzentrieren sich auf spezielle Aufgaben wie das Lehren von Mathematik und Programmierung oder die Unterstützung der Benutzer bei Online-Transaktionen.

Unser VLC-System ist als Chrome-Plugin und als simulierte Version von Instagram mit modularen Funktionalitäten konzipiert. Dies ermöglicht eine flexible Anwendung in verschiedenen Szenarien. Wir haben auch eine Reihe von möglichen Aussagen im xAPI-Standardformat entworfen, die speziell darauf abzielen, Benutzerinteraktionen zu protokollieren.

Wir haben empirische Studien in Deutschland und Italien durchgeführt, die sich auf tägliche Probleme sozialer Netzwerke wie Hassrede, Diskriminierung, Rassismus, Verschwörungstheorien, Fake News und insbesondere Bildmanipulation konzentrieren. Das VLC-Browser-Plugin verwendet verschiedene Chat-Dialoge, um interaktive Lernerfahrungen zu erleichtern. Es bindet die Lernenden aktiv mit wissensaktivierenden Fragen ein und bietet in bestimmten Fällen vordefinierte, aufschlussreiche Antworten.

Im ersten Szenario wurde eine qualitative Studie mit einer begrenzten Benutzergruppe durchgeführt. Das Feedback der Benutzer führte zur Optimierung sowohl des VLC-Systems als auch der simulierten Umgebung und beeinflusste das Design und die Benutzerinteraktion für bessere qualitative Ergebnisse. Anschließend wurde eine zweite Studie mit einer größeren Stichprobe durchgeführt, die quantitative Daten lieferte, die unsere ersten Erkenntnisse bestätigten. Hier konzentrierten wir uns auf Bildmanipulationen, die durch Reverse-Image-Suchmaschinen (RIS) angesprochen werden konnten, und zugleich bedeutende Einblicke ermöglichten. Die Ergebnisse der Analyse für eine Schule in Deutschland zeigen, dass die Zeit, die Lernende mit den Empfehlungen des VLC aus der umgekehrten Bildsuche und den Metadaten verbrachten, mit einer verbesserten Beurteilungsfähigkeit korreliert.

Vergleichende Studien wurden auch in Italien durchgeführt, um den Einfluss kultureller Variablen zu ermitteln. Diese verstärkten die Wirksamkeit des VLC in verschiedenen Bildungsumgebungen für das Szenario der Bildmanipulation. Diese Untersuchungen bestätigten den konstanten Einfluss des VLC auf die Förderung kritischer Bewertungsfähigkeiten, unabhängig vom kulturellen Hintergrund.

Darüber hinaus haben wir auch eine gezielte Studie zum Thema Rassismus durch das VLC-System in einer simulierten soziale Medienplattform durchgeführt. Diese Studie enthüllte sowohl das Potenzial also auch die Anpassungsfähigkeit des VLC, junge Menschen über die negativen Auswirkungen des Rassismus aufzuklären und damit zu einer inklusiveren Gesellschaft beizutragen.

Schlüsselwörter: Learning Companion Systems (LCS), Intelligent Tutoring System (ITS), Empfehlung, Widerstandsfähigkeit, Bewusstsein steigern, Falschmeldungen, soziale Medien, KI-basierte Lernunterstützungssysteme, Chatbot, xAPI, KI-basierte Lernunterstützungssysteme, Empfehlung, Bewusstseinstool, Desinformation, Falschmeldungen, Bildmanipulation, Chatbot, xAPI, Gamification, COVID-19.

Table of Contents

List of Figures

List of Tables

1 Introduction

In this chapter, we discuss the motivation behind the development of a Virtual Learning Companion (VLC) and an initial mini-game. These tools are designed to heighten awareness and build resilience against toxic content on social media, with a particular focus on serving as research-based awareness tools for adolescents. Following the discussion on motivation, the section on the context and structure of the thesis provides a detailed roadmap for navigating the remainder of the document.

1.1 Motivation

Social media serves as a powerful platform for knowledge-sharing and co-creation, but it also presents considerable risks for adolescent users. These risks include fake news, false identities, conspiracy theories, hate speech, discrimination, and racism.

Social media as a primary information source has seen a significant increase, especially among the younger generation (Shearer, 2018), who use it to access and disseminate information (Jin & Liu, 2010). A study conducted by the European Commission revealed that 83% of individuals within the European Union perceive fake news on the Internet as a significant issue affecting the union's democratic values. The study further underscored the importance of reliable media sources, with respondents indicating traditional media as the most trusted platform for news (radio 70%, TV 66%, and print 63%). In contrast, online news sources and video-hosting websites garnered the lowest levels of trust, with only 26% and 27% of respondents expressing confidence in them, respectively (Eurobarometer, 2018).

The need for digital support measures stems from societal issues like racism, which involves marginalizing or harming individuals based on physical appearance, origin, or language (Kirkinis et al., 2018). Racism combines the ideology of racial superiority with social structures and behaviors that propagate dominance and oppression (Pieterse and Powell, 2016). It comprises prejudice, stereotyping, and discrimination (Levy et al., 2016). Prejudice or bias refers to negative feelings or evaluations towards a specific racial group, while stereotyping involves associating negative characteristics with the discriminated group (Allport, 1979). Racial discrimination can occur directly or indirectly through various activities or comments (Kirkinis et al., 2018; Combs, 2018).

Given these challenges, there is an urgent need for tools that enhance awareness and resilience concerning social media content and its associated risks for adolescents. The goal is to move beyond mere filtering or censorship methods on social media platforms.

One of the tools designed to address these issues in the field of education is the VLC, which can be applied in different educational and social settings. By conducting empirical studies in Germany and Italy with this tool, this thesis aims to offer insights, conclusions, and recommendations.

To achieve these objectives, the development of the VLC tool and environment is essential. This development requires a modular and adaptable architecture, enabling the application to work across various platforms and in different social contexts. The backend should be designed for easy integration or removal of various services in response to changing scenarios, without the need to alter the core system. This core system plays a crucial role in managing communications with the frontend and interfacing with both local and remote services.

The research findings will contribute to advancing our understanding of AI-based learning support systems, awareness tools, and recommendation mechanisms in the context of teenagers' well-being in social media environments.

1.2 Context and Structure of the Thesis

In this section, we provide a comprehensive outline of the thesis. We begin with an introduction to the 'Courage' project and its objectives, followed by the specific aims and research questions of this thesis. We also summarize key publications related to this thesis and conclude with an overview of its chapters.

1.2.1 Courage Project Description and Objectives

The 'Courage' project[1] is a collaborative initiative led by a group of European academics. This effort seeks to build an educational environment aimed at providing adolescents with the skills they need to navigate social media responsibly. Integral to this project is the creation of a companion for social media users, designed as both an educational guide and an interactive assistant. The companion addresses a variety of online threats that teenagers often face, such as discrimination, hate speech, cyberbullying, and false information dissemination.

Beyond the objectives of 'Courage,' the companion uses innovative gamification methods and intelligent content curation algorithms to offer an engaging user experience. The system integrates interactive counter-narratives and machine-learning models to understand the spread of toxic content across social media platforms. This multifaceted

[1] Courage EU project, http://www.couragecompanion.eu/, July 2023

approach allows the companion to provide personalized guidance tailored to each user's social context.

The customization is driven by two primary goals: Firstly, the project aims to foster healthier social interactions among users, their peers, and those who are targeted by various forms of prejudice. Secondly, it strives to increase users' awareness of the real-world ramifications of harmful content and their role in its spread.

The 'Courage' project emphasizes the importance of resilience and self-care as proactive defenses against harmful online content. Rather than solely relying on external mechanisms like censorship or filtering, 'Courage' is committed to developing critical thinking skills, particularly in the fight against fake news and disinformation.

This thesis contributes to the 'Courage' project by designing and developing a Virtual Learning Companion (VLC) in the form of a Chrome plugin. Created by Farbod Aprin under the supervision of the Rhine-Ruhr Institute for Applied System Innovation[2] (RIAS), this versatile chatbot is designed to adapt to various environments with the primary aim of tackling harmful content. The VLC is further enhanced through the development of various scenarios (Aprin, et al., 2022; Donabauer et al., 2023).

Other institutions contributing to the 'Courage' project, in addition to the RIAS Institute, include Hochschule Ruhr West[3] (HRW), Pompeu Fabra University[4] (UPF), The National Research Council (CNR)[5], Università Milano-Bicocca[6] and University of Regensburg[7].

1.2.2 Objectives of the Thesis

Aligned with the goals of the 'Courage' project, which aims to enhance the well-being of teenagers in social media environments, this thesis explores the design, implementation, and effectiveness of a Virtual Learning Companion (VLC) for specific innovative scenarios.

This thesis emphasizes the need for human-centric methods to recognize and counteract fake news on social media, as opposed to AI-powered techniques designed for automatic fake news detection. Given the educational focus of the 'Courage' project, such human-centric approaches are particularly relevant. The thesis discusses systems that support learners in fake news detection, content manipulation, and heightened awareness. The

[2] Rhine-Ruhr Institute for Applied System Innovation http://rias-institut.de/, Aug 2023

[3] Hochschule Ruhr West , https://www.hochschule-ruhr-west.de/, Aug 2023

[4] Pompeu Fabra University https://www.upf.edu/, Aug 2023

[5] The National Research Council, https://www.cnr.it/en, Aug 2023

[6] The University of Milano-Bicocca https://www.unimib.it/, Aug 2023

[7] Regensburg University, https://www.uni-regensburg.de/, Aug 2023

base of this thesis methodology is on the Intelligent Tutoring System (ITS) tradition of learning companion systems.

The VLC is designed to function across various scenarios, including social media platforms where image editing and enhancement are common. A key feature of this VLC is an embedded tool for detecting fake content, specifically manipulated images that could serve as sources of misinformation. The VLC integrates 'Reverse Image Search' (RIS) to provide contrasting contextual data for similar images. Offering RIS links is a novel educational strategy to combat fake news by providing insights from multiple contexts. Utilizing Natural Language Processing (NLP), the VLC extracts and presents key phrases from various websites containing specific toxic or specialized terms, based on a predefined dictionary.

The VLC also stimulates cognitive engagement by posing questions that prompt knowledge activation and encourages learners to participate in fact-checking exercises and maintain a questioning attitude toward social media content (Aprin, Chounta, et al., 2022).

We are particularly interested in exploring the role of recommendations provided through VLC interactions, based on data obtained from school trials overseen by an in-person instructor and a chatbot companion. Specifically, we aim to calculate and compare disagreement scores before and after a recommendation is given by the companion.

1.2.3 General Research Questions of the Thesis

The thesis aims to address the following research questions:

1. How can a learning companion support be orchestrated for a social media environment addressing young learners?

2. How can a modular VLC be designed and developed to support adolescents within different scenarios in a simulated social media environment (Adaptability)?

3. Can young learners benefit from access to reverse image search through VLC to make better decisions on the quality (fake or real) of certain social media items?

4. Can young learners benefit from access to VLC to make better decisions about the quality (racism, discrimination) of certain social media items?

5. Does a revision of the initial judgment about image manipulation guided by the VLC lead to a higher agreement with the expert opinion and other participants' reviews (convergence)?

6. Does different cultural backgrounds impact on the user’s judgment about image credibility?

1.2.4 Overview of Publications

Below is a list of publications included in the chapters, along with the authors' contributions:

[1] CSCL 2021 (Malzahn et al., 2021)

Malzahn, N., Aprin, F., Hoppe, H. U., Eimler, S. C., & Moder, S. (2021). Measures of disagreement in learning groups as a basis for identifying and discussing controversial judgements. In Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning-CSCL 2021. International Society of the Learning Sciences.

In collaboration with Hochschule Ruhr West (HRW), a partner in the 'Courage' project, the RIAS institute conducted a study using socio-cognitive conflict as a learning tool. RIAS contributed to the development of a game called SwipeIt, which measures disagreement among students learning about social media discrimination. Despite the challenges posed by COVID-19, the results demonstrate the potential of this approach in various educational settings, promoting critical thinking through disagreement.

Own contributions:

- Main developer of the SwipeIt application (front-end and back-end, designed the data structure, user token management and server-side functionalities and maintenance).

- Selected the technology stack and facilitated user authentication through a unique user-link solution architecture.

- Implementation of analytic measures as part of SwipeIt application.

- Contributed to the first draft of the mathematical sections and basic relations.

- Engaged in the paper revision process.

[2] ICWL 2021 (Aprin et al., 2021)

Aprin, F., Manske, S., Chounta, I. A., & Hoppe, H. U. (2021, November). Is this fake or credible? a virtual learning companion supporting the judgment of young learners facing *social media content. In International Conference on Web-Based Learning (pp. 52-60). Cham: Springer International Publishing.*

The ICWL research paper presents a web-based learning environment with a VLC to help young learners navigate social media responsibly. The VLC, implemented as a browser plugin, uses natural language processing and student modeling to provide personalized support. This system includes controlled and open social network environments for strategy development against the negative influences of Social Media. The VLC also logs user interactions for analysis, fostering pedagogical reflection. The approach emphasizes resilience, critical thinking, and awareness, offering a practical solution for responsible social media navigation.

Own contributions:

- First and leading author for conference submission.
- Responsible developer and designer of the VLC and its various scenarios.
- Introduced the concept of reverse image search to assist users with recommendations.
- Worked on the technical architecture and developed a VLC plugin.
- Adapted PixelFed for the 'Fake or Fact' scenario.
- Designed and conducted the first unofficial preliminary user study, with a focus on friendship and classmate zones.

[3] ITS 2022 (Aprin, Chounta, et al., 2022)

Aprin, F., Chounta, I. A., & Hoppe, H. U. (2022, June). 'See the image in different contexts': Using reverse image search to support the identification of fake news in Instagram-like social media. In International Conference on Intelligent Tutoring Systems (pp. 264-275). Cham: Springer International Publishing [Overall Best Paper Award].

This article discusses the impact of social media and fake news on youth, introducing the VLC as a tool to differentiate between credible and fake information. The VLC uses image analysis and Reverse Image Search (RIS) to foster critical thinking. The study emphasizes human-oriented approaches over AI-based techniques to combat fake news and conspiracy theories especially in pandemic situations. It also highlights games and initiatives that improve fact-checking skills. The results from a study involving 45 test users show that the VLC effectively helps users revise their initial judgments.

Own contributions:

- First and leading author for conference submission.
- Participated in designing the VLC and its various scenarios.
- Main technical developer (VLC plugin ‘Fake or Fact’ scenario, xAPI logging component and server side).
- Played a key role in designing xAPI statements to capture user interactions, providing valuable insights into user behavior.
- Introduced the innovative concept of utilizing reverse image search to assist users with recommendations, thereby aiding in the detection of image manipulation and fake content.
- Coordinated and designed the first unofficial preliminary user study with 45 participants.

[4] RPTEL 2023 (Malzahn et al., 2023)

Malzahn, N., Schwarze, V., Eimler, S. C., Aprin, F., Moder, S., & Hoppe, H. U. (2023). How to measure disagreement as a premise for learning from controversy in a social media context. Research and Practice in Technology Enhanced Learning.

In a joint study with HRW, the researchers in RIAS used SwipeIt to assess its impact on students’ personality traits and socio-emotional competencies. Participants showed high self-management skills, low authoritarianism, and strong self-awareness. The use of the ‘none’ button and the label ‘sexism’ significantly influenced their responses. The team introduced a ‘dispersion index’ for assessing controversy in collaborative learning. Despite gender representation limitations and the need for diverse samples, the study underscores SwipeIt’s influence on personality traits and its potential as a tool for analyzing controversy in education.

Own contributions:

- Developed the SwipeIt application, including adaptation, testing, and documentation.

- Handled technical orchestration: Provided user tokens for each trial and contributed to both front-end and back-end development.

- Transferred all implementation and theoretical responsibilities assigned as contributions from the CSCL paper [1] to this paper.

- Participation in the revision procedure of the textual content

Note: Did not play a role in the calculation of psychological indexes, correlations, or their interpretation to answer research questions.

[5] EAIT 2023 (Aprin, Peters, et al., 2023)

Aprin, F., Peters, P. & Hoppe, H. U. (2023, Nov). The Effectiveness of a Virtual Learning Companion for Supporting the Critical Judgment of Social Media Content. ***.*** *Education and Information Technologies.*

A web-based learning environment featuring a VLC was created to combat harmful social media content. It works in a simulation of Instagram named InstaCour, delivering real and manipulated content and aiding students in image evaluation and critical thinking enhancement. The project introduces a VLC that encourages skepticism and fact-checking, examining AI-based learning support systems' role in tackling misinformation. The empirical findings in this paper indicate that the VLC effectively engaged learners, and a difficulty measure was introduced to gauge the learners' ability to judge an image's credibility. Moreover, spending more time on results improved judgment accuracy, and guided revision led to a higher agreement with expert opinions which had positive improvements. The tool showed promising outcomes in enhancing students' sensitivity to social media issues and their critical thinking skills.

Own contributions:

- First and leading author for the journal submission.

- Designed the Scenario of the study ('Fake or Real').

- Engineered the technical architecture of the VLC, including updates to the VLC plugin.

- Developed a controlled and simulated version of Instagram (InstaCour), and managed the content specifically for the 'Fake or Real' scenario.

- Coordinated collaborative efforts for school trials conducted in Germany Managed the technical orchestration.

- Actively engaged in data analysis, extracting data from Learning Locker and MongoDB databases.

- Produced data visualizations based on the analysis employing the R programming language.

- Developed methods for data extraction, used for calculating correlations among different analytical indices.

[6] HELMETO 2023 (Aprin, Malzahn, et al., 2023)

Aprin, F., Malzahn, N., Lomonaco, F., Donabauer, G., Ognibene, D., Kruschwitz, U., ... & Hoppe, H. U. In Fulantelli, G., Burgos, D., Casalino, G., Cimitile, M., Bosco, G. L., & Taibi, D. (Eds.). (2023) The 'Courage Companion'–An AI-Supported Environment for Training Teenagers in Handling Social Media Critically and Responsibly. Higher Education Learning Methodologies and Technologies Online: 4th International Conference, HELMeTO 2022, Palermo, Italy, September 21–23, 2022, Revised Selected Papers. Springer Nature.

The HelmeTo paper is mostly about the architecture of the VLC. This paper explains that the RIAS and 'Courage' group collaboration have developed the VLC, a technologically advanced educational tool that provides adaptive tutoring and can detect harmful online content. It uses machine learning and gamification to enhance digital media literacy. The VLC, which integrates with an AI system for threat detection in social media, can be used on both controlled and open platforms. Its modular design enables the addition of new features to further enhance learners' online safety and media literacy.

Own contributions:

- Coordinating first author, authored the majority of the text up to section 5.1, technical architecture of the VLC and laying the groundwork for the paper's focus and structure.

- Managed the submission procedure as the first author, coordinating with co-authors and ensuring timely submission.

- Provided the conceptual and technical architecture of the Virtual Learning Companion system in sections 3 and 5.

- Handled the final revision process, ensuring all feedback was incorporated and the paper met academic standards.

[7] Frontiers 2023 (Ognibene et al., 2023)

Ognibene, D., Wilkens, R., Taibi, D., Hernández-Leo, D., Kruschwitz, U., Donabauer, G., Theophilou, E., Lomonaco, F., Bursic, S., Lobo, R. A., Sánchez-Reina, J. R., Scifo, L., Schwarze, V., Börsting, J., Hoppe, U., Aprin, F., Malzahn, N., & Eimler, S. (2023). Challenging social media threats using collective well-being-aware recommendation algorithms and an educational virtual companion. Frontiers in Artificial Intelligence, 5, 654930.

This paper introduces an adaptive framework for a ‘Social Media Virtual Companion’ aimed at educating and supporting teenagers in their interactions with social media. The companion combines automatic processing with expert intervention and guidance, powered by a recommender system that optimizes for Collective Well-Being (CWB) metrics, rather than platform profit or engagement. With this focus, the framework aims to combat social media threats such as toxic content, disinformation, beauty stereotypes, and bullying. The companion also functions as an Intelligent Tutor System (ITS), selecting suitable content and effective interventions. The paper emphasizes enhancing users’ digital literacy and highlights the need for new forms of literacy like ‘digital citizenship’ or ‘new media literacy.’ It also discusses the role of educators and behavioral economics strategies such as nudging and boosting to improve the social media experience. The paper concludes by advocating for a more democratic and transparent objective for social media platforms, aligning with the Collective Well-Being of the community rather than platform-centric goals.

Own contributions:

- Conducted research and developed VLC scenarios focused on combating fake content and conceptual development.

- Contributed to section 7.2, which aims at raising learners' awareness of fake content in an Instagram-like social media environment through reverse image search, based on the implementation of the VLC and environment platform.

- Provided foundational contributions that shaped the paper's educational focus.

- Had a significant impact on the VLC conceptual and technical architecture, analysis and conclusions drawn from the VLC scenarios, impacting the overall findings and implications of the research.

Note: Did not contribute to the writing and revision procedure of this paper.

1.2.5 Chapters Overview

Chapter 2 provides an in-depth look at the risks and threats teenagers face on social media. It covers psychological and pedagogical strategies for mitigating these risks and introduces the concept of Virtual Learning Companions (VLCs). The chapter concludes with an overview of existing educational technologies, games, and tools.

Chapter 3 focuses on developing the judging Social Media Content (SwipeIt) application and its study, drawing on insights from paper [1] and journal [4] in the list of the overview of publications. We describe the research and methods to calculate agreement among raters.

Chapter 4 in this approach presents the initial design of the Virtual Learning Companion (VLC), including its architecture and potential scenarios [2,3]. We also discuss limited user interactions and gather valuable feedback.

Chapter 5 discusses the evolution of the VLC's design and scenario, as well as initial user tests and their feedback.

Chapter 6 focuses on Study [3] and its 'Fake or Fact' scenario, which is primarily related to the COVID-19 situation as well as the distinction between fake and factual news. In this chapter, we will also discuss the analysis of small-scale user interactions through both qualitative and quantitative evaluations.

Chapter 7 highlights the updates made to the VLC interface and dialogue, including the switch from PixelFed to InstaCour for technical and conceptual reasons. Based on the journal [5], this chapter describes the VLC's role in image manipulation scenarios, specifically 'Fake or Real'. It covers classroom experiments conducted in both a German high school and an Italian university. It also analyses the learner's interaction with the system, explores the impact of cultural differences on user assessments and is guided by specific research questions.

Chapter 8 outlines the VLC's application in a Racism scenario, detailing the technical challenges and adaptability of the VLC. It also discusses the sample conditions and trial results.

Chapter 9 offers a comprehensive conclusion, summarizing the thesis's key findings and insights and answering the general research question of the thesis. It also outlines future directions for combating misinformation and toxic content on social media and some vision for further technologies for image and content authentication.

1.2.6 Correspondence between the General Research Questions and Research Questions Addressed in the EAIT Article

The general research questions of this thesis, along with those specifically addressed in Chapter 6 from the EAIT journal, both focus on the functionalities of the VLC and its related hypotheses. These are crucial for managing scenarios and guiding learners in their exploration to extract insights from the analysis. To elucidate the connection between the research questions in Chapter 6 and the general research questions addressed in section 8.3, the following Table 1-1 is provided. This table also indicates the corresponding chapters in this thesis that are related to these questions:

Table 1-1 Relation of research questions and corresponding chapters

General Research Questions	**Corresponding EAIT Journal Questions**	**Chapter Addressing the Question**
1. Support orchestration for social media environment.	Not directly addressed in EAIT questions.	Chapter 3,4,5: VLC Design and Implementation
2. Design and development of a modular VLC (Adaptability).	Not directly addressed in EAIT questions.	Chapter 4: VLC Architecture, Chapter 7 and 8 test this adaptable architecture
3. Benefit from reverse image search in VLC for decision-making on content quality.	**RQ2:** Was the accuracy of the judgment positively influenced by the time spent visiting or reading web pages related to the target picture?	Chapter 4,5,6,7: Implementation and User Interaction
4. Benefit from VLC for decisions about content quality (racism, discrimination).	Not directly addressed in EAIT questions.	Chapter 7: addressed but not directly evaluated or answered
5. Revision of judgment and higher agreement with experts (convergence).	**RQ3**: initial judgment guided by the VLC lead to higher agreement.	Chapter 6: Evaluation Studies
6. Impact of cultural background on image credibility judgment.	**RQ 4:** Influence of personality traits on tagging behavior.	Chapter 6: Evaluation Studies, VLC in Different Cultural Contexts Italy and Germany

2 Background and Related Work

In this chapter, we start by examining the risks and threats that social media poses, with a special focus on teenagers considering content from the Frontiers paper [7]. We then explore strategies from both psychological and pedagogical perspectives aimed at mitigating the negative impacts of social media. Subsequently, we delve into the foundational principles of Virtual Learning Companions (VLC), which can serve as supplementary tools to alleviate the adverse effects of social media. We conclude the chapter by evaluating existing solutions that employ educational technologies, including tools and mini-games, to tackle these challenges.

2.1 Social Media Threats in the Perspective of Teenagers

Research has shown that while enhancing social connections and overall well-being, social media can pose significant risks to individuals, particularly adolescents (Costello et al., 2019; Mladenović et al., 2021). These risks can be mitigated by fostering digital literacy and citizenship (Fedorov, 2015; Jones & Mitchell, 2016; Xu et al., 2019), which can counterbalance the problems arising from users' unawareness and excessive dependency on algorithm-based suggestions (Banker & Khetani, 2019; E. M. Meyers et al., 2013; K. L. Walker, 2016). The spread of false information, or fake news, is a major problem in online environments. It can lead to misinformation and disinformation, which can have serious societal consequences (Kozyreva, Lewandowsky, et al., 2020). The threats posed by social media can be broadly categorized into four types: content-based, algorithmic, dynamics-induced, and cognitive and socio-emotional threats and risks for teenagers and young adults (Ognibene et al., 2023)[7].

Content-based social media threats:

These threats include exposure to harmful content such as toxic content, disinformation or fake news, beauty stereotypes, and bullying. Hate instigators were, in fact, among the first to harness the potential of the internet (J. Chan et al., 2016; Gerstenfeld et al., 2003; Schafer, 2002). A study conducted in Norway involving high school graduates aged between 18 and 21 found that those engaged in cyberbullying reported significantly higher levels of anxiety, depression, self-harm, suicide attempts, and antisocial behavior compared to their non-involved counterparts (Skilbred-Fjeld et al., 2020). The study underscores the importance of understanding these psychological characteristics for early detection, prevention, and treatment of cyberbullying victims.

Racism on Social Media is a content-based threats. We should be concerned also about the problem of racism and the manifestation of racist viewpoints on social media (Matamoros-Fernández & Farkas, 2021), particularly about the impact it has had on young users in recent years during COVID (Rideout et al., 2021).

The impact of racism on social media platforms is also evident, with users experiencing hate comments and detrimental effects on mental health (Brandt, 2017; Layug et al., 2022; NRW, 2022; Tao & Fisher, 2022; Ybarra et al., 2011). Victims, bystanders, and perpetrators can suffer negative emotions and long-term consequences (Baldry et al., 2019). Without increased attention to everyday racism, racial prejudice may persist and be perpetuated. Innovative technology can play a role in raising awareness of racism and its consequences and fostering empathy (Aprin et al., 2023 [6]; Taibi et al., 2022; Theophilou et al., 2022).

Fake news is a debatable content-based social media threats that can alter users' perceptions of current events. News posts on social media often combine text and images, with images significantly influencing users' beliefs. Unfortunately, images can be easily manipulated using software or AI to spread false information (Zannettou et al., 2019). Different kinds of false and misleading information online are included in this definition of 'False information' (Kozyreva et al., 2020; Wardle & Derakhshan, 2017):

- **False Information:** This refers to any data or details that are untrue or factually incorrect[8].

- **False or Fake News:** Media and publications that are purposefully and provably inaccurate, with the potential to deceive readers (Allcott & Gentzkow, 2017).

- **False Rumors:** This involves the spread of general talk or hearsay that is widely disseminated and not based on factual knowledge.

- **Satire and Parody** Utilizing humor and satire without the intent to harm, yet with the possibility of deceiving and misleading.

- **Factitious Information Blends** Partial truths and conjectures that blend genuine data with misinformation (Rojecki & Meraz, 2016).

- **Deepfakes and Cheap Fakes:** Deepfakes are AI-based hyper-realistic digital falsifications of images, video, and audio (Chesney & Citron, 2018). Cheap fakes involve audiovisual manipulations using conventional techniques like speeding, slowing, cutting, re-staging, or re-contextualizing footage (Paris & Donovan, 2019).

[8] False information, https://www.lawinsider.com/dictionary/false-information, July 2023

Information Disorders:

- **Misinformation:** Inaccurate or deceptive content disseminated 'without harmful intentions.

- **Disinformation:** False, concocted, or distorted content disseminated 'with the purpose' to deceive or inflict damage.

- **Malinformation**: Authentic information distributed with the intention to cause damage, like spreading hate or revealing private details.

Conspiracy and Propaganda:

- **Propaganda:** This refers to data or details, often slanted or deceptive, utilized to advocate for a political agenda or perspective (NATO Strategic Communications Centre of Excellence, 2017). It can be political or industrial, such as the tobacco industry health benefits late 50-60th. This involves the spread of false or fake news. These are news articles that are deliberately and verifiably false, potentially misleading readers (Allcott & Gentzkow, 2017).

- **Systemic Lies:** These are carefully constructed fabrications or obfuscations intended to protect and promote material or ideological interests with a coherent agenda (McCright & Dunlap, 2017).

- **Conspiracy Theories:** These are different interpretations of conventional news events, suggesting that such events are orchestrated by a secretive, typically malevolent, elite group (Roozenbeek & van der Linden, 2019).

Deep fakes, created using advanced machine learning, present a major threat due to their hyper-realistic manipulation of images, video, and audio. These AI-generated forgeries can mislead human perception, potentially deceiving countless users. (Chesney & Citron, 2018). One of the most successful viral deep fake images was created by a 31-year-old construction worker who enjoyed experimenting with the AI image generator, Midjourney[9]. This deep fake depicted Pope Francis wearing a white puffer coat, reaching ankle-length, but it was later revealed to be a fake (Fetters Maloy & Branigin, 2023).

[9] AI image generator system, https://www.midjourney.com/ , Aug 2023

Algorithmic Social Media Threats

Social media algorithms can introduce additional threats. For instance, the selective exposure of users to news sources can lead to a continuous state of isolation from diverse ideas and perspectives, creating 'filter bubbles' (Geschke et al., 2019; Nikolov et al., 2015), and forming polarized social structures or 'echo chambers' (Del Vicario et al., 2016; Gillani et al., 2018). Another potential issue is gerrymandering (Stewart et al., 2019) where users are exposed to unbalanced neighborhood configurations. AI-assisted Information architectures are systems that use AI to personalize the information that users see, which can lead to echo chambers and filter bubbles (Kozyreva et al., 2020).

Threats Induced by Social Media Dynamics

The dynamics of social media, driven by the rapid interaction between their algorithms, common social tendencies, and stakeholders' interests, can also pose threats (Anderson & McLaren, 2012; Milano et al., 2021). These dynamics can amplify the acceptance of harmful beliefs (Neubaum & Krämer, 2017; Stewart et al., 2019) make social media users' opinions vulnerable to phenomena such as the spread of hateful content, and trigger large-scale violent outbreaks of fake news (Del Vicario et al., 2016; Webb et al., 2016).

The obscured regulations governing commercial activities have come to the forefront due to controversies involving Facebook's questionable transactions with user data (Guardian, 2018). Regulators and the general public now recognize the scope of privacy violations and information control by digital technologies and tech corporations. Moreover, these controversies have exposed the manipulative potential of strategies such as 'dark ads' (advertisements seen only by their intended targets) and microtargeting (tailoring ads for specific individuals). These techniques aim to influence people's decisions and voting habits by exploiting their psychological weaknesses and personal identities (Matz et al., 2017)

Social Media Cognitive and Socio-emotional Threats

Despite numerous studies exploring the mechanisms of content propagation on social media, there remains ambiguity around the modeling of users' emotional and cognitive states' impact on the dissemination of harmful content, particularly given their cognitive limitations (Allcott & Gentzkow, 2017; Pennycook & Rand, 2019; Weng et al., 2012). Crucial factors include limited attention span and error-prone information processing, often amplified by the emotional nature of the messages (Brady et al., 2017; Kramer et al., 2014). Lack of non-verbal communication and limited social presence often heighten carelessness and misbehavior, as users perceive themselves as anonymous (Diener et al., 1980; Postmes & Spears, 1998).

Over time, user behavior can degrade, showing signs of impulsivity and addictive tendencies, with social media usage exhibiting many neurocognitive features typical of recognized forms of addiction (Kuss & Griffiths, 2011; Lee et al., 2019). This so-called Digital Addiction (Almourad et al., 2020; Lavenia, 2012; Nakayama & Higuchi, 2015) has harmful effects, including hasty decisions, particularly impacting teenagers' academic performance and mood (Aboujaoude et al., 2006). Social media companies face issues related to keeping users on their platforms. Unfortunately, they present or filter posts that align with prevailing opinions to increase user engagement and exposure to advertisements (Bucher, 2016). It is now understood that diagnosing social media addiction is not solely based on the amount of time spent online, but also on users' behavior (Musetti & Corsano, 2018; Taymur et al., 2016).

Social media environments, largely controlled by recommender systems, could play a significant role in managing this condition, using strategies such as introducing delays for impulsive users and identifying triggers like Fear of Missing Out (Alutaybi et al., 2019). The connection between cognitive threats and digital addiction illustrates the multifaceted challenges posed by social media, especially in the context of teenagers.

The primary focus of this dissertation is on misinformation and disinformation, particularly conspiracy theories related to COVID-19, and predominantly on image manipulation classified under the categories of deepfakes and cheap fakes, which are created using common image editing software.

Summary: In conclusion, social media offers numerous benefits but also presents significant risks to teenagers, ranging from exposure to harmful content to algorithmic biases. These threats can be categorized into content-based, algorithmic, dynamics induced, and cognitive and socio-emotional threats. The spread of misinformation, especially through deep fakes and other false content, is particularly concerning. The complex interplay between cognitive limitations, emotional responses, and addictive behaviors underscores the urgency of the situation. It is crucial to continue research and development of tools and strategies to mitigate these risks and promote safe and responsible use of social media.

2.2 Strategies to Mitigate the Adverse Effects of Social Media: A Psychological and Pedagogical Perspective

The challenges that social media presents to teenagers and young adults are substantial and diverse. However, a variety of psychological and pedagogical methods can be effectively utilized to mitigate these challenges. Building on the idea of digital literacy, another method is to utilize collective well-being-aware recommendation algorithms. These algorithms can be programmed to endorse content that is advantageous to the user's well-being, while reducing exposure to harmful or toxic content. This can be particularly effective in combating the challenges posed by algorithmic biases, filter bubbles, and echo chambers (Ognibene et al., 2023) [7].

The primary focus of this thesis is to address the issue of misinformation, particularly concerning image manipulation on social media. In the following sections, we will provide literature that deals specifically with misinformation and disinformation problems and fight bias, filter bubbles and pre-judgments by recommending different opinions.

The theoretical model formulated by Victoria L. Rubin identifies three critical elements in the propagation of disinformation and misinformation: a **Vulnerable Pathogen**, a **Susceptible Host**, and a **conducive Environment** (V. L. Rubin, 2019). According to the model, the proliferation of fraudulent information can only occur when all these factors coexist (Figure 2-1).

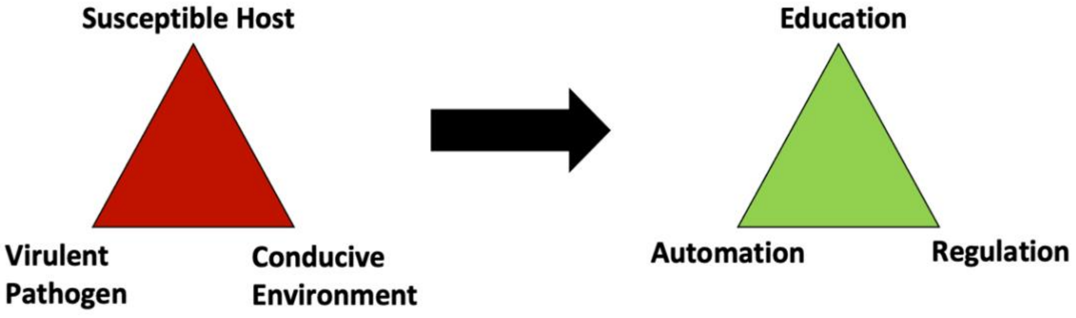


Figure 2-1: Victoria L. Rubin defines a conceptual model for fake news in the form of a triangle with the vertices susceptible host, virulent pathogen, and the conducive environment.

To counteract this phenomenon, the model suggests three interventions: using technology and automation to combat the deceptive element, educational initiatives to safeguard vulnerable recipients, and regulatory measures to ensure a safer environment.

The first intervention is **'Automation',** which can be achieved through the use of AI-based systems. This is an effective method for identifying and automatically labeling misinformation. An example of such an application is the LiT.RL news verification browser. Intended for use by news consumers, journalists, editors, or information professionals, this search tool scrutinizes the language on digital news platforms to determine whether the content is clickbait, satirical news, or counterfeit news. Despite its effectiveness, it is only sometimes accurate and ill-equipped for public use, primarily since it does not support multimedia specifications (Chen et al., 2015; V. Rubin et al., 2019). Further, Natural Language Processing (NLP)-based systems can benefit content creators, helping them discern common misinformation traits efficiently. These systems can also aid information professionals in filtering out and highlighting dubious content, thereby easing information overload on news consumers, or helping educational institutions impart critical content evaluation skills (Chen et al., 2015).

However, detecting fraudulent information, such as those embedded in images, remains challenging for both human and machine efforts due to advanced falsification techniques (Nguyen et al., 2022).

The second proposed intervention is '**Education'**. Nevertheless, it is critical to understand what obstacles may hinder the efficacy of teaching and why individuals resist altering their beliefs, even when confronted with logical and factual contradictions. Badke (2018) argues that confirmation bias, the tendency to favor information that aligns with one's existing beliefs, results in individuals not critically examining the facts. Instead, they focus on what they wish or expect to see (Badke, 2018).

The Dunning-Kruger effect, a cognitive bias where individuals mistakenly believe they're more intelligent and competent than they are, further compounds the problem. This overestimation, stemming from a lack of self-awareness and lower cognitive ability, is another educational barrier (Pennycook et al., 2017).

According to Lawson (2006), the illusion of explanatory depth exacerbates the challenge. Studies indicate that a significant proportion of non-experts find it difficult to accurately address or explain abstract queries. Education is proposed as a solution to this problem. A notable observation from this research was that while participants believed they understood the mechanics of a bicycle, when asked to illustrate a bike with its chain, over 40% of the non-experts committed errors in both the drawing and the forced-choice tasks (Lawson, 2006).

To combat this, the suggestion of implementing fact-checking awareness tools and providing unrestricted access to accurate, uncensored information from various contexts has been proposed. This approach is also championed by the American Library Association (ALA) as the most effective strategy for fighting disinformation and media manipulation. The stance of the ALA is rooted in the principle that individuals should have the freedom to make well-informed choices, even if it involves encountering viewpoints they don't agree with. The ALA further argues that censorship and content filtering are ineffective and can drive people to seek information from untrustworthy sources. The surest route to censorship is to first cast both fact and opinion as equal, and equally suspect," said Office for Intellectual Freedom Director James LaRue. "Accurate public information — scientific, medical, statistical, journalistic — is one foundation of our democracy and our freedoms" (Diaz Eleanor, 2017; McDonald & Levine-Clark, 2017).

Kyza et al. (2020) suggest various strategies to combat misinformation, including critically reviewing the information, acting responsibly, declining to share or like misleading content, flagging suspicious posts for review, and scrutinizing a post's justification and misinformation potential. They emphasize that ordinary users should also participate in this endeavor (E. A. Kyza et al., 2020).

The third vertex of Rubin's triangle pertains to '**Regulation'**. As briefly discussed in this dissertation, particularly in the section on threats induced by social media dynamics, 'Regulation' refers to the area where coordinated legislative efforts are essential for curbing the dissemination of false information. The Internet Society[10] (ISOC) and the European Commission have pushed for greater regulation, and initiatives have been launched to tackle pandemic-related disinformation.

The Digital Services Act (DSA[11]) is one regulation proposed by the EU that mandates major online platforms combat disinformation by adopting more responsible and transparent practices. While a total ban on misinformation is often impractical and undesired, the EU has focused on limiting the impact of misinformation while preserving freedom of speech. An example of such an initiative is the EU's 2018 Action Plan against Misinformation, which focuses on promoting media literacy, supporting independent media, and detecting disinformation (Z. Meyers, 2022).

Regulatory measures can, for example, establish clear guidelines for data protection, such as the EU's GDPR (EU parliament, 2016), or set standards for political advertising on social media platforms. Violations of these rules can result in substantial penalties. Additionally, these regulations can offer both rewards and penalties to tech companies and media outlets to guarantee the accuracy of shared information and promote respectful online discourse. It is essential for regulatory efforts to aim for a unified framework for user protection, rather than the existing piecemeal legislative approach, as seen in Germany and the EU, referenced by (Jaursch, 2019). At the Web Foundation conference, Tim Berners-Lee, the creator of the World Wide Web, underscored the need for governments to create appropriate digital-era regulations to maintain an open, competitive, and innovative online space while preserving individual rights and freedoms[12].

Kozyreva and her colleagues mentioned the crucial role of psychological science in comprehending and tackling the complexities of the digital world. They mentioned in their article similar solutions and interventions to Rubin's triangle solutions. They also elaborated on some other categories of psychological and social sciences. These categories indicate behavioral and cognitive interventions could be practical to empower

[10] Internet Society, https://www.internetsociety.org/, Feb 2023

[11] Digital Services Act, https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation, Jan 2023

[12] 30 years on, what's next #ForTheWeb Tim Berners-Lee, https://bit.ly/3LzacLS, Jan 2022

people and steer their decision-making toward the greater individual and public good. They propose that insights derived from psychological research can significantly contribute to the development of interventions across various domains, including legal, technological, educational, and cognitive and behavioral policymaking. In their work, they delineate three primary methodologies for intervention design: nudging, boosting, and technocognition (Kozyreva et al., 2020). Check in the below Figure 2-2:

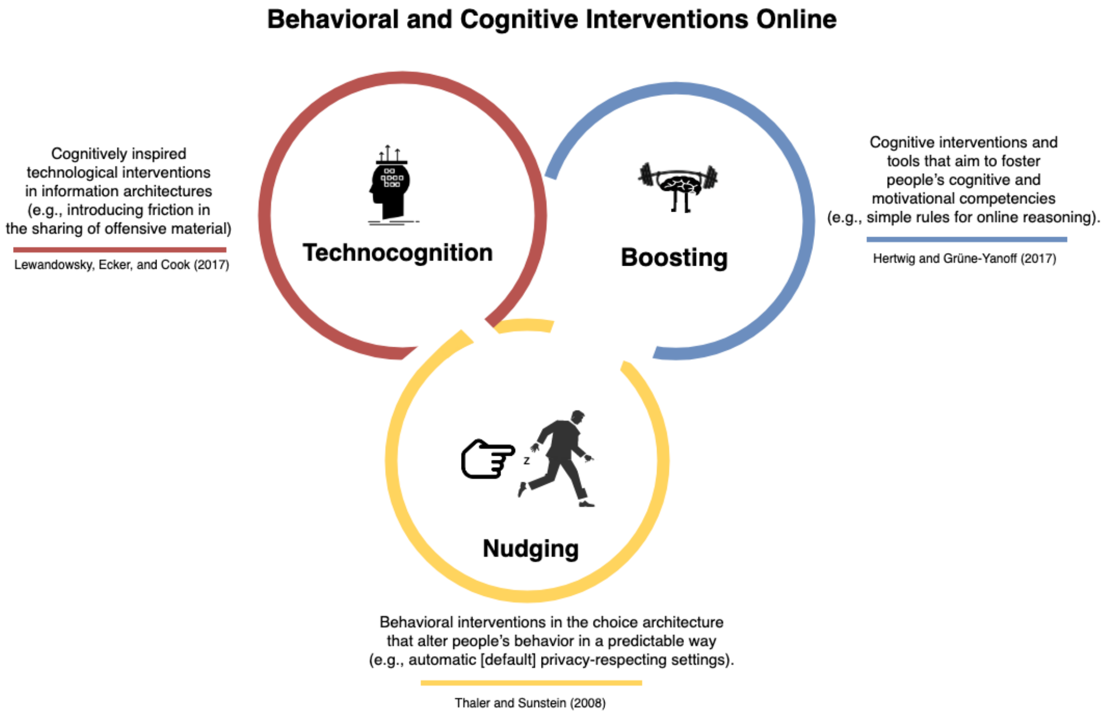

Behavioral and Cognitive Interventions Online
Cognitively inspired technological interventions in information architectures (e.g., introducing friction in the sharing of offensive material)
Lewandowsky, Ecker, and Cook (2017)
Technocognition
Boosting
Cognitive interventions and tools that aim to foster people's cognitive and motivational competencies (e.g., simple rules for online reasoning).
Hertwig and Grüne-Yanoff (2017)
Nudging
Behavioral interventions in the choice architecture that alter people's behavior in a predictable way (e.g., automatic [default] privacy-respecting settings).
Thaler and Sunstein (2008)

Figure 2-2: The image, inspired by the similar diagram in Kozyreva et al. (2020), illustrates various behavioral and cognitive interventions designed for the digital realm. It includes an icon representing the concept of 'nudging,' which is used under the Attribution 3.0 Unported (CC BY 3.0) license provided by Luis Prado from thenounproject.com. Additionally, other icons in the image are utilized with licensing from Adobe Stock.

The potential effects of any behavioral intervention, such as nudging, boosting, or technocognition-depend on the specific context and individual characteristics of the users. Therefore, one should be careful about proposing 'one-size-fits-all solutions'. The vision is that of a toolbox of interventions that would reach different people and tackle different existing and emerging challenges.

'Nudging' is a behavioral policy approach that uses knowledge of human psychology and the power of choice environments to guide people's decisions toward maximizing individual or public benefit (Thaler & Sunstein, 2008). The concept is based on the idea that people's behavior can be changed via their environment without changing their minds. Nudging does not block or significantly burden choices, but rather proposes interventions that are easy, reversible, and cheap to implement.

Nudging can be achieved in various ways, such as by changing the order in which options are presented, thus altering their physical and cognitive accessibility. For example, rearranging food options in a cafeteria to make healthier foods more accessible is a nudge intended to increase healthy food consumption (Broers et al., 2017; Bucher et al., 2016). Another common nudging technique is the preselected default option, which has a significant impact on decisions. People are more likely to accept a preselected option than to select a different one (Jachimowicz et al., 2019).

Nudge intervention in combating COVID-19 misinformation on social media has been experimentally validated by Pennycook et al. (2020). Their research provides evidence that a scalable accuracy-nudge intervention can effectively reduce the spread of false information. Published in Psychological Science, the study underscores the potential of such interventions in enhancing the quality of information shared on online platforms, especially during critical times like the COVID-19 pandemic (Pennycook et al., 2020). However, the success and ethical permissibility of nudging largely depends on the goals of the choice architects and their alignment with the goals and values of individuals. Commercial choice architects often use nudging to maximize benefits for the service provider rather than the consumer (social media user), which can lead to undesirable outcomes for the consumer such as displaying advertisements and retaining users on the platform (Thaler, 2018).

There are also ‘educative nudges’, which involve some form of education, such as additional information. These interventions are transparent to people, engage their deliberate faculties, and preserve the autonomy of choice. People generally prefer educative nudges over non-educative ones. The example showing statistics that smoking causes cancer is more effective than labeling the cigarette as a warning (Sunstein, 2016).

‘Technocognition’ is an approach proposed by Lewandowsky, Ecker, and Cook (Lewandowsky et al., 2017) that combines insights from cognitive science with interventions in digital architectures to design technological safeguards against the spread of false information or targeted adversarial manipulation. It considers the design context of digital environments through the lens of cognitive science. Technocognition can introduce friction into the process of commenting on or sharing information. For instance, the Norwegian broadcaster NRK launched an experiment where readers had to pass a brief comprehension quiz on an article before posting a comment. This friction was meant to foster deliberate thinking and discourage toxic commenting (Lewandowsky, 2020).

Other examples of technocognition involve combining friction with prompts to engage in analytical thinking. Fazio showed that making people pause and think to explain why a certain headline is true or false can reduce their intention to share false headlines (Fazio, 2020). Introducing reminders about accuracy before sharing can reduce people’s intention to share false headlines also by subtly prompting participants to think about accuracy, the quality of news they chose to share improved. This suggests that when people’s attention is directed towards accuracy, they are less likely to share misinformation (Pennycook et al., 2019).

However, similar techniques can also be used to restrict freedom of choice and communication on the Internet, as seen in authoritarian regimes that use friction (slowing down internet speeds or legal, such as requiring multiple steps of verification to access a site) to limit citizens' access to information (Roberts, 2018). Therefore, it is important to ensure that technocognitive interventions are designed with people's best interests in mind and with public oversight.

'Boosting' is a cognitive intervention that aims to foster lasting and generalizable competencies in users. It targets individual cognitive and motivational competencies rather than immediate behavior (which, as explained, is the target of nudges), intending to empower people to make better decisions in accordance with their own goals and preferences. Boosting interventions can be directed at domain-specific competencies (e.g., understanding health information) and domain-general competencies (e.g., statistical literacy). They can target human cognition (e.g., decision strategies), the environment (e.g., information representation), or both. Unlike nudges, boosts aim to preserve, foster and extend human agency and autonomy. Boosts are transparent because they require an individual's active cooperation.

Kozyreva and her colleagues advocate for 'boosting' as the intervention of choice, arguing that it not only empowers users but also promotes a less paternalistic approach. Furthermore, they assert that it cultivates enduring competencies (Kozyreva et al., 2020). This will also be the main strategy for our educational approach in this thesis. One example of boosting is a risk-literacy boost that can be applied to quickly educate people about relative versus absolute risks in, for instance, the health domain (Gigerenzer et al., 2007). Boosting cognitive competencies online by redesigning the environment might involve changing the way information is presented to users or providing additional cues to existing information to improve the epistemic quality of online content (Lorenz-Spreen et al., 2020).

Inspired by the strategies employed by fact-checkers, researchers have developed straightforward guidelines to enhance abilities in civic online reasoning. This skill set encompasses three main areas: assessing the source of information, evaluating the evidence presented, and engaging in lateral reading. These competencies can be encapsulated in three fundamental questions: (a) Who is responsible for this information? (b) What evidence supports it? (c) What do other sources have to say about it? (Breakstone et al., 2018).

A study by McGrew demonstrated that after receiving two 75-minute lessons focused on the credibility assessment of online sources (an elaboration of the three aforementioned questions), students in the experimental group (n = 29) were over twice as likely to improve their scores in post-testing compared to pre-testing. In contrast, the control group (n = 38) showed no significant difference in their scores. This outcome signifies the success of the intervention. These rules, designed as a rapid boosting intervention, can be conveyed as uncomplicated tips for verifying claims, such as those found in users' social media feeds (McGrew et al., 2019).

Simple decision aids serve as tools that boost digital information literacy, offering straightforward strategies for people to effectively evaluate online information. Acquiring skills in digital information literacy is crucial for navigating the online landscape, encompassing abilities such as obtaining, evaluating, and understanding data (Sparks et al., 2016). While media literacy education has a positive impact, especially on young people's ability to assess online news (Caulfield, 2018), these skills can be difficult to acquire, particularly for seniors. To make this process easier, simple decision aids aim to create easy, automatic habits for evaluating online information. Techniques like lateral reading, commonly used by professional fact-checkers, have proven effective for online evaluation (Wineburg & McGrew, 2019). Other decision aids like Fast and Frugal Trees (FFTs) offer comprehensive guides for real-world decision-making (Hafenbrädl et al., 2016). Inoculation strategies have been shown to bolster cognitive resilience against misinformation (Cook et al., 2017; der Linden et al., 2017), and educational games like *Bad News* which we will describe in detail in the following section could further train individuals in this skill set.

Transitioning from algorithmic solutions, the utilization of a companion system in a daily social media scenario aimed at raising teenagers' awareness of fake content in an online environment offers a more interactive approach. An interactive digital companion is proposed that encourages activities designed to teach individuals skills that aid them in making informed decisions. These skills include selecting, reading, and trusting articles from credible sources over those that merely reflect individual opinions. This approach employs both boosting and nudging strategies, operating on the assumption that people are not inherently 'irrational' but may need nudging towards better decision-making (Ognibene et al., 2023) [7]. Building on this, virtual agents can further enhance this educational framework. These agents offer guided, step-by-step fact-checking, which is particularly useful for young learners. As we shift our focus from behavioral strategies like 'nudging' and 'boosting' to 'technocognition,' it's crucial to explore their integration into educational settings. Learning Companion Systems (LCS) and Virtual Learning Companions (VLC) serve as prime examples of this integration. These intelligent systems, designed to assist and engage learners, can benefit significantly from behavioral insights. For instance, principles of 'nudging' could inform the design of adaptive feedback mechanisms within LCS, gently guiding students toward more effective learning strategies. Similarly, 'technocognition' could offer valuable frameworks for building digital safeguards within VLCs, ensuring that learners engage with accurate and beneficial information. One example could be a VLC giving a prompt asking the user if they are confident about an action before performing it, such as sharing news online.

The embedded form of the companion into the learning environment also has the potential to guide teenagers through various examples in a chatbot-like dialogue. Additionally, learners are provided with access to other instances of the embedded images that are found through Google Reverse Image Search. The idea is that seeing the image in multiple contexts helps learners make more informed decisions about the image's veracity (Aprin, Chounta, et al., 2022) [3].

2.3 Virtual Learning Companion(s) Background

VLC designed with ‘boosting’ principles can be transformative. Such VLCs could include features that teach students how to evaluate online sources, thereby enhancing their critical thinking skills. As we delve into the upcoming sections, we will explore how VLCs have evolved to aid both teaching and learning. We’ll also discuss their integration with other cognitive tools to navigate the complexities of the digital world.

2.3.1 History and Concept of Learning Companion Systems (LCS)

Learning Companion Systems (LCS) constitute a unique subset of Intelligent Tutoring Systems (ITS). These educational agents, assumed to play a non-authoritative role in social learning environments, offer tailored support and adaptive feedback through explicit agents (Chou et al., 2003). By fostering competition and collaboration, learning companions contribute to effective learning dynamics (T.-W. Chan & Baskin, 1990). Incorporating multimedia elements, chatbot dialogues, speech input/output, animation, virtual reality, and other interactive techniques enhances the implementation of LCS.

ITSs distinguish themselves from traditional computerized learning systems by infusing intelligence to elevate the learning experience (Brusilovsky, 2003). Agents that mimic real-life communication cues and cultivate positive relationships with students aim to enhance motivation, thereby achieving improved learning outcomes (Moreno, 2001). The advent of virtual agents is poised to revolutionize learning in multifaceted ways (Clarebout et al., 2002). Research supports the idea that pedagogical agents, especially those with expressive attributes, play a crucial role in assisting students with problem-solving and motivation (Johnson et al., 2000; Lester et al., 1999) Benefits of LCS.

Integrating pedagogical agents in e-learning offers promising prospects for enhancing emotion (Krämer & Bente, 2010). Implementing LCS frequently taps into machine learning and natural language processing techniques, facilitating seamless communication between the system and learners. The process of logging interactions contributes to student modeling. Applying self-explanation techniques, in alignment with self-regulated learning strategies, enhances comprehensive understanding and aids in clarifying misconceptions (Chi et al., 1989). A pivotal challenge in developing LCS lies in their adaptability to learners' responses and activities, requiring contextual responsiveness (Aleven et al., 2013). LCS, in their varied roles, can alternate between challenging ideas as critics or introducing novel concepts as leaders (Goodman et al., 2016).

2.3.2 Virtual Learning Companion (VLC) Overview

A VLC can be defined as a companion that accompanies the student throughout their learning journey, assuming the role of a knowledgeable peer rather than an expert in the subject domain (Pezzullo et al., 2017). Typically, it embodies a computer-generated character that engages in interactive conversations with the student through a chat interface (Chou et al., 2003). Hietala and Niemirepo noted that consistent interaction with a flawlessly knowledgeable companion might decrease learners' motivation. This prompts an inquiry into a learning companion agent's optimal knowledge level, which is necessary to fulfill learner expectations and sustain their motivation for active engagement. Intriguingly, a companion that initially errs, akin to human fallibility, can offer greater advantages when confronting complex tasks or tackling unfamiliar issues (Hietala & Niemirepo, 1998).

The use of VLCs in educational settings is not new; their benefits in enhancing student engagement and motivation have been well-documented (Hsu et al., 2007). Prior research has successfully integrated VLCs into classroom (Shearer, 2018), who utilized it to access and disseminate information (Jin & Liu, 2010), settings, particularly in subjects such as science and mathematics (Wu et al., 2012) and mathematics (Hsu et al., 2007). Beyond the acquisition of cognitive skills, VLCs have also been proven to positively impact affective learning (Woolf et al., 2010) and foster curiosity in learning (Wu et al., 2012). Consequently, it is apparent that the integration of VLCs in education can yield benefits across various domains.

Summary: Addressing the challenges posed by social media requires a multifaceted approach. Enhancing digital literacy, leveraging well-being-aware algorithms, and introducing companion systems can help mitigate risks. Rubin's triangle model suggests interventions in technology, education, and regulation to combat misinformation. Approaches like nudging, boosting, and technocognition offer potential solutions, with boosting being particularly favored for its empowerment and less paternalistic nature. The VLC could serve as a practical educational tool, promoting critical evaluation and a deeper understanding of online content.

2.4 Existing Solutions Using Educational Technologies

In the digital age, combating misinformation is paramount. Educational technologies have emerged as powerful tools to address this challenge. This section delves into various technological solutions designed to educate users about misinformation and promote critical thinking. In the end, we provide a comprehensive comparative analysis of these educational technologies, highlighting their strengths, limitations, and potential for future development.

2.4.1 Co-Inform

Kyza et al. (2021) introduced *Co-Inform*, a toolkit designed to tackle misinformation on the internet and social media platforms. This system encompasses detection, awareness enhancement, fact-checking, correction, and bolstering resistance to misinformation. Utilizing a Chrome extension, it operates on Twitter by merging AI technology with user input. The team investigated how various factors influence users' decisions to 'like' or 'share' misinformation (E. Kyza et al., 2021).

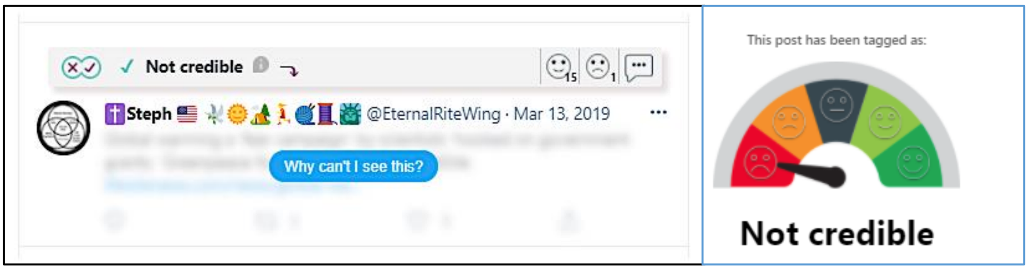

Not credible
15
1
Steph
@EternalRiteWing · Mar 13, 2019
Why can't I see this?
This post has been tagged as:
Not credible

Figure 2-3: On the left, non-credible text is blurred to draw user attention. On the right, there's a credibility scale for users to reference image sources[13].

The extension combats misinformation by using AI and rule-based algorithms to evaluate tweets for credibility. It then displays credibility scores and explanations to users through a blurring feature, as depicted in Figure 2-3.

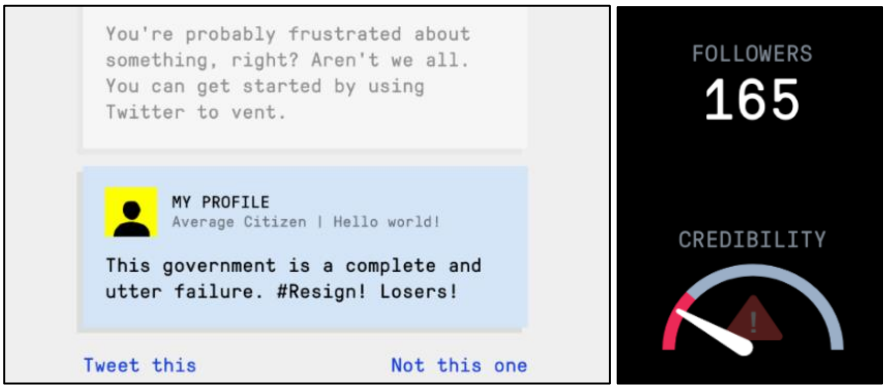
In their study, 80 participants were split into two groups: one with credibility labels (n=40) and one without (n=40). Both groups were exposed to a Twitter feed containing both credible and non-credible posts. The findings highlighted a strong correlation between the absence of the plugin and increased acceptance of misinformation. Those who trusted the plugin were less inclined to disseminate false information. Trust in the

[13] Image from Co-inform, an H2020 project that received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 770302 H2020-EU.3.6. - SOCIETAL CHALLENGES - Europe In A Changing World - Inclusive, Innovative And Reflective Societies.

technology was also linked to its acceptance. The study concluded that co-designed technological solutions could effectively prevent users from disseminating misinformation on genuine social media platforms. Additionally, a dashboard was designed for journalists and policymakers to track and analyze identified misinformation (Zschache, 2022).

2.4.2 Harmony Square and Bad News

Rather than just eliminating fake news, there are interactive tools and games aimed at enhancing users' discernment. Titles such as *Harmony Square*[14] and *Bad News*[15] games, depicted in Figure 2-4, educate players on the prevalent tactics of disseminating misinformation. Within these games, participants encounter various disinformation strategies, and the most successful headlines are those that garner the most followers. Players earn badges as they familiarize themselves with each tactic. An empirical study on *Bad News* revealed that players' proficiency in identifying deceptive methods was notably better than that of a gamified control group. Furthermore, this game not only bolsters individuals' trust in their own judgments but also equips them with a mental defense against common online deceptive maneuvers across diverse cultural backgrounds (Basol et al., 2020).



You're probably frustrated about something, right? Aren't we all. You can get started by using Twitter to vent.

MY PROFILE
Average Citizen | Hello world!

This government is a complete and utter failure. #Resign! Losers!

Figure 2-4: Players adopt the role of a tweeter, employing tactics to attract followers. Techniques include Impression, Emotion Polarization, conspiracy theories, discrediting, and trolling. (Bad News game)[12].

2.4.3 Reality Check

Media Smarts, formerly known as MNet[16], has developed a platform featuring mini-games and a booklet to help teenagers critically analyze digital advertising.

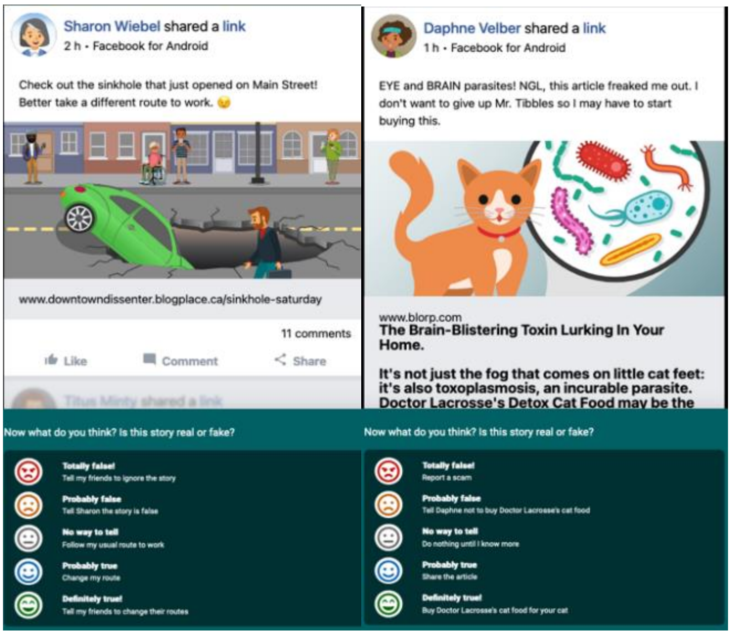
[14] Harmony Square: https://harmonysquare.game/en, Jan 2022

[15] Bad-news-game: https://www.sdmlab.psychol.cam.ac.uk/research/bad-news-game, Jan 2022

[16] Canada's center for digital and media literacy https://mediasmarts.ca/ 22-01-2022

These games, available for free, are tailored to address biases stemming from information deficits, promoting critical thinking. Media Smarts' mission is to enlighten teenagers, urging them to think critically.

The *Reality Check*[17] game by Media Smarts aims to educate users on how to critically analyze online news sources. It provides guidance on verifying evidence, comparing different news outlets, and using fact-checking tools. Our approach aligns with that of the *Reality Check* game, emphasizing the importance of teaching fact-checking skills and encouraging users to compare various sources to identify questionable news and biases (Figure 2-5).



Sharon Wiebel shared a link
2 h · Facebook for Android

Check out the sinkhole that just opened on Main Street! Better take a different route to work. 😉

www.downtowndissenter.blogplace.ca/sinkhole-saturday

11 comments

Titus Minty shared a link

Now what do you think? Is this story real or fake?

Totally false!
Tell my friends to ignore the story

Probably false
Tell Sharon the story is false

No way to tell
Follow my usual route to work

Probably true
Change my route

Definitely true!
Tell my friends to change their routes

Daphne Velber shared a link
1 h · Facebook for Android

EYE and BRAIN parasites! NGL, this article freaked me out. I don't want to give up Mr. Tibbles so I may have to start buying this.

www.blorp.com
The Brain-Blistering Toxin Lurking In Your Home.

It's not just the fog that comes on little cat feet: it's also toxoplasmosis, an incurable parasite. Doctor Lacrosse's Detox Cat Food may be the

Now what do you think? Is this story real or fake?

Totally false!
Report a scam

Probably false
Tell Daphne not to buy Doctor Lacrosse's cat food

No way to tell
Do nothing until I know more

Probably true
Share the article

Definitely true!
Buy Doctor Lacrosse's cat food for your cat

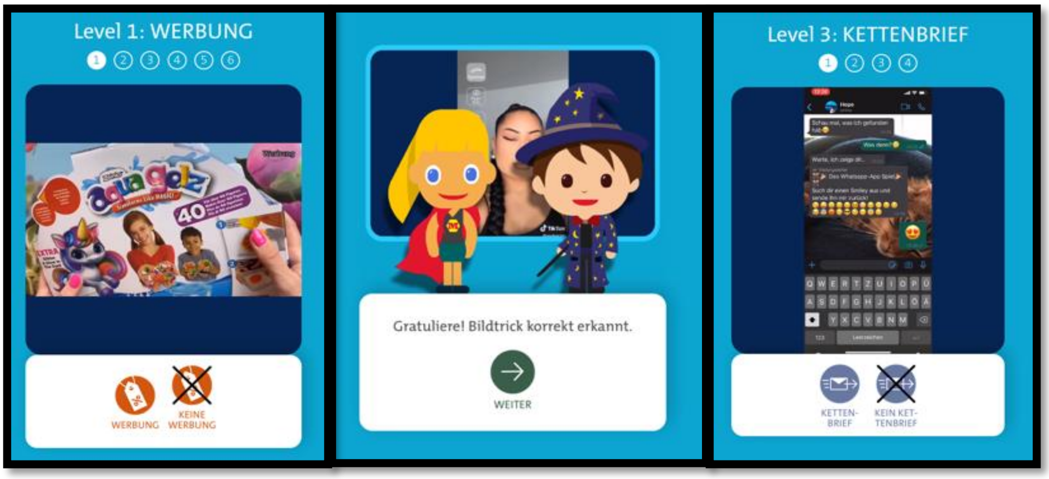
Figure 2-5: In *Reality Check*, the user must categorize the sample news into five levels.

2.4.4 Fake Finder

Gabriel highlighted the initiative by Südwestrundfunk (SWR) that offers a set of four games centered on addressing misinformation (Gabriel, 2021). These games aim to encourage players to delve deeper into specific topics, enhancing their understanding of facts and scenarios. Through these games, participants are trained to assess the reliability of sources and comprehend the dynamics of fake news dissemination. Initial observations from a study conducted at an Austrian university teacher college revealed that participants often struggled to rationalize their choices, with some inadvertently turning to unreliable websites. Despite being adept at digital platforms, there is a clear need to reinforce critical information literacy.

[17] Reality Check: The Game https://mediasmarts.ca/digital-media-literacy/educational-games/reality-check-game, 12-02-2021.

To address this, SWR introduced two distinct online games for children and adults to counter digital misinformation and promote analytical thinking. The children's version focuses on educating them about potential misinformation triggers in social media, such as image manipulations, distinguishing between ads and genuine content, and understanding chain messages. Throughout the game, two virtual avatars as shown in Figure 2-6, representing a boy and a girl, guide the players, presenting scenarios that require discernment and providing feedback on their decisions.



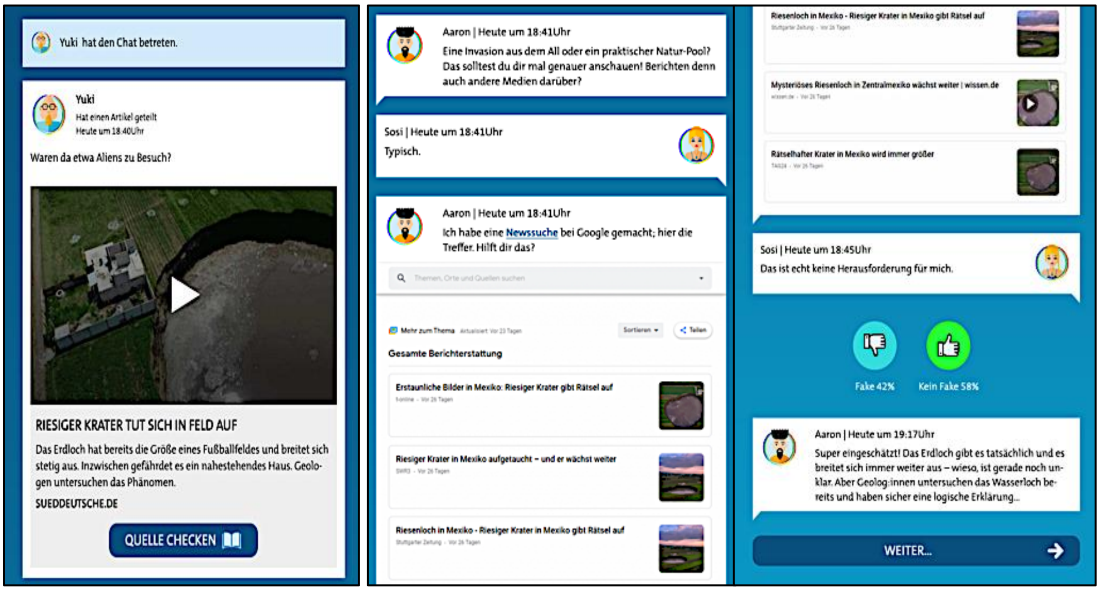
Gratuliere! Bildtrick korrekt erkannt.

Figure 2-6: Two moderators in the coaching role come after the user's answers. In this case, the user could successfully recognize the image trick on TikTok social media.

The mature-oriented version, dubbed *Fakefinder*[18], presents a simulated environment featuring a mix of genuine and fabricated social media posts. Users begin by choosing a snapshot from popular platforms like Twitter or Instagram. The game's interface allows educators to customize the content and duration of the exercise. Participants then select a virtual persona, complete with an avatar and username. As they progress, they're presented with social media posts, as illustrated in Figure 2-7. Players are tasked with determining the authenticity of these posts, often requiring them to undertake independent research. This might involve resource validation, reverse image searches, text quality assessment, and more. Once a decision is made, players are shown how their judgment aligns with others in percentage terms. If players falter, the virtual companion steps in, offering guidance and corrective feedback. The *Fakefinder* game provides teachers with a robust mechanism to evaluate students' comprehension of fake news. As they navigate the game, teachers can actively monitor students' decisions in real-time, observing their interactions with various news items. Post-game, students are tasked with completing a worksheet that includes specific questions, such as "Which news surprised you the most

[18] Fakefinder game https://www.swr.de/unternehmen/medienkompetenz/unterrichtseinheit-fakefinder-100.html, screenshot taken Jan 2023.

and why?" and "What advice would you give younger students about how to deal with fake news?". These questions not only gauge their understanding of the game's content but also their ability to articulate and reflect upon their experiences. To ensure individual performance tracking without compromising privacy, students use pseudonyms during the game, which they later correlate with their real names on the worksheet. The document's provision of expected answers further aids teachers in assessing the depth and accuracy of students' understanding.



Yuki hat den Chat betreten.

Yuki
Hat einen Artikel geteilt
Heute um 18.40Uhr

Waren da etwa Aliens zu Besuch?

RIESIGER KRATER TUT SICH IN FELD AUF

Das Erdloch hat bereits die Größe eines Fußballfeldes und breitet sich stetig aus. Inzwischen gefährdet es ein nahestehendes Haus. Geologen untersuchen das Phänomen.

SUEDDEUTSCHE.DE

Aaron | Heute um 18:41Uhr
Eine Invasion aus dem All oder ein praktischer Natur-Pool? Das solltest du dir mal genauer anschauen! Berichten denn auch andere Medien darüber?

Sosi | Heute um 18:41Uhr
Typisch.

Aaron | Heute um 18:41Uhr
Ich habe eine Newssuche bei Google gemacht; hier die Treffer. Hilft dir das?

Mehr zum Thema Aktualisiert: Vor 23 Tagen

Gesamte Berichterstattung

Erstaunliche Bilder in Mexiko: Riesiger Krater gibt Rätsel auf
t-online · Vor 26 Tagen

Riesiger Krater in Mexiko aufgetaucht – und er wächst weiter
SWR3 · Vor 26 Tagen

Riesenloch in Mexiko - Riesiger Krater in Mexiko gibt Rätsel auf
Stuttgarter Zeitung · Vor 26 Tagen

Mysteriöses Riesenloch in Zentralmexiko wächst weiter | wissen.de
wissen.de · Vor 26 Tagen

Rätselhafter Krater in Mexiko wird immer größer
TAG24 · Vor 26 Tagen

Sosi | Heute um 18:45Uhr
Das ist echt keine Herausforderung für mich.

Fake 42%
Kein Fake 58%

Aaron | Heute um 19:17Uhr
Super eingeschätzt! Das Erdloch gibt es tatsächlich und es breitet sich immer weiter aus – wieso, ist gerade noch unklar. Aber Geolog:innen untersuchen das Wasserloch bereits und haben sicher eine logische Erklärung...

Figure 2-7: The '*Fakefinder Game*' by SWR is the adult iteration of the *'Fake or Not'* game.

2.4.5 Troll Simulator and Wisdom of the Crowd

Gökçe Komaç and Kür¸sat Ça˘gıltay developed a game titled *'Troll Simulator'* to raise awareness about toxic online behaviors, commonly referred to as trolling. This game is a part of the broader research into the use of video games for purposes beyond entertainment, specifically as tools for social and political change. The game focuses on various trolling behaviors such as spamming, flaming, trash-talking, misdirection, and using offensive language. The study employed a single-group pre-test and post-test design to gauge the game's impact on players' perceptions of trolling. Data were collected from (n=111) participants through an online questionnaire. The findings indicated that while there wasn't a significant shift in self-reported knowledge about trolling post-gameplay, participants did exhibit a more negative perception of trolling behaviors after playing. The game successfully represented all five types of trolling in a realistic setting. The game's design is unique, placing the player in the role of the troll, and exposing them to various toxic behaviors. The aim is to provide an unsettling experience, potentially fostering antipathy towards trolling behaviors and raising awareness about the issue. The game's setting is a simulated pre-game lobby of a fictional multiplayer online game, a common place where trolling can initiate. The research underscores the potential of persuasive games in raising awareness and influencing perceptions about real-world issues. The game ‘Troll Simulator’ serves as a testament to the power of games in educating and bringing about social change (Komaç & Ça˘gıltay, 2021).

Lomonaco and his colleagues introduced a game-oriented educational experience aimed at increasing students’ awareness of complex phenomena such as information personalization, social influence, filter bubbles, and echo chambers. This innovative approach was inspired by the *‘wisdom of the crowd’ (WOC)* concept (Lorenz-Spreen et al., 2020). The experiment included questionnaires, a digital media literacy talk, and the WOC educational game that simulated social media’s influence. Participants were shown an image, gave their estimations, received aggregated social feedback, and then provided a second estimation. The aim was to contrast the effects of unbiased versus biased social information. While initial results showed no major shift in students’ perceptions, the full experiment increased their awareness of social media’s influence. However, the number of trials and participant involvement were limited due to COVID-19 restrictions. (Lomonaco et al., 2022).

2.5 Limitation of the tools

The current educational technologies for combatting misinformation, while beneficial, are not without their limitations. One significant shortcoming is their static nature; they often fail to adapt to the rapidly evolving techniques of misinformation. Games and tools like Bad News and Reality Check are designed to educate users through a one-time interaction, which may not be sufficient for the nuanced, continual learning required to navigate the complexity of real-time information flows.

Additionally, these tools tend to utilize a one-step process for classifying information (Such as Reality Check and SWR games), which may oversimplify the critical evaluation process. In reality, discerning misinformation often requires a more iterative approach, where users must repeatedly analyze and reassess information as new evidence comes to light via factchecking tools. Existing solutions may also lack comprehensive mechanisms to simulate the breadth of fake news, relying on simplified models that do not fully capture the intricacies of social media environments.

Moreover, while browser extensions like Co-inform aim to mitigate misinformation, the effectiveness of these solutions can be hindered by their limited integration into users' daily online activities. To enhance their impact, educational tools must be seamlessly integrated into the user experience, ensuring they are both accessible and engaging in the context of everyday online navigation.

3 Judging Social Media Content (SwipeIt Study)

In this chapter, we explore the design, development, objectives, and results of the SwipeIt game, as documented in two academic papers: CSCL [1] and RPTEL [4]. The SwipeIt study was a collaborative endeavor between RIAS and HRW, and it employed both a mobile and desktop application. This app was designed to prompt users to categorize internet images based on criteria such as hate speech.

Before the COVID-19 pandemic, SwipeIt aimed to foster classroom study and discussion, with a particular focus on controversial cases, to stimulate meaningful dialogues among students. Due to the constraints imposed by the pandemic, we adapted the SwipeIt trials to an online format to ensure that educational objectives could still be met in a remote learning environment.

3.1 SwipeIt Game Design and Goals

SwipeIt, an application developed by RIAS, is designed to foster critical thinking among users by prompting them to categorize a wide range of internet images into categories such as hate speech, cyber mobbing, verbal violence, and discrimination. The application serves as a tool for data collection, promoting discussions, and encouraging responsible online behavior. The data collected from school trials are utilized for classroom discussions through a specialized teacher dashboard, as well as for research purposes by RIAS and HRW.

The study presented in the CSCL paper [1] utilizes controversy as a catalyst for learning about the harmful and discriminatory impacts of social media, issues common among the target demographic of middle-school students (Schultze-Krumbholz et al., 2012). The envisaged ‘educational workflow’ for our classroom setting begins with individual students classifying potentially problematic social media content. These individual assessments are recorded in a class repository that informs a teacher’s dashboard, where the entries are sorted and grouped based on their controversy level. This enables the teacher to select examples for group or whole-class discussions, guided by the level of controversy or disagreement. The study explores the feasibility of evaluating disagreement among a group of students by systematically and statistically comparing each student's judgments.

The design of the learning scenario was originally aimed at a secondary high school level, focusing on hate speech and cyberbullying. The goal was to help students develop strategies to counteract these issues, improve social relationships, and understand the social effects of toxic content. The approach builds understanding and resilience rather than avoidance.

In this context, empathy plays a crucial role, in influencing individuals' reactions to others' experiences. It has both a cognitive component (recognizing and understanding others' emotions) and an affective component (experiencing others' emotions). Empathy is linked to the likelihood of displaying discriminatory and hateful behavior, with cyberbullying perpetrators and victims often showing less empathy (Schultze-Krumbholz et al., 2012).

Other socio-emotional competencies, such as social awareness, are also important and were considered in the SwipeIt study. For instance, adolescents with higher social awareness often report being victims of cyberbullying more frequently (Yang et al., 2021) on the other hand, competencies like self-management and responsible decision-making are associated with less frequent reports of cyberbullying victimization.

The concept of social closeness, or familiarity with another person, also influences decision-making. Decisions related to oneself and members of the ingroup are evaluated differently than decisions related to outgroup members (Linke, 2012).

In the context of community, the concept of universal-diverse orientation, which refers to the awareness and acceptance of similarities and differences in others, plays a crucial role in facilitating communication within a diverse social environment (Miville et al., 2004).
Lastly, authoritarianism, a personality trait associated with intolerance and a lack of openness to experience, could pose a challenge for interventions aiming to effect attitudinal change (Nicol & De France, 2016; Vasilopoulos & Lachat, 2018).

Research Questions:

1. How do users characterize the examples, and can these judgments be considered as meaningful and adequate [4]? (rather than arbitrary, implying that they entail mental effort)

2. Is there a correlation between the time taken to answer and the level of disagreement among users [4]?

3. Is there a correlation between the agreement of user ratings and expert tagging and the agreement among participant ratings [1], [4]?

4. Does rating behavior deteriorate as participants progress through items [1]?

5. Given that high levels of empathy and low levels of authoritarianism predict sensitivity towards harm experienced by other people, are these personality traits reflected in the user's tagging behavior [4]?

3.2 Methodology for Calculating Disagreement Measure

Disagreement and agreement measures have applications in various contexts, such as collective statistics, opinions, and social applications. In the SwipeIt game, we aimed to determine the level of disagreement between student judgments as follows: We have a particular set of items (images adorned with text akin to those on Instagram) that are categorized (i.e., tagged) by a specific group of evaluators (students in the learning group) by selecting one of several predefined labels or tags. This indicates that we must compare multiple evaluators who are rating numerous items. The ratings or tags are defined on a nominal scale, that is, they lack a built-in order, eliminating the use of most dispersion measures from descriptive statistics and leaving only a few options. The ‘dispersion index’ (J. T. Walker, 1999) is among these, and we have also identified a measure of disagreement (‘group disagreement’) that was originally devised from the collaboration research perspective (Whitworth, 2007). There is a clear correlation between measures of disagreement (D) and agreement (A). If these measures are normalized on a scale from 0 to 1, the relationship is depicted by the equation D = 1 – A. This hints that well-known measures of agreement, like those used to calculate inter-rater reliability, could be used inversely. Considering we must handle multiple raters and a nominal scale, Fleiss’ kappa (Fleiss, 1971) is a potential candidate. We have looked at details in each formula, including Shannon’s entropy (Shannon, 1948), and compared them in RPTEL paper [4].

Upon a mathematical analysis and comparison of these measures, which can’t be fully detailed here due to constraints in this thesis document, we discovered that Fleiss’ kappa is the A-measure that precisely corresponds to Whitworth’s group disagreement (GD), meaning it equals 1 - GD. As Whitworth (2007) previously noted, the maximum value of GD tends to near (K – 1) / K for a high number of raters, where K represents the number of categories. This cap is smaller when there are fewer categories and equates to 0.5 when K = 2. The dispersion index (DI) employs a different normalization factor that adjusts for this cap in the range of values. Intriguingly, these measures, despite their differing origins, only vary in the normalization factor, and we have opted for DI as a disagreement measure due to its superior scaling properties. The measure is computed as follows:

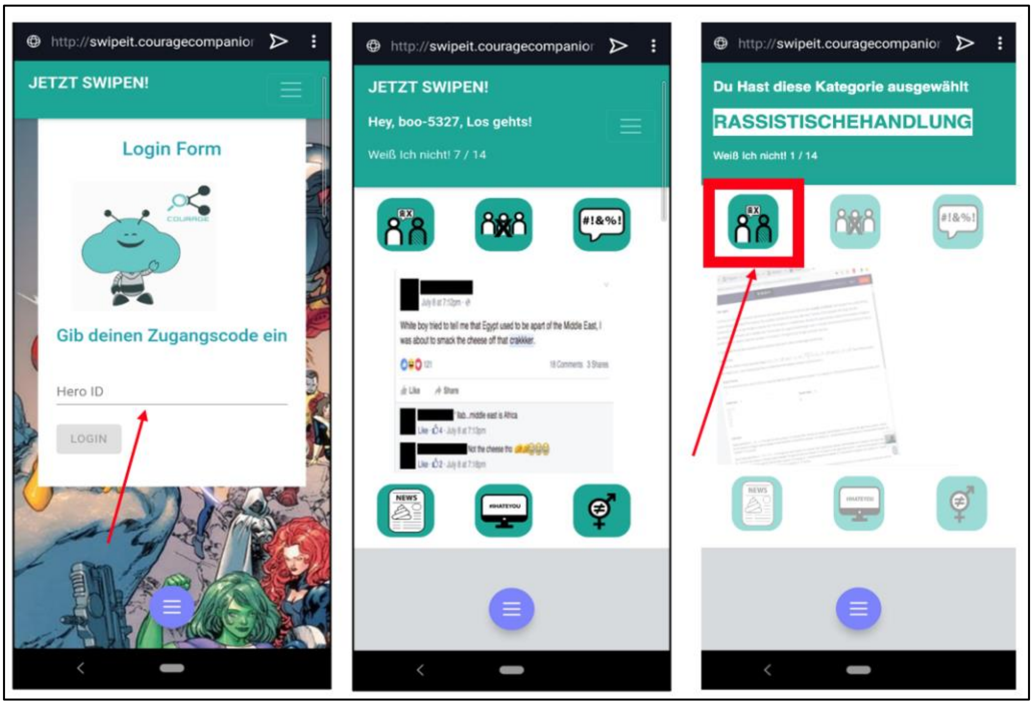
$$\boldsymbol{DI} = \frac{k\left(n^2 - [\,\sum_{k=1}^{k} fk]^2\right)}{n^2\,(k-1)}$$

n: Number of raters
k: Sum of categories
fk: Number of ratings (frequencies) for each category
∑ fk2: Sum of squared frequencies/ratings

3.3 SwipeIt Game Interface

SwipeIt is an application that presents users with a set of 30 images sourced from social media platforms like Facebook and Instagram. These images were independently categorized into four labels: ‘verbal violence’, ‘hate speech’, ‘discrimination’, and ‘cyber mobbing’ by two psychology experts in the first study. The final set of images included six from each category, plus six that did not fit into any category with a ‘none of them’ label. The app displays these images in a fixed order, asking users to select the label that best describes each image. All interactions are stored in a database, with users distinguished by IDs not connected to real names.

The choice of 30 images assumed that a larger number could tire users and affect their concentration. The design of SwipeIt, as shown in Figure 3-1, allows users to see the upcoming content and judge them by classification. This design was used in both studies [1,4]. The research questions were centered around understanding user interactions with the app and their perceptions of the categorized content.


http://swipeit.couragecompanior
JETZT SWIPEN!
Login Form
COURAGE
Gib deinen Zugangscode ein
Hero ID
LOGIN
http://swipeit.couragecompanior
JETZT SWIPEN!
Hey, boo-5327, Los gehts!
Weiß Ich nicht! 7 / 14
#!&%!
NEWS
#IHATEYOU
http://swipeit.couragecompanior
Du Hast diese Kategorie ausgewählt
RASSISTISCHEHANDLUNG
Weiß Ich nicht! 1 / 14

[redacted]
July 8 at 7:12pm ·

White boy tried to tell me that Egypt used to be apart of the Middle East, I was about to smack the cheese off that crakkker.

121 18 Comments 3 Shares

Like Share

[redacted] lab...middle east is Africa
Like · 4 · July 8 at 7:13pm

[redacted] Not the cheese tho
Like · 2 · July 8 at 7:18pm

Figure 3-1: The interface of the SwipeIt Game application. Left: authentication without user-token URL (initial approach); Middle: the initial application interface with initial toxic categories displayed as labels; Right: the interface when the user chooses the ‘Hate Speech’ label and swipes the image to see the next one.

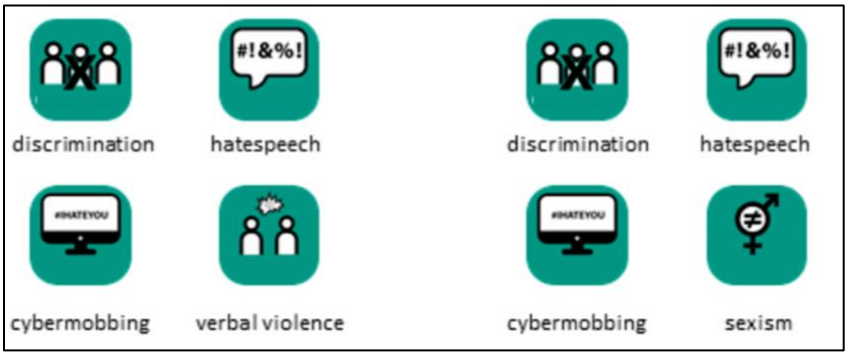


Figure 3-2: Icons that were used in SR1 (left) and SR2 (right) to support the labels.

In the second version of SwipeIt, we selected the last four labels as depicted in Figure 3-2 above, and we bypassed the login page by providing an anonymous user link for each user, randomly delivered by the SociServer[19] questionnaire. Students simply had to click on each user link to access SwipeIt.

In the following Figure 3-3, we illustrate the initial scenario before the COVID situation. The process involved students scanning a QR code in the classroom to choose a favorite animation character and answer demographic questions. These responses were then directed to SociServer, and users were automatically redirected to the SwipeIt application. The teacher would provide an explanation of the game, after which the students would engage in playing the SwipeIt games. After 30 iterations of this item-based classification process for all students, the dashboard would highlight controversial cases characterized by a low level of agreement. These cases would be discussed in the classroom under the guidance of the educators. Following the discussion, users would be guided to answer additional questions related to levels of authoritarianism and empathy.

[19] Sociserver questionnaire platform, https://www.soscisurvey.de/ taken 19 Jul 2023

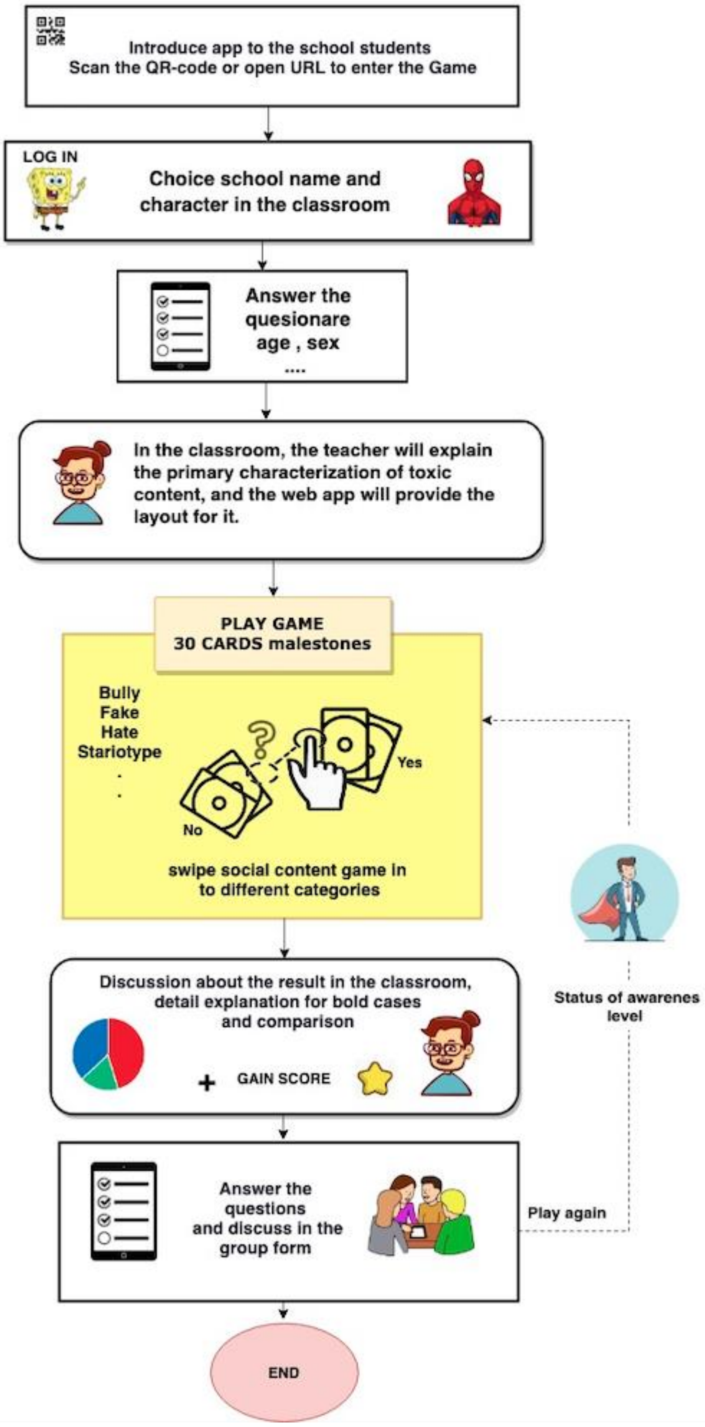
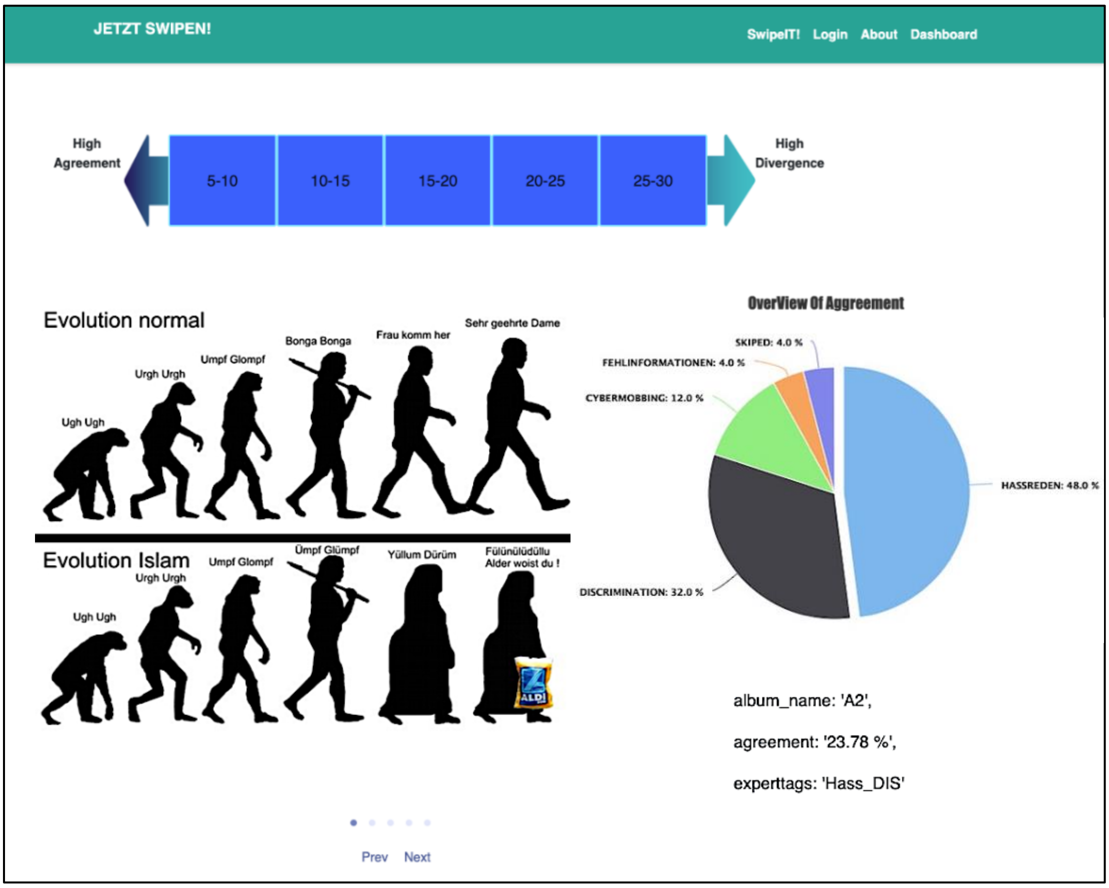

Introduce app to the school students
Scan the QR-code or open URL to enter the Game
LOG IN
Choice school name and character in the classroom
Answer the quesionare age , sex ....
In the classroom, the teacher will explain the primary characterization of toxic content, and the web app will provide the layout for it.
PLAY GAME
30 CARDS malestones
Bully
Fake
Hate
Stariotype
Yes
No
swipe social content game in to different categories
Status of awarenes level
Discussion about the result in the classroom, detail explanation for bold cases and comparison
GAIN SCORE
Answer the questions and discuss in the group form
Play again
END

Figure 3-3: The cognitive process flow of the SwipeIt application, including classroom discussion, as originally planned before the COVID-19 situation.

Teacher Dashboard:

The SwipeIt teacher dashboard was developed to empower moderators, such as teachers, with valuable insights and tools for effective management of the study. From a user perspective, the dashboard offers a comprehensive view of participant interactions and responses, providing a clear understanding of the overall dynamics.

One of the key features of the teacher dashboard is the presentation of data in a visually appealing and easily understandable format as shown in Figure 3-4. For instance, the total number of votes for each image is displayed in a pie chart, allowing moderators to quickly grasp the distribution of labels assigned by the class. This visual representation enhances the user experience by providing a quick overview of the participants' categorizations.



album_name: 'A2',

agreement: '23.78 %',

experttags: 'Hass_DIS'

Figure 3-4: Teacher dashboard allows the moderator (e.g., teachers) to survey the spectrum of all thirty images ranked by degree of controversy (Disagreement) and to inspect the distribution of labels assigned by the class for each image in the pie-chart. On the left is the target picture, and on the right is the analysis based on the labeling. Right-down is the Agreement level in percentage.

Moreover, the teacher dashboard enriches the interface by incorporating a calculation of the agreement level for each item. This agreement level is derived using the dispersion index (DI) based on the participants' rankings and labels. By integrating this calculation into the dashboard, moderators can assess the level of consensus or disagreement among participants for each image, thus facilitating informed discussions and analysis.

In the pre-COVID scenario, one of the primary goals of the SwipeIt teacher dashboard was to facilitate classroom discussions by showcasing the results of other votes, especially those with the lowest agreement levels. By highlighting the images or instances that generated contrasting opinions and had the lowest level of agreement, the dashboard aimed to encourage meaningful discussions among students.

3.4 Technologies in SwipeIt

In the SwipeIt project, the ReactJS[20] library was used for the front-end development, along with additional libraries such as React Router and Material-UI[21] for navigation of different pages and UI design. The front-end was responsible for rendering the user interface and handling user interactions. As shown in Figure 3-5, we chose ReactJS to create SwipeIt as a responsive online application, allowing users easy access without requiring local installations. ReactJS's component-based design enhances user engagement and adaptability across devices. Material-UI integration enables seamless navigation and a visually appealing user interface, thereby elevating the overall experience.

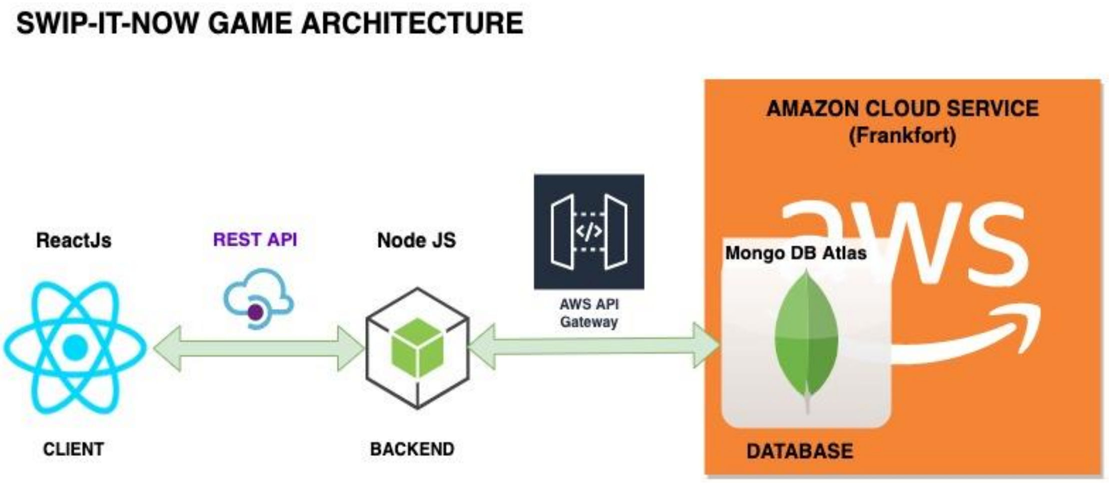


Figure 3-5: SwipeIt Architecture includes a cloud base MongoDB database.

For user authentication, an anonymous user-token was used. This token was obtained from the URL of the user-token link. This approach allowed users to access certain features or data without requiring a full user account.

On the back end, Node.js[22] and Express[23] were utilized. Node.js provided the runtime environment for server-side JavaScript execution, while Express was used as a web application framework to handle HTTP requests and route them to the appropriate endpoints.

[20] The library for web and native user interfaces, https://react.dev/, July 2023

[21] ReactJS UI tools, https://mui.com/, July 2023

[22] Node.js® is an open-source, cross-platform JavaScript runtime env. https://Node.js.org/en, July 2023

[23] Fast, unopinionated, minimalist web framework for Node.js , https://expressjs
.com/, July 2023

For the database, MongoDB[24] was used as the back-end database solution. MongoDB is a NoSQL database that offers flexibility and scalability, making it suitable for storing and retrieving data in the SwipeIt application. In the SwipeIt game, we utilized cloud-based MongoDB on Amazon Web Services (AWS) cloud[25].

3.5 Design Decisions

In our effort to streamline the transition from the Soci-questionnaire to the application, it was imperative to establish a direct link connecting the two platforms while maintaining user privacy and ensuring the traceability of interactions across different systems. This necessity is what informed our decision to develop a web application for SwipeIt that is compatible across various platforms and to implement a unique user-token for each learner. This approach allows for a seamless user experience and meticulous tracking of user engagement without compromising individual identities.

3.6 Data Structure

In the SwipeIt study, MongoDB was used to store data. We collected data based on user labeling. The teacher dashboard is generated based on stored data. MongoDB's data structure included the following fields:

1. **Image ID:** Each image in the study was assigned a unique identifier represented by this field.

2. **Image URL:** Each image in the study was assigned a unique URL that can be represented in SwipeIt Album.

3. **Expert Tag:** This field represents the tag or label assigned to an image by the expert e.g., Image1 has ‘discrimination’.

4. **Student Tags:** This field stores an array of tags assigned by students. Each student’s tag is represented as an object with two properties: the student identifier (e.g., Student 1, Student 2) and the tag assigned to that student.

5. **Album Name:** This field represents the name of the album or collection in which the image is categorized. It is based on the experiment being conducted.

[24] MongoDB database JSON based data structure, https://www.mongodb.com/, July 2023

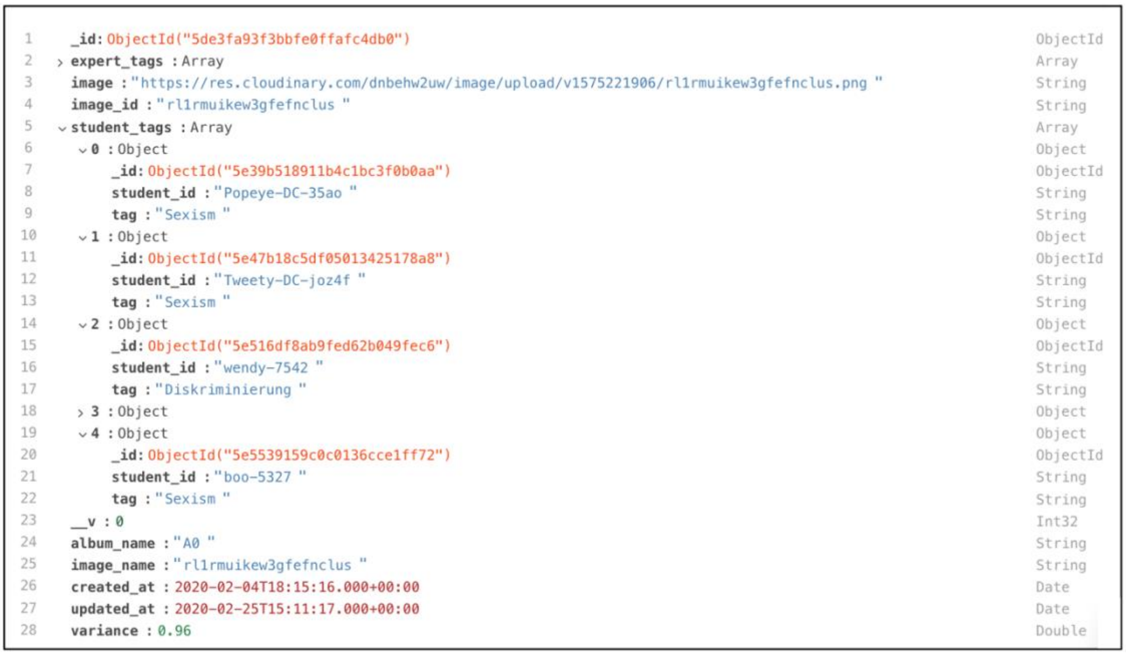
[25] Amazon Web Services (AWS) cloud, https://aws.amazon.com/, Aug 2023

6. **Image Name:** This field stores the name or identifier of the image.

7. **Created At:** This field captures the timestamp indicating when the data entry was created.

8. **Updated At**: This field captures the timestamp indicating when the data entry was last updated.

9. **Dispersion Measure Amount (Variance):** This field represents a measure of dispersion or variability, ranging from 0 to 1. It quantifies the extent of disagreement or variation in the tags assigned to the image by different students.



```
  _id: ObjectId("5de3fa93f3bbfe0ffafc4db0")                                                      ObjectId
> expert_tags : Array                                                                            Array
  image : "https://res.cloudinary.com/dnbehw2uw/image/upload/v1575221906/rl1rmuikew3gfefnclus.png "  String
  image_id : "rl1rmuikew3gfefnclus "                                                              String
⌄ student_tags : Array                                                                           Array
  ⌄ 0 : Object                                                                                   Object
      _id: ObjectId("5e39b518911b4c1bc3f0b0aa")                                                  ObjectId
      student_id : "Popeye-DC-35ao "                                                             String
      tag : "Sexism "                                                                            String
  ⌄ 1 : Object                                                                                   Object
      _id: ObjectId("5e47b18c5df05013425178a8")                                                  ObjectId
      student_id : "Tweety-DC-joz4f "                                                            String
      tag : "Sexism "                                                                            String
  ⌄ 2 : Object                                                                                   Object
      _id: ObjectId("5e516df8ab9fed62b049fec6")                                                  ObjectId
      student_id : "wendy-7542 "                                                                 String
      tag : "Diskriminierung "                                                                   String
  > 3 : Object                                                                                   Object
  ⌄ 4 : Object                                                                                   Object
      _id: ObjectId("5e5539159c0c0136cce1ff72")                                                  ObjectId
      student_id : "boo-5327 "                                                                   String
      tag : "Sexism "                                                                            String
  __v : 0                                                                                        Int32
  album_name : "A0 "                                                                             String
  image_name : "rl1rmuikew3gfefnclus "                                                           String
  created_at : 2020-02-04T18:15:16.000+00:00                                                     Date
  updated_at : 2020-02-25T15:11:17.000+00:00                                                     Date
  variance : 0.96                                                                                Double
```

Figure 3-6: Sample MD document of SwipeIt data structure in JSON format based on image ID.

For the study with ID A0 (Album name ‘0’), the data structure as described above was used details can be found in Figure 3-6. The same set of images was used in SR1 and SR2, representing two different stages or iterations of the study. Please note that the provided information only outlines the general data structure used in the SwipeIt study and does not include specific details or implementation aspects of the MongoDB database setup for the server side.

3.7 SwipeIt Game Trials and Restrictions

We managed to have two studies for SwipeIt. Due to the COVID-19 restrictions, both the first Study Run (SR1) and the second Study Run (SR2) had to adapt their originally planned face-to-face classroom scenarios to an online format. This transition was made to prioritize the safety and well-being of the participants due to the pandemic situation. Instead of in-person interactions, the studies utilized the Moodle learning platform[26] for participant recruitment and as a central hub for study materials and communication.

The studies made use of introductory texts and a video presentation to provide instructions and examples of the SwipeIt game content and functionality. These materials were delivered to the participants through the Moodle platform. To ensure the tracking of participants' answers and maintain consistency across tools, pre-defined tokens were used to link the questionnaire data with the game results.

Recruitment for both studies was conducted online through a call for participation, specifically targeting university students. Participants were informed about the study details and the credits they would receive for their participation, as required by their study program. The data collection periods for SR1 and SR2 were specified as late July to late August 2020 and mid to end of November 2020, respectively Figure 3-7.

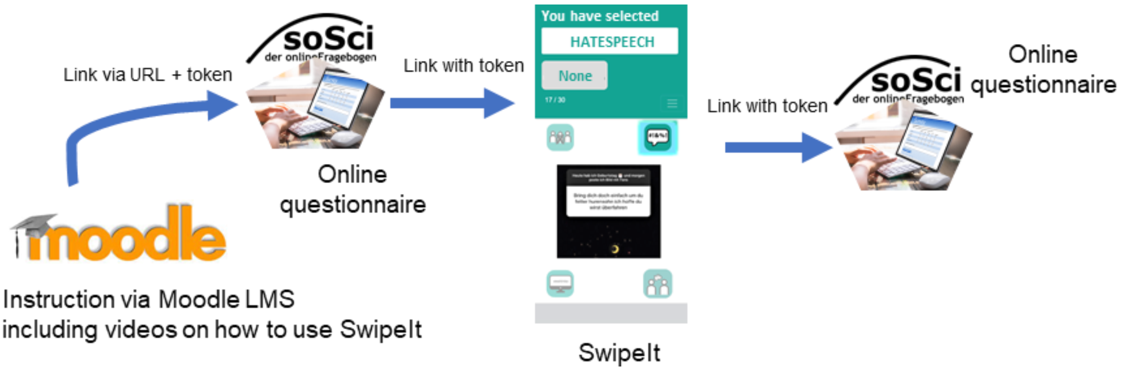


Figure 3-7: The study setting adopted by COVID-19 restrictions discussion in the classroom was changed to the non-collaborative online questionnaire.

SR1 and SR2 are compared in terms of their study settings, participants, measures, procedures, and evaluation results. Here are the key differences between the two studies:

1. **Study Setting:** SR1 was originally planned as a face-to-face classroom scenario with teenagers, but due to COVID-19 restrictions, it had to be adapted to an online scenario with university students. SR2 also took place online with university students.

2. **Labels Used:** In SR1, the labels provided to the participants for tagging the images were discrimination, hate speech, cyberbullying (cyber mobbing), and verbal violence. In SR2, 'verbal violence' was replaced by 'sexism' due to reported difficulties distinguishing hate speech from verbal violence.

[26] Moodle learning platform , https://moodle.org/, June 2023

3. **Participants:** SR1 included 42 university students, mostly in their second semester, with a mean age of 21.76. SR2 had 89 participants, including 72 males and 16 females, with a mean age of 21.83. Both studies recruited participants from a bachelor's program in Human-Computer Interaction.

4. **Measures:** SR1 measured empathy using the Basic Empathy Scale (BES), while SR2 used the Social Emotional Competencies Questionnaire (SECQ). The measurement of authoritarianism was consistent across both studies using the KSA-3 scale. Other measures, such as universal-diverse orientation and social closeness, were also consistent between the studies.

5. **Evaluation Results:** SR1 participants showed a high level of empathy, a universal-diverse orientation, and a tendency towards low levels of authoritarianism. SR2 participants displayed similar traits. However, SR2 showed a negative mood change after playing the SwipeIt game, as indicated by the participants' self-reported emotional states.

3.8 Analysis and Research Questions

The results of the studies indicated that participants demonstrated high levels of socio-emotional competencies, universal orientation, and self-management skills. They also tended to have low authoritarianism and social closeness towards their peers. Moreover, participants exhibited strong self-awareness, responsible decision-making, and relationship management. Introducing the label 'sexism' reduced ambiguity in distinguishing verbal violence from hate speech, leading to increased agreement among raters.

The assessment of the SwipeIt experience indicated that participants typically found both the preliminary questionnaire and the image labeling task to be straightforward. As observed in Figure 3-1 located in the right section pertinent to SwipeIt, participants were able to complete the task with ease, requiring no prior practice or supplementary instructions, thanks to the intuitive nature of the user interface. The images displayed in the app represented typical social media content to model scenarios as close to reality as possible. Additionally, on average, participants tended to agree that images of this type appeared frequently in their own social media channels. Also, most participants felt confident in their ability to navigate the SwipeIt interface. On the downside, there was a perception that the pictures used in the game were somewhat difficult to read (Text size was for some cases small).

Both studies had fewer females than males, with computer science students being the primary participants, who are stereotypically male. All of these findings were statistically significant, indicating that the results are reliable. For 25 out of the 30 examples, the concrete expert ratings were available (5 of the images labeled as “not sure” by the experts). We found a significant correlation between the agreement of user ratings with the expert tagging and the agreement between the participant ratings. This suggests that inter-user agreement can be relied upon even in the absence of an expert ground truth. The Spearman correlation of dispersion indices between SR1 and SR2 indicates a strong positive correlation ($\rho = .71$, $p < .0001$). This indicates that there is consistency in the level of disagreement among users across both studies. The high correlation coefficient suggests that when users disagreed on an image in SR1, another group of users were also likely to disagree on the same image in SR2, and vice versa as shown in images Figure 3-8 and Figure 3-9. This highlights the reliability and consistency of user judgments in characterizing the examples.

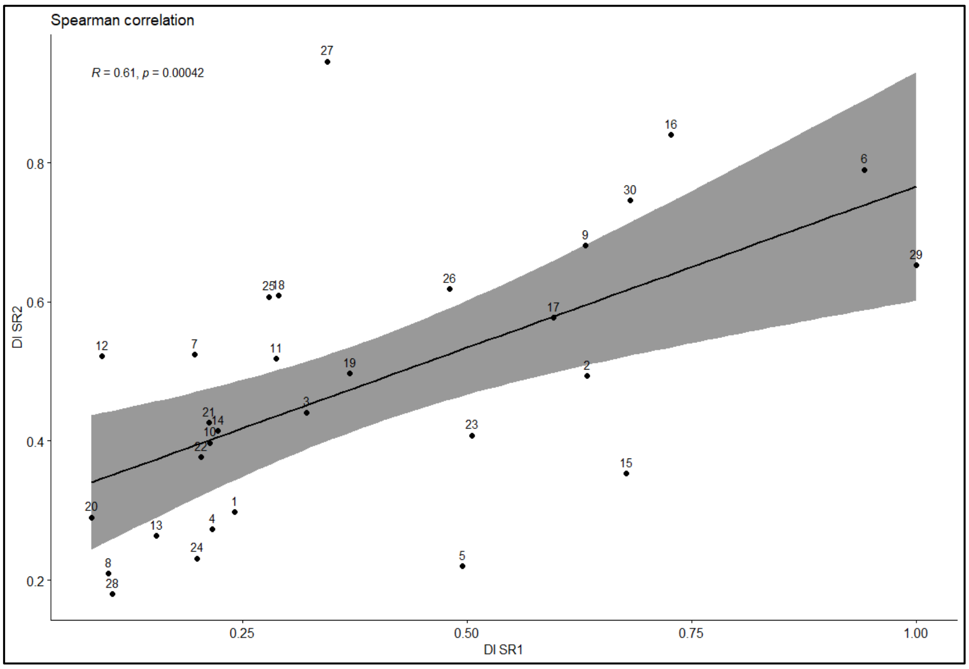


Figure 3-8: Spearman correlation of dispersion indices (DI) between SR1 and SR2

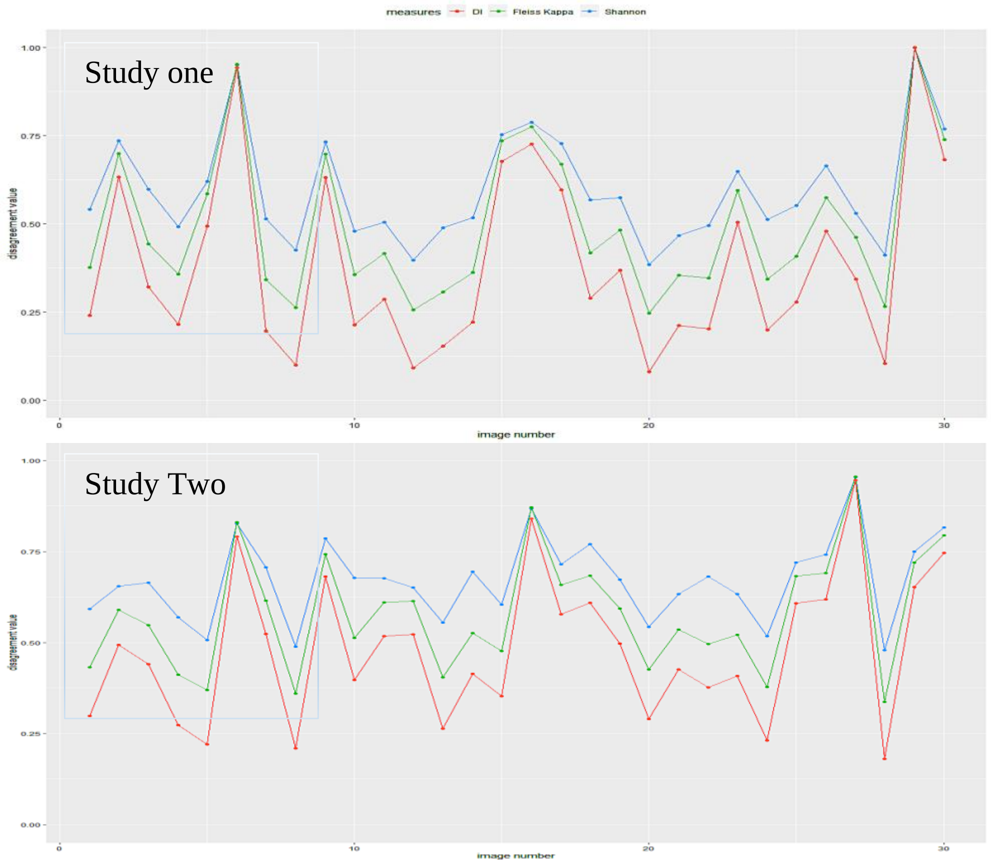


Figure 3-9: the disagreement value of 30 images for both studies. Red line: Dispersion measure, Green line: Fleiss' kappa, blue line: Shannon's entropy

Here are answers to the research questions regarding the analysis:

1: How do users characterize the examples, and can these judgments be considered as meaningful and adequate? (rather than arbitrary, implying that they entail mental effort)

We found that the user judgments were not arbitrary, indicating that the participants made an effort to characterize the examples meaningfully and adequately. This was evidenced by a significant correlation between the agreement of user ratings with expert tagging and the agreement between the participant ratings.

2: Is there a correlation between the time taken to answer and the level of disagreement among users?

We found no significant correlation between answer time and disagreement (r = .14, p = .5), ruling out the option to use answer time as an indicator of controversiality.

3: Is there a correlation between the agreement of user ratings and expert tagging and the agreement among participant ratings?

The analysis indicates a strong concordance between user agreements and expert evaluations, suggesting that when users reach a consensus, it often aligns with expert opinions. This alignment validates the reliability of user agreements as a proxy for expert evaluations in their absence. Furthermore, the study revealed a uniform pattern of user disagreement across various study segments, reinforcing the consistency and dependability of user judgments.

4: Does rating behavior deteriorate as participants progress through items?

We found no significant deterioration of the rating behavior when progressing through the sequence of items, which is a positive message indicating that there is no significant decrease in response time and agreement rates as participants progress through the items.

5: Given that high levels of empathy and low levels of authoritarianism predict sensitivity towards harm experienced by other people, are these personality traits reflected in the user's tagging behavior?

The study explored the potential relationship between personality traits, specifically high levels of empathy and low levels of authoritarianism, and user tagging behavior. The results from qualitative questionnaire confirmed the assumption that these personality traits influence the user's tagging behavior. This suggests that individuals with higher empathy and lower authoritarianism tend to exhibit greater sensitivity towards harm experienced by others, influencing their tagging behavior accordingly.

3.9 Summary

The SwipeIt study illuminated the potential of a mobile application in nurturing critical thinking and responsible online conduct among students. The results showcased participants' high socio-emotional competencies, self-management skills, and sensitivity toward harm experienced by others, reflected in their tagging behavior. The study also revealed the consistency and reliability of user judgments, even in the absence of expert ground truth, and the influence of introducing specific labels like 'sexism' in reducing ambiguity as the average of the agreement level increased. Despite the challenges brought about by the COVID-19 pandemic, the study's success in meeting its objectives lays the groundwork for future research, emphasizing the importance of empathy and awareness in online interactions.

The SwipeIt Game demonstrated essential requirements and experiences for developing web applications, which were informed by the cited studies involving learners. It detailed methods of logging data, structuring learner information across various collections, and applying agreement metrics and other measures for analyses. These methodologies were instrumental for subsequent developments in virtual learning companion applications and scenarios which will be describe in the next chapters.

4 Approach

In this chapter, we will explain the conceptual architecture of our Virtual Learning Companion (VLC) system. This evolved from December 2020 to June 2023 to meet the requirements of each scenario. Subsequently, we will delve into the technical background and each detailed module.

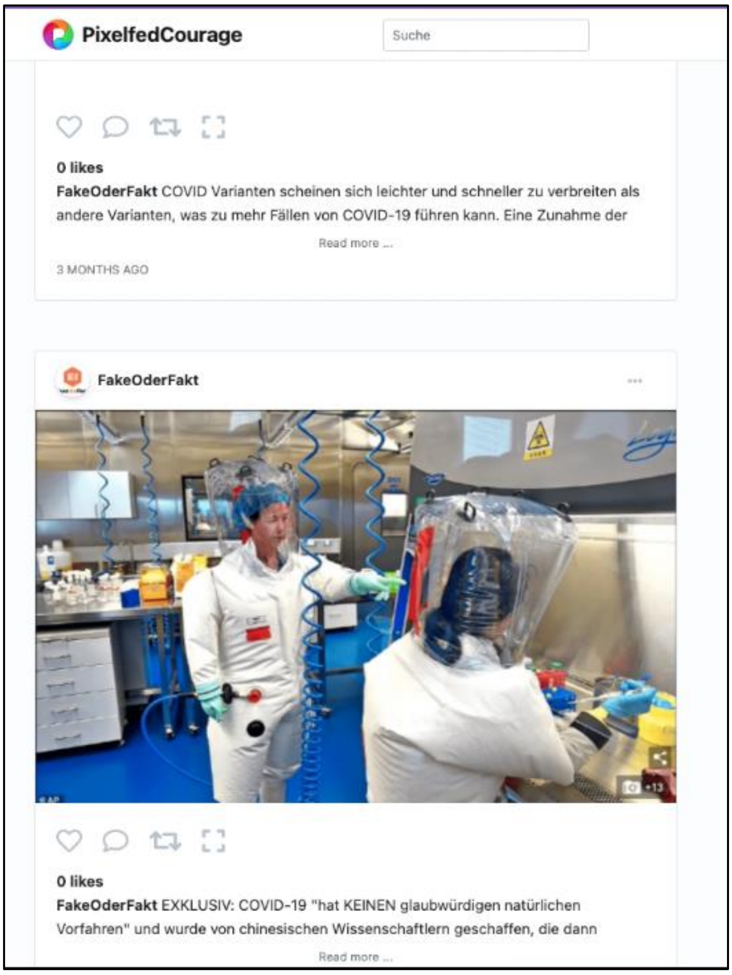
4.1 Courage VLC Design

The initial work involved implementing the VLC in social media environments. The architecture is designed to be applicable in real social media, open environments, such as Instagram, and Facebook. However, there was a need to implement a closed version of social media (a controlled closed environment) with preselected stimuli. This had two reasons: First, to monitor user interactions within a limited timeframe for specific stimuli and cases. Second, was the necessity of controlling exposition of the target group to possibly toxic content in free and open environment for reasons of pedagogical responsibility.

within a simulated social media environment, similar to a basic form of Instagram, known as PixelFed[27]. PixelFed is a social network service for sharing images that is free and open source. It operates in a decentralized manner, ensuring that user data isn't stored on a central server, setting it apart from other platforms. This controlled environment used in the proof-of-concept scenario (Aprin et al., 2021) [2], enables participants to engage in interactions by sharing, commenting, and expressing content approval. Considering the specific context, slight adjustments were made to the platform to facilitate the inclusion of educational materials.

After conducting the initial scenario and testing with limited users, we identified certain limitations and complexities in PixelFed based on our user feedback. For instance, unlike Instagram, where users can view comments without clicking on a separate button, the version of PixelFed used in June 2021 required users to click on the comments button to access others' comments (Figure 4-1). Also, as in some scenarios, we did not need the collaborating feature of the closed environment, such as commenting and liking interaction with the profile page and other pages, to reduce complexity.

[27] PixelFed, https://PixelFed.org/, 12 Aug 2023



0 likes

FakeOderFakt COVID Varianten scheinen sich leichter und schneller zu verbreiten als andere Varianten, was zu mehr Fällen von COVID-19 führen kann. Eine Zunahme der

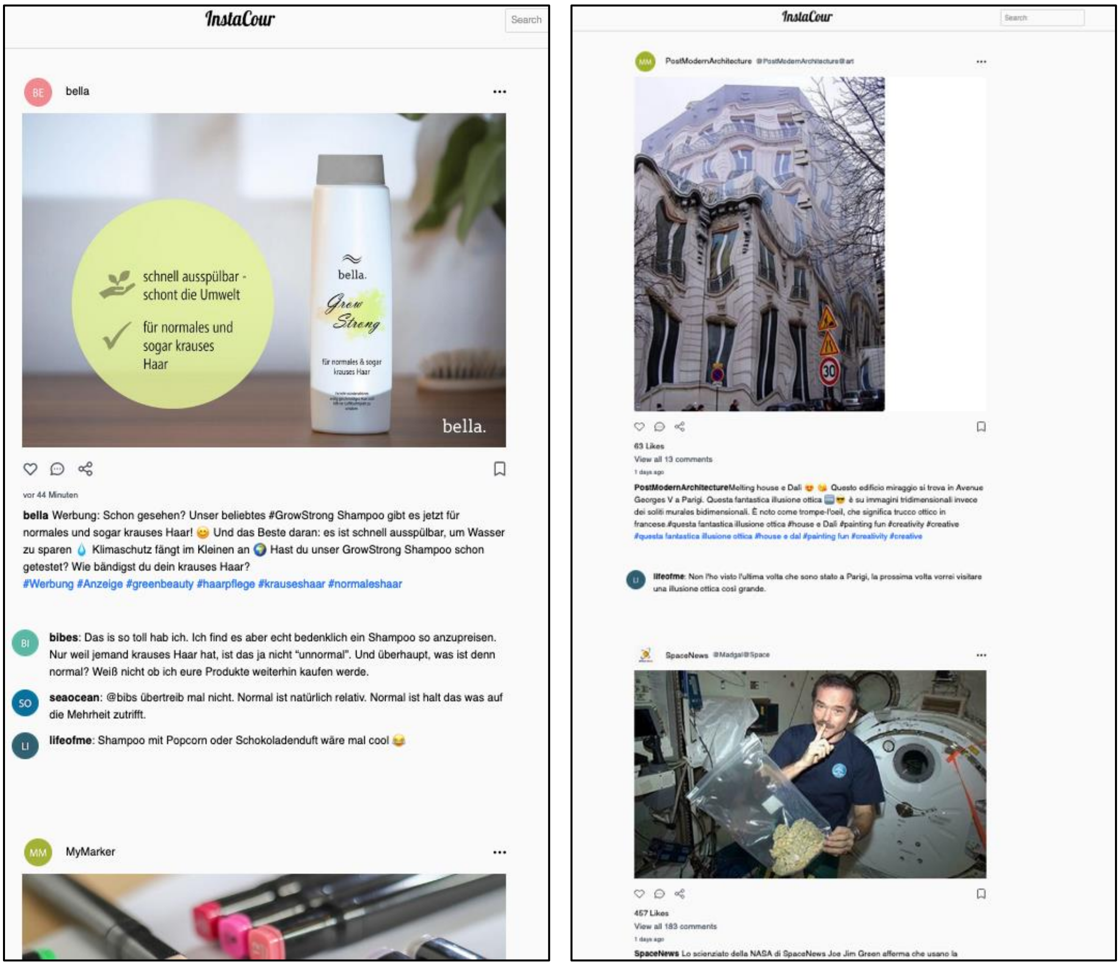
0 likes

FakeOderFakt EXKLUSIV: COVID-19 "hat KEINEN glaubwürdigen natürlichen Vorfahren" und wurde von chinesischen Wissenschaftlern geschaffen, die dann

Figure 4-1: The PixelFed environment was used in the 'Fake or Fact' Scenario 2020-2021 as a control environment.

Additionally, we observed that features like authentication and the home button were unnecessary in our testing scenarios. Furthermore, as Instagram updated, there was a noticeable difference in the user interface between PixelFed and the current Instagram version.

Due to the theoretical reasons mentioned, as well as technical considerations that we will explain in the following chapter, we developed InstaCour—a simplified mimic of Instagram—as an alternative to meet the needs of our designed scenario. This controlled social network environment served as the platform for our subsequent trials in June 2022 and June 2023. It facilitated the controlled distribution of content, including randomizing images for each user and their captions. It also enabled fundamental social media interactions. This environment was designed to handle predetermined images, custom images for each user, and apply captions and comments. Administrators and researchers could introduce new Instagram-like content as needed for their respective scenarios.



InstaCour
Search
schnell ausspülbar - schont die Umwelt
für normales und sogar krauses Haar
bella.
Grow Strong
für normales & sogar krauses Haar
bella.

bella

vor 44 Minuten

bella Werbung: Schon gesehen? Unser beliebtes #GrowStrong Shampoo gibt es jetzt für normales und sogar krauses Haar! 😊 Und das Beste daran: es ist schnell ausspülbar, um Wasser zu sparen 💧 Klimaschutz fängt im Kleinen an 🌍 Hast du unser GrowStrong Shampoo schon getestet? Wie bändigst du dein krauses Haar?
#Werbung #Anzeige #greenbeauty #haarpflege #krauseshaar #normaleshaar

bibes: Das is so toll hab ich. Ich find es aber echt bedenklich ein Shampoo so anzupreisen. Nur weil jemand krauses Haar hat, ist das ja nicht "unnormal". Und überhaupt, was ist denn normal? Weiß nicht ob ich eure Produkte weiterhin kaufen werde.

seaocean: @bibs übertreib mal nicht. Normal ist natürlich relativ. Normal ist halt das was auf die Mehrheit zutrifft.

lifeofme: Shampoo mit Popcorn oder Schokoladenduft wäre mal cool 😂

MyMarker

PostModernArchitecture @PostModernArchitecture@art

63 Likes
View all 13 comments
1 days ago

PostModernArchitectureMelting house e Dalì 😍 🤩 Questo edificio miraggio si trova in Avenue Georges V a Parigi. Questa fantastica illusione ottica 🆒😎 è su immagini tridimensionali invece dei soliti murales bidimensionali. È noto come trompe-l'oeil, che significa trucco ottico in francese.#questa fantastica illusione ottica #house e Dalì #painting fun #creativity #creative
#questa fantastica illusione ottica #house e dal #painting fun #creativity #creative

lifeofme: Non l'ho visto l'ultima volta che sono stato a Parigi, la prossima volta vorrei visitare una illusione ottica così grande.

SpaceNews @Madgal@Space

457 Likes
View all 183 comments
1 days ago

SpaceNews Lo scienziato della NASA di SpaceNews Joe Jim Green afferma che usano la

Figure 4-2: left: InstaCour for racism scenario school trial September 2022 - Right: InstaCour for 'Fake or Real' scenario June 2022 Germany / June 2023 Italy.

In the subsequent 'Fake or Real' scenario [5] conducted in June 2022 and 2023, we intentionally excluded the development of features such as 'like,' 'share,' and 'comment' buttons. This omission aligned with the scenario's objective, which focused on allowing trial users to compose comments capable of influencing others' judgments (Figure 4-2).

4.1.1 Conceptual Architecture

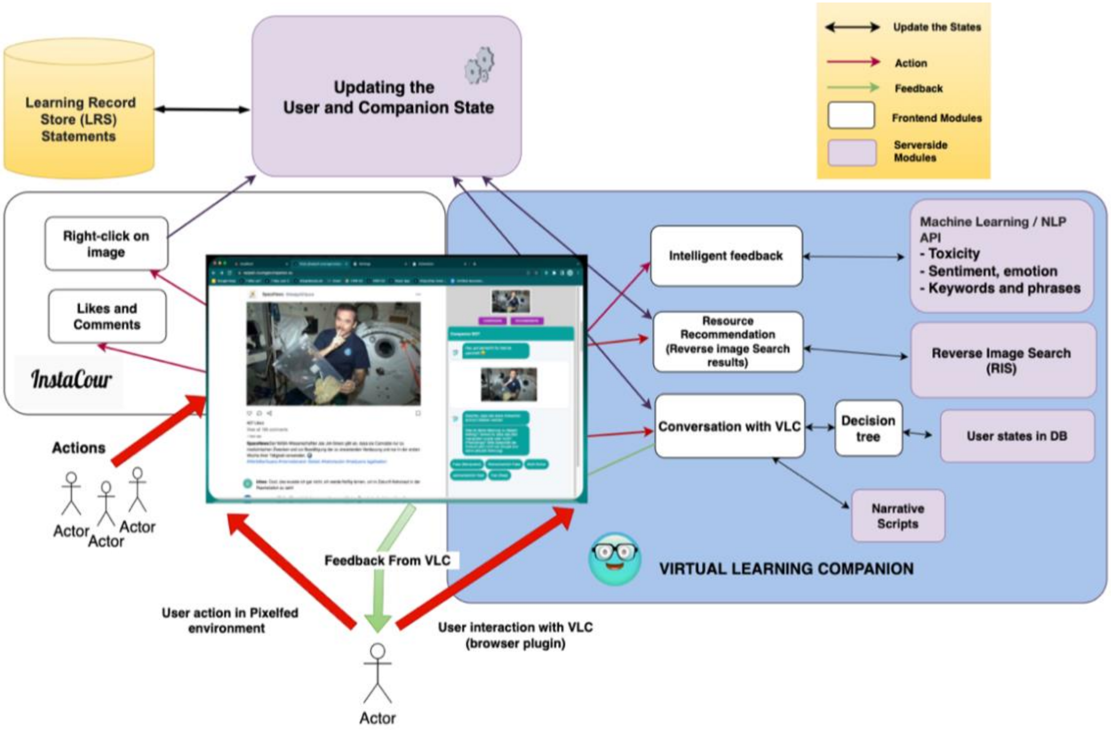


Figure 4-3: Architectural design of the Virtual Learning Companion (VLC) system and the InstaCour setting in a conceptual framework (2022) an updated version of the diagram [6].

The conceptual architecture presents an overview of the VLC system and the InstaCour environment (Figure 4-3). In a single browser tab, learners access both the VLC interface and InstaCour as interactive elements. The VLC, functioning as a Chrome browser plugin, collaborates with InstaCour's elements. It includes a chatbot that responds to user inputs. Through the chatbot, the VLC system moderates interactions, providing queries and suggestions, allowing learners to engage with the content in the environment.

The system maintenance and development are facilitated by maintaining independence between the Chrome extension and InstaCour, enhancing modularity. Moreover, the Chrome extension offers the advantage of monitoring user browsing behavior, tracking tab interactions, unique tab names, time spent on each tab, and more. The Chrome extension's Background.js handles the information relay, communicating directly with the middleware through the browser via REST API[28] endpoints.

Data structures conforming to the Standard Experience API (xAPI)[29] statement format were successfully developed to record chatbot dialogues, user interactions with VLC and InstaCour, as well as interactions with other websites. These records are simultaneously

[28] RESTful API, https://www.redhat.com/en/topics/api/what-is-a-rest-api, July 2023

[29] What is standard learning xAPI statements, https://xapi.com/overview/, March 2023

stored in both Learning Locker and MongoDB. However, it is important to note that using the Chrome extension has limitations, such as constraining the trial environment to the Chrome browser and necessitating extension setup and configuration on each PC before school trials (Aprin, Chounta, et al., 2022) [3].

For larger-scale trials, such as those involving more than 1,000 participants, the VLC architecture needs to be adapted and embedded directly into the environment, rather than relying on a Chrome Extension. This is because installation requires specific prerequisites like time and lab facilities. At the very least, a tutorial on how to install the Chrome browser and the VLC plugin. Despite these challenges, the Chrome Extension approach has the advantage of being applicable to real-life scenarios. It can read the content of applications like Instagram and Facebook and deliver it to the companion system.

4.2 Implementation

In this section, we delve into the practical implementation, focusing on two vital elements: the front-end and back-end of the VLC system, including how we seamlessly embedded the companion's front-end into the Chrome extension plugin. We also explore our data storage intricacies, detailing data models in MongoDB and the architecture of the Learning Locker for statement storage. This in-depth exploration offers a practical understanding of our system's effective implementation, as well as advancing our study's objectives.

4.2.1 Technical Architecture

The companion system is divided into two primary technical components. On the front-end, a Chrome extension interfaces with the InstaCour environment as two independent components. Meanwhile, as depicted in Figure 4-4 on the back-end, a combination of internal and cloud-based microservices[1] communicate with middleware (Express) which listens to the data from the plugin's Background.js. This middleware analyzes and integrates different APIs. This communication occurs via a REST API (Aprin, Chounta, et al., 2022) [3].

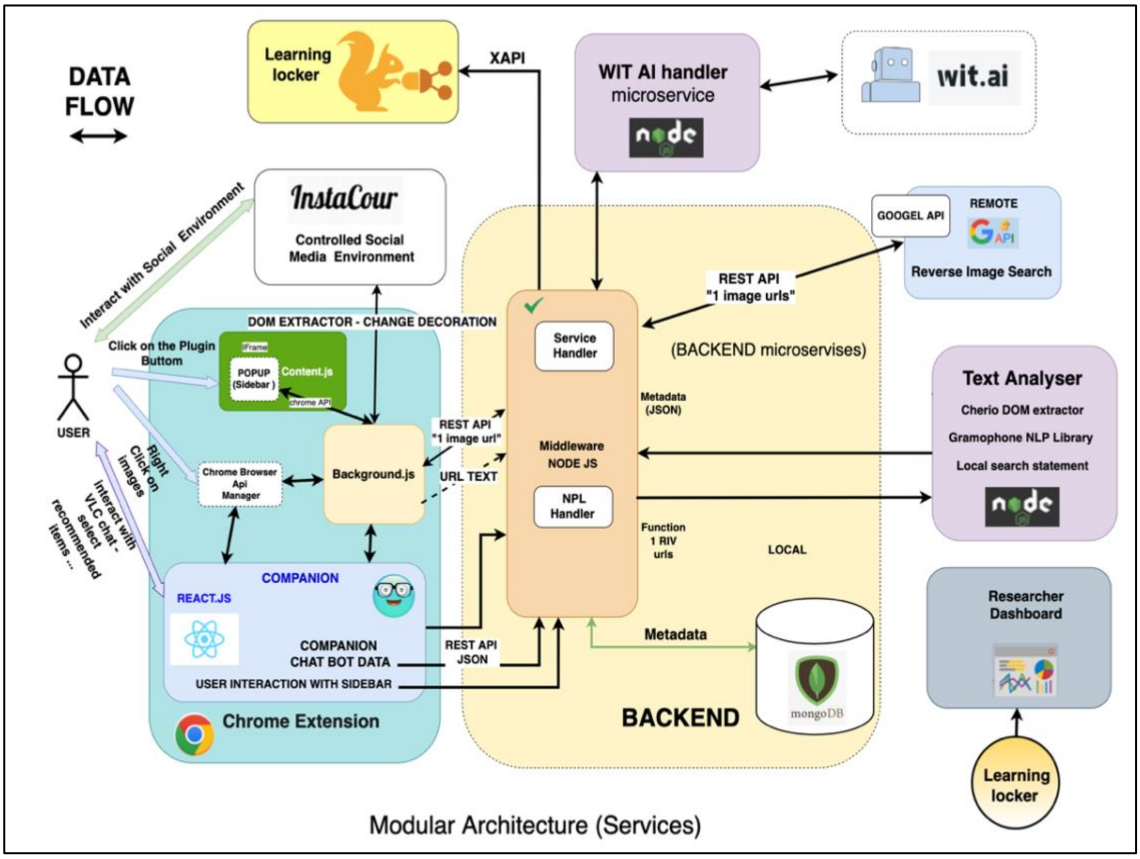


Figure 4-4: Navigating the technical landscape: a visual representation of our system's architectural framework, blending virtual learning companion and InstaCour environment.

Utilizing a cloud-based Wit.ai[30] service, the chatbot's interaction is extended with AI-powered functionality that interprets the user's intent in free-text dialogues, guided by a trained model based on manual entry by the admin. Notably, this feature was implemented and tested but not utilized in the present chatbot experiment as in our scenarios flexible answers were not considered.

InstaCour manages content such as image links and captions (stimulus) using a static JSON-formatted file. Additionally, a function was developed to randomize content upon user refresh or when InstaCour is accessed. The middleware manages tabs and their time slots, filters tabs based on analysis goals, and stores data (e.g., unique tab information in a JSON structure) in MongoDB collections.

Metadata, user credentials, chat history, and models are stored in MongoDB's document-oriented database. The Reverse Image Search (RIS) module communicates with Google's API and stores relevant RIS links which contain details information about similar selected image. The local text analysis module accesses the RIS list, scrapes text content from

[30] WIT-AI is a tool to Build Natural Language Experiences https://wit.ai, Nov 2022

remote websites, and analyzes it. The output includes significant keyword sets, encompassing predefined keywords (e.g., fake, fact, credible, evidence) and keywords derived from the TF-IDF (Term Frequency-Inverse Document Frequency)[31] algorithm. Learning actions are recorded in the xAPI format and stored in the Learning Locker[32] as the Learning Record Store (LRS)[33].

Now that we understand the overarching technical framework, let's examine the specific components starting with the front-end and the Chrome plugin architecture.

4.2.2 Front-end Details and Chrome Plugin Architecture

The architecture of our Chrome extension encapsulates various integral components, each contributing to the extension's robust functionality. Among these, Background.js Scripts/Event' pages play a pivotal role. Functioning discreetly in the background, they skillfully manage the extension's behavior, even when popup or content scripts are not actively executed. This orchestration encompasses event handling, timers, and state management.

In our implementation, we adopted a tailored approach where we embedded the VLC front-end using the ReactJS library. Once compiled, the VLC front-end is seamlessly integrated into the Chrome browser through an *I-frame* in the Content.js file. Notably, Background.js operates as a watchful sentinel, monitoring activities such as user interactions with components like images and Chrome tabs. These data points are efficiently transmitted to the back-end server via a REST API, ensuring real-time communication and data synchronization.

By harnessing the command `chrome.contextMenus.onClicked.addListener` within the Background.js of the Chrome extension, we can accurately extract the URLs of images selected within the InstaCour environment.

In the Content.js of the Chrome plugin, the ReactJS application is embedded within an *iframe*, this integration enables the dynamic rendering of content. Through the `chrome.runtime.onMessage.addListener` function, messages are received from the extension's background script, triggering actions such as toggling the I-frame's visibility. By toggling the I-frame's width, we effectively control its display, optimizing the user experience.

[31] Term Frequency - Inverse Document Frequency, https://rb.gy/5sltmx, Feb 2023
[32] Learning Locker tool, https://learninglocker.atlassian.net/wiki/spaces/DOCS/overview, July 2023
[33] Learning Record Store (LRS), https://xapi.com/learning-record-store/, July 2023

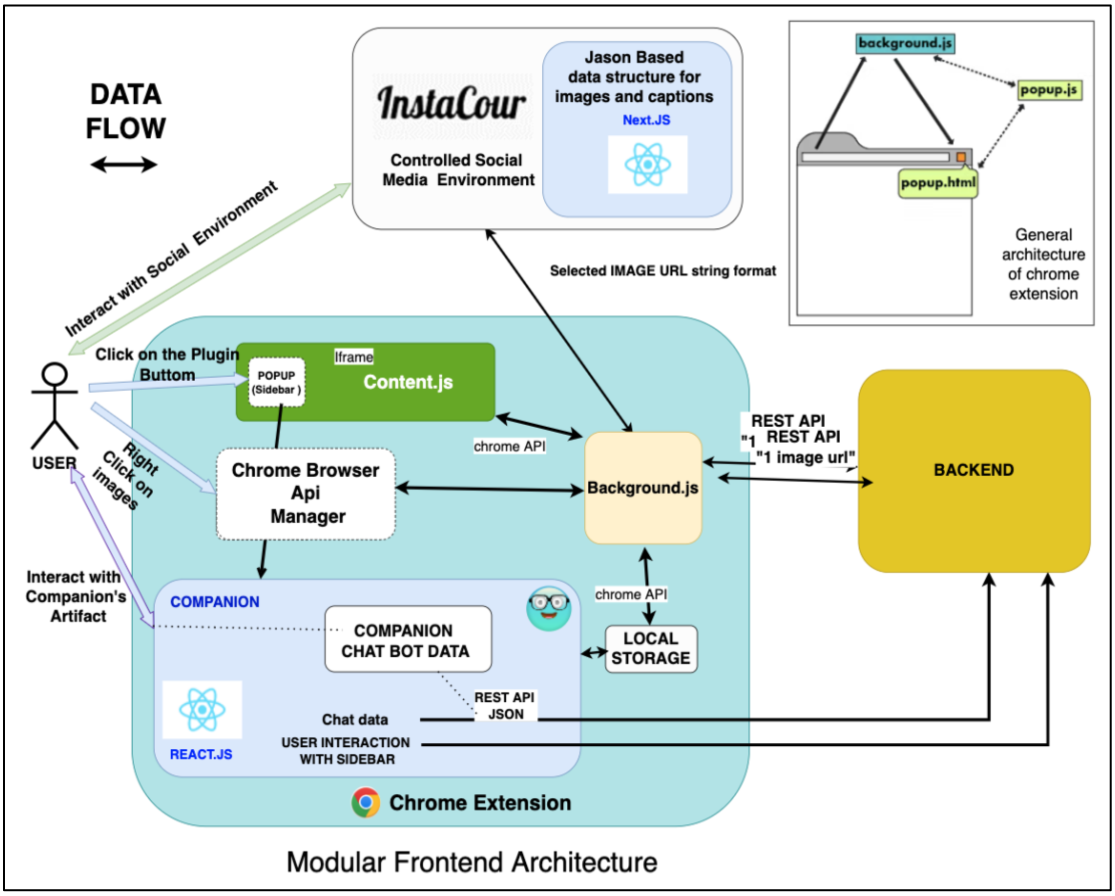

DATA FLOW
InstaCour
Controlled Social Media Environment
Jason Based data structure for images and captions
Next.JS
background.js
popup.js
popup.html
General architecture of chrome extension
Interact with Social Environment
Selected IMAGE URL string format
Click on the Plugin Buttom
USER
POPUP (Sidebar)
Iframe
Content.js
Right Click on images
chrome API
Chrome Browser Api Manager
Background.js
REST API
"1 REST API
"1 image url"
BACKEND
Interact with Companion's Artifact
chrome API
COMPANION
COMPANION CHAT BOT DATA
LOCAL STORAGE
REST API JSON
Chat data
USER INTERACTION WITH SIDEBAR
REACT.JS
Chrome Extension
Modular Frontend Architecture

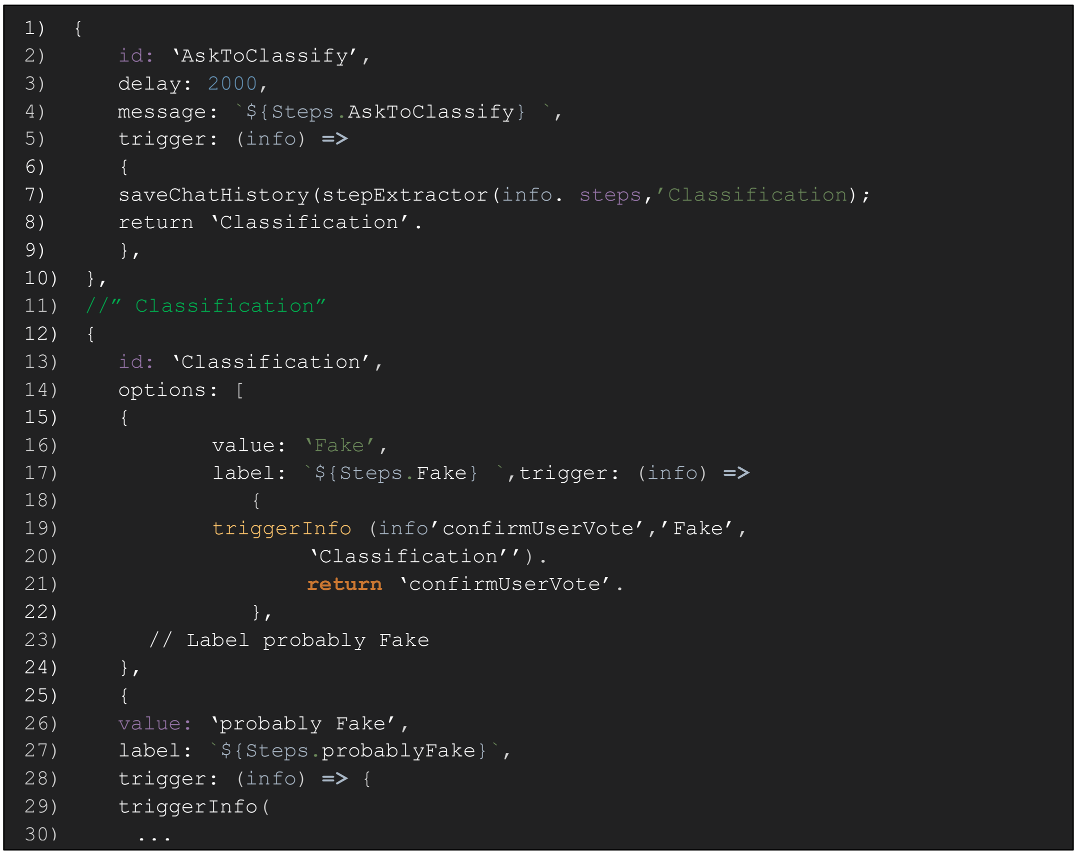
Figure 4-5: Illustrating Data Flow: InstaCour, Chrome Extension, and Embedded React App Interaction.

The App.js file of the VLC front-end showcases its core functionalities. It manages various components, including '*Analysis*', '*Recommended*', and '*CompanionCore*[34]'. The front-end interacts with the server-side using REST API to obtain crucial data, facilitating informed user interactions (Figure 4-5).

The *'Analysis component'* is responsible for handling metadata generated by the back-end, such as keywords and important sentences. *Recommended component:* Responsible for fetching the RIS link that is generated and requested in back-end microservices, and displaying it as a list for users. In the '*CompanionCore*' component we handled chat dialogues based on predefined scenarios.

Here we are referring to technical details to describe how 55 predefined dialogues are implemented within a dynamic chat-dialogue. In this example, the companion asks the learner to classify the image according to whether it is fake (manipulated) or other categories:

[34] Companions Core component, https://github.com/Farbod29/VLC-OpenAccess/blob/master/VLC-Front-end-REACT/src/container/Companion/CompanionCore.js, Aug 2023



```
1)  {
2)      id: ‘AskToClassify’,
3)      delay: 2000,
4)      message: `${Steps.AskToClassify} `,
5)      trigger: (info) =>
6)      {
7)      saveChatHistory(stepExtractor(info. steps,’Classification);
8)      return ‘Classification’.
9)      },
10)  },
11)  //” Classification”
12)  {
13)     id: ‘Classification’,
14)     options: [
15)     {
16)             value: ‘Fake’,
17)             label: `${Steps.Fake} `,trigger: (info) =>
18)                {
19)             triggerInfo (info’confirmUserVote’,’Fake’,
20)                     ‘Classification’’).
21)                     return ‘confirmUserVote’.
22)                },
23)        // Label probably Fake
24)     },
25)     {
26)     value: ‘probably Fake’,
27)     label: `${Steps.probablyFake}`,
28)     trigger: (info) => {
29)     triggerInfo(
30)       ...
```

Figure 4-6: code snippet example from the core element of the chatbot from ‘CompanionCore.js’.

In the above code snippet from Figure 4-6, the ‘id’ in line 2 refers to the step-ID of the chatbot conversation, which in this case is ‘AskToClassify’. In line 3, a delay time of 2000 milliseconds is applied to the chatbot, representing the thinking time of the VLC to make the chatbot appear more like a natural agent. The delays in dynamic response not only enhance users’ sense of the chatbot’s humanness and social presence but also contribute to a higher level of satisfaction with the overall interaction with the chatbot (Gnewuch et al., 2018). We calculate the delay dynamically based on the length of chatbot responses.

The companion step message is returned in a predefined JSON file on line 4 which was written by the researchers. Here is the key also ‘AskToClassify’. On line 19, the ‘triggerInfo’ function is called to send the data to the related function. This then sends it back to the back-end as a log to be stored in the user’s history. The first parameter of ‘trigger vote,’ which is also returned in the function, is the key to the next step in the companion’s response to the user, in this case is ‘confirmUserVote’. Line 21 includes a check to see if the user has voted for ‘Fake’, and line 26 performs a similar check to see if the user has chosen the image as ‘probably Fake’, with the procedure continuing as explained before.

User interactions generate data including chat logs and other engagement-related information, securely transmitted to the back-end server via a REST API. This streamlined communication ensures that the user experience is seamlessly synchronized with the back-end system, enhancing the extension's efficacy.

In this architecture, the VLC Chrome extension integrates packages from the ReactJS module. This combination enables the dynamic use of modern JavaScript libraries like ReactJS, along with its component-based, modular architecture, to provide flexibility for the chat scripts.

Controlled Social Media

With the front-end intricacies explained, it's crucial to address how we managed controlled social media challenges in our implementation. In addressing the challenges stemming from PixelFed's PHP-based architecture, which hindered seamless modifications aligned with our Instagram-based scenarios, we found that modifying PixelFed required more effort than creating a new modular system. As a result, building an environment from scratch emerged as a more feasible approach. This allowed us to design and implement according to our specific scenarios within a shorter timeframe.

Our resulting solution, InstaCour, was developed using Next.js[35] framework for server-side rendering. We designed a function in InstaCour that allowed us to efficiently convert the JSON file containing images and captions into HTML components. To ensure that the application would dynamically render random content each time the Chrome browser is refreshed, we integrated a simple random function to shuffle the components in the JSON file.

Consequently, a controlled environment was established, optimized for unique user experiences while accommodating data logging requirements. InstaCour streamlined the process by omitting unnecessary sections such as registration and login, making access easier for users in VLC scenarios through a link. InstaCour addressed the complexities associated with PixelFed's intricate setup and provided a more tailored and efficient solution for our controlled social media environment.

User Authentication via Anonymized User URL

Having tackled the challenges of social media, let's move on to how we ensure secure and seamless user authentication, could work well here. User authentication is seamlessly orchestrated through the utilization of specially prepared URLs, each containing unique tokens. Leveraging the capabilities of the Chrome API, we effortlessly access the URLs

[35] What is Next.js? https://nextjs.org/learn/foundations/about-nextjs/what-is-nextjs, Sep 2023

that users click on, enabling a secure, anonymized, and streamlined authentication process. To extract the URL and populate it into the local storage, the following command was employed:

Javascript:

```
1) chrome.tabs.query({ active: true, lastFocusedWindow: true }, …
```

As exemplified by the sample anonymized user-token:

JavaScript:

```
2) https://swipeit.couragecompanion.eu/mailto:TEST143g@ItalyRIAS1June2023.eu
```

For instance, in the case of the June 2022 (Germany) and June 2023 tests (Italy), a repository of 500 distinctive user-tokens was generated. From this reservoir, users were provided with a user-URL-token selected randomly. This user-URL-token, when clicked, facilitated automatic authentication into the VLC system. The companion promptly initiated chat interactions, tailoring them based on the user-token’s unique profile and chat history. This innovative approach to user authentication ensures a seamless entry into the VLC system, enhancing both security, privacy, and user experience.

Highlighter Component

For a specific scenario that we will explain in Chapter 7, in collaboration with HRW University, the requirement for highlighting text was introduced. In this scenario, the VLC presents users with certain advertisements within the text. The users are then tasked with highlighting the sections of the text that they believe contain discrimination or racism. To implement this functionality, we utilized an embedded React Highlighter within the companion chatbot interface; in Figure 4-7 below you can see the Front-end Architecture:

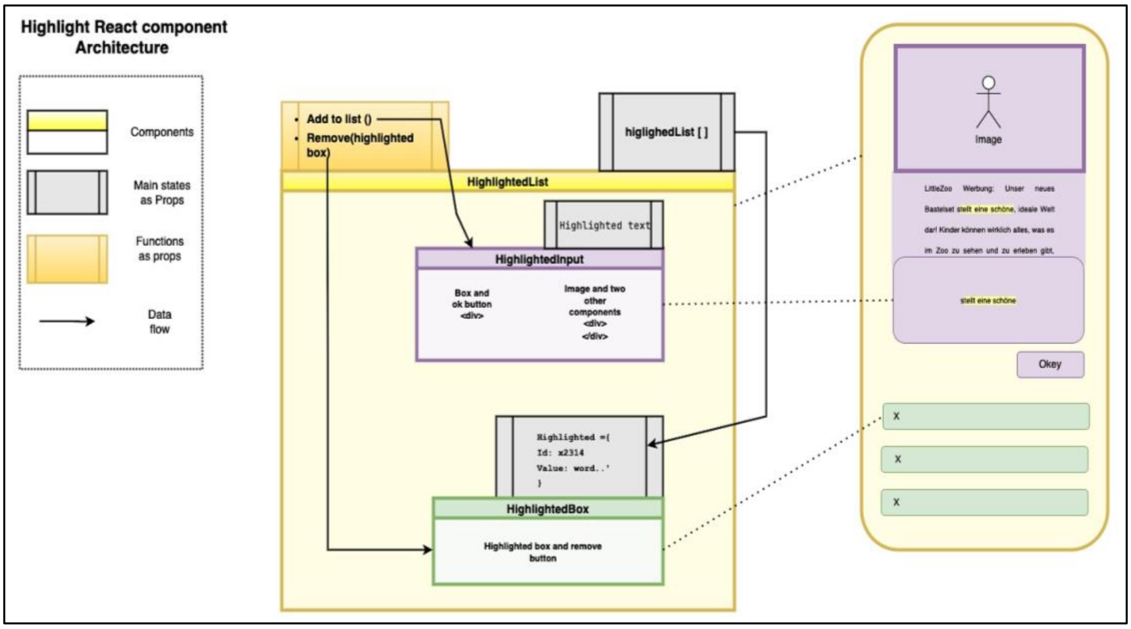

Highlight React component Architecture
Components
Main states as Props
Functions as props
Data flow
Add to list ()
Remove(highlighted box)
higlighedList [ ]
HighlightedList
Highlighted text
HighlightedInput
Box and ok button <div>
Image and two other components <div> </div>
Highlighted ={
Id: x2314
Value: word..'
}
HighlightedBox
Highlighted box and remove button
Image
LittleZoo Werbung: Unser neues Bastelset stellt eine schöne, ideale Welt dar! Kinder können wirklich alles, was es im Zoo zu sehen und zu erleben gibt,
stellt eine schöne
Okey
X
X
X

Figure 4-7: The architecture of the embedded highlighter in the companion chatbot.

This *highlighter* module is an embedded component that was developed into the chatbot interface. It allows the user to highlight the text selection and approve it. Think of this as the control center, it keeps track of what you've highlighted and whether the user finished or not. It has two main parts: Highlighted Input and Boxes as depicted in Figure 4-7.

Highlighted Input Component: The Highlighted Input component serves as the interactive interface where the user engages in text highlighting. Upon encountering a segment of text deemed significant, the user has the capability to select it. This selected text is subsequently displayed in a designated area for review. Should the user find the selection to be accurate, a confirmation action—executed via a button click—facilitates the addition of this text to a personalized repository of highlighted segments.

Highlighted Boxes Component: The Highlighted Boxes component functions as the user's personalized repository for highlighted text segments. Each confirmed highlight is systematically cataloged and displayed in this component, providing the user with a consolidated view of all selected text deemed important. Moreover, the user retains the flexibility to reconsider the relevance of each highlighted segment. An integrated removal option is available for each entry, allowing the user to effortlessly eliminate specific highlights from the repository**.**

Collectively, the Highlighted Input and Highlighted Boxes components constitute a comprehensive system that empowers the user to effectively identify, confirm, and manage text segments of interest within a digital reading environment.

4.2.3 Back-end of VLC

As mentioned, the VLC-middleware service of the back-end is the Node Express module. This middleware handles communication between the VLC and local or remote services.

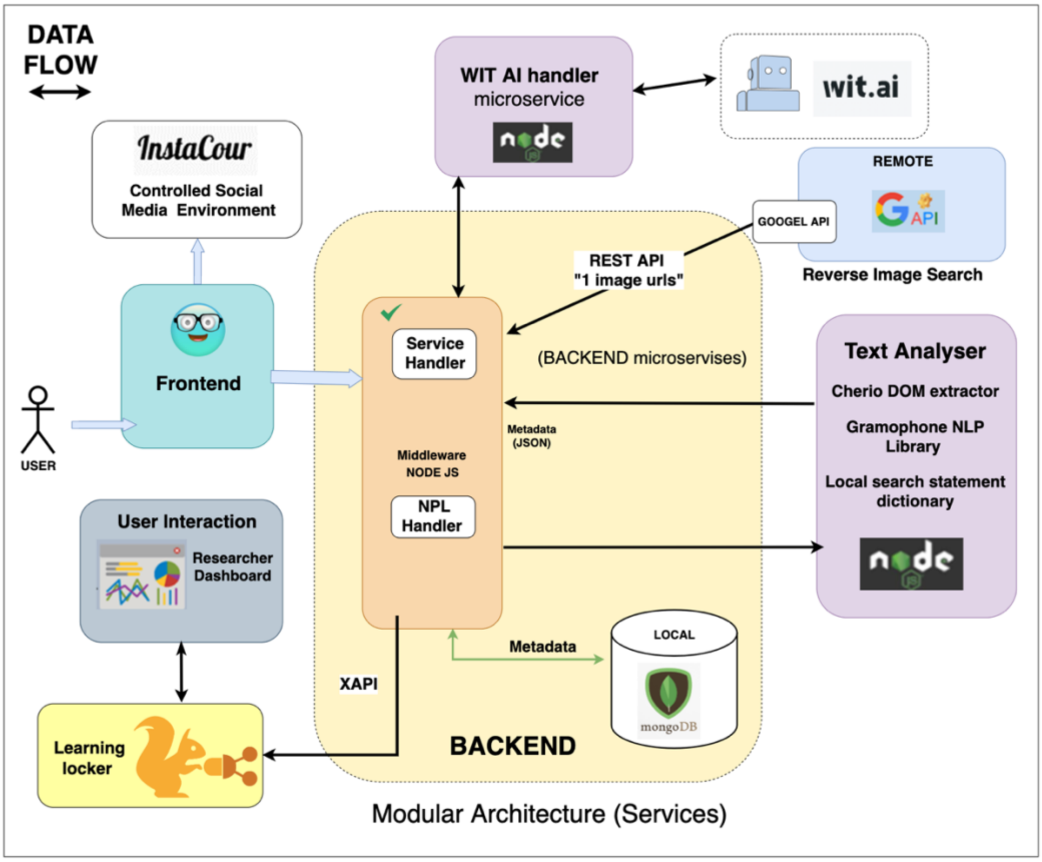


Figure 4-8: Back-end Architecture of VLC system

This middleware plays a crucial role in managing communication. It serves as a bridge between various services and the front-end. These services include VLC-NLP-processing microservice (Natural Language Processing), VLC-RIV-OCR microservice, and VLC-WitAI microservice. The middleware accomplishes this through REST API, data processing, and request handling within the application[36]. The procedure waits for the image-API from the Background.js file on the Chrome extension side. When it triggers the URL that the user has chosen, it queries the database to check if the user has already examined this image. If the user has previously checked the image, the conversation with the companion is retrieved from the user's history and sent back to the front-end in a structured JSON format for parsing (Figure 4-8). If it is the first time that the target unique

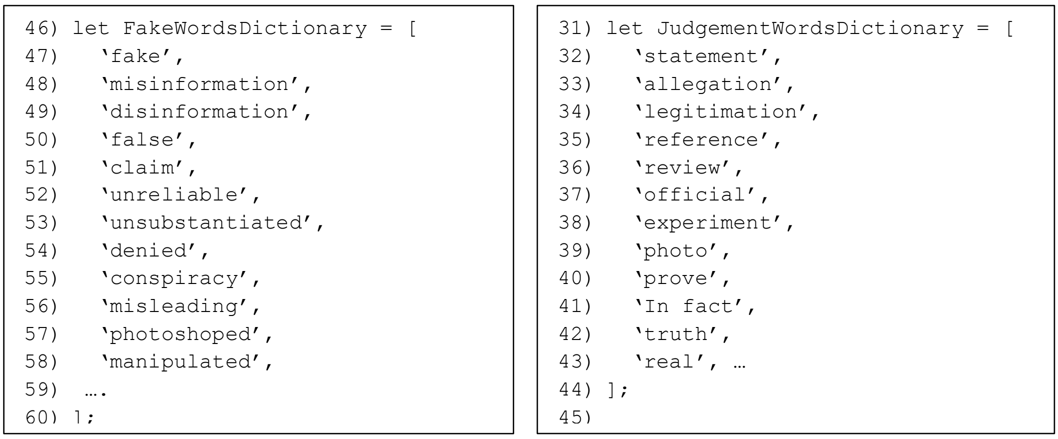
[36] VLC services, https://github.com/Farbod29/VLC-OpenAccess Aug 2023

user-token has chosen that image, the URL is sent to the VLC-RIV-OCR service. This service requests the RIS engine to select the six most similar images along with their websites and sends them to the middleware. In the next step, the middleware formats the URLs in JSON format and sends them one by one to the VLC-NLP-processing microservice, and waits for the response.

Text Analyzer Module

In the NLP service, we implemented a scraper that opens each URL’s page, reads the HTML content, and cleans it by removing redundant stop words. It then extracts the keywords based on TF-IDF methodologies. We utilized the Cheerio[37] and Gramophone Node.js[38] libraries, which have their stop-word lists and term frequency functions.

Additionally, we created our own dictionary, ‘FakeWordsDictionary,’ because we wanted to extract sentences containing special words and present them to the user. To display sentences from extracted webpages that make claims or judgments about statements, we used a dictionary as shown in Figure 4-9 below:



```
let FakeWordsDictionary = [
  ‘fake’,
  ‘misinformation’,
  ‘disinformation’,
  ‘false’,
  ‘claim’,
  ‘unreliable’,
  ‘unsubstantiated’,
  ‘denied’,
  ‘conspiracy’,
  ‘misleading’,
  ‘photoshoped’,
  ‘manipulated’,
 ….
];
```

```
let JudgementWordsDictionary = [
  ‘statement’,
  ‘allegation’,
  ‘legitimation’,
  ‘reference’,
  ‘review’,
  ‘official’,
  ‘experiment’,
  ‘photo’,
  ‘prove’,
  ‘In fact’,
  ‘truth’,
  ‘real’, …
];
```

Figure 4-9: Left - Special words in the field of false information. Right - Special words that might represent a claim and judgment about the found image on the web.

Database Structures

We employed MongoDB and Learning Locker, functioning as a Learning Record Store (LRS), to store all interaction data. This includes all interactions users have with the system, ranging from communication with the chatbot to interactions with images in controlled environments, as well as actions on browser tabs like opening new ones or

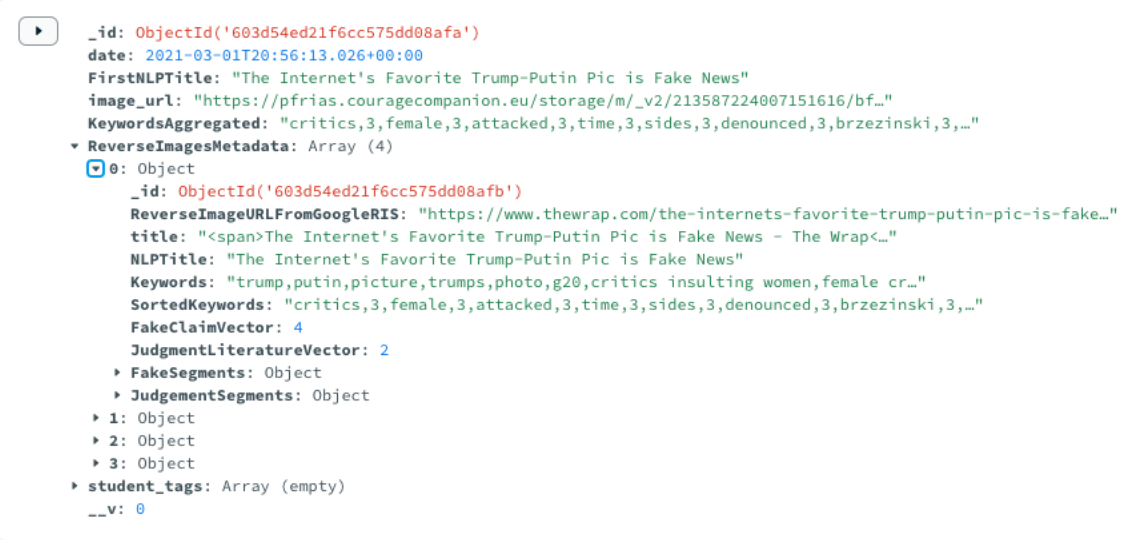
[37] The fast, flexible, and elegant library for parsing and manipulating HTML and XML. https://github.com/cheeriojs/cheerio, Aug 2023

[38] Extracts most frequently used keywords and phrases from text, https://github.com/bxjx/gramophone, Aug 2023

closing existing ones. These interactions are simultaneously logged in both MongoDB and Learning Locker, as illustrated in Figure 4-8

Image Metadata Storage Architecture

After extracting keywords and important sentences from all URLs, we store them in the database under the image metadata collection (Figure 4-10):

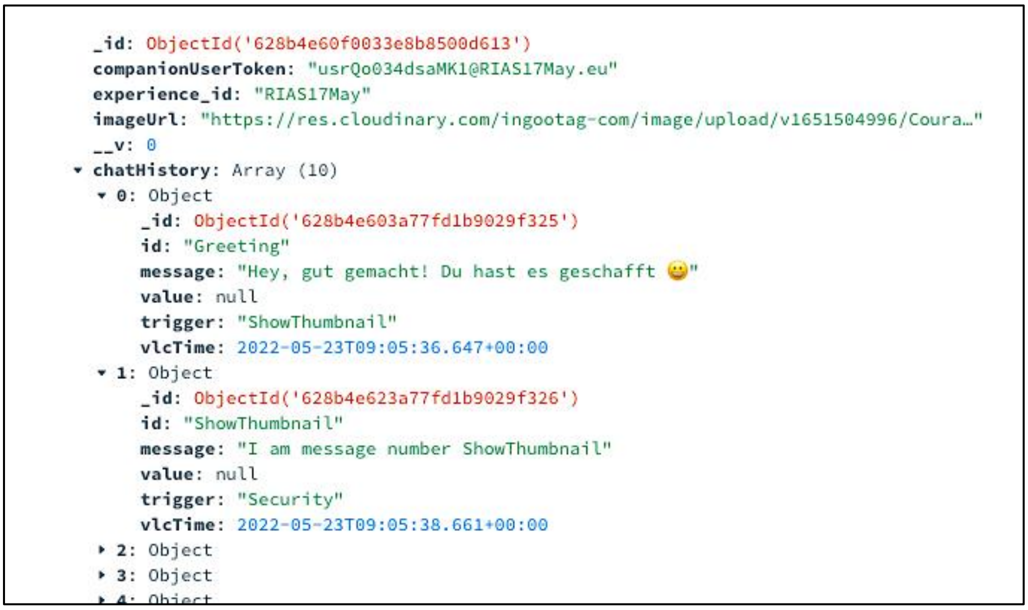


```
_id: ObjectId('603d54ed21f6cc575dd08afa')
date: 2021-03-01T20:56:13.026+00:00
FirstNLPTitle: "The Internet's Favorite Trump-Putin Pic is Fake News"
image_url: "https://pfrias.couragecompanion.eu/storage/m/_v2/213587224007151616/bf…"
KeywordsAggregated: "critics,3,female,3,attacked,3,time,3,sides,3,denounced,3,brzezinski,3,…"
ReverseImagesMetadata: Array (4)
  0: Object
    _id: ObjectId('603d54ed21f6cc575dd08afb')
    ReverseImageURLFromGoogleRIS: "https://www.thewrap.com/the-internets-favorite-trump-putin-pic-is-fake…"
    title: "<span>The Internet's Favorite Trump-Putin Pic is Fake News - The Wrap<…"
    NLPTitle: "The Internet's Favorite Trump-Putin Pic is Fake News"
    Keywords: "trump,putin,picture,trumps,photo,g20,critics insulting women,female cr…"
    SortedKeywords: "critics,3,female,3,attacked,3,time,3,sides,3,denounced,3,brzezinski,3,…"
    FakeClaimVector: 4
    JudgmentLiteratureVector: 2
  FakeSegments: Object
  JudgementSegments: Object
  1: Object
  2: Object
  3: Object
student_tags: Array (empty)
__v: 0
```

Figure 4-10: metadata format that includes a list of RIS pages and their titles, URLs of the images, sorted keywords based on term frequencies, and their special sentences in MongoDB.

Chat History of the VLC and the Users

We have another collection in the database where we store the user’s conversations around the selected image in the closed social media environment. This collection includes the school trial-ID, user-token, main image URL, and chat histories, as shown in Figure 4-11:

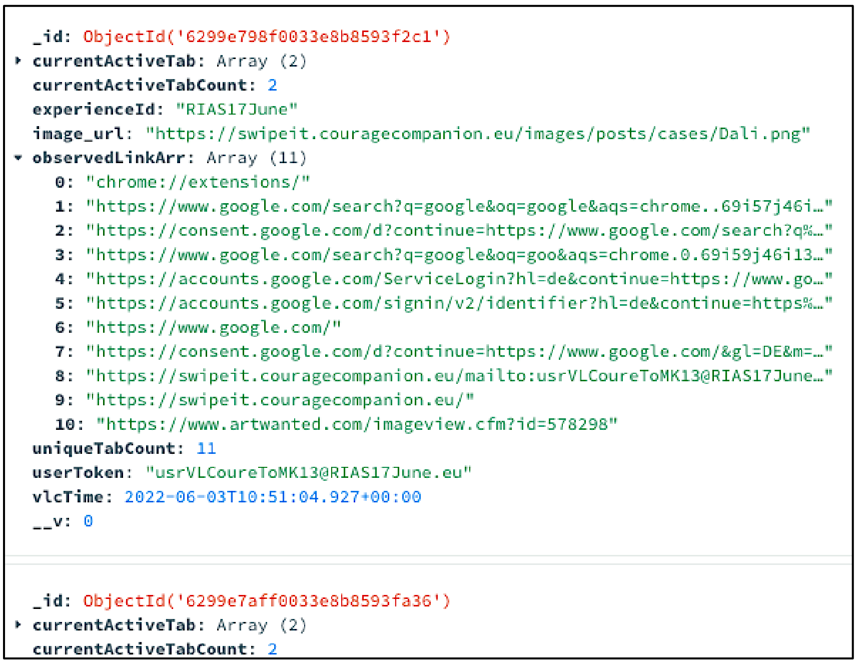


```
  _id: ObjectId('628b4e60f0033e8b8500d613')
  companionUserToken: "usrQo034dsaMK1@RIAS17May.eu"
  experience_id: "RIAS17May"
  imageUrl: "https://res.cloudinary.com/ingootag-com/image/upload/v1651504996/Coura…"
  __v: 0
▾ chatHistory: Array (10)
  ▾ 0: Object
      _id: ObjectId('628b4e603a77fd1b9029f325')
      id: "Greeting"
      message: "Hey, gut gemacht! Du hast es geschafft 😀"
      value: null
      trigger: "ShowThumbnail"
      vlcTime: 2022-05-23T09:05:36.647+00:00
  ▾ 1: Object
      _id: ObjectId('628b4e623a77fd1b9029f326')
      id: "ShowThumbnail"
      message: "I am message number ShowThumbnail"
      value: null
      trigger: "Security"
      vlcTime: 2022-05-23T09:05:38.661+00:00
  ▸ 2: Object
  ▸ 3: Object
  ▸ 4: Object
```

Figure 4-11: A sample MongoDB document in the companion chat collection.

Logging User Activity Through Chrome Extension: Capturing URL Data and Timestamps

As explained in the front-end section, one of the benefits of using the Chrome extension is to have control over the browser properties. We were able to log all the observed URLs. For example, if the user searches for a topic on Google and opens different tabs or opens the recommended links from the companion, we are able to store the unique visited websites along with their timestamps, as shown here in Figure 4-12:



```
  _id: ObjectId('6299e798f0033e8b8593f2c1')
▸ currentActiveTab: Array (2)
  currentActiveTabCount: 2
  experienceId: "RIAS17June"
  image_url: "https://swipeit.couragecompanion.eu/images/posts/cases/Dali.png"
▾ observedLinkArr: Array (11)
    0: "chrome://extensions/"
    1: "https://www.google.com/search?q=google&oq=google&aqs=chrome..69i57j46i…"
    2: "https://consent.google.com/d?continue=https://www.google.com/search?q%…"
    3: "https://www.google.com/search?q=google&oq=goo&aqs=chrome.0.69i59j46i13…"
    4: "https://accounts.google.com/ServiceLogin?hl=de&continue=https://www.go…"
    5: "https://accounts.google.com/signin/v2/identifier?hl=de&continue=https%…"
    6: "https://www.google.com/"
    7: "https://consent.google.com/d?continue=https://www.google.com/&gl=DE&m=…"
    8: "https://swipeit.couragecompanion.eu/mailto:usrVLCoureToMK13@RIAS17June…"
    9: "https://swipeit.couragecompanion.eu/"
    10: "https://www.artwanted.com/imageview.cfm?id=578298"
  uniqueTabCount: 11
  userToken: "usrVLCoureToMK13@RIAS17June.eu"
  vlcTime: 2022-06-03T10:51:04.927+00:00
  __v: 0

  _id: ObjectId('6299e7aff0033e8b8593fa36')
▸ currentActiveTab: Array (2)
  currentActiveTabCount: 2
```

Figure 4-12: In this example, the user opened eleven unique tabs.

We log all surfed URLs with their timestamps, even if they are repeated, as xAPI (Experience API) statements simultaneously in the LRS (Learning Record Store).

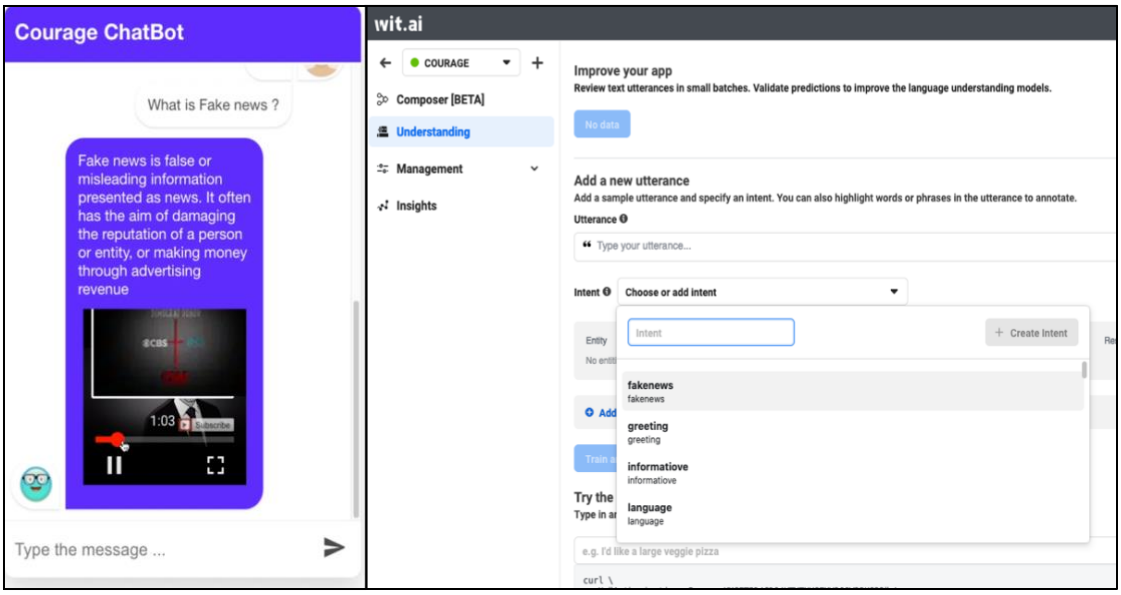
WitAI Module and VLC

Wit.ai is a natural language processing (NLP) service that enables developers to build applications that can understand and interact with human language. It offers a platform to create models that can recognize various entities, intents, and contexts within a user's input, making it suitable for developing conversational interfaces like chatbots. In the context of our VLC, we have leveraged the capabilities of Wit.ai to enhance the interaction and understanding of user intentions. The process of utilizing Wit.ai's API in our chatbot can be described in the following stages:

Detection of User Intention: We define specific intents and associate them with keywords and example phrases within Wit.ai. This approach helps the model recognize the user's intention based on the input. Unlike multiple-choice interfaces, Wit.ai allows users to interact with the chatbot using natural language. This enhances the user experience by providing a more conversational and flexible interaction, aligning with the needs of a virtual learning environment.

Integration with Chatbot: Wit.ai provides an API that we have integrated into our VLC. By sending user input to the Wit.ai API, the chatbot can receive a structured response containing the detected intent and entities. Based on the detected intent, the chatbot can generate custom responses or actions, providing a tailored interaction for the user. This integration enables our virtual learning companion to respond intelligently to various queries and commands.

Training and Adaptation: One of the key features of Wit.ai is the ability for continuous training and adaptation of the model. The model can improve its understanding of various user inputs over time by providing feedback and additional examples. This constant learning aligns with the evolving needs of learners and ensures that the virtual learning companion remains responsive and adaptive. The interface of WitAi and the VLC application can be shown in the following Figure 4-13:



What is Fake news ?

Fake news is false or misleading information presented as news. It often has the aim of damaging the reputation of a person or entity, or making money through advertising revenue

Improve your app

Review text utterances in small batches. Validate predictions to improve the language understanding models.

Add a new utterance

Add a sample utterance and specify an intent. You can also highlight words or phrases in the utterance to annotate.

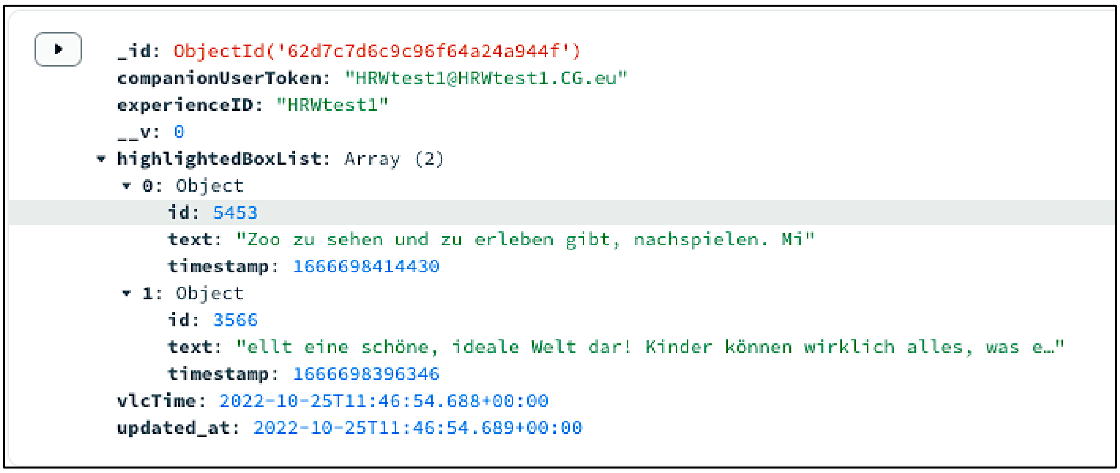
Try the

Type in ar

Figure 4-13: On the right side of the panel, Wit.ai is used for the training set, where intents are manually defined and configured. On the left side, the VLC is equipped with the Wit.ai API. When a user asks in free input text about a specific topic, the VLC processes the query through Wit.ai, recognizes the intent, and recommends video resources related to that topic.

Data Structure of Highlighter Component

The data structure that we used for highlighter component is as follows:



```
_id: ObjectId('62d7c7d6c9c96f64a24a944f')
companionUserToken: "HRWtest1@HRWtest1.CG.eu"
experienceID: "HRWtest1"
__v: 0
▾ highlightedBoxList: Array (2)
  ▾ 0: Object
      id: 5453
      text: "Zoo zu sehen und zu erleben gibt, nachspielen. Mi"
      timestamp: 1666698414430
  ▾ 1: Object
      id: 3566
      text: "ellt eine schöne, ideale Welt dar! Kinder können wirklich alles, was e…"
      timestamp: 1666698396346
vlcTime: 2022-10-25T11:46:54.688+00:00
updated_at: 2022-10-25T11:46:54.689+00:00
```

Figure 4-14: A sample of the data structure in the highlighted collection

Above Figure 4-14 shows the MongoDB document including the fields: unique ID of the trial and User as *experimentID* and *companionUserToken* and a list of highlighted text as an embedded JSON object.

4.2.4 Design of Statements and xAPI Structure

The Experience API (xAPI) is a standard for tracking learning experiences and activities. It is designed to be flexible and can be used in various programming languages, making it more versatile compared to older standards like SCORM[39]. Here's an overview of the key aspects of xAPI:

- **Statements**: xAPI uses 'statements' to dictate the format for specific moments in a stream of activity. These statements are the core of xAPI, and developers working with xAPI will become very familiar with them.

- **Statement Structure**: At a basic level, xAPI statements are expressed in a structure that includes an actor (the learner), a verb (the action), and an object (what the action is performed on).

- **Tools**: Tools like Learning Locker aid developers in creating and sending statements to a Learning Record Store (LRS).

To store the interactions between the VLC and the learner in the standard xAPI format, we first need to design the verbs that are compatible with our scenarios. To accomplish this goal, we made documentation and discussed the different statements that follow the actor, verb, and object format[40].

[39] The older standards of SCORM (Sharable Content Object Reference Model) have evolved over time, and you can find detailed information about these versions on the following link https://scorm.com/scorm-explained/business-of-scorm/scorm-versions/, June 2023.

[40] Courage xAPI GitHub documentation provided by Farbod Aprin under supervisor of RIAS experts. https://github.com/Farbod29/CORAGE-xAPI/blob/main/LearningLockerreceipt.md , July 2023

```
{
  ‘actor’: {
    ‘mbox’: ‘mailto:usr838fh3@courage.eu’,
    ‘name’: ‘usr838fh3’
  },
  ‘verb’: {
    ‘id’: ‘http://adlnet.gov/expapi/verbs/answered’,
    ‘display’: {
      ‘en-US’: ‘answered’
    }
  },
  ‘object’: {
    ‘id’: ‘http://couragecompanion.eu/activities/Image143fg-COVID19-5G’,
    ‘definition’: {
      ‘name’: {‘en-US’: ‘Image143fg-COVID19-5G’}
    }
  },
  ‘context’: {
    ‘extensions’: {
      ‘http://nextsoftwaresolutions.com/xapi/extensions/user-info’: {
        ‘image-id’: ‘http://couragePixelFed.eu/activities/2334’,
        ‘chat-Id’: ‘FakeOrNotQuestion’,
        ‘id’: ‘http://couragePixelFed.eu/artifact/chat/FakeOrNotQuestion’,
        ‘environment’: ‘VLC’,
        ‘vlc-time’: ‘162312432545’,
        ‘experienceID’: ‘5OCTOBER2021HRW’
      }
```

Figure 4-15: An example of the design of a Statement for the VLC system

This example, illustrated in Figure 4-15, shows a user (actor) who has answered (verb) a question related to an image (object) within the VLC environment. The context provides additional details about the interaction, including the image ID, chat ID, environment, and timestamps, all of which are in the extension field of the xAPI structure.

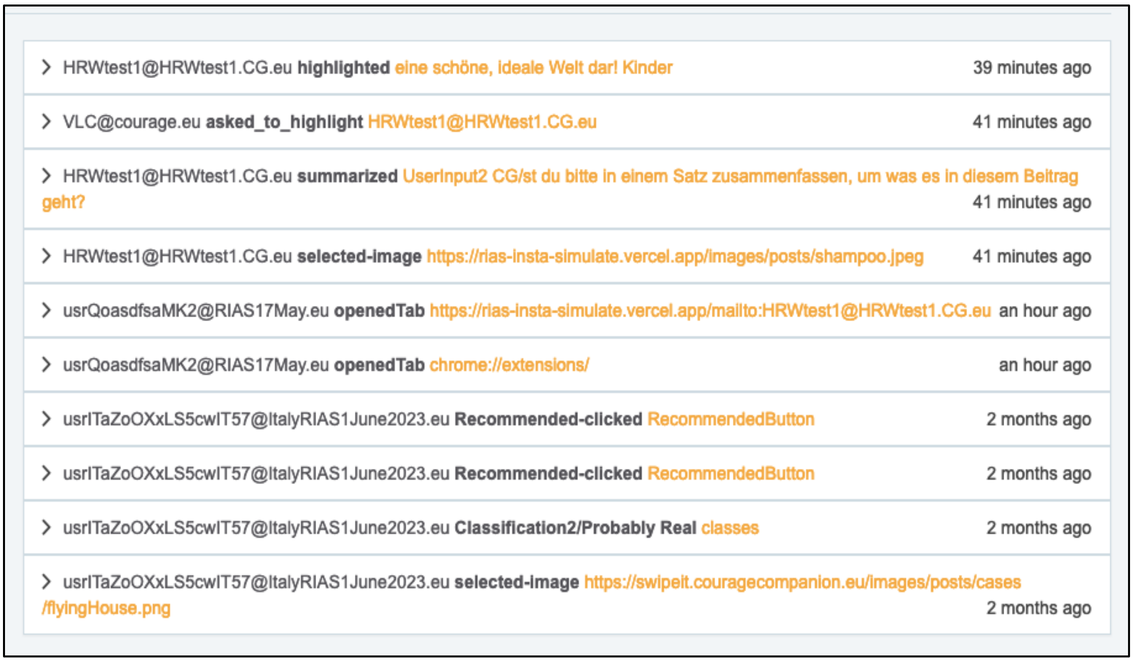
In the context of xAPI communication, the standardization of verbs and URIs plays a crucial role in defining and accessing various learning activities. Each verb within the system is associated with a unique description and URI, allowing for precise identification and categorization. A prominent example of this standardization can be found in the domain of TinCan[41], an organization responsible for xAPI statement standardization.

The xAPI specification further emphasizes the importance of vocabulary development and standards. Unlike rigid data models that enforce uniformity, xAPI adopts a more flexible approach. There is no requirement to adhere to a universal data model or agree on every type of learning activity that can be recorded. Instead, Activity Providers have the liberty to publish different statements about the same activity, utilizing their unique profiles and vocabulary.

[41] TinCan statement architecture “http://id.tincanapi.com/activitytype/source.”, July 2023

This approach offers significant advantages in terms of flexibility and resilience, allowing for a wide range of learning activities to be captured and represented. However, it also presents a potential challenge, as the lack of standardized practices can lead to disorderly or inconsistent implementations. The balance between flexibility and standardization thus becomes a critical consideration in the effective use of xAPI, reflecting both the potential and the complexity of this versatile framework.

The references and URIs for the statements, along with additional resources and further support for the development and application of these standards, contribute to a robust and adaptable system for tracking and analyzing learning experiences.



HRWtest1@HRWtest1.CG.eu **highlighted** eine schöne, ideale Welt dar! Kinder — 39 minutes ago

VLC@courage.eu **asked_to_highlight** HRWtest1@HRWtest1.CG.eu — 41 minutes ago

HRWtest1@HRWtest1.CG.eu **summarized** UserInput2 CG/st du bitte in einem Satz zusammenfassen, um was es in diesem Beitrag geht? — 41 minutes ago

HRWtest1@HRWtest1.CG.eu **selected-image** https://rias-insta-simulate.vercel.app/images/posts/shampoo.jpeg — 41 minutes ago

usrQoasdfsaMK2@RIAS17May.eu **openedTab** https://rias-insta-simulate.vercel.app/mailto:HRWtest1@HRWtest1.CG.eu — an hour ago

usrQoasdfsaMK2@RIAS17May.eu **openedTab** chrome://extensions/ — an hour ago

usrITaZoOXxLS5cwIT57@ItalyRIAS1June2023.eu **Recommended-clicked** RecommendedButton — 2 months ago

usrITaZoOXxLS5cwIT57@ItalyRIAS1June2023.eu **Recommended-clicked** RecommendedButton — 2 months ago

usrITaZoOXxLS5cwIT57@ItalyRIAS1June2023.eu **Classification2/Probably Real** classes — 2 months ago

usrITaZoOXxLS5cwIT57@ItalyRIAS1June2023.eu **selected-image** https://swipeit.couragecompanion.eu/images/posts/cases/flyingHouse.png — 2 months ago

Figure 4-16: Interface of the Learning locker that was used as LRS for all VLC scenarios.

As depicted in Figure 4-16 above, the short form of statements is summarized in the base form of actor, verb, and object. For VLC scenarios, we had to generate new statement structures, as they did not explicitly exist in known xAPI statement dictionaries like the Tin-can Registry and Activity Stream Standard. For example, in the statement [userX, highlighted, 'ideal word!'], the verb 'highlighted' was added according to our scenario.

4.2.5 Data Analysis Dashboard

We developed a dashboard for analyzing and creating diagrams based on the LL data. This dashboard was primarily developed by master's students working at the RIAS Institute (Figure 4-17). Meanwhile, we utilized Excel for specific analyses. We converted the JSON data to CSV files based on specific queries to MongoDB and Learning Locker. We then fed them into Excel to generate various diagrams. For the calculation of correlations and other mathematical computations, we used R Studio[42], employing the R programming language[43].

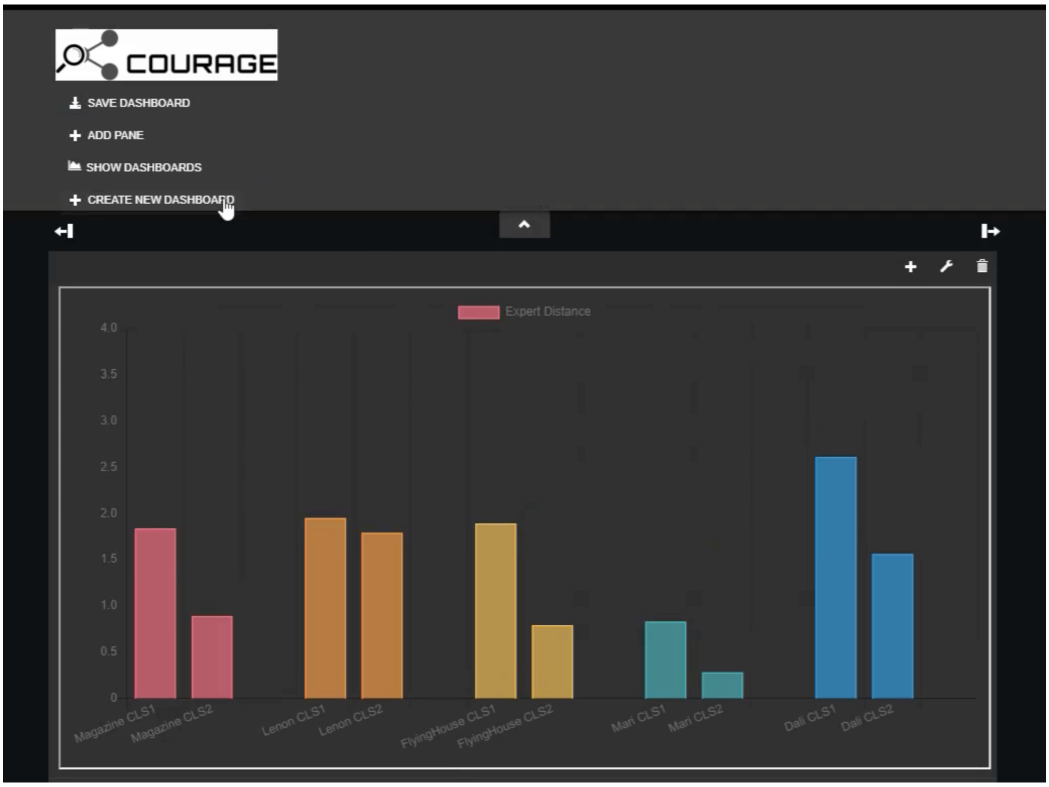
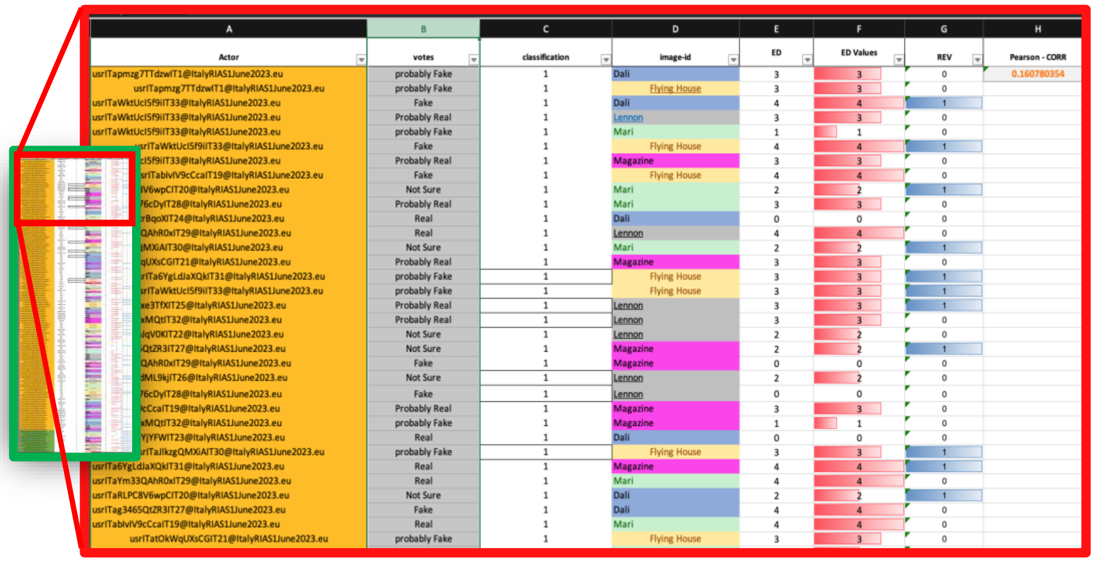


Figure 4-17 The image displays a user interface for a data analysis dashboard where the user can create a new dashboard, add diagrams, or save their current dashboard configuration.

[42] R-Studio https://posit.co/download/rstudio-desktop/, Aug 2023

[43] R language https://www.r-project.org/, Aug 2023



A	B	C	D	E	F	G	H
Actor	votes	classification	image-id	ED	ED Values	REV	Pearson - CORR
usrITapmzg7TTdzwIT1@ItalyRIAS1June2023.eu	probably Fake	1	Dali	3	3	0	0.160780354
usrITapmzg7TTdzwIT1@ItalyRIAS1June2023.eu	probably Fake	1	Flying House	3	3	0	
usrITaWktUcI5f9iIT33@ItalyRIAS1June2023.eu	Fake	1	Dali	4	4	1	
usrITaWktUcI5f9iIT33@ItalyRIAS1June2023.eu	Probably Real	1	Lennon	3	3	0	
usrITaWktUcI5f9iIT33@ItalyRIAS1June2023.eu	probably Fake	1	Mari	1	1	0	
srITaWktUcI5f9iIT33@ItalyRIAS1June2023.eu	Fake	1	Flying House	4	4	1	
cI5f9iIT33@ItalyRIAS1June2023.eu	Probably Real	1	Magazine	3	3	0	
srITablvIV9cCcaIT19@ItalyRIAS1June2023.eu	Fake	1	Flying House	4	4	0	
V6wpCIT20@ItalyRIAS1June2023.eu	Not Sure	1	Mari	2	2	1	
76cDyIT28@ItalyRIAS1June2023.eu	Probably Real	1	Mari	3	3	0	
rBqoXIT24@ItalyRIAS1June2023.eu	Real	1	Dali	0	0	0	
QAhR0xIT29@ItalyRIAS1June2023.eu	Real	1	Lennon	4	4	0	
MXiAIT30@ItalyRIAS1June2023.eu	Not Sure	1	Mari	2	2	1	
qUXsCGIT21@ItalyRIAS1June2023.eu	Probably Real	1	Magazine	3	3	0	
rITa6YgLdJaXQkIT31@ItalyRIAS1June2023.eu	probably Fake	1	Flying House	3	3	1	
srITaWktUcI5f9iIT33@ItalyRIAS1June2023.eu	probably Fake	1	Flying House	3	3	1	
xe3TfXIT25@ItalyRIAS1June2023.eu	Probably Real	1	Lennon	3	3	1	
xMQtIT32@ItalyRIAS1June2023.eu	Probably Real	1	Lennon	3	3	0	
JqV0KIT22@ItalyRIAS1June2023.eu	Not Sure	1	Lennon	2	2	0	
QtZR3IT27@ItalyRIAS1June2023.eu	Not Sure	1	Magazine	2	2	1	
QAhR0xIT29@ItalyRIAS1June2023.eu	Fake	1	Magazine	0	0	0	
dML9kjIT26@ItalyRIAS1June2023.eu	Not Sure	1	Lennon	2	2	0	
76cDyIT28@ItalyRIAS1June2023.eu	Fake	1	Lennon	0	0	0	
cCcaIT19@ItalyRIAS1June2023.eu	Probably Real	1	Magazine	3	3	0	
xMQtIT32@ItalyRIAS1June2023.eu	probably Fake	1	Magazine	1	1	0	
YjYFWIT23@ItalyRIAS1June2023.eu	Real	1	Dali	0	0	0	
rITaJlkzgQMXiAIT30@ItalyRIAS1June2023.eu	probably Fake	1	Flying House	3	3	1	
usrITa6YgLdJaXQkIT31@ItalyRIAS1June2023.eu	Real	1	Magazine	4	4	1	
usrITaYm33QAhR0xIT29@ItalyRIAS1June2023.eu	Real	1	Mari	4	4	0	
usrITaRLPC8V6wpCIT20@ItalyRIAS1June2023.eu	Not Sure	1	Dali	2	2	1	
usrITag3465QtZR3IT27@ItalyRIAS1June2023.eu	Fake	1	Dali	4	4	0	
usrITablvIV9cCcaIT19@ItalyRIAS1June2023.eu	Real	1	Mari	4	4	0	
usrITatOkWqUXsCGIT21@ItalyRIAS1June2023.eu	probably Fake	1	Flying House	3	3	0	

Figure 4-18 Logged data from the Learning record store in structural components, here the left is A: column is user Tokens, B: labels, C: First Judge or Second judgment (revision) D: Image-Id F: Distance From Experts judgments G: Revision participation status of the user for that case (0,1), H: Pearson correlation.

In Figure 4.18 above, we have logged data from the Learning Record Store presented in structured components. On the left, column A, lists of anonymous user tokens. Column B contains labels that user have chosen after VLC request like Fake, Real, probably/Fake Not sure, and column C shows whether it is the First Judge (1) or a Second Judgment (revision=2). Column D represents the Image-Id Distance from Expert Judgments. Additionally, column F indicates distance from experts judgments and column G indicates the revision participation status of the user for that case marked as 0 or 1 (not participated 0 or 1 participated), and column H details the Pearson correlation between columns F and G.

4.3 Chronological Improvements of the VLC

The below Table 4-1 presents a concise overview of the Virtual Learning Companion (VLC) system's evolution, with each version signifying a leap forward in functionality and scope, from initial fake news detection to sophisticated modules for identifying image manipulation and racism. These upgrades, influenced by user input and research insights, have successively broadened the VLC's pedagogical impact. The specifics of each scenario, along with the incremental developments and user experience enhancements, will be elaborated upon in the subsequent sections of this thesis.

Table 4-1 chronological improvements made to the Virtual Learning Companion (VLC) system.

Version	Release Date	Features/Versions	Improvements	Objective/ Technical Updates
1.0	2021	Initial release of Fake News scenario (Duisburg Essen University).	Base functionality for VLC- reverse image search – keywords -forum and diagram for other’s vote	Initial feedback from users about UX UI – Chrome Extension interface – English version with 3 COVID related images
2.0	2021-2022	Update to Fake News scenario (Duisburg Essen University).	Enhanced detection algorithms and user interface improvements.	Test backend of system and user logs – primitive analysis based on votes – Update the labels (Fake-Probably/Fake/Fact-Not sure)
3.0	2022-2023	Release of Image Manipulation scenario (Wolfskuhle ESSEN, Germany and LUMSA University, Palermo Italy LAM).	Expanded features for image manipulation detection and analysis.	Update in German and Italian language – 5 Images – manipulation and authentic- comprehensive analysis based on LL – calculation of agreement – improvement
4.0	2023	Release of Detect Racism scenario (Wolfskuhle ESSEN, Germany and Freiherr-vom-Stein-Gymnasium Oberhausen, Germany).	Introduced new modules for racism detection and incorporated user feedback from previous versions.	Add new InstaCour content and chatbot dialogue-implement imbedded text highlighter in chatbot

4.4 Challenges in Deployment Time

During the development phase, we faced notable challenges, particularly during the testing of the Chrome extension. The need for full compilation, building, and execution for even minor functional or user interface tests was a significant hurdle. This was largely because the core functions, developed in ReactJS, had to be compiled into the Chrome extension structure. The absence of hot reloading meant that a considerable amount of development time was devoted to the selection and refinement of the application. To enhance efficiency in future development, it would be advantageous to implement an engine that enables hot reloading of ReactJS within the Chrome extension environment.

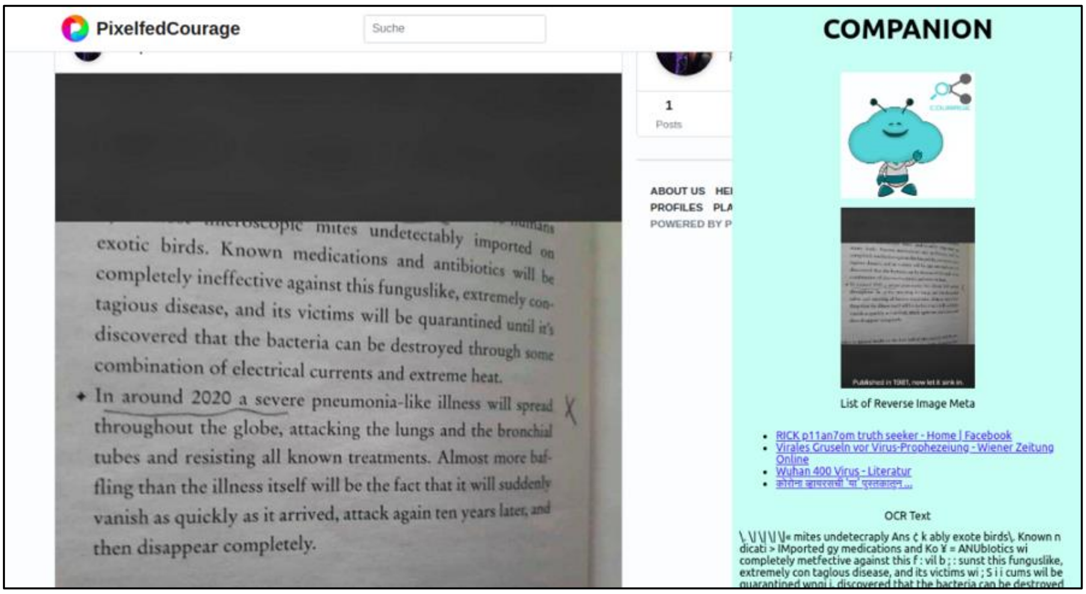
Another complex task was the crafting of xAPI statements, as the Learning Locker system is quite sensitive to syntax accuracy. For instance, any errant character in a randomly generated userToken within the xAPI actor statement could prevent the action from being logged in the cloud-based Learning Locker. To address this, we had to write test functions to retrieve the last statement and cross-verify it with a simultaneous MongoDB, or meticulously monitor the statement functionality via Postman or live application testing after each development cycle.

We also faced challenges with Google and other remote APIs, which do not offer free services for research purposes. This limitation forced us to utilize trial versions or services from third parties, which could potentially incur additional costs or raise privacy concerns later on. However, as our work operated within a simulated environment with a limited number of images for the Reverse Image Search (RIS) service, we were able to use the third-party trials effectively for school tests.

4.5 Evolution of VLC Design and Scenarios

In this section, we explore the evolution of the design and scenarios for the Virtual Learning Companion (VLC) from 2019 to 2023, covering the span of this Ph.D. program. This includes two related publications, [2] and [3], which also follow the general ‘Courage’ project’s goals.

The initial version of the companion was a simple standard popup page of the Chrome extension, implemented using basic HTML in the ‘popup.js’ file of the companion app. It reads the image of the PixelFed environment and passes the link of the selected image to the Google reverse image search API. This returned a list of similar images. This functionality allowed for straightforward integration with the PixelFed environment, leveraging Google’s image search capabilities to enhance the user’s experience within the virtual learning scenario (Figure 4-19).



microscopic mites undetectably imported on exotic birds. Known medications and antibiotics will be completely ineffective against this funguslike, extremely contagious disease, and its victims will be quarantined until it's discovered that the bacteria can be destroyed through some combination of electrical currents and extreme heat.

- In around 2020 a severe pneumonia-like illness will spread throughout the globe, attacking the lungs and the bronchial tubes and resisting all known treatments. Almost more baffling than the illness itself will be the fact that it will suddenly vanish as quickly as it arrived, attack again ten years later, and then disappear completely.

X

List of Reverse Image Meta

- RICK p11an7om truth seeker - Home | Facebook
- Virales Gruseln vor Virus-Prophezeiung - Wiener Zeitung Online
- Wuhan 400 Virus - Literatur
- कोरोना व्हायरसची 'या' पुस्तकातून...

OCR Text

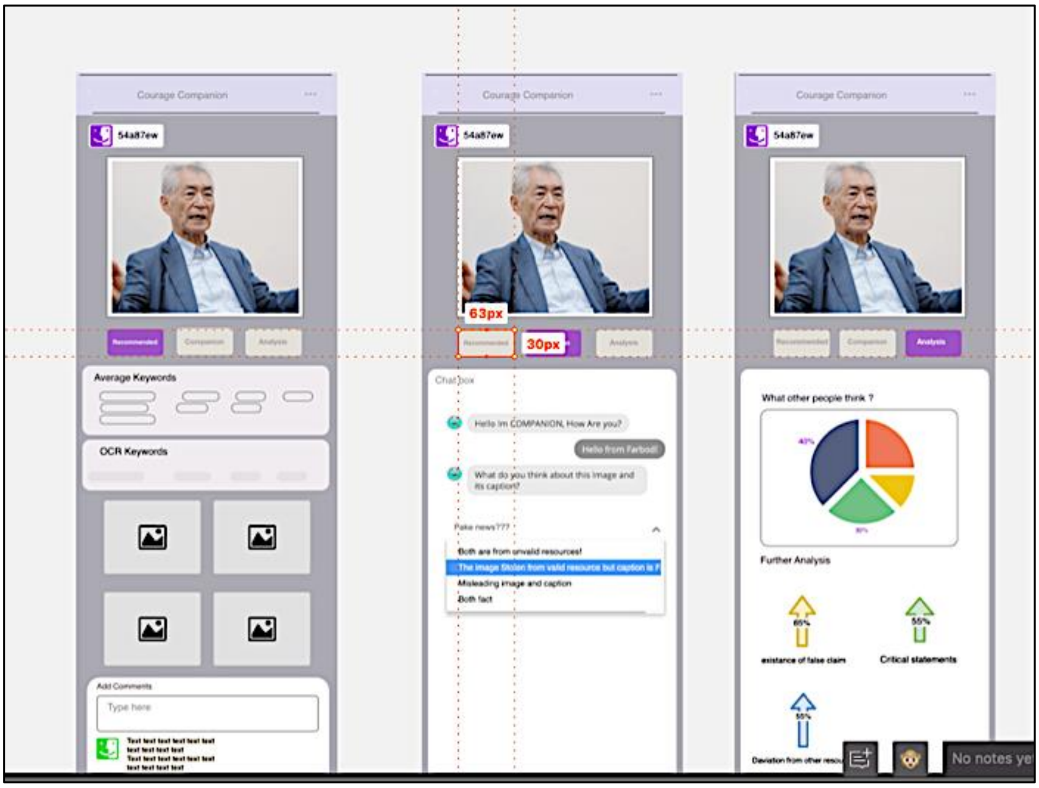
\. \/ \/ \/ \/« mites undetecraply Ans ¢ k ably exote birds\. Known n dicati > IMported gy medications and Ko ¥ = ANUblotics wi completely metfective against this f : vil b ; : sunst this funguslike, extremely con taglous disease, and its victims wi ; S i i cums wil be quarantined wnqi i. discovered that the bacteria can be destroyed

Figure 4-19: shows the initial version of the companion. The image content refers to a book that claimed to forecast the COVID-19 pandemic in 1991. This topic generated significant debate among conspiracy theorists and skeptics on social media in 2020.

Displayed below in Figure 4-20 is the initial design of the companion, crafted in Adobe XD[44]. This design encompasses three primary sections accessible via distinct buttons: **Recommendation/collaboration**, **Analysis**, and the **Companion** itself. In the

[44] Adobe XD is a vector design tool for web and mobile applications, developed and published by Adobe Inc, https://www.adobe.com/products/xd/learn/get-started/what-is-adobe-xd-used-for.html, Aug 2023

recommendation section, we planned to present the generated keywords and vital sentences extracted from the target image. Furthermore, we organized the keywords obtained from the image in the controlled environment, utilizing Optical Character Recognition (OCR) algorithms.



Text text text text text text text text text text Text text text text text text text text text text

Hello Im COMPANION, How Are you?

Hello from Farbod!

What do you think about this image and its caption?

Both are from unvalid resources!

The image Stolen from valid resource but caption is f

Misleading image and caption

Both fact

Figure 4-20: Initial UX/UI design of the companion, created in Adobe XD.

In the section **‘Recommendation/collaboration’** that can be accessed through the **‘Recommendation’** button, we considered implementing a collaboration forum where students can view others’ ideas at a meta-level, vote for comments, or reply to them. The best-voted comments would be ranked at the top, allowing others to easily read the comments that have received the most agreement or ‘like’ points.

The reason we sought to design the meta-level comments was to emulate real open environments like Instagram, where we could show the same discussion based on image similarity. If a user encounters the same image in a different context, they could find the same meta-information and discussion surrounding it. Essentially, this would require sending the images to a central system to find similarities, and if the similarity was high, the previous discussions and analyses would be displayed.

Another section is the **‘Analysis’** section, which displays others’ opinions based on their votes in different illustrations and diagrams. For this design, we planned that the companion gathers this data, asking users to answer critical questions during interaction with social media artifacts and other indexes that show the user’s time spent on the target social media content, usually an image and its caption. Additionally, the companion asks users to classify chosen categories (such as disinformation, misinformation, fact, fake, etc.) and demonstrate the result in pie charts and bar charts.

These features are designed with the aim of helping users break free from the filter bubble and echo chamber they may have inadvertently created on social media platforms (Figure 4-21). As discussed in the background section, many platforms nudge users to create filter bubbles, keeping them constantly engaged with content that aligns with their existing views based on the platforms' business model algorithms. Through this meta-level approach, users can see opinions outside their social media feeds. This provides a broader perspective alongside their regular social media usage.

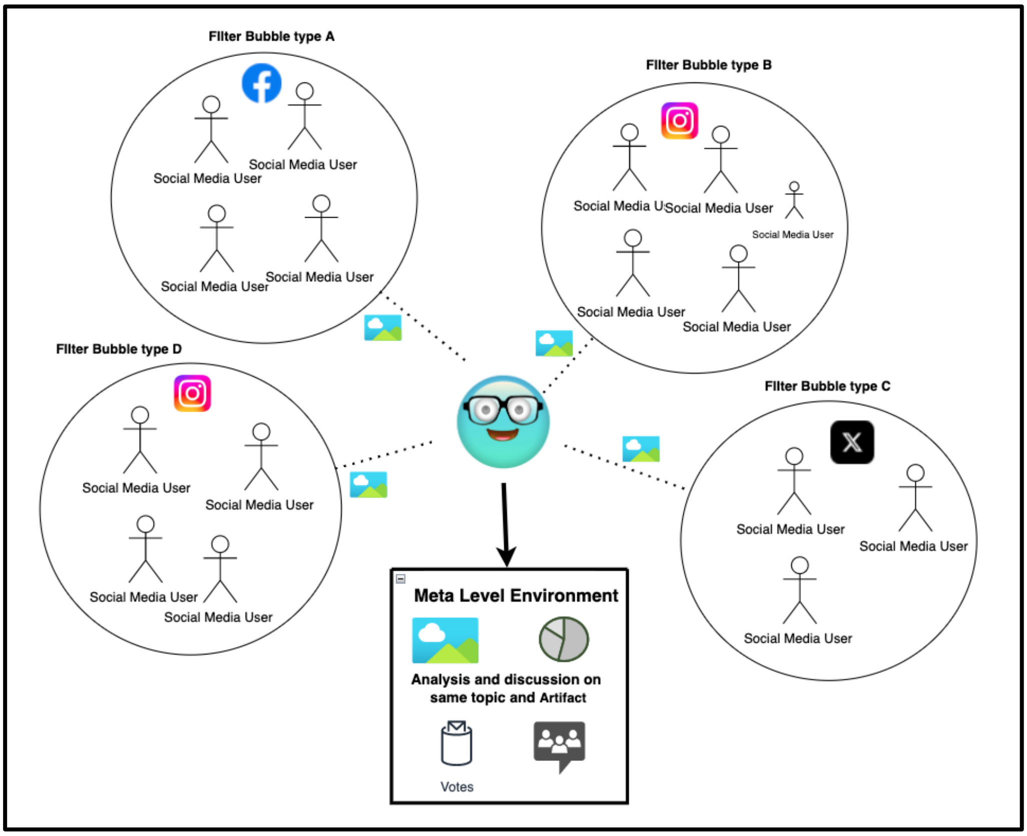


Figure 4-21: Conceptual architecture of the VLC, providing a meta-level environment where people with different mindsets, biases, and filter bubbles can discover others' opinions from various platforms on the same topic (similar images with different keywords)

The other button serves as a companion that interacts with the user through a chat dialogue. The plan was to use this feature in school trials so that VLC could ask demographic questions and activate prior knowledge, warn users about biases, and so on. It also prompts users to check recommendations while they are browsing content and provides additional ideas, analyses, and resources.

The discussed concept is implemented to assist users with fake content during the COVID using the discussed technologies stack (2021 version) which will be described in more detail.

4.6 Preliminary Scenario with VLC and User Feedback

In this section, we describe the preliminary scenario and report on user feedback, as detailed by Aprin et al. (2021) [2]. The test runs are conducted on desktop machines using Chrome as the web browser. Users can interact with PixelFed platforms that feature a limited number of images and captions, including both factual and fake news. Additionally, the arrangement incorporates a Chrome extension that serves as the browser-based virtual learning companion system (Aprin et al., 2021).

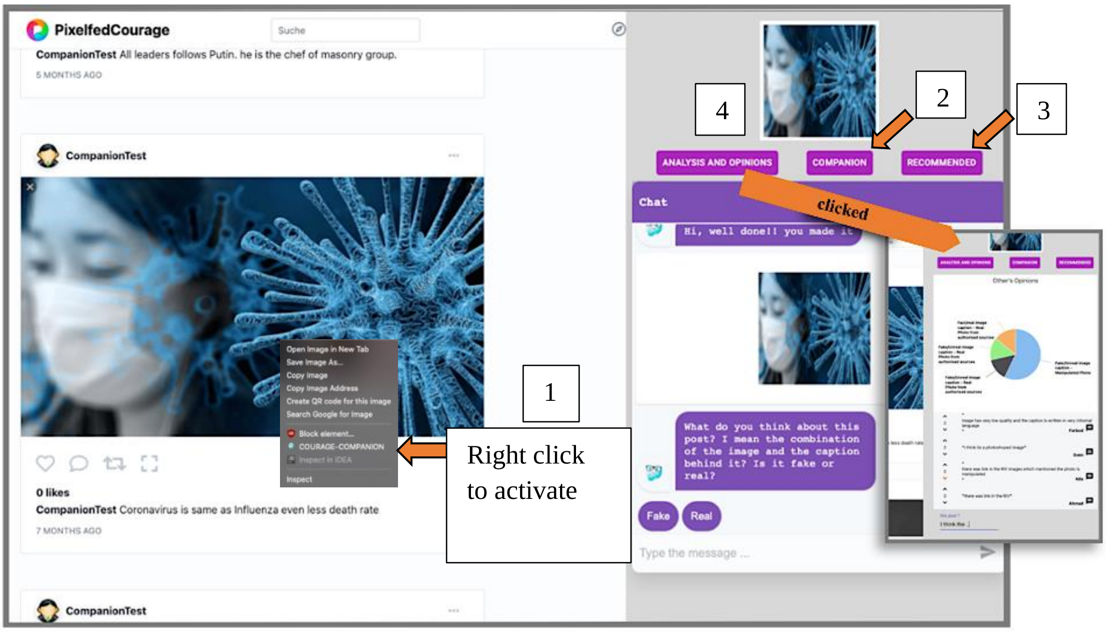
In the preliminary scenario planned for subsequent school trials, four pictures related to pandemic situations, including fake news and conspiracy theories, were selected for the PixelFed environment. The goal was to foster critical thinking and awareness. After interacting with the four pictures, the companion offers expert hints, provides answers, and solicits user feedback.

The technical goal behind the initial test was to assess the initial functionality of the Chrome Extension and chatbot interface and to gather feedback from the user regarding functionality and interface.

4.6.1 Example Walkthrough

Upon accessing the PixelFed environment via the provided web link, users initially encounter a brief guide on how to open and use the Courage Chrome extension. Following this, they log into the companion system by clicking an anonymous user-token link. They are then presented with a user-guide video on the sidebar featuring a virtual tutor that explains how to engage with the environment's artifacts to activate the companion.
The companion initiates communication with a chatbot-style conversation, asking general introductory questions (e.g., "Hey, how are you?", sex, age, etc.). After a brief exchange, the companion prompts the user to right-click on an image, followed by questions to activate knowledge and solicit the user's opinion about the selected image in the social media environment. The user is then asked reflective questions to explain their choice and whether they believe the image represents fake news or fact. The user's options are limited to predefined choices (Fake or Fact), and the companion further stimulates reflection with questions like "How sure are you?" based on a predetermined decision tree.

Next, the companion unlocks a 'Recommended' tab containing Reverse Image Search (RIS) links from the web, displaying the same image in various contexts. Learners can explore these links, comparing keywords, metadata, and abstract texts for the selected post. They can then return to the conversation to write their opinions, receiving adaptive feedback from the chatbot. The system unlocks the collaboration forum, chart option, and Analysis Tab for the selected artifact. Suppose the user judges or answers too quickly without checking the recommended RIS or exhibits poor participation in collaboration sections. In that case, the companion may provide alerts or feedback as a pedagogical intervention.



CompanionTest All leaders follows Putin. he is the chef of masonry group.

5 MONTHS AGO

CompanionTest Coronavirus is same as Influenza even less death rate

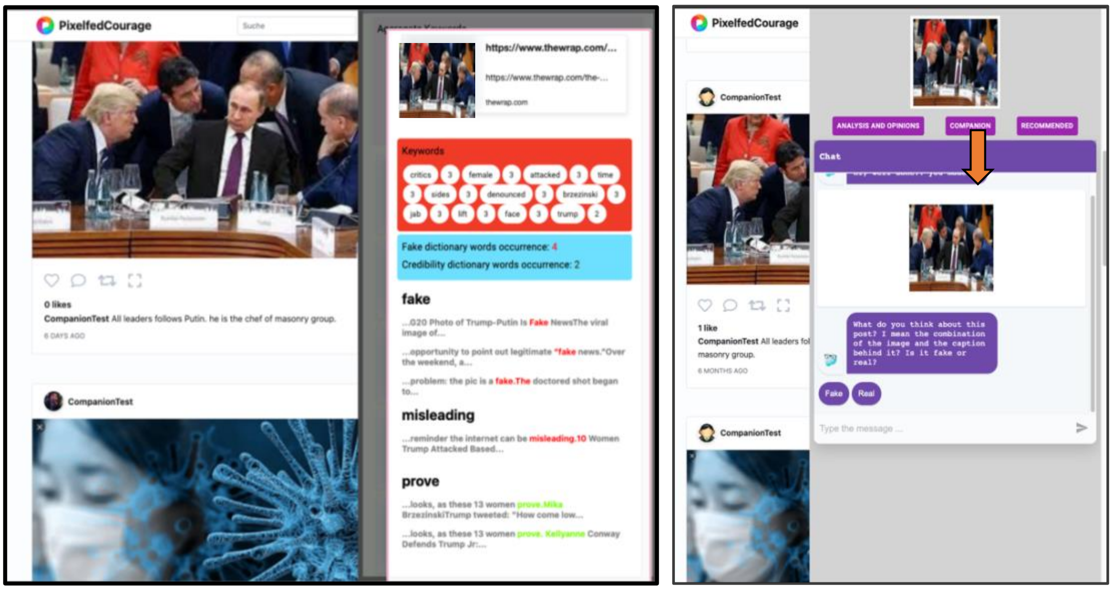
7 MONTHS AGO

Hi, well done!! you made it

What do you think about this post? I mean the combination of the image and the caption behind it? Is it fake or real?

Figure 4-22: Left: Upon right-clicking on a picture, the companion activates and responds by asking the user’s opinion about the selected image. Right: The companion unlocks the collaboration tab and inquires if the user has changed their mind after viewing comments and the votes of others in the diagram.

In Figure 4-22, illustrates how learners can view others’ opinions in a collaborative forum and compare other answers. The top-rated answers, according to user votes, are highlighted. The learner can also find better links from the comments and compare answers via pie charts. After observing the recommended metadata, the companion inquires if learners wish to reconsider their judgments.



0 likes
CompanionTest All leaders follows Putin. he is the chef of masonry group.
6 DAYS AGO

Fake dictionary words occurrence: 4
Credibility dictionary words occurrence: 2

...020 Photo of Trump-Putin Is Fake NewsThe viral image of...

...opportunity to point out legitimate "fake news."Over the weekend, a...

...problem: the pic is a fake.The doctored shot began to...

...reminder the internet can be misleading.10 Women Trump Attacked Based...

...looks, as these 13 women prove.Mika BrzezinskiTrump tweeted: "How come low...

...looks, as these 13 women prove. Kellyanne Conway Defends Trump Jr.:...

1 like
CompanionTest All leaders fol masonry group.
6 MONTHS AGO

What do you think about this post? I mean the combination of the image and the caption behind it? Is it fake or real?

Figure 4-23 Left: The PixelFed environment containing fake/fact stimuli, esp. related to the COVID-19 Pandemic. Right: The companion section and ask the user's opinion.

In Figure 4-23, we extracted keywords for each retrieved image, sorted by term frequency, and identified influential sentences based on a predefined dictionary. The companion is activated by selecting individual images within the PixelFed environment.

For researchers, it is intriguing to observe if learners alter their thinking or judgment after receiving feedback from the companion and interacting with recommended learning instructions. Therefore, all significant user interactions, including timestamps, are logged. The companion may also warn or encourage learners, such as alerting them to the 'spiral of silence' phenomenon if they change their idea based on the wrong, false decision of majority analysis. In this case, the VLC will present a tutorial video as a pedagogical intervention (Glynn et al., 1995).

4.6.2 Study and Feedback

Prior to finalizing the VLC, a preliminary study was undertaken with 14 participants. Their ages varied from 19 to 35, and they hailed from a broad spectrum of cultural and educational settings. The cohort included bachelor's students and postdoctoral scholars from various fields of study, such as business, information technology, and psychology. The participants originated from several countries, including Germany, Iran, India, Palestine, and the USA. They were presented with a controversial report on Coronavirus in a simple environment, including a picture shared on social media, and asked to classify it as 'fake' or 'fact'. The aim of the research was to assess user comprehension and execution of the task, as well as to pinpoint any potential misconceptions.

Users determine whether a source is fake or factual based on the website's appearance, its level of seriousness, or the quality of its language. They make this assessment based on search engine results and the links provided in a simplified environment containing examples. According to a qualitative questionnaire featuring six major oral questions, when we asked users, "Why do you think it is fake?" two respondents mentioned, "If it has a lot of advertisements, it's not credible." However, this is not always the case. Private news outlets like 'DailyMail[45]' are not state-funded and often rely solely on advertising for support.

Also, the study revealed that over half of the participants were uncertain about their answers and desired more nuanced options like 'probably fake,' 'not sure,' or 'probably fact.' Only three participants utilized external fact-check tools or search engines. As a result, none of them were aware of Reverse Image Search tools for fact-checking, as they had not used them during the fact-checking process, a feature considered in the VLC tool and scenario that was considered. This finding supports the vision that a real-time VLC can assist learners in making more accurate judgments and utilizing web resources more effectively [2].

[45] New Platformm, DailyMail https://www.dailymail.co.uk/news/germany/index.html, Aug 2023

5 Testing the VLC Distinguishing Fake from Fact Scenario

This section, based on the third study by Aprin, Chounta, et al. (2022) [3], elaborates on the enhancements made to the Virtual Learning Companion (VLC), including the ability to log users' browser tabs.

5.1 Technical improvements

In the front-end interface, like the last iteration, learners can freely select an image on the social media platform. The companion system, implemented as a Chrome extension, sends the image URL to a middleware microservice. Along with the extracted semantic keywords, it searches the extracted text for bold sentences corresponding to a predefined dictionary. This includes sentences containing words like 'Claim,' 'Fake,' etc. The text analyzer micro-service utilizes the Cheerio library to perform web scraping in JavaScript (Dom Tree extractor[46]). Additionally, to analyze the scraped text from each Reverse Image Link (RIS) website, the system extracts keywords based on term frequency and a filtering algorithm we implemented. The Gramophone can be configured to extract phrases (n-grams) of any length, not just keywords.

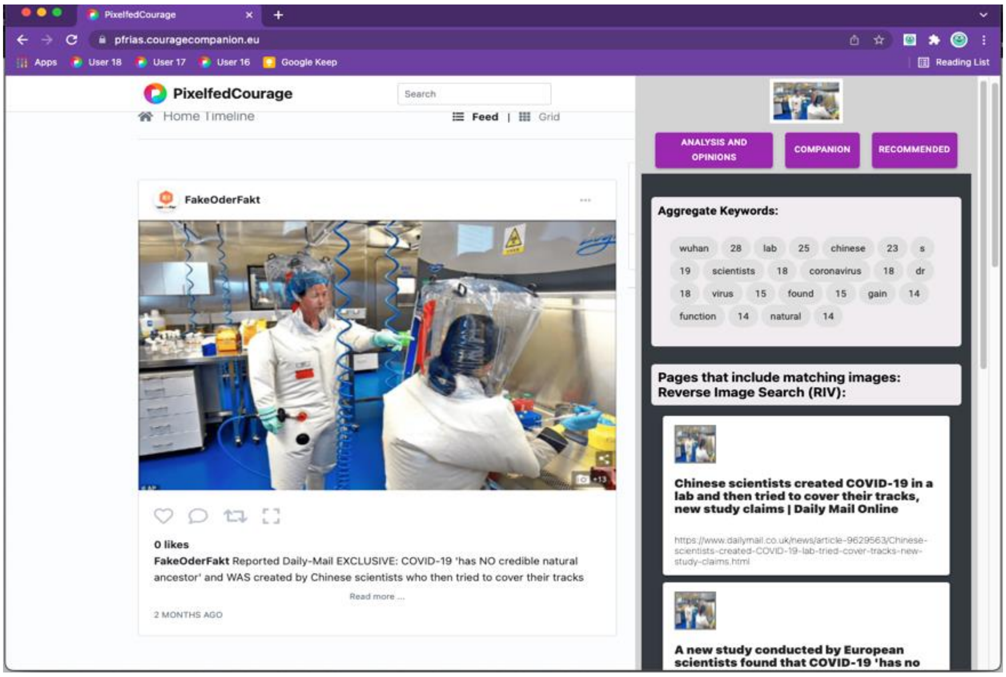
5.2 'Fake or Fact' Scenario for Trials

The learning environment operates through a Chrome browser, accessing PixelFed with a curated collection of news (related to the COIVD-19 situation) images and their captions. The VLC is integrated as a Chrome extension or plugin.

Learners enter the shared PixelFed environment by clicking a unique, anonymous user-token link. They receive brief instructions (video and oral) on how to open the Chrome extension, and the companion image in the sidebar guides them to select an image on PixelFed to activate the chatbot. After right-clicking an image and selecting the VLC button, the companion automatically initiates a chatbot-style conversation (e.g., "Hey, you made it!"), asking demographic questions (gender, age, etc.), followed by knowledge activation questions. Learners are required to give their opinion on the selected image and categorize it into one of five options: Fake, Probably Fake, Not Sure, Probably Fact, or Fact. Reflective questions are then posed to understand the reasoning behind their choice. The companion further stimulates reflection with questions like "How sure are you?" and

[46] What is Document Object Model (*DOM*) tree, https://shorturl.at/hsuzF , July 2023.

unlocks a ‘Recommended’ tab that includes Reverse Image Search (RIS) links. These display the same image in different contexts, allowing learners to compare keywords, metadata, and bold phrases see Figure 5-1. The VLC encourages and guides learners to utilize external tools for gaining additional context on the topic.



FakeOderFakt

0 likes

FakeOderFakt Reported Daily-Mail EXCLUSIVE: COVID-19 'has NO credible natural ancestor' and WAS created by Chinese scientists who then tried to cover their tracks

Read more ...

2 MONTHS AGO

Aggregate Keywords:

wuhan	28	lab	25	chinese	23	s
19	scientists	18	coronavirus	18	dr	
18	virus	15	found	15	gain	14
function	14	natural	14			

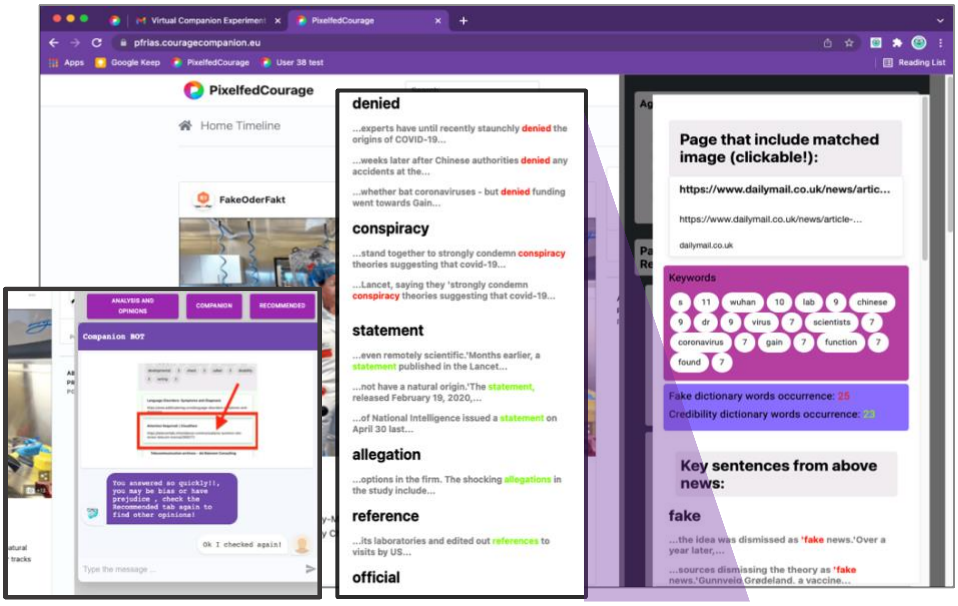
Pages that include matching images: Reverse Image Search (RIV):

Chinese scientists created COVID-19 in a lab and then tried to cover their tracks, new study claims | Daily Mail Online

https://www.dailymail.co.uk/news/article-9629563/Chinese-scientists-created-COVID-19-lab-tried-cover-tracks-new-study-claims.html

A new study conducted by European scientists found that COVID-19 'has no

Figure 5-1: The PixelFed environment displays an image and its caption on the left, while the right side features the VLC add-on, providing contextual information. The caption claimed the COIVD-19 virus was made in the Wuhan lab.



FakeOderFakt

Companion BOT

You answered so quickly!!, you may be bias or have prejudice , check the Recommended tab again to find other opinions!

Ok I checked again!

denied

...experts have until recently staunchly denied the origins of COVID-19...

...weeks later after Chinese authorities denied any accidents at the...

...whether bat coronaviruses - but denied funding went towards Gain...

conspiracy

...stand together to strongly condemn conspiracy theories suggesting that covid-19...

...Lancet, saying they 'strongly condemn conspiracy theories suggesting that covid-19...

statement

...even remotely scientific.'Months earlier, a statement published in the Lancet...

...not have a natural origin.'The statement, released February 19, 2020,...

...of National Intelligence issued a statement on April 30 last...

allegation

...options in the firm. The shocking allegations in the study include...

reference

...its laboratories and edited out references to visits by US...

official

Page that include matched image (clickable!):

https://www.dailymail.co.uk/news/artic...

https://www.dailymail.co.uk/news/article-...

dailymail.co.uk

Keywords

s	11	wuhan	10	lab	9	chinese
9	dr	9	virus	7	scientists	7
coronavirus	7	gain	7	function	7	
found	7					

Fake dictionary words occurrence: 25

Credibility dictionary words occurrence: 23

Key sentences from above news:

fake

...the idea was dismissed as 'fake news.'Over a year later,...

...sources dismissing the theory as 'fake news.'Gunnveig Grødeland. a vaccine...

Figure 5-2: showcases learner interactions with the VLC: Left, advice to examine generated metadata more closely; after clicking on a RIS link, the corresponding keyword analysis is displayed (right); Center: Learners can scroll through ‘key sentences’ with highlights.

Figure 5-2 also details how learners can explore each retrieved RIS link in-depth, reviewing titles and clicking links to see bold sentences with highlighted words. They can compare these to the caption on PixelFed and determine if the image is manipulated, the text is distorted, indicative of misinformation, or based on credible news sources.
Researchers find it valuable to know if learners change their minds after interacting with additional artifacts or recommended learning instructions. Understanding the direction of learners' thinking and any changes in judgment after feedback is crucial. All meaningful learner interactions are logged, including timestamps and related websites visited, in the standard xAPI format.

If learners quickly assess or respond without checking the recommended RIS, they receive a chat notification, e.g., "Be careful not to be biased! "After browsing the RIS links, the companion provides adapted feedback and asks if they want to revise their vote. After explaining and justifying any vote changes for all three established images in PixelFed, the companion unlocks the 'Analysis Tab and Collaboration' to view others' votes as a bar chart.

The system also records learners' responses, companion chat reactions, interactions with environmental artifacts, and browser tabs opened during the experiment. Three selected images related to pandemic situations may be used in future school trials, containing fake or credible news and conspiracy theories. Critical thinking, awareness, and using the RIS tool for better judgment are the goals. Finally, the companion provides expert feedback and determines if learners know the RIS method to detect fake news.

The technical goal behind the Fake or Fact scenario was to test the functionality of the Chrome Extension with new features and chatbot interface, functionality of the backend and modular services like keyword extraction, RIS search API, Forum chat, check user authentication via URL and to control and monitor the data logs that generated in the MongoDB. and to get feedback from the user about functionality and interface.

5.3 User Evaluation

In the initial trial with the Chrome extension and PixelFed environment, forty-five invitations were sent to volunteer test users from the University of Duisburg-Essen. Twelve participants, aged 19-65, actively participated. They were presented with three images containing fake, factual, or controversial COVID-19 items and instructed to interact with the companion.

The findings suggest that the metadata from retrieved RIS links, along with guidance from the VLC, positively influences users to revise their initial judgments. After RIS and companion instructions, one in four participants changed their decision at least once for the provided social media post (6/24 votes).

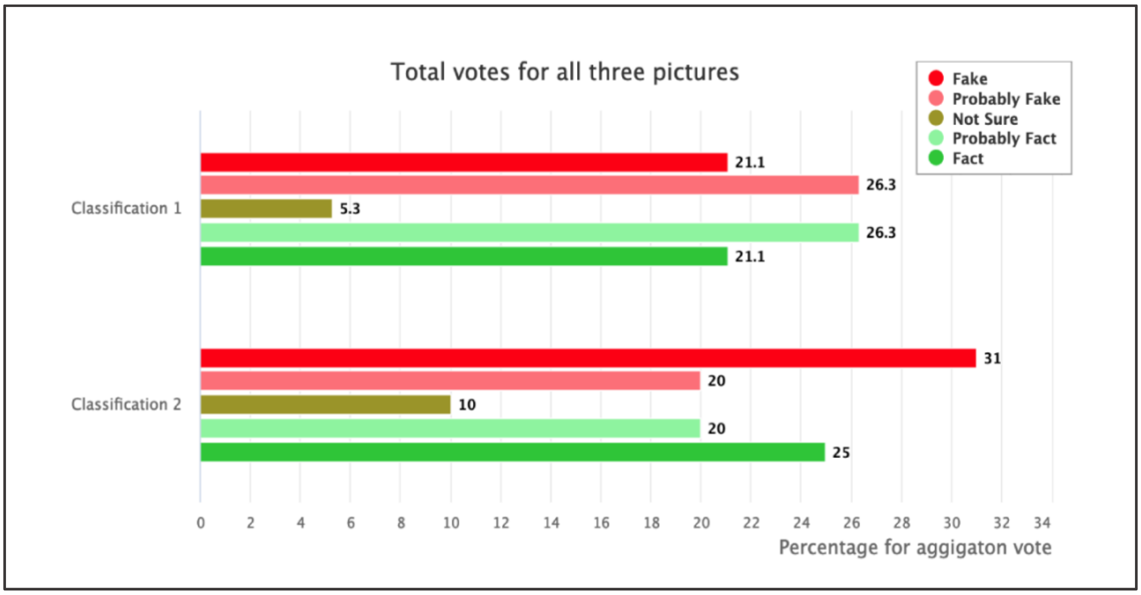


Figure 5-3: displays the distribution of votes for all three images before and after guidance by the companion for 45 test users (classification 1, 2).

The chart in Figure 5-3 shows a shift in votes, with deterministic classifications of 'Fake' or 'Fact' increasing, but also a rise in the neutral category 'Not Sure.' Five user judgments were changed after receiving additional information, with most changes (3) decreasing the difference from the expert judgment. Seven matches with expert judgment were found initially, increasing to ten in the final classification. The item that had the least agreement between user judgments and expert ratings was classified as 'Not Sure' by the experts. Comments from users revealed that the additional information helped them rethink their judgment more accurately. Some even changed their vote to match the expert's judgment, explaining their reasoning based on discrepancies between the image's caption and other resources.

Overall, the trial demonstrated the system's practicality and the potential of the VLC to support learners in making more informed judgments about fake and real news.

6 The Effectiveness of a Virtual Learning Companion for Supporting the Critical Judgment of Social Media Content

This chapter draws on the journal article and related studies (Aprin, et al., 2023) [5]. We updated the Virtual Learning Companion (VLC) interface and dialogue, transitioning from PixelFed to InstaCour for enhanced user experience and technical feasibility, as detailed in chapter (4.2). The VLC was specifically designed to tackle the issue of image manipulation. We conducted a pilot 'Fake or Real' study in a German high school in June 2022 and replicated the study at an Italian university in June 2023. The following research questions guided our classroom experiments and are addressed in this initial evaluation:

RQ1: Was the VLC effective for engaging learners in a high school classroom in meaningful interactions with stimulus material in a simulated social media environment?

RQ2: Was the accuracy of the judgment positively influenced by the time spent visiting or reading web pages related to the target picture?

RQ3: Did a revision of the initial judgment guided by the VLC lead to higher agreement with the expert opinion and with other participants' judgments (convergence)?

Research questions for the replicated study (Italy):

RQ4: Are the judgments from Italy aligned with those from Germany? Does one's cultural background and native language have an impact on how they assess the credibility of an image?

RQ5: Are the classifications made by participants arbitrary, considering that the credits are not based on the quality of their answers? Or do the participants genuinely strive to categorize the stimuli in a meaningful and appropriate manner?

6.1 Scenario

In this scenario, the content of InstaCour primarily focused on image manipulation, diverging from the previous scenarios that dealt with pandemic situations, fake news, and conspiracy theories as we had in [2] and [3]. For our classroom evaluation, we utilized stimulus material containing both authentic and manipulated images. The status of these images (real or manipulated) had been previously determined based on provenance information and expert judgments; we have chosen controversial cases that detect their manipulation without using search engines is difficult, even for adults.

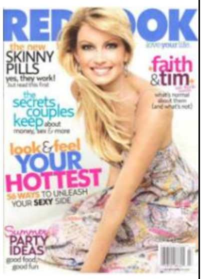
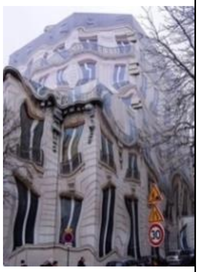
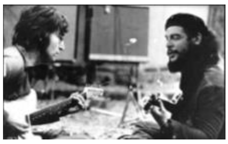
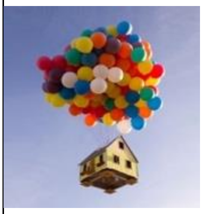
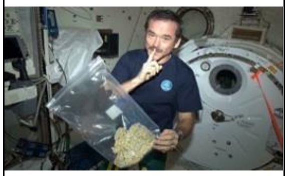
In this trial, each learner is linked to an anonymous user-token managed by the browser plugin, allowing them to access the computer conveniently. Learners can enter the learning environment with the Chrome browser open on each PC.

The companion guides users to right-click on an image within the simulated social media platform (InstaCour) and select the companion tool icon. The images, which may be either manipulated or real, are displayed on InstaCour, arranged like Instagram posts, allowing users to scroll through them.

6.2 Stimulus Materials

We selected five specific cases or stimuli (images and their captions), categorized as either ‘Fake’ or ‘Real’ in Table 6-1 based on expert judgment. ‘Fake’ refers to images altered with software like Adobe Photoshop, while ‘Real’ denotes unaltered images captured directly from a camera lens.

Table 6-1: Heading: Identification of Stimulus (Images and accompanying captions) along with their color coding. Middle: Representative images as displayed on InstaCour. Bottom: Categorization according to expert assessment (label).

Magazine	Dali	Lennon	Flying House	Mari
				
Fake	Real	Fake	Real	Fake

The ‘Magazine’ case emphasizes stereotypes, featuring a digitally altered woman who doesn’t exist in reality[47]. The ‘Dali’ case explores human bias judgments, showcasing a real but seemingly impossible image[48]. The ‘John Lennon’ case involves a manipulated

[47] Before and after: Photoshopped celebrity pictures, https://joshbenson.com/before-and-after-photoshopped-celebrity-pictures, March 2023

[48] The melting Building, https://www.artwanted.com/imageview.cfm?id=578298, Feb 2023

photo, aiming to entrap followers of Lennon and Che Guevara within a specific filter bubble[49]. The ‘Flying House’ case presents a real but fantastical project by National Geographic[50]. The ‘Mari’ case falsely implies that astronaut Chris Hadfield endorsed marijuana use in space, whereas he was merely holding colorful eggs for an Easter ceremony[51].

6.3 Scenario Description

The teacher, acting as a conductor, introduces the scenario through a short video, demonstrating how to use the environment and the VLC tool. After right-clicking on an image and selecting the VLC button, the companion initiates a chatbot-style conversation, asking demographic questions and prompting the learner to express their opinion about the selected image. The learner must then vote on the image’s authenticity and justify their choice. The companion then reveals a ‘Recommended’ tab with links from Reverse Image Search (RIS) and Google Lens, guiding the students step by step.

If students evaluate or respond hastily without consulting the suggested RIS, they receive a cautionary message ‘Try not to be Biased!’. After exploring the RIS links, the companion provides additional input and asks if the user wants to revise their judgment. If the learner changes their vote or decides to participate in the second vote, the companion seeks an explanation.

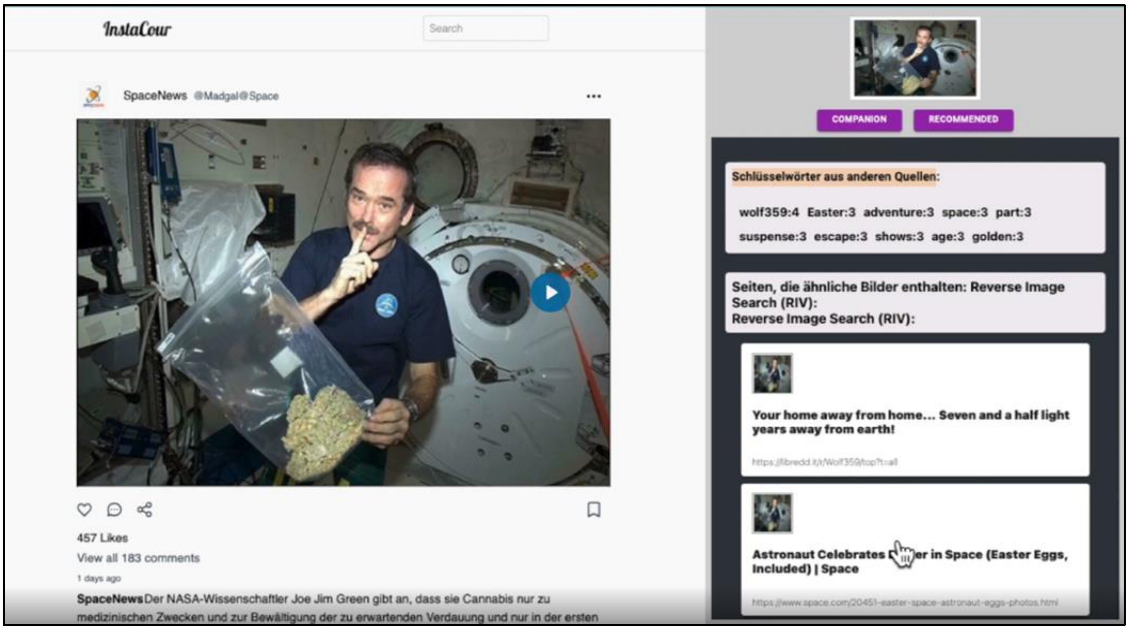
Following Figure 6-1 and Figure 6-2 illustrate various aspects of the InstaCour environment, the companion chat, and the retrieved images from Google Lens. In the next stage, the companion poses reflective questions and awards badges, presenting expert judgment and credible sources for each case. Finally, the VLC requests feedback on the system and any recommendations for improvements.

All the important sections of chat dialogue are illustrated in Flowchart of Figure 6-3.

[49] Urban Legends, https://www.pinterest.com/pin/361906520027529025/, Feb 2023

[50] National Geographic - Flying House,1. 5 Dec 2022 https://rb.gy/qcjxw, May 2023,

[51] Happy easter in Nasa, https://www.universetoday.com/101124/happy-easter-sunday-from-the-iss-crew-hunts-easter-eggs-goodies/, Dec 2022



SpaceNews Der NASA-Wissenschaftler Joe Jim Green gibt an, dass sie Cannabis nur zu medizinischen Zwecken und zur Bewältigung der zu erwartenden Verdauung und nur in der ersten

Schlüsselwörter aus anderen Quellen:

wolf359:4 Easter:3 adventure:3 space:3 part:3 suspense:3 escape:3 shows:3 age:3 golden:3

Seiten, die ähnliche Bilder enthalten: Reverse Image Search (RIV):
Reverse Image Search (RIV):

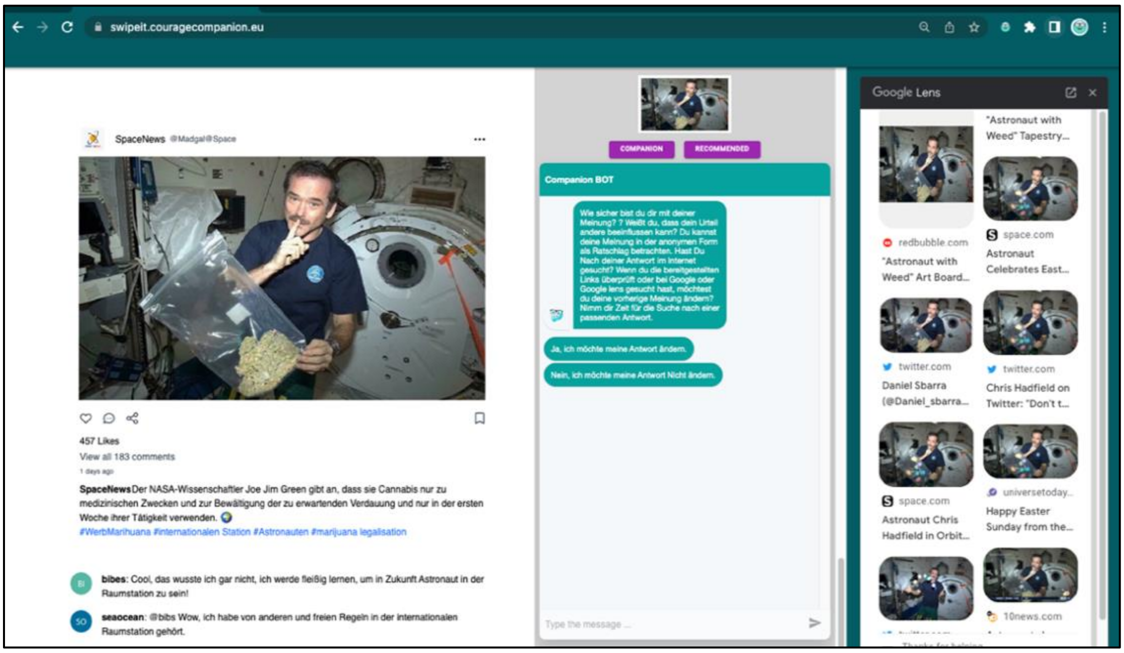
Your home away from home... Seven and a half light years away from earth!

https://libredd.it/r/Wolf359/top?t=all

Astronaut Celebrates Easter in Space (Easter Eggs, Included) | Space

https://www.space.com/20451-easter-space-astronaut-eggs-photos.html

Figure 6-1: The InstaCour environment displays an image along with its caption on the left side, while the right side features the VLC add-on. This add-on provides contextual information and a summary of keywords from various resources, sorted and organized based on the Reverse Image Search (RIS) stimulus case: Scientists use marijuana on the International Space Station (ISS)



SpaceNews Der NASA-Wissenschaftler Joe Jim Green gibt an, dass sie Cannabis nur zu medizinischen Zwecken und zur Bewältigung der zu erwartenden Verdauung und nur in der ersten Woche ihrer Tätigkeit verwenden. 🌍
#WerbMarihuana #internationalen Station #Astronauten #marijuana legalisation

bibes: Cool, das wusste ich gar nicht, ich werde fleißig lernen, um in Zukunft Astronaut in der Raumstation zu sein!

seaocean: @bibs Wow, ich habe von anderen und freien Regeln in der internationalen Raumstation gehört.

Wie sicher bist du dir mit deiner Meinung? ? Weißt du, dass dein Urteil andere beeinflussen kann? Du kannst deine Meinung in der anonymen Form als Ratschlag betrachten. Hast Du Nach deiner Antwort im Internet gesucht? Wenn du die bereitgestellten Links überprüft oder bei Google oder Google lens gesucht hast, möchtest du deine vorherige Meinung ändern? Nimm dir Zeit für die Suche nach einer passenden Antwort.

Ja, ich möchte meine Antwort ändern.

Nein, ich möchte meine Antwort Nicht ändern.

"Astronaut with Weed" Tapestry...

"Astronaut with Weed" Art Board...

Astronaut Celebrates East...

Daniel Sbarra (@Daniel_sbarra...

Chris Hadfield on Twitter: "Don't t...

Astronaut Chris Hadfield in Orbit...

Happy Easter Sunday from the...

Figure 6-2: The InstaCour environment showcases an image and its accompanying caption on the left side. On the right side, the companion chat is displayed, along with images retrieved from Google Lens, each representing a different website. As revealed in the credible links on the right side, the image was manipulated, and this was disclosed through the Easter eggs the astronauts carried.

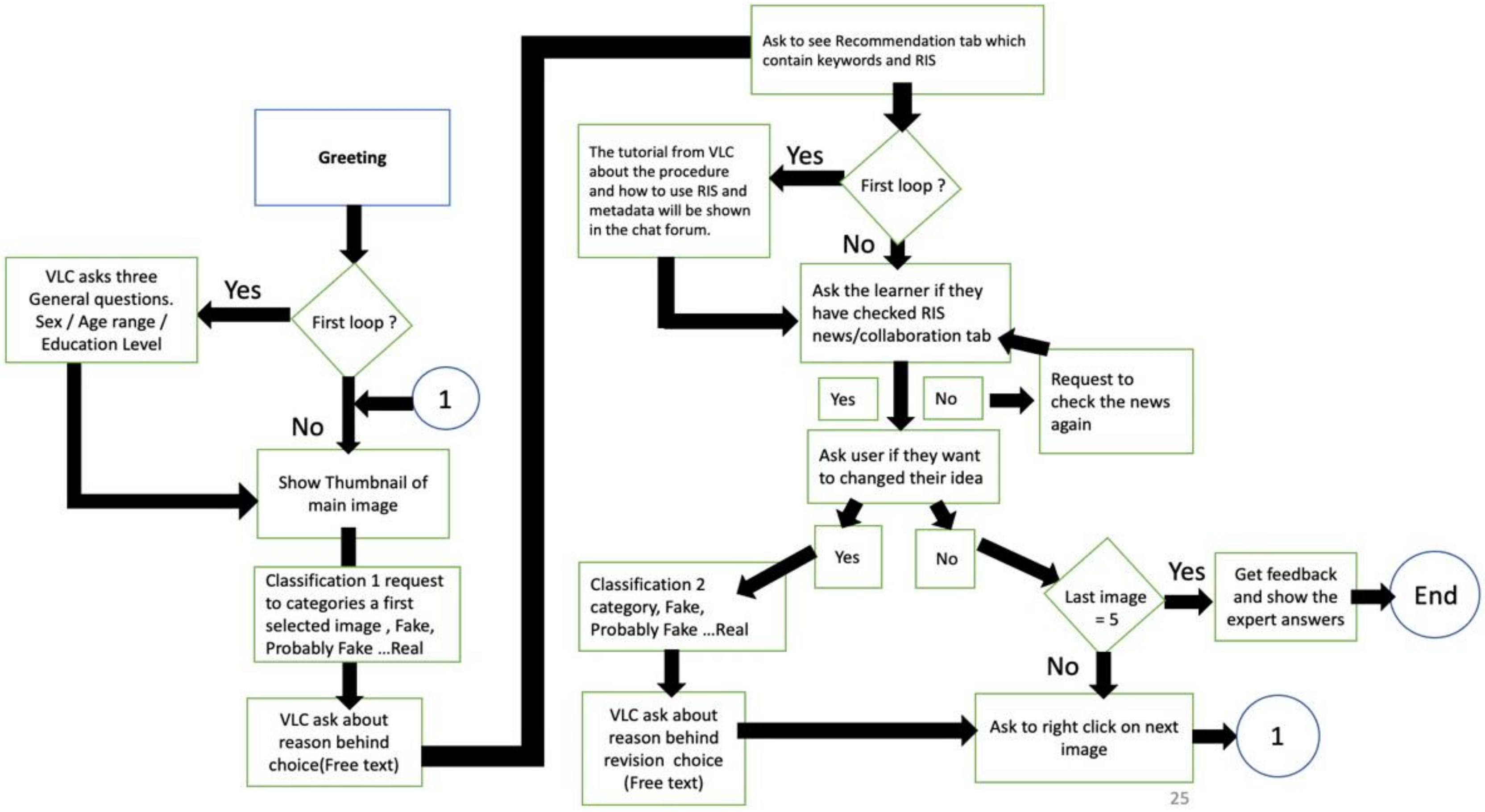
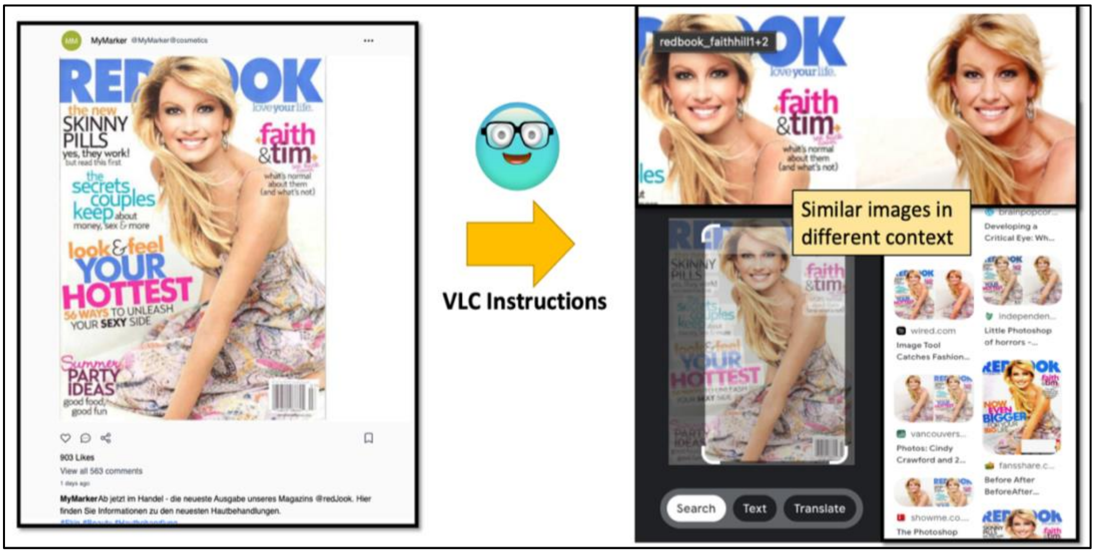

Greeting
First loop ?
Yes
VLC asks three General questions. Sex / Age range / Education Level
No
1
Show Thumbnail of main image
Classification 1 request to categories a first selected image , Fake, Probably Fake ...Real
VLC ask about reason behind choice(Free text)
Ask to see Recommendation tab which contain keywords and RIS
First loop ?
Yes
The tutorial from VLC about the procedure and how to use RIS and metadata will be shown in the chat forum.
No
Ask the learner if they have checked RIS news/collaboration tab
Yes
No
Request to check the news again
Ask user if they want to changed their idea
Yes
No
Classification 2 category, Fake, Probably Fake ...Real
VLC ask about reason behind revision choice (Free text)
Last image = 5
Yes
Get feedback and show the expert answers
End
No
Ask to right click on next image
1
25

Figure 6-3 Flowchart illustrating the key elements of the companion scenario in the Fake/Real image identifications.



MyMarker Ab jetzt im Handel - die neueste Ausgabe unseres Magazins @redJook. Hier finden Sie Informationen zu den neuesten Hautbehandlungen.

Figure 6-4: Find similar images with their link by the Google lens tool guided by the companion; users can see the truth behind the manipulation – stereotype case on social media.

In Above Figure 6-4 the companion guides the user to find similar images via the help of the Google Lens tool. Throughout the experiment, the VLC system records various data, including learner responses, interactions with environmental objects, and browser activity like opening Google Lens links connected to similar images. After the VLC trial, the constructor reveals the truth behind each stimulus and conducts an oral examination with researchers and organizers.

6.4 Evaluation

This section provides an overview of a school trial experiment's design. It covers classroom conditions, data collection methods, the nature of the data collected, analytical techniques, and terminology. The chapter then explains how logs and mathematical measurements will be utilized to address research questions.

6.5 Classroom Experiment

In June 2022, Gymnasium Wolfskuhle in Essen, Germany, hosted a single-day trial using the 'FakeOrReal' setting within the InstaCour environment. The experiment included 22 students, 15 boys and seven girls, aged between 13 and 15 (Figure 6-5). They were selected based on a data privacy agreement, signed by their parents, that allowed for their anonymous participation through individual access tokens. However, five students from the same course were excluded due to a lack of a signed agreement. Additionally, three more students were later excluded for failing to follow standard instructions; they either collaborated or searched the web for answers before classifying the stimulus.

The InstaCour environment is designed to randomize the presentation of stimuli, forcing each learner to engage with the task independently. This is because neighboring computers might display different image orders due to shuffling. This randomization ensures an even distribution and prevents any specific stimulus from always appearing first. The experiment resulted in 95 conversations with the VLC and recorded 2,056 unique interactions in the chatbot collection database. During the 50-minute experiment, 228 unique tabs were opened, averaging 12 tabs per user, or roughly 10 minutes per image.

Figure 6-5: Some images showcase a classroom experiment in featuring a companion. In the top left, there is a depiction of live control over anonymous logs within a cloud-based Learning Locker environment (Gymnasium Wolfskuhle in Essen).

Feedback from Learners

After finishing a scenario run, the companion asked users to comment on the system's features and suggest improvements. Nine students answered the initial question regarding what they learned from the system, with six affirming and three denying any learning. When asked about potential improvements or criticisms, nine students responded again, with seven offering positive remarks like "All fine, you are a very helpful companion", “No, I have no suggestions,” and “Not at all, you are perfect!” Two students, however, found fault with the VLC’s chat process, as it shouldn’t always repeat the same questions for each stimulus. They also noted that conversations could be briefer, particularly in the second to fifth rounds. The students pointed out that the VLC dialogue and the recurring companion window obstructed swift research among RIS links and replies. Following the experiment, an oral session was conducted with the students, who praised the VLC and scenario as both engaging and educational. They struggled to verify the authenticity of an image and caption of John Lennon due to the absence of text explanations in the content provided by RIS resources. Even with the information at hand, the students found it challenging to discern the truth, lacking historical context on the cases of *Che Guevara* and *John Lennon* for the younger generation could be the reason for that.

6.6 Analytical Measures for School Trial Experiment

This section outlines the analytical tools used to evaluate the school trial experiment. It begins with the definition of key terms and concepts necessary for the analysis, followed by an explanation of the methods and calculations involved.

Voting Scores: Voting scores are the pre-established numerical values corresponding to each label, as illustrated in Table 6-2:

Table 6-2: Credibility categories and their respective scores

Fake	Probably Fake	Not Sure	Probably Real	Real
-2	-1	0	1	2

Expert Score (ES): The Expert Score is allocated to social media images that have been examined and classified by specialists. For example, the NASA-Marijuana image (referred to as the Mari stimulus) has been marked as fake by experts, resulting in an expert score (ES) of (-2).

Vote Records and Assessments: User-tokens and votes are grouped into separate categories based on their involvement in image assessment (stimulus index), forming a distinct record/tuple as follows:

Record/Tuple: {stimulus (index), user-token, vote}

To address the specific research inquiries, examining the number of users and their votes for each stimulus is necessary. Users can take part in the initial voting round (Vote1), which occurs before students observe the recommendations given by VLC. A second round of voting (Vote2 or revision) takes place after engaging with the RIS, recommendations, and metadata for each stimulus through companion dialogue.

As noted, 22 students participated in the experiment. However, the final analysis included only 19 students after outliers were excluded. In total, there are 129 distinct records or tuples formed by combining the stimulus, user-token, and vote data. The majority of users took part in each stimulus assessment for the first vote, resulting in 92 records. For the revision vote, there are 37 records. The count of records (tuples) for each stimulus and their assessments can be found in below Table 6-3:

Table 6-3: Count of records (tuples) during the first vote and subsequent revisions.

Stimulus / Index	Lennon	Mari	Magazine	Dali	Flying House	Total
First votes	18	18	19	18	19	92
Revision votes	3	6	9	9	10	37

Expert Distance (ED): Expert Distance refers to the absolute difference between the expert's evaluation score and the learner's voting score for a specific stimulus.

$$\boldsymbol{ED} = |\, \boldsymbol{ES_{in}} - \boldsymbol{Learner's\ Vote\ score_{in}}\, |$$

Max ED = 4 and Min ED = 0 (based on defined scores), 0 < ED < 4

$\boldsymbol{i}$: the index of the image (stimulus)
$\boldsymbol{n}$: the index of the learner's token

Change per User (Improvement/Deterioration): IPU is a per-user, per-image metric that can be either positive (improvement) or negative (deterioration) for each stimulus. It is calculated by subtracting the second user's vote score expert distance ($\mathbf{ED_2}$) from the first vote score expert distance ($\mathbf{ED_1}$).

$$\boldsymbol{IPU} = \boldsymbol{ED_{2ni}} - \boldsymbol{ED_{1ni}}$$
$$-\mathbf{4} < \boldsymbol{IPU_{ni}} < \mathbf{4}$$

$\boldsymbol{n}$: the index of the learner's token
$\boldsymbol{i}$: the index of the image (stimulus 1-5)

Improvement per Image for All Users ($\boldsymbol{IPI_i}$):
This is the sum of **IPU** across all participants for a particular stimulus.

$$\boldsymbol{Improvements\ per\ image}\ (\boldsymbol{IPI_i}) = \sum_{n=1}^{N} \boldsymbol{IPU_{ni}}$$

$\boldsymbol{i}$: the index of the image (stimulus)
$\boldsymbol{n}$: the index of the learner's token
$\boldsymbol{N}$: the total number of participants in the simulated environment (in this experiment, N_{max}=19)

Relative Improvement Per Image (RIPI): RIPI is based on the sum of improvements per image, considering the number of participants and votes for each stimulus:

$$\boldsymbol{Relative\,IPI\ (RIPI)\ =\ IPI\ *\ \frac{r_i}{R}}$$

i: the index of the image (stimulus)
***r*:** sum of records (tuple) per stimulus for revision vote,
R: total of records in revision vote **(R=**37 total,).

Improvement for All Images for a User ($\boldsymbol{IAI_n}$): This is the sum of the IPU of a user for all images.

$$\boldsymbol{Improvments\ for\ all\ images\ (IAI_i)\ = \sum_{i=1}^{5} IPU_{ni}}$$

i: the index of the image (stimulus)
n: the index of the learner's token

Total Improvement (TI): TI can be calculated from different aspects that lead to the same result. It includes the sum of Improvements Per Image (IPI) for five images, the sum of the Improvement Per User (IPU) for all 92 user's votes, and the sum of Improvements for All Images (IAI) for all users (19), all leading to the same total improvement.

$$\boldsymbol{TI = \sum_{i=1}^{5} IPI_i\ = \sum_{v=1}^{92} IPU_l = \sum_{n=1}^{N} IAI_n}$$

v: the index of the learner's vote
i: the index of the image (stimulus)
n: the index of the user
N: the total number of participants in the simulated environment (in this experiment, N_{max}=19)

Correlation Measures: The analysis employs the Pearson correlation coefficient as the main method for calculating the correlation between indexes. However, the Spearman correlation coefficient is used in some cases, as it is more efficient for small sample sizes.

Analysis Conclusions: In the initial judgment or vote, there were 92 votes, while the second judgment (revision) after VLC's suggestion to consult other sources resulted in 37 votes. Among these 37 revision votes, 12 were influenced by the metadata from the Google Lens sidebar, 22 were based on RIS, metadata, and keywords from the companion, and the remaining three votes did not have a specified reason for the revision.

IPI Calculation and Outcomes: The total IPI for all five images was 71, yielding an average improvement of 14.2 per image.

Table 6-4: Votes, IPI, and Relative IPI (RIPI) per Image**,** 19 participants total, 129 interactions, 92/37 = 2.4, 24% (Participation Rate in revision).

Stimulus Index	Lennon	Mari	Magazine	Dali	Flying House	Total
First votes	18	18	19	18	19	92
Revision votes	3	6	9	9	10	37
IPI	1	12	18	19	21	71 (TI)
RIPI	0.08	1.945	4.37	4.62	5.67	16.702

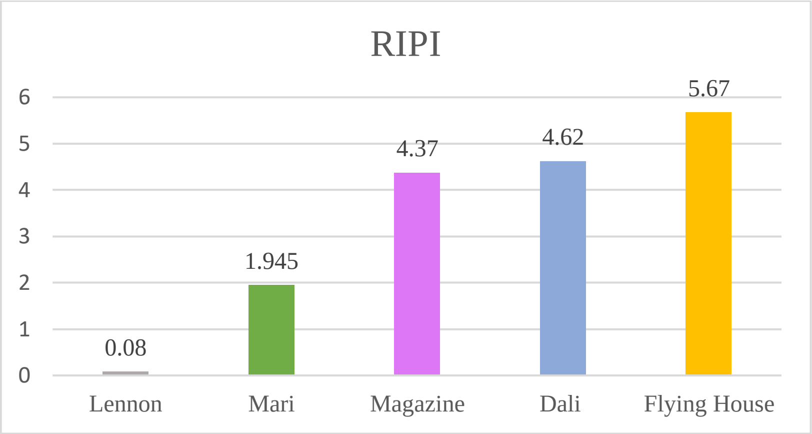


Figure 6-6: Relative Improvements Per Image Index (RIPI) as a Bar Chart.

RIPI is considered the primary method to categorize the improvement per image. Since RIPI also considered the number of participants for each stimulus, it was chosen over the IPI index (Table 6-4 and Figure 6-6).

Most students found the ‘Lennon’ case to be the most difficult to judge, as it had less credible information. This aligns with the calculated RIPI value.
Alternative Approach to Calculating Improvement: Another way to calculate improvement is to consider the average of expert distances for each stimulus, denoted as AED:

$$\boldsymbol{Average\ of\ Expert\ Distances_{in} = AED = \frac{1}{N} * \sum_{n=0}^{N} ED_{in}}$$

i: The index of the image (stimulus)

N: The total number of participants in the target stimulus

The AED for the first vote describes the stimulus difficulty. The analysis shows that the order of AED difference for 92 tuples is consistent with the improvement seen in the analysis for RIPI, indicating that both approaches yield the same result in terms of improvement calculation.

Table 6-5: Average of expert distance per image, REV = 1 subgroup participated in the revision, and REV=0 decided not to participate in the revision and other related indexes from the analysis.

Subgroups Index	AED (First vote), difficulty	AED (Revision)	AED First vote	AED First vote	AED Revision
Revision status (REV 0 or 1)/ All records	All records	All records	Subgroup REV=0	Subgroup REV=1	Subgroup REV=1
Total Count of Unique Records	92	92	55	37	37
Lennon	1.88	1.79	2.20	3	2.00
Mari	0.83	0.28	1	1.6	0
Magazine	1.84	0.89	1.82	2	0
Dali	2.61	1.56	2.27	2.66	0.56
Flying House	1.89	0.79	1.38	2.70	0.60
Overall Share of Deterministic Votes (Fake, Real)	40 43%	61 66%	31 56%	9 24%	30 81%

Results of Participant Judgments and Interaction with Web Content: This section outlines the findings from the analysis of participants' judgments and their engagement with web content, guided by the VLC in image classification tasks.

- **REV=0 Subgroup:** This subgroup includes participants who were confident in their initial judgment (vote) and did not partake in revisions for that specific stimulus. This group contains 55 unique records.

- **REV=1 Subgroup:** This subgroup consists of participants who were uncertain about their initial judgment and participated in revisions after interacting with the VLC recommendations. This group encompasses 37 unique records.

In Table 6-5 column AED (revision) with the status 'All records' is used to maintain a constant tuple amount for AED calculation. For those who did not participate in the revision, their first votes were applied as their revision votes, keeping the total vote count at 92. In the same (Table 6-5), column AED with 'REV=1' considered the group of records that participated in the revision (REV=1, a total of 37 unique records).

Analysis of REV=1 Subgroup: For the cases ‘Mari’ (six votes) and ‘Magazine’ (receiving nine votes), all second choices aligned with the expert vote, resulting in an Average Expert Distance (AED) of zero. For other stimuli, there was a reduction in the Expert Distance (ED) value, indicating an improvement.

The average expert distance score was calculated for five stimuli based on the total number of records. Those who participated in the revision (37 votes) had varying AED averages depending on whether their first or second vote was considered. The results showed that participants who altered their vote following the VLC interaction achieved better alignment with expert judgment compared to those who retained their initial vote.

Total Improvement (TI) Analysis:

- For the REV=0 subgroup, TI = 1.43.
- For the REV=1 subgroup, TI for their second vote is 0.45, indicating a nearly one step improvement rate (0.98) towards expert judgment for those engaging in the revision.

The total improvement for learners who only participated in the initial judgment was 1.43 (REV=0), while for those who participated in the revision (REV=1), the TI for their first vote was 2.37.

Statistical Correlation: The analysis indicated a statistically significant positive correlation between ‘Revision Participation status (REV=0 or 1)’ and ‘Expert Distance’ (ED, 0 - 4) in the initial judgment across all records, r(90) = .313, p = .002. This pattern suggests that learners who did not revise their initial judgments (REV=0) had a better average score (average ED = 1.43) compared to those who engaged in revisions (average ED = 2.37). Conversely, individuals with a higher expert distance in their initial judgments (REV=1) were more inclined to revise their judgments after receiving VLC recommendations, unlike those who did not revise for a specific stimulus.

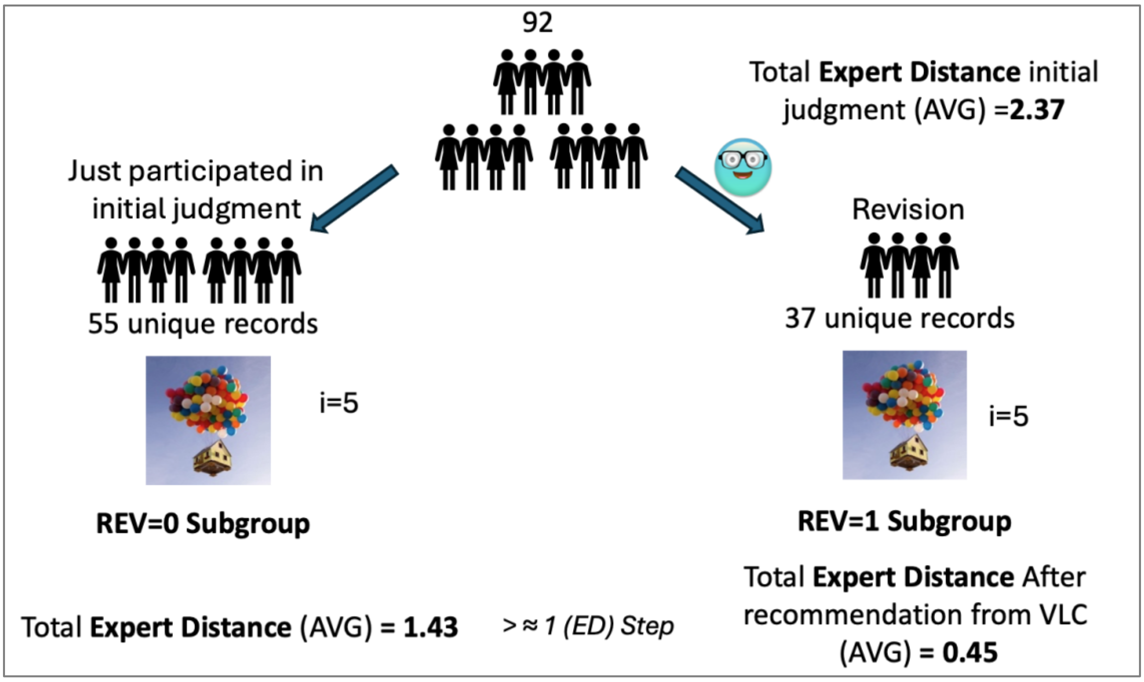

92
Total Expert Distance initial judgment (AVG) =2.37
Just participated in initial judgment
55 unique records
i=5
REV=0 Subgroup
Revision
37 unique records
i=5
REV=1 Subgroup
Total Expert Distance (AVG) = 1.43
> ≈ 1 (ED) Step
Total Expert Distance After recommendation from VLC (AVG) = 0.45

Figure 6-7 Correlation Between Revision Participation and Judgment Accuracy. Learners confident in their initial judgments (REV=0) had a lower average Expert Distance (ED) of 1.43, while those revising (REV=1) had a higher initial ED of 2.37. A significant positive correlation (r (90)= .31, $p < 0.05$) suggests that learners with less accurate initial judgments are more inclined to revise after VLC recommendations.

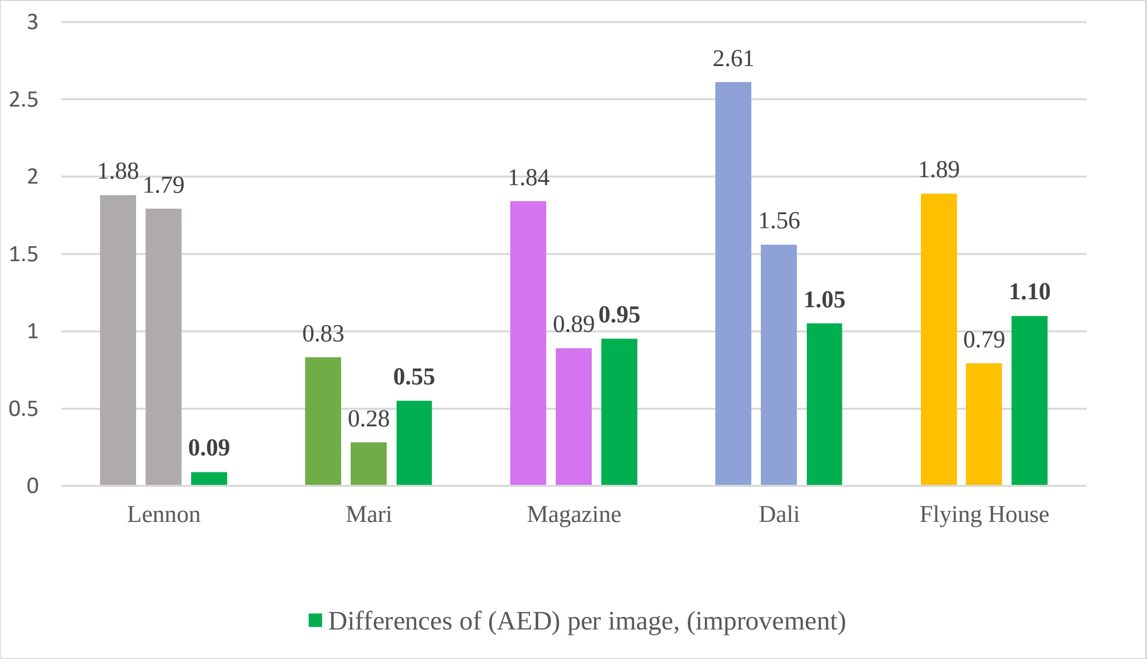


Figure 6-8: Average of Expert Distances (AED) for All Records (92), Range 0<AED<4.

Average of Expert Distances: Figure 6-8 illustrates the AED for the learners before and after interaction with the VLC recommendations and instructions, as well as the difference for each of the five images. The dark green portion in Figure 6-8 shows the difference in AED between the initial vote and the revision. It reveals that for two images, ‘Flying House’ and ‘Dali,’ it was challenging to judge them as ‘real’ before accessing information from RIS and metadata. However, due to abundant metadata online from credible sources, most learners who revised their answers identified the correct response, which was ‘Real’ for both images.

Statistical Data and Observations: The data for each image demonstrate that VLC’s web-based RIS recommendations enhance learners’ judgments by bringing them closer to expert opinions. Key observations from this analysis include:

- Out of 37 revision votes, 35 were in line with expert judgment, showing improvement.
- On average, each user’s revision improved by a score of 3.84 according to expert judgments.
- 17 of the 19 users revised their judgment for at least one image, with maximum participation in revisions for two users across four images, and minimum participation for five users with one image. On average, users revised 2.1 out of 5 images.

Deterministic Votes Count and Changes: Table 6-5’s last row reveals the frequency of users choosing deterministic categories (‘Fake’ and ‘Real’) versus uncertain categories (‘probably fake,’ ‘not sure,’ ‘probably real’). In the revision subgroup (REV=1), the use of definitive votes increased from 24% to 81%, marking a 57% growth in the use of definitive votes during revision.

Relationship Between RIPI and Time Consumption: The case with the lowest RIPI score, ‘Lennon,’ had the least total time spent by learners, at 2696 seconds. Conversely, ‘Flying House’ had the most time spent, at 4295 seconds (Table 6-6, Figure 6-9), and also showed the most significant improvement. This suggests that investing more time in evaluating each stimulus may lead to more accurate answers, as gauged by the expert distance ES in this study.

Correlation Between Improvements and Time: If we consider the five images and calculate the correlation between improvements in IPU and time per user (with a sample size of 19), the following results are obtained:

Table 6-6: The correlation between IPU (Improvement Per User) and time spent per user is analyzed based on user logs. In this analysis, the green-labeled p-values meet the criteria for statistical significance, leading to the rejection of the null hypothesis.

Stimulus Index	Lennon	Mari	Magazine	Dali	Flying House
Total Time (sec)	2696(minimum)	3369	4240	3556	4295(max)
RIPI	0.04(minimum)	1.94	4.37	4.62	5.67(max)
Corr. IPU and time of sample size = 19	-0.06 Weak Negative	0.1 Weak Positive	0.48 Moderate Positive	0.58 Strong Positive	0.74 Strong Positive
P-value	0.7994	0.6701	0.03666	0.008751	0.0002277

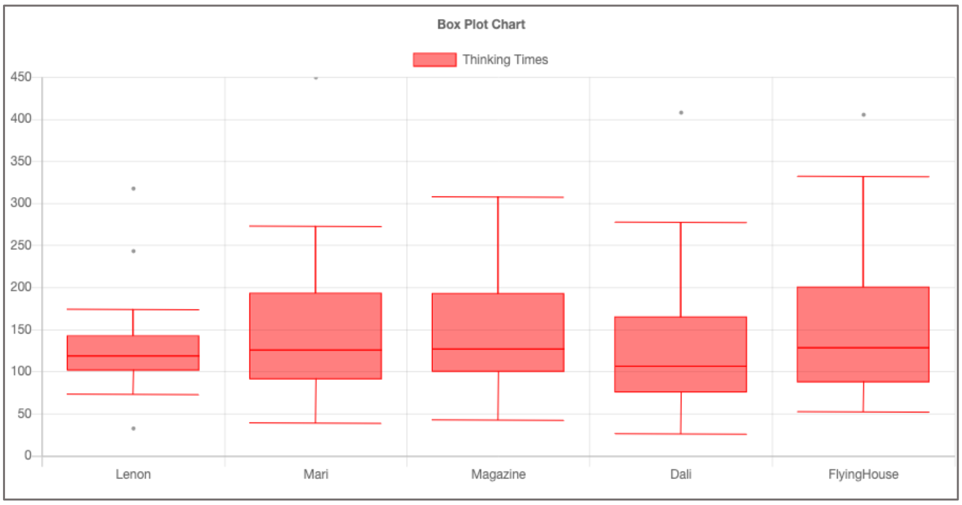


Figure 6-9: Time consumption as recorded in the Learning Locker data store, based on xAPI statements logs.

The analysis of learners' logs, comprising 92 records (Figure 6-9), revealed a moderate and positive correlation between improvement (IPU) and time spent on tasks, r(90) = .48, p < .000001. This significant association is depicted in Figure 6.10, which includes a regression chart illustrating the relationship between improvement and time.

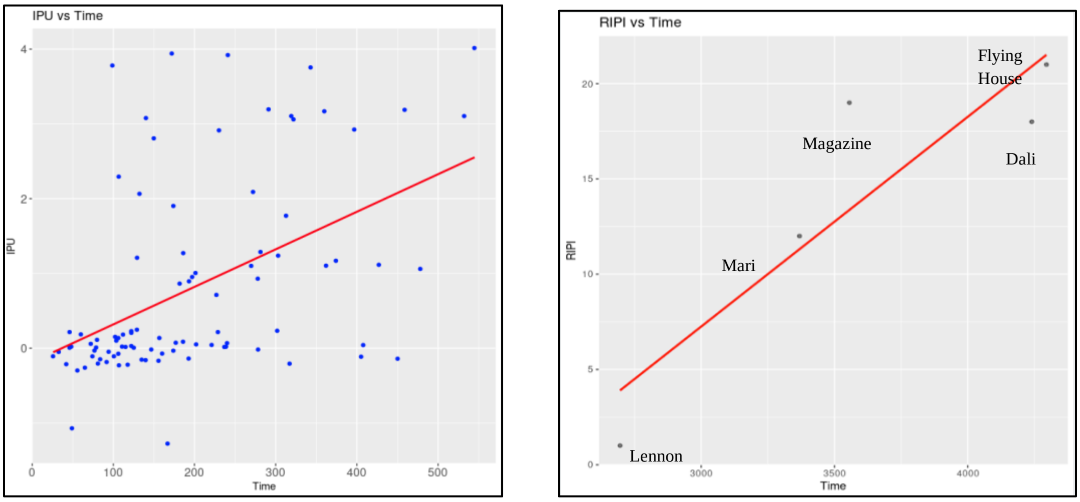


Figure 6-10: The regression chart illustrating improvement and time consists of two plots. The left plot displays the data for individual user logs, encompassing 92 records (IPU), while the right plot contrasts the improvement (IPI) for each image against the total time spent on that image.

In summary, the data reveal a positive correlation between the time invested in a task and the rate of improvement. This is a relationship that has been found to be statistically meaningful. Additionally, spending more time on each stimulus, particularly when interacting with the VLC instructions based on RIS, contributes to greater refinement in the revision process. As for the level of agreement, the findings indicate that guidance from the VLC in interacting with web content leads to more uniform judgments and a reduction in conflicting decisions (indicating higher agreement), as seen in the right section of Figure 6-11. This pattern was consistent across different stimuli and was evident in both initial votes and revision votes (REV = 0,1).

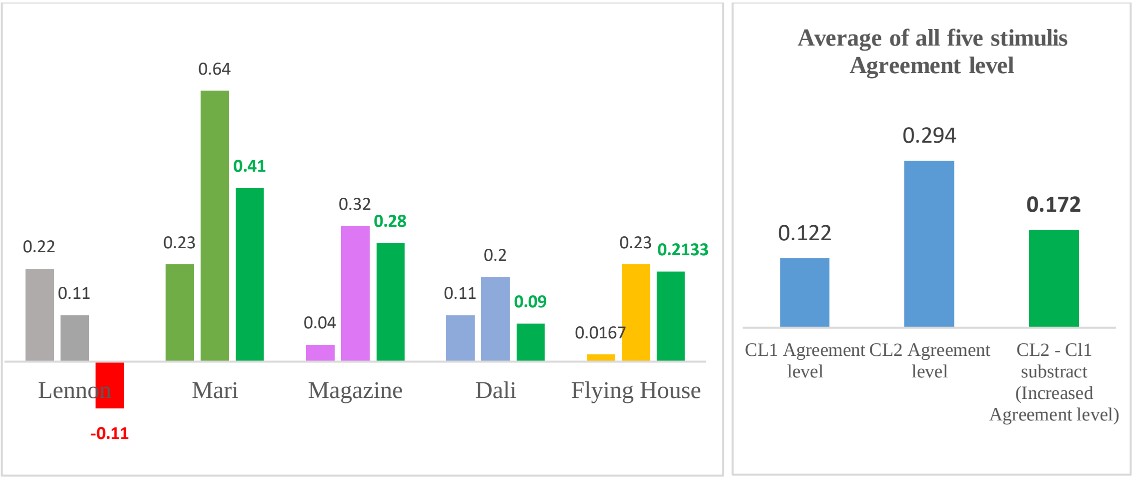


Figure 6-11: Left: the outcome showing the agreement level (1 – DI the dispersion index) for all records, both initial and revised (REV 0,1). Right: a comparison between the first and second votes, highlighting their differences.

A notable observation from the previous calculations reveals that the most difficult task, labeled as 'Lennon,' led to a decrease in the agreement index by (-0.11) following interaction with RIS. Conversely, tasks that were simpler in finding credible content through RIS experienced an increase in the agreement index. Specifically, the 'Flying House' scenario saw an increase of (+0.21) in the agreement level, and the 'Mari' case increased by (+0.41).

On average for all five the agreement index increased by **17.2%**. Additionally, when looking at the average agreement across different subgroups, those who participated in the revision (REV=1, consisting of 37 records) demonstrated a significantly higher agreement level (0.60) compared to the subgroup that did not engage in the revision (REV=0, consisting of 55 records), which had an agreement level of (0.15), as detailed in Table 6-7 below:

Table 6-7: Showcases the differences in agreement measures and levels across various subgroups.

Stimulus Subgroups	Lennon	Mari	Magazine	Dali	Flying House	Avg
First vote, Agreement REV=0 (55 records)	0.32	0.25	0.01	0.11	0.07	0.15
First vote, Agreement REV=1 (37 records)	1	0.016	0.3	0.13	0.07	0.30
Revision, Agreement REV=1 (37 records)	0.16	1	1	0.54	0.275	0.60
Difference of Agreements REV=1 (37 records)	-0.84	0.984	0.7	0.36	0.67	0.30

An analysis of the correlation between the number of opened tabs and the time spent on each tab was conducted using data from 95 observations. The findings indicated a non-significant correlation between these two variables, r(93) =.13. This negligible correlation may be due to the variance in content types across the opened tabs; some tabs could be text-heavy, while others may be predominantly visual. Such differences in content type could influence the amount of time a learner requires to process the information in each tab.

Table 6-8: Comparison of the duration, enhancement index, and the quantity of opened tabs (OT) for each stimulus.

Stimulus Index	Lennon	Mari	Magazine	Dali	Flying House
Time (Seconds)	2696 (Min Sec)	3369	4240	3556	4295 (Max Sec)
Opened tab (OT)	54 (Max OT)	45	36	50	29 (Min OT)

Consequently, the analysis recommends evaluating the total duration and the complete count of opened tabs individually for every stimulus. Among the participants, 214 distinct tabs were opened, averaging (11.26) per user, with a similar distribution between male and female participants. If learners find trustworthy and credible information in the initial RIS and Google Lens tab, they are less likely to open additional tabs to seek the answer. Conversely, the study shows that when RIS content is vague and lacks textual information, more external tabs are likely to be opened.

As detailed in Table 6-8, the ‘Lennon’ stimulus had the most opened tabs and the least total time and (RIPI) improvement rate, mainly because the RIS for this stimulus lacked sufficient textual information. In contrast, the ‘Flying House’ stimulus, which had credible resources recommended, had the fewest opened tabs and the highest total time and improvement rate.

An examination of the data for each image revealed a negative correlation between the changes in the number of opened tabs and the time spent on all stimuli, except for ‘Dali.’ The study concludes that the count of opened tabs serves as a valuable measure of the abundance and quality of online information for a specific stimulus. A higher number of opened tabs may signal a lack of textual information in the RIS, possibly resulting in a diminished improvement rate.

6.7 Scientific Contribution and Relevance for ‘Fake or Real’ Study

Critical Thinking Across Disciplines: The VLC nurtures essential skills for interpreting scientific and other claims propagated on social media.

Information / Media Literacy: In the digital age, VLC offers a robust method for students to verify the credibility and relevance of data and develop media literacy skills.

Empowering Independent Learning: VLC supports modern pedagogy’s emphasis on self-directed learning, aiding students in information verification during independent study.

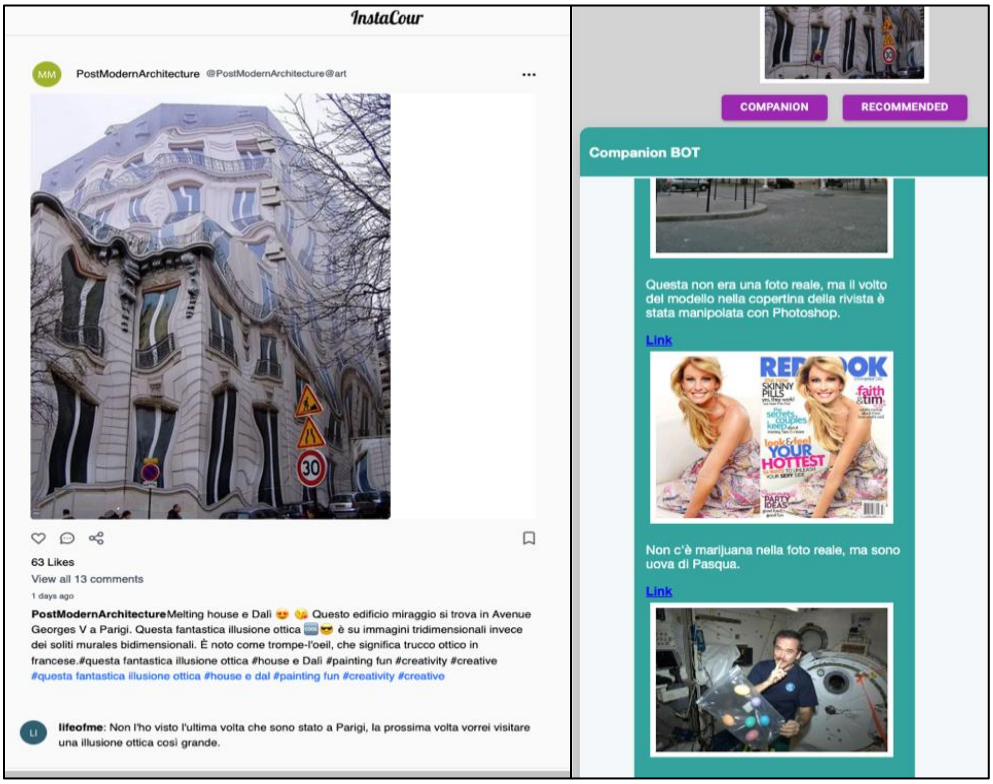
Ethical and Responsible Behavior: The tool equips students to be responsible digital citizens, facilitating ethical decision-making and informed debates.

Interdisciplinary Applications: VLC can be adapted for specific academic requirements, enhancing the rigor of student research across disciplines.

In summary, the research cultivates an informed, critically thinking learner, fulfilling both media literacy objectives and broader educational goals across various subjects.

6.8 University Classroom Study Fake/Real (Italy)

As explained in the introduction of this chapter in June 2023, we collaborated with CNR, the Italian partner of the Courage Project, to test the VLC at LUMSA[52] University in Italy. The study involved a total of 32 active participants. The stimuli in the InstaCour environment replica of the trial in Germany [5] and the VLC dialogue and scenario were the same as in Germany but in Italian language (Figure 6-12).



PostModernArchitecture @PostModernArchitecture@art

63 Likes

View all 13 comments

1 days ago

PostModernArchitecture Melting house e Dalì 😍 😘 Questo edificio miraggio si trova in Avenue Georges V a Parigi. Questa fantastica illusione ottica 🆒😎 è su immagini tridimensionali invece dei soliti murales bidimensionali. È noto come trompe-l'oeil, che significa trucco ottico in francese.#questa fantastica illusione ottica #house e Dalì #painting fun #creativity #creative #questa fantastica illusione ottica #house e dal #painting fun #creativity #creative

lifeofme: Non l'ho visto l'ultima volta che sono stato a Parigi, la prossima volta vorrei visitare una illusione ottica così grande.

Questa non era una foto reale, ma il volto del modello nella copertina della rivista è stata manipolata con Photoshop.

Non c'è marijuana nella foto reale, ma sono uova di Pasqua.

Figure 6-12 Italian version of the VLC and InstaCour. On the right side, the user has surfed and finished with all cases on InstaCour, and the companion explains the correct answer based on experts' votes for each image and the reference link, ensuring that participants are aware of the correct answers.

6.9 Trial Setup at LUMSA University

The study was conducted in two sessions, with 15 participants on June 8, 2023, and the remaining 17 on June 27, 2023. The participants were students aged 18 to 24, majoring in psychology and educational science.

During our study at LUMSA University, we encountered a limitation with installing the VLC plugin on the university's PCs due to security policies. To overcome this challenge, we provided Chromium Portable for Windows and installed the VLC plugin on this

LUMSA University, LUMSA's website: https://www.lumsa.it/, Aug 2023 portable version of the browser as we managed for trial in Germany. This solution allowed us to bypass the installation restrictions on the university's computers, ensuring that the study could proceed as planned without violating any institutional guidelines. It demonstrated our ability to adapt and problem-solve within the constraints of the research environment.

The study yielded 449 reactions in the chat. Of these, 192 interactions were related to the classification of images. This included 152 initial judgments and 44 revisions, as depicted in Table 6-9 compared to Germany, Italy had a 10% higher participation rate in revisions compared to Germany, with 34% in Italy and 24% in Germany, as shown in Table 6-4. Improvement was observed in all cases except for the 'John Lennon' stimulus, which had weak metadata and naive resources on the web, making judgment challenging.

Table 6-9: Votes, IPI, and Relative IPI (RIPI) per Image**,** 32 participants, total= 192 interactions, 152 (first vote)/44(REV) = 34, 34% (Participation Rate in REV).

Stimulus Index	Lennon	Mari	Magazine	Dali	Flying House	Total
First votes	29	30	30	31	32	152
Revision votes	9	7	10	6	12	44
IPI	-5	13	24	10	30	72 (TI)
RIPI	0.08	1.945	4.37	4.62	5.67	16.7

6.10 Evaluation and Comparative Analysis of Both Studies

The analysis of the comparison of the effect of VLC on each country, Germany and Italy, revealed the following results in Figure 6-13:

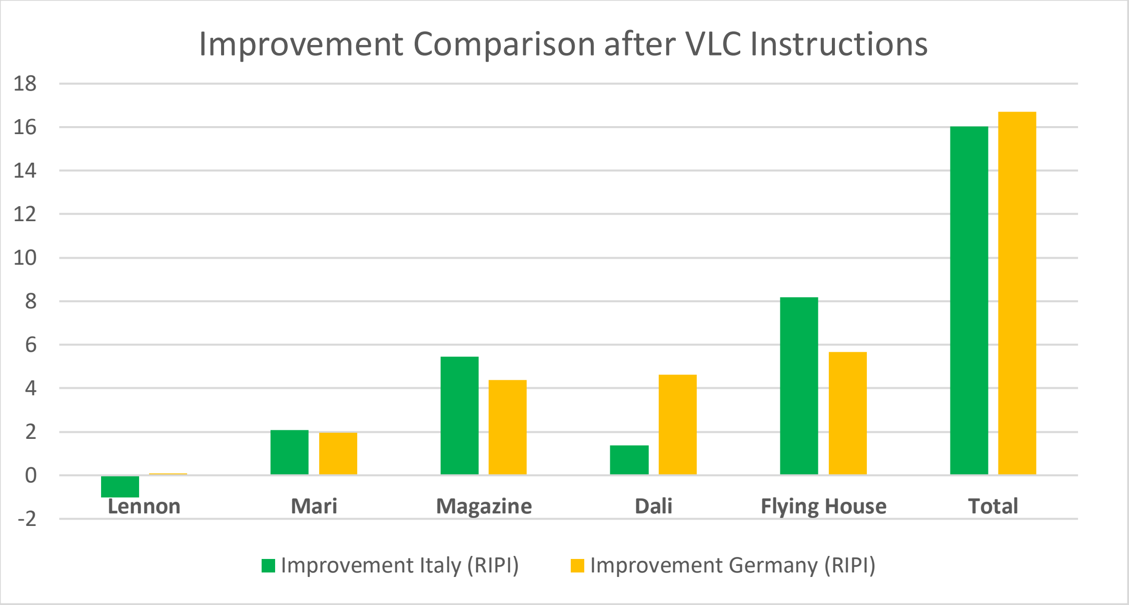


Figure 6-13: Relative Improvement Per Image (RIPI) Amount for Each Stimulus, Respectively for Italy and Germany. On the Right Side, the sum of the Improvements.

Table 6-10: Exact improvement values (RIPI index) after interaction with VLC recommendations in Germany and Italy.

Comparison	Lennon	Mari	Magazine	Dali	Flying House	Total
Improvement Italy (RIPI)	-1.02	2.06	5.45	1.36	8.18	16.04
Improvement Germany (RIPI)	0.08	1.94	4.37	4.62	5.67	16.70

As we can see in Table 6-10 and Figure 6-13, Germany showed a higher improvement rate after normalization with the population rate (RIPI =16.70), with the most significant difference observed in the 'Dali' stimulus.

In the analysis, we noted the similarity of the results and differences in total improvement, which was almost the same at around 16. The tendency to improve among stimuli in all cases, except 'Dali', was different. The news provided in RIS for 'Dali' was mostly in English, and in Germany, we had better results after encountering this case.

Statistical Analysis: A Pearson correlation analysis between 'Revision Participation status (REV=0 or 1)' and 'Expert Distance' (ED) for the initial judgment, using 153 records, revealed a weak, positive linear relationship, r(151) = 0.161, p = .04784. Although this correlation is statistically significant at the 5% level, the strength of the relationship is low, indicating that while revision participation and expert distance are statistically related, the practical significance of this relationship may be limited depending on the context.

6.11 Summary of Results

In summary, the empirical evidence gathered provides answers to the initial research questions as follows:

RQ1: Was the VLC effective for engaging learners in a high school classroom in meaningful interactions with stimulus material in a simulated social media environment?

Based on feedback from learners [Classroom Experiment] also the oral questionnaire via teacher after the tests, both the instructor and the learners indicated that the trial worked as planned and expected. They were engaged with the companion and the procedure based on the oral discussion. In addition, interaction with the stimuli in the provided InstaCour and the VLC worked as planned in the school lab. According to the log files, all participants were able to complete the initial judgments for their chosen stimuli. The total and average time spent indicates that each learner spent 16 minutes engaging in VLC conversations and making judgments.

The system could successfully manage both the VLC and the learners' interactions according to the scenario simultaneously. It stored their interaction logs in dedicated databases, cloud-based Learning Locker, and MongoDB in different predefined structures.

RQ2: Was the judgment accuracy positively influenced by the time spent visiting or reading web pages related to the target picture?

Before considering the impact of time spent on improvement, we need to determine the improvement based on the expert judgment score (ES), which serves as a ground truth. This is especially relevant for challenging stimuli. To determine the difficulty of the stimulus based on learners' classification, we looked at their initial judgment scores. These were then compared to ES. The difference between the ES and the learners' initial votes was calculated to represent the 'difficulty per image'. This metric offers insights into how challenging it is for learners to correctly assess an image's credibility.
According to these assumptions and calculations, we found that spending more time evaluating results from the RIS feature offered by the VLC can improve judgment. This is likely because credible resources often contain more detailed and richer textual content.

This Our findings provide support for the hypothesis, indicating a strong correlation between the time spent and improvement in judgment for tasks with high difficulty indices, such as those involving the 'Dali' and 'Flying House' stimuli. Further analysis of 92 logs revealed a moderate, positive correlation between the improvement per user

(IPU) and the time spent on the task, r(90) = .48. This suggests that investing additional time in providing recommendations is associated with more accurate outcomes.

RQ3: Did a revision of the initial judgment guided by the VLC lead to higher agreement with the expert opinion and with other participants' judgments (convergence)?

Yes, the analysis indicates that leveraging the VLC to revise initial judgments has led to a marked increase in agreement with expert opinions and among participants themselves. In five cases studied, there was an observed enhancement in consensus by an average of 17.2%. Additionally, there was a notable shift towards the adoption of definitive labels such as 'Fake' or 'Real', as opposed to more uncertain options like 'Probably Real', 'Probably Fake', or 'Not Sure'. This change suggests a 65% average increase in the use of deterministic labels, indicating a higher level of agreement in classifications among participants.

RQ4: Are the judgments from Italy aligned with those from Germany? Does one's cultural background and native language have an impact on how they assess the credibility of an image?

The judgments from Italy and Germany appear to be largely aligned, as indicated by the comparative analysis. Both countries showed improvement in judgments, with similar total improvement rates (RIPI index amount was around 16). However, there were some differences, most notably in the 'Dali' stimulus, where Germany had better results. This could be attributed to the news provided in RIS for 'Dali' being mostly in English. The study does not provide explicit evidence that cultural background and native language have a significant impact on how participants assess the credibility of an image. However, the discrepancy in the 'Dali' stimulus suggests that language could be a contributing factor.

In general, we can assert that search engine results have a more significant influence on learners' judgments during the fact-checking process than do cultural backgrounds in both Italy and Germany. In general, we can assert that search engine results have a more significant influence on learners' judgments during the fact-checking process than do cultural backgrounds in both Italy and Germany.

RQ5: Are the classifications made by participants arbitrary, considering that the credits are not based on the quality of their answers? Or do the participants genuinely strive to categorize the stimuli in a meaningful and appropriate manner?

The judgments made by participants in Italy do not appear to be arbitrary. The improvement rate amount for stimuli is similar to that observed in Germany as explained, which suggests that the answers are not given by chance. Furthermore, qualitative answers to the VLC and discussions in the classroom indicate that learners are genuinely engaged in the VLC ‘Fake or Real’ image scenario. According to empirical evidence, the learners were actively participating in oral discussions and interacting with the stimuli in the provided InstaCour and the VLC environment. This level of engagement suggests that participants are making a concerted effort to categorize the stimuli in a meaningful and appropriate manner.

6.12 Symmetries and Perspectives

We executed school trials for our specially designed scenario, which included key components like InstaCour, a simulated Instagram variant featuring both authentic and altered images, and a VLC as a browser extension. The VLC offered RIS to students, summarized suggestions, and guided them through a two-stage scenario. In the first stage, students classified images before receiving advice from the companion. In the second stage, they had the option to revise their judgments after engaging with VLC’s suggestions and recommendations. The trial in Germany proceeded as planned, with interactive responses successfully gathered and stored in cloud-based databases.
Our analysis revealed that the provision of RIS by the VLC companion significantly enhanced students’ ability to discern image manipulation. Notably, the time spent on evaluating results from the RIS feature positively influenced the accuracy of judgments, corroborating our findings in RQ2. The revised evaluations were more accurate than the initial ones and aligned more closely with expert judgments. When a recommended image lacked sufficient related textual information online, the revision exhibited less improvement, and students struggled to reach the correct conclusion. In these challenging cases, we observed that users opened multiple tabs quickly to locate trustworthy web content.

Our limited study in a German school showed that students exposed to various internet images with diverse rich contexts and credible text were more likely to agree with the image’s authenticity. This led to more definitive responses that aligned with our categorization labels and classes. With the insights gained from our school trial, we expanded our research to include a university in Italy (junior bachelor students), enabling us to assess the effectiveness of our VLC and InstaCour tools in a distinct linguistic and cultural environment. Our goals were to gather extensive data and confirm our tools' efficiency across diverse student demographics through expansive trials.

The study in Italy further confirmed that the VLC effectively engages learners and enhances their ability to discern manipulated images. The comparison with Germany provided valuable insights into cultural and linguistic factors, which had a minimal impact on outcomes. This offers a comprehensive understanding of VLC’s applicability across different educational settings and cultural backgrounds. The overall assessment impact of the VLC on stimulus judgment was similar in both studies, emphasizing that search engine results (RIV) directly impact learners' judgments more than cultural background.

Given the effectiveness of the VLC tool in both Germany and Italy, there are broader practical implications to consider. The tool could potentially be adapted for use in other educational settings or for different types of media literacy. Additionally, the mathematical formulas provided for calculating improvement agreement in scenarios with two phases, like ours with pre and post recommendations, are applicable.

7 Adaptation of the VLC for the Racism Scenario

Our Virtual Learning Companion (VLC) was adapted for a racism study and its requirements that were organized by researchers at Hochschule Ruhr West (HRW), aimed at raising awareness about racism. As we mentioned in the implementation chapter (4.2) we designed the VLC and chose the technologies package to be flexible based on new requirements.

In this context, the role of the thesis's author, Farbod Aprin, was multifaceted. He provided technical support for the project. His focus was specifically on the VLC development and its modification. He also worked on finding solutions to challenges that arose, based on new scenario chatbot conversations and component adaptation. Lastly, he ensured the system's functionality during school trials.

In September 2022, the adaption of VLC for Racism scenario was deployed in two targeted educational institutions in Germany: Gymnasium Wolfskuhle in Essen and Freiherr-vom-Stein-Gymnasium in Oberhausen. To align with the study’s objectives, we enriched the InstaCour platform with content addressing daily discrimination and racism among young adults.

The VLC’s interface and functionality were adapted based on a variety of factors, such as available technologies and technical limitations. The study goal was to evaluate how effectively the VLC could enhance understanding and combat racism within predefined parameters.

7.1 Adapted Version of the VLC based on Racism Scenario

The study design was meticulously planned to offer students a valuable educational experience. Participants engaged with an adopted version of the VLC tailored to help them recognize and understand instances of everyday racism, specifically within the context of advertisements on the InstaCour platform.

The add-on chatbot system elaborated in Section 4.2 of the implementation underwent specific technical modifications to enhance its functionality. The RIS and NLP microservices are excluded in this version as it was not needed for this scenario. We updated the chatbot's aesthetics, altering its color scheme and logos, and introduced a new dialogue in JSON format as the VLC was designed to handle complex dialogues. This new dialogue was structured with keys and ideas that corresponded to the steps outlined in Figure 4-6, enabling a more intuitive and contextually relevant interaction for the users.

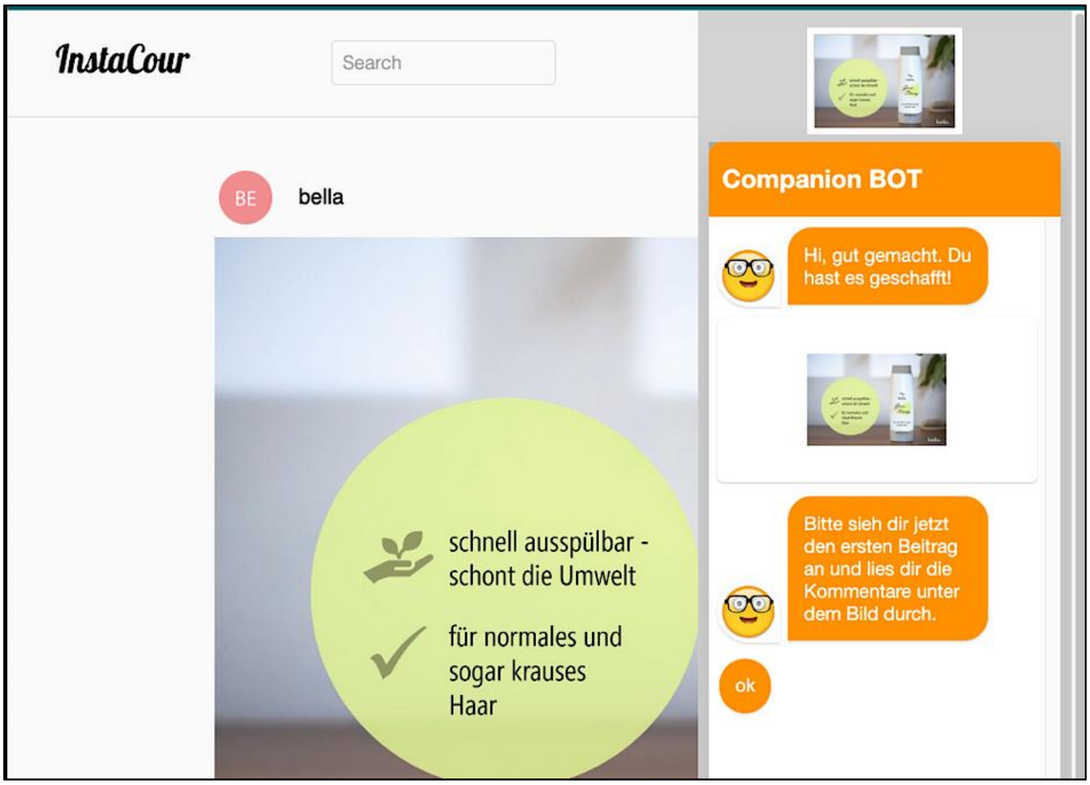
This refined capability of the VLC empowered students to delve into meaningful discussions on racism, elevating their engagement from simple recognition to an in-depth comprehension of the underlying issues. The HRW researchers utilized this tool to analyze and compare the experiences of the Control and Experimental Groups, aiming to derive valuable insights into the effectiveness of targeted anti-racism educational interventions.

As mentioned, the study participants were divided into two distinct groups: the Control Group (CG) and the Experimental Group (EG). The CG with a basic version of the VLC, which lacked specialized features for identifying racism. As a result, their experience was more generalized and did not offer targeted guidance for recognizing and understanding racist content. On the other hand, the EG used an advanced version of the VLC, specifically designed to identify and explain racist elements. This version was equipped with specialized tools and dialogues that guided students in recognizing racist content and understanding why it is considered as racist.

To enhance the user experience, we developed a flexible system that could adapt its chatbot dialogue based on user-tokens. This innovation enabled the VLC to provide personalized guidance tailored to the group type, whether it be the CG or the EG. To maintain the integrity and confidentiality of user interactions, we embedded an anonymous user-token link within the forms used by both groups:

*Example CG user-token: XXXX@HRW23Aug2022.**CG**.eu*

For the updates in the racism scenario within InstaCour, it was necessary to transition from the previous JSON configurations to a new, singular JSON content format. This file contains all the relevant components, including links to images, captions, comments, and like counts, ensuring a cohesive content structure for an improved user experience.

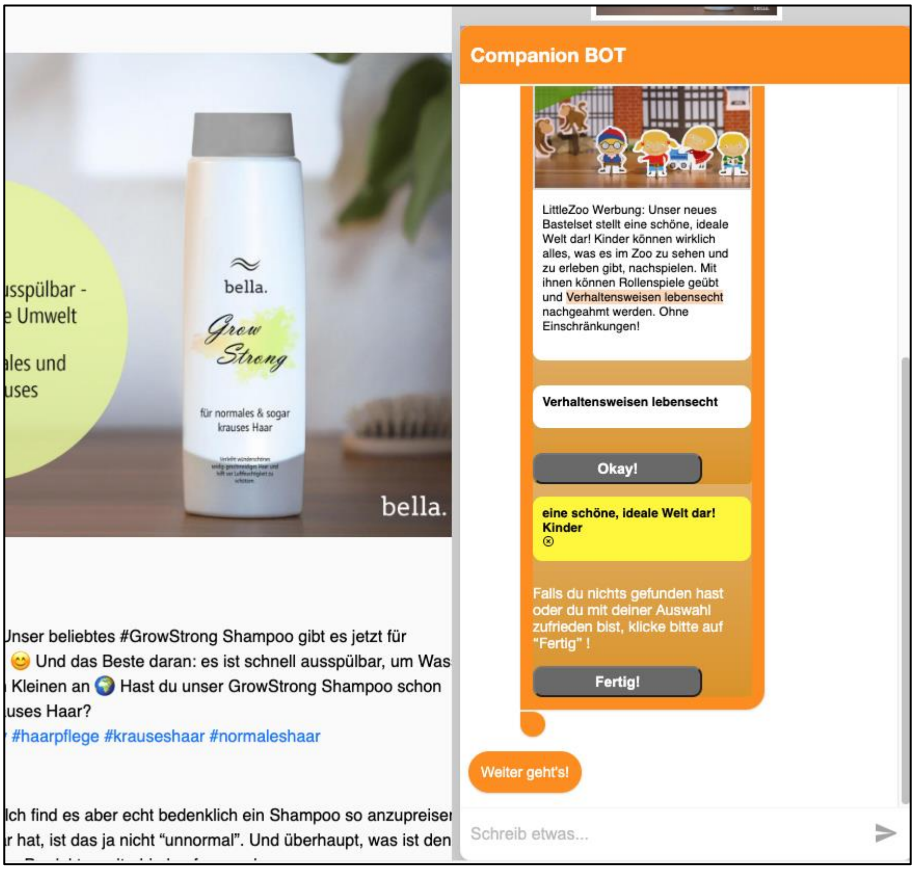

InstaCour
Search
BE
bella
schnell ausspülbar - schont die Umwelt
für normales und sogar krauses Haar
Companion BOT
ok

Hi, gut gemacht. Du hast es geschafft!

Bitte sieh dir jetzt den ersten Beitrag an und lies dir die Kommentare unter dem Bild durch.

Figure 7-1: The VLC, as utilized in the racism scenario within the InstaCour platform, prompts users to read the comments beneath the advertisement image.

In the study’s context, the companion feature within the InstaCour platform serves a crucial role, as shown in Figure 7-1. It guides users to focus on the comments section beneath an advertisement image, which includes a diverse range of opinions and reactions. To align with the study’s requirements, we disabled the image shuffle option.



Unser beliebtes #GrowStrong Shampoo gibt es jetzt für
😊 Und das Beste daran: es ist schnell ausspülbar, um Was
Kleinen an 🌍 Hast du unser GrowStrong Shampoo schon
uses Haar?
#haarpflege #krauseshaar #normaleshaar

Ich find es aber echt bedenklich ein Shampoo so anzupreiser
r hat, ist das ja nicht “unnormal”. Und überhaupt, was ist den

LittleZoo Werbung: Unser neues Bastelset stellt eine schöne, ideale Welt dar! Kinder können wirklich alles, was es im Zoo zu sehen und zu erleben gibt, nachspielen. Mit ihnen können Rollenspiele geübt und Verhaltensweisen lebensecht nachgeahmt werden. Ohne Einschränkungen!

Verhaltensweisen lebensecht

eine schöne, ideale Welt dar! Kinder
⊗

Falls du nichts gefunden hast oder du mit deiner Auswahl zufrieden bist, klicke bitte auf "Fertig" !

Figure 7-2: The VLC’s user interface features a text highlighter that allows users to select portions of text they deem toxic. Users have the option to add or remove selected parts and ultimately confirm their choices.

The companion aims to deepen the user’s analysis of public reactions to the advertisement, potentially identifying instances of racism, stereotyping, or other forms of discrimination. As illustrated in Figure 7-2, the VLC enables users to highlight text they consider racist and add it to a designated section for further analysis. These data inform our evaluation of the VLC’s impact on the accuracy and performance of both the EG and CG. At the end of the interaction, the VLC directs users to an embedded quiz button. The architecture of Highlighter component in the Chatbot conversation were described in Highlighter Component for implementation section.

We have also designed and implemented new statements based on evolving requirements. For instance, in the highlighting section, actions such as 'highlighted', 'highlighted deleted', and 'highlight submitted' were incorporated into the xAPI statements. These enhancements were successfully implemented and applied in school trials, demonstrating the adaptability and practical application of our system in educational settings.

7.2 Sample Size, Age and Gender Specification of Racism Study

The study conducted by HRW researchers initially included 130 students and was an essential component of workshops focused on combating racism through meaningful dialogues with students. After eliminating 30 cases due to incomplete or inaccurate data, the final sample size was reduced to 100 students. The age range of the participants was between 13 and 16 years, with a mean age of 13.45 and a standard deviation of 0.62. Since the students were minors, informed consent was obtained from their parents or legal guardians. The sample consisted of 44 females, 52 males, and 4 individuals identifying as diverse; no participants opted to leave the gender field blank. All participants were active users of social media.

7.3 Data Storage and Highlights from Analysis

We successfully stored all user inputs and interactions for the two aforementioned studies involving the VLC and InstaCour, using the same organizational methods as in previous scenarios. Data were stored simultaneously in both a JSON-based database and Learning Locker statements, formatted in standard xAPI. We filtered and converted the JSON files into CSV format, enabling HRW researchers to easily utilize the data in standard Excel spreadsheet software. After the data analysis, the highlights related to the VLC system's efficiency and user feedback are as follows:

Evaluation of the VLC functionality: Participants generally rated the VLC favorably, with 66 positive evaluations, 19 neutral, and 7 negative evaluations. The experimental condition (CG, EG) did not significantly affect the overall evaluation of the VLC functionality.

Perceived Target Group The average age group that participants deemed most suitable for the virtual companion ranged from 13 to 15 years.

7.4 Evaluating the Adaptive VLC's Role in Anti-Racism

The findings highlight the VLC's adaptability in combating racism and its suitability for different educational settings, laying the groundwork for future research and the development of more effective anti-racism strategies.

While the tool received positive feedback, its functionality varied among users. In future work, the VLC's flexible design offers promising avenues, including its application to other forms of discrimination and the integration of advanced technologies for personalized guidance. Overall, the study sets the stage for leveraging technology in the ongoing fight against racism and discrimination.

8 Conclusion

This dissertation explored the orchestration and evolution of Virtual Learning Companion (VLC) support within simulated social media environments. It focused on the impact of the companion on users' judgments, critical thinking, resilience, level of agreement, and awareness of both non-toxic and toxic content.

We designed and implemented the VLC with a modular microservice-based architecture. This architectural strategy endows the VLC with the flexibility to adapt to various scenarios. The modular capability empowers the server side of the VLC system to connect or disconnect from different modules, such as AI-based user intention detection, Reverse Image Search (RIS) service, or text analysis server, depending on the requirements of the trial.

It can address specific types of toxic content on social media and can also be integrated into a simulated version of Instagram with modular functionalities. The selected scenarios concentrated on addressing toxic content prevalent in everyday social networks, including hate speech, discrimination, racism, conspiracy theories, fake news, and image manipulation. The VLC's design enables the addition or removal of components for each specific scenario with minimal technical effort.

Before designing the architecture of the VLC, we conducted a comprehensive literature review. We developed a minigame that enabled users to categorize internet images according to various dimensions, including hate speech, cyber mobbing, verbal violence, and discrimination. Overarching goal of the minigame was to promote learning through controversy.

Learning Companion Systems (LCS) enhance traditional Intelligent Tutoring Systems (ITS) by incorporating a computer companion that provides personalized guidance and interactivity. The VLC, a virtual form of companion, has applications across various domains, including business and education. VLC can be an innovative approach to enhancing digital literacy and combating misinformation. One such VLC that we have developed is called ‘Courage Fake or Fact’ VLC.

VLC offers features such as role-playing and adaptive feedback based on prior interactions and learner responses. It also incorporates Reverse Image Search (RIS) links. It delivers educational recommendations, analytical artifacts for evaluating environmental images, and knowledge activation questions. These questions are designed to prompt users to reflect on the information they’ve consumed, ensuring a deeper understanding and awareness of the content. By interacting with the VLC and answering these questions, users are not only educated but also encouraged to critically evaluate the information they encounter. This added layer of interaction serves as a ‘boost’ to their cognitive competencies, ensuring that they are not passively consuming information but actively engaging with it. The VLC system's primary functions include displaying learning material and processing input.

We were able to successfully add and implement new xAPI statements tailored to our scenarios, which are not covered by standard dictionaries. The generated statements, including their descriptions, are accessible as part of an open-source project on GitHub. Utilizing the statements and data logs in the database and learning stores, we applied them in mathematical formulas for analysis and to generate various meaningful diagrams. Additionally, we employed statistical formulas to address specific research questions.

The VLC Objectives

The primary goal of the VLC that introduced in this thesis was not to optimize users' prejudice but to boost their resilience and increase their awareness of the available tools for factchecking and practicing that through chat dialogue. By engaging learners in interactive learning, the VLC actively prompted learners with knowledge-activating questions and provided insightful responses, enhancing their ability to evaluate social media content critically.

Empirical Studies in Germany and Italy for the Image Manipulation Scenario

We conducted empirical studies to test the VLC scenarios and optimized the design of the VLC and the environment based on user experience. The results indicated that the time learners spent engaging with the VLC recommendations correlated with improved judgment. We also explored cultural and linguistic elements, providing insights into the VLC's suitability across diverse educational environments.

Focused on Image Manipulation: The research substantiated the efficacy of the VLC in engaging learners and enhancing their ability to recognize image alterations.

8.1 Comparative Analysis

This section serves as a comprehensive reference by juxtaposing various technologists for educators, policymakers, and researchers. It offers insights into described tool in section 2.4 with each tool's strengths and potential areas of improvement. It underscores the importance of continuous innovation and adaptation in the fight against misinformation, emphasizing the need for tools that are effective, engaging, and relevant to the target audience. The landscape of educational technologies designed to combat misinformation is vast and varied. For instance, the SWR *Fakefinder* game offers an interactive platform where students navigate fake news scenarios, with educators assessing their understanding through in-game choices and post-game worksheets.

In contrast, our VLC operates within a simulated Instagram environment. This unique approach is particularly relevant given the platform's popularity among teenagers. VLC's two-phase system first asks teenagers to judge content based solely on their existing knowledge. After this initial judgment, the VLC provides various resources and recommendations to aid in discerning the integrity of the content. A subsequent assessment gauges the impact of these resources on their judgment. This two-tiered approach educates the user and measures the efficacy of resource recommendations in real time, offering a dynamic and responsive learning experience.

To combat toxic content in simulation form, we have developed a scenario called 'Fake or Real,' for our companion focusing on a specific type of fake news related to image manipulation. Our goal is to engage learners by providing them how to use accessible tools and encouraging them to view images in different contexts and find answers based on provided clues, rather than making an instant judgment based on a superficial glance or prejudgment. While many approaches, such as the *Bad News* game, emphasize practicing critical thinking and revealing the motives behind fake producers, we aim to provide a unique perspective through the VLC (Vaccination against Social Media content and strengthen the prejudge) it also considered evolution in different cultures as our system did. In *Fake and Real* we want to also emphasize to our target learners that the human visual system can be unreliable in many cases, especially when interpreting ambiguous stimuli (Wallisch, 2017). By providing controversial image manipulation samples in a simulated environment, we highlight the potential for misunderstandings or incorrect interpretations. Therefore, consulting different resources before forming a judgment is vital.

Our VLC use real time RIS tool an ML API to find similar images but Reality Check and SWR Games used mockup of the RIS, furthermore we have two step initial and revision, but they provide one step stimulus classification.
Co-inform and VLC both uses chrome extension to control the browser variable and to applicable in open environments.

In Table 8-1 below, we have illustrated various educational technologies designed to combat misinformation and promote digital literacy. The table encompasses different types, target audiences, methodologies, effectiveness, and limitations of each technology. We have added 'Courage' VLC' in the last row to provide a comprehensive comparison with existing solutions. This inclusion allows us to highlight its unique methodology, effectiveness in enhancing digital literacy and critical thinking, as well as its limitations.

Table 8-1: Comparative Analysis of Educational Technologies to Combat Misinformation.

Educational Technology	**Type**	**Target Audience**	**Methodology**	**Effectiveness**	**Limitations**
Co-Inform	Browser Extension	General public using Twitter on Chrome	Uses AI and rule-based algorithms to evaluate tweets for credibility. Displays credibility scores and explanations through a blurring feature	Strong correlation between the absence of the plugin and increased acceptance of misinformation	Limited to Twitter; Requires user trust
Harmony Square & Bad News	Interactive Games	General public	Educates players on prevalent misinformation tactics	Improved proficiency in identifying deceptive methods	Effectiveness varies based on engagement
Media Smarts Game series	Educational Platform	Teenagers	Mini-games and booklets to critically analyze digital advertising and biases	Promotes critical thinking	Limited to digital advertising; Specific age group
SWR Games	Interactive Games	Children, Teenagers and Adults	Educates about misinformation triggers in social media	Trains participants to assess the reliability of sources	Requires active engagement
Troll Simulator	Interactive Game	General public	Exposes players to toxic online behaviors	More negative perceptions of trolling behaviors	Focuses on trolling
WOC Educational Game	Game-Oriented Experience	Students (high school)	Simulates influence of social media on perceptions	Increased awareness of social media influence	Limited participant engagement due to COVID
'Courage' 'Fake or Real' VLC	Virtual Learning Companion As chrome plugin	Students (high school) University Students And can be applied for General public,	Role-playing, adaptive feedback, Reverse Image Search RIS links, and educational recommendation Two phase stimulus classification	Enhances digital literacy, critical thinking, and combats misinformation	Requires active engagement; Current scenarios limited to specific educational context.

We can admit that this research has not only traversed the academic landscape of VLC within the challenging terrains of simulated social media environments but has also heralded a new era of technical adaptability in educational technology. The modular microservice-based architecture underscores a significant milestone, showcasing remarkable flexibility in adapting to different social media scenarios and user interactions. The technical prowess of the VLC, demonstrated through its adaptability in technical perspective and its proficient role in enhancing users' critical engagement with digital content, reflects the innovative spirit of this work. Empirical studies across diverse cultural landscapes further validate the robust, dynamic capabilities of the VLC system, emphasizing its effectiveness in fostering digital literacy and combating misinformation. As we reflect on the journey and its outcomes, we see a future where learners are not only digitally literate but also technically empowered to critically navigate and shape the evolving digital ecosystem. The provided modular system can be used in different studies and scenarios that demanded to have a VLC and the learner interaction with web environment.

8.2 Future Directions for Awareness Tools

The digital age continually grapples with the challenge of misinformation, underscoring the need for ongoing advancements in educational technologies. Emerging tools like augmented reality (AR) and virtual reality (VR) offer immersive learning experiences by placing users in real-time misinformation scenarios. The success of VLC suggests a future where educational experiences are personalized to meet individual needs, leveraging machine learning for dynamic adaptation. As misinformation spreads across various platforms, the focus shifts toward the development of universal tools that can be applied to both VR and AR, featuring cross-platform functionality to ensure seamless operation across a myriad of social media sites.

While initiatives like *Co-Inform* and SWR *Fakefinder* represent significant strides, the horizon is replete with prospects, such as real-time fact-checking systems that harness the swift data processing capabilities of AI. As technological advancements give rise to novel misinformation techniques, like deep fakes, it is imperative for our educational tools to remain a step ahead, equipping users to tackle these nascent challenges. Implementing watermarking on media and maintaining a history of metadata changes can aid AI tools in tracking alterations, and alerting users to potential inconsistencies.

In our research, as mentioned we prioritize exploring viable methods for the education sector to safeguard vulnerable recipients against misinformation rather than relying on social media content filtering. Naturally, regulations enforcing watermarking and the storage of cloud-based blockchain metadata for every media piece would be essential to streamline this process. External challenges, such as those posed by COVID restrictions, emphasize the need for versatile solutions that cater to both online and offline modes of engagement. Despite these challenges, there are abundant opportunities for groundbreaking innovations and research aimed at countering digital misinformation.

8.3 Research Questions and Answers

1. How can a learning companion support be orchestrated for a social media environment addressing young learners?

A learning companion, like the Virtual Learning Companion (VLC), can be integrated into a social media environment to provide real-time guidance and educational support to young learners. This could include tutorials, prompts, and interactive dialogues that help them navigate and understand the social media landscape, thereby enhancing users' ability to analyze and evaluate online content.

2. How can a modular VLC be designed and developed to support adolescents within different scenarios in a simulated social media environment (Adaptability)?

The front-end of the VLC was designed as a Chrome plugin, featuring modular functionalities. This design allows for its application across various scenarios and social environments, as demonstrated in the different user trials conducted for this thesis. Such a modular VLC system can be adapted with minimum effort and can comprise various components and services, or 'modules,' which can be activated or deactivated depending on the specific scenario. For example, in the Racism scenario, the dialogue was updated and adapted to suit different target groups. The provided chatbot has a flexible architecture that allows for dialogue adjustments based on predefined scripts, AI models, or even the addition of new components, such as a text highlighter, as seen in the VLC-Racism study.

3. Can young learners benefit from access to reverse image search through the VLC to make better decisions on the quality (fake or real) of certain social media items?

Yes, evaluations from the ‘Fake or Real’ scenario studies suggest that integrating a reverse image search feature within the VLC can empower young learners to verify the authenticity of images they encounter on social media. This could enable more informed decisions regarding the credibility of items, based on the presence of similar content and images in various contexts.

4. Can young learners benefit from access to VLC to make better decisions about the quality (racism, discrimination) of certain social media items?

The VLC could include modules designed to educate users about the harmful effects of racism and discrimination. These modules have the potential to assist users in making more informed judgments about the content they encounter on social media. However, as elaborated in Chapter 8, the effectiveness of these modules for detecting racism remains a subject for debate and falls beyond the scope of this thesis.

5. Does a revision of the initial judgment about image manipulation guided by the VLC lead to a higher agreement with the expert opinion and other participants' reviews (convergence)?

Based on the study design for the 'Fake or Real' scenario, the VLC's guidance via Reverse Image Search (RIS) recommendations could have assisted young learners in revising their initial judgments about image manipulation. This could have potentially led to greater agreement with expert opinions, particularly in cases where abundant content was available on the internet. Additionally, it led to increased agreement for most of the cases (evidenced by a decrease in the dispersion index (DI) 4/5) among the learners themselves, especially for cases that had rich online content.

6. Does different cultural backgrounds impact on the user's judgment about image credibility?

As explained in the research questions for section 6.10 and its conclusion, the analysis comparing Italy and Germany indicates a general alignment in judgments. Both countries showed similar rates of improvement, as evidenced by a RIPI index score of around 16. While the study does not offer conclusive evidence that cultural background and native language significantly influence assessments of image credibility, there were some variations that suggested language could be a contributing factor. Overall, search engine results for similar information or images appear to have a greater impact on learners' judgments during the fact-checking process than do cultural backgrounds in both Italy and Germany.

Final Thoughts:

This dissertation marked a significant advancement in understanding and combating misinformation and harmful content on social media, with a specific focus on adolescents. Through the design and development of the Virtual Learning Center (VLC), offered as a Chrome plugin, we pioneered a new approach to bolstering learner resilience and awareness against specific types of disinformation and image manipulation. Our empirical studies conducted in Germany and Italy served as proof of concept. The VLC's seamless integration into social media platforms offers real-time guidance, information, and tools, such as reverse image search, making it a promising resource for supporting young learners. The insights garnered from this research laid the groundwork for a more inclusive society by promoting empathy, awareness, and critical thinking. Future research should delve into the long-term effects and ethical implications of this intervention, as well as its applicability across diverse cultural and educational settings.

8.4 Discussion

The exploration of the VLC in this dissertation has opened new horizons at the intersection of education, technology, and social media. The VLC's design as a Chrome plugin serves as a starting point. However, future expansions could include various browsers and platforms, such as mobile devices. Such an expansion would broaden the VLC's reach and impact, making it accessible to learners across different technological landscapes. Personalization and adaptation are significant areas for growth. The VLC's modular design offers flexibility, but there is potential for further personalization and multi-platform functionality, as illustrated in Figure 4-21 in the abstract. By incorporating machine learning and AI, the VLC could adapt to individual learners' needs and preferences, offering a more tailored and engaging experience. This customization could revolutionize young learners' interactions with technology, social media, and education.

Cultural and linguistic diversity also offers exciting opportunities. Studies conducted in Germany and Italy have provided valuable initial insights, but there is a need to explore the VLC's effectiveness in diverse cultural and linguistic contexts. To make the VLC inclusive and globally relevant, careful consideration and adaptation to various cultural norms and values are necessary. Ethical considerations will be crucial in the VLC's future development. Given that it collects and analyzes user data, issues of privacy, consent, and data security must be prioritized. Developing clear ethical guidelines and ensuring transparency with users will be vital for maintaining trust and integrity in this innovative educational tool.

Understanding the VLC's long-term effects on resilience, awareness, and critical thinking will require longitudinal studies. Such comprehensive research would provide deeper insights into the VLC's sustained impact and its potential to shape responsible digital citizens over time. Integrating the VLC into formal educational curricula could further enhance its reach and effectiveness. Collaboration with educators to align the VLC with educational standards and goals would bridge the gap between technology and traditional education, fostering a more holistic learning experience.

One hypothesis will test the VLC in an open environment and analyze the data for a larger community, such as that on X platform or Instagram. The we will be able to answer if the integration of Reverse Image Search (RIS) and recommendations from different search engines can potentially reduce bias, bringing together people with different mindsets on a common platform to discuss and vote. As mentioned, this thesis conclusion based on limited number of learners suggest that search engine results influence agreement on a subject more than the dichotomy of fake or real, which could also apply to larger groups. Challenges arise as platforms like Instagram create barriers for third-party content reading, but as mentioned in the technical section, image and text extraction is possible via a Chrome plugin for X, Facebook and Instagram.

This modular VLC work can serve as a foundation for developing companion assistants for open scenarios as well. New AI models as external APIs can be integrated to facilitate free-text conversation. However, the credibility of the content generated by AI models like ChatGPT [53]can be the subject of another scenario, demonstrating to users that they should not always trust the outcomes of these models as fact-checking tools without considering different neutral references.

8.5 Outlook

As described, we recommend blockchain technology for preserving image metadata. This is a strategy that can combat various challenges, including media manipulation and conspiracy theories. A unique identifier can be assigned to each image embedded identifier as a watermark before to sharing on social platforms. Key metadata, including capture time, the photographer's identity, and pixel details in a compressed format, can be immediately stored on a cloud-based blockchain after the image is taken. AI can assist in generating textual metadata for images base on image proceeding and recognition, offering detailed descriptions that can be employed to improve indexing procedures. This facilitates more efficient indexing by search engines, leading to quicker and more accurate query responses.

This method of storing metadata on a blockchain allows us to detect modified versions of pictures and track ownership changes. This functionality provides media consumers with access to related images, complete with their metadata and historical background. It empowers them to assess the image, locate it in various contexts, and more accurately gauge its authenticity. Additionally, it enhances awareness of the consumption of social media content if it is combined with educational tools like VLC.

A VLC at the meta-level, serving as an intermediary, can educate users, particularly the younger generation, about this metadata framework. In the foreseeable future, as we have explained, the deployment of blockchain-based metadata and watermarking will be vital in fighting content manipulated by humans or AI on the internet, contributing to a more enlightened online community. Naturally, this feature requires legal support and regulation. It mandates that manufacturers of phones and cameras, as well as producers, adhere to the practice of storing image metadata following its creation. This is in the process of enabling sharing. The ever-evolving landscape of social media presents new challenges, such as deep fakes and sophisticated misinformation and disinformation campaigns. The VLC must continue to evolve to address these emerging threats, requiring ongoing research and development. Staying ahead of the curve will ensure that companion tools remain a relevant and powerful tool in the fight against misinformation.

[53] Open AI ChatGPT is AI based chatbot, https://openai.com/, December 2023

Community engagement will also be essential to shaping the VLC’s future direction. By fostering collaboration between users, educators, and technologists, the VLC can continue to innovate and respond to the real-world needs and challenges of young learners. To mitigate search engine biases, it is recommended that fact-checking and companion systems employ a variety of search engines when retrieving content.

Our study shows that users’ opinions are directly impacted by the richness and coverage of search engine results. Increasing awareness of this fact within society can further enhance information consumption accuracy and authenticity. Furthermore, such a VLC platform that operates at a meta-level and is independent of traditional social media platforms, has the capacity to bridge the gap between diverse communities. This high-level platform allows individuals to acquaint themselves with voices and perspectives from various communities that they might not otherwise encounter. By transcending filter bubbles, VLC offers a broader understanding and appreciation of the multifaceted world around us.

In conclusion, the VLC outlook is both promising and complex. The work presented in this dissertation represents a significant starting step. However, it is just the beginning of a journey towards creating a more responsible and educational social media environment. The challenges and opportunities that lie ahead provide a roadmap for continued exploration, innovation, and impact. The future of the VLC holds the promise of not only enhancing individual learning experiences but also contributing to a broader societal shift towards empathy, critical thinking, and responsible digital engagement. The journey ahead is exciting and filled with possibilities, and this dissertation lays a strong foundation upon which to build.

Acknowledgments

Data Accessibility: The information and datasets created or examined during this study can be obtained from the research author upon reasonable request. You can find a link to a Microsoft Word document hosted on the cloud, which includes the data logs and various analyses in an Excel sheet related to the stated analysis, and a demo of the VLC here:

https://1drv.ms/w/s!Ar1ufk0VxMrmhi_wMc-iRMBKACZW?e=Tu71lX

Bibliography

Aboujaoude, E., Koran, L. M., Gamel, N., Large, M. D., & Serpe, R. T. (2006). Potential markers for problematic internet use: a telephone survey of 2,513 adults. *CNS Spectrums*, *11*(10), 750–755.

Aleven, V., Beal, C. R., & Graesser, A. C. (2013). Introduction to the special issue on advanced learning technologies. *Journal of Educational Psychology*, *105*(4), 929–931. https://doi.org/10.1037/A0034155)

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/JEP.31.2.211

Allport, G. W. (1979). The Nature of Prejudice. Unabridg. *Reading: Perseus Books*.

Almourad, M. B., McAlaney, J., Skinner, T., Pleya, M., & Ali, R. (2020). Defining digital addiction: Key features from the literature. *Psihologija*, *53*(3), 237–253.

Alutaybi, A., McAlaney, J., Arden-Close, E., Stefanidis, A., Phalp, K., & Ali, R. (2019). Fear of Missing Out (FoMO) as really lived: Five classifications and one ecology. *2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC)*, 1–6.

Anderson, S. P., & McLaren, J. (2012). Media mergers and media bias with rational consumers. *Journal of the European Economic Association*, *10*(4), 831–859.

Aprin, F., Chounta, I., Intelligent, H. H.-I. C. on, & 2022, undefined. (2022). "See the image in different contexts": Using reverse image search to support the identification of fake news in instagram-like social media. *SpringerF Aprin, IA Chounta, HU HoppeInternational Conference on Intelligent Tutoring Systems, 2022•Springer*, *13284 LNCS*, 264–275. https://doi.org/10.1007/978-3-031-09680-8_25

Aprin, F., Malzahn, N., Lomonaco, F., Donabauer, G., Ognibene, D., Kruschwitz, U., Hernández-Leo, D., Fulantelli, G., & Hoppe, H. U. (2022). The "Courage Companion"–An AI-Supported Environment for Training Teenagers in Handling Social Media Critically and Responsibly. *International Workshop on Higher Education Learning Methodologies and Technologies Online*, 395–406.

Aprin, F., Malzahn, N., Lomonaco, F., Donabauer, G., Ognibene Dimitri, Kruschwitz, U., Hernández-Leo, D., Hoppe, H. U., & Hoppe, H. U. (2023). The "Courage Companion" – An AI-Supported Environment for Training Teenagers in Handling Social Media Critically and Responsibly. *In: Fulantelli et al. Higher Education Learning Methodologies and Technologies Online. HELMeTO 2022. Communications in Computer and Information Science, Springer, Cham*.

Aprin, F., Manske, S., Chounta, I. A., & Hoppe, H. U. (2021). Is This Fake or Credible? A Virtual Learning Companion Supporting the Judgment of Young Learners Facing Social Media Content. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13103 LNCS*, 52–60. https://doi.org/10.1007/978-3-030-90785-3_5/COVER

Aprin, F., Peters, P., & Hoppe, H. U. (2023). The Effectiveness of a Virtual Learning Companion for Supporting the Critical Judgment of Social Media Content. *Education and Information Technologies EAIT*.

Badke, W. (2018). Fake News, Confirmation Bias, the Search for Truth, and the Theology Student. *Theological Librarianship*, *11*(2), 4–7. https://doi.org/10.31046/TL.V11I2.519

Baldry, A. C., Sorrentino, A., & Farrington, D. P. (2019). Post-traumatic stress symptoms among Italian preadolescents involved in school and cyber bullying and victimization. *Journal of Child and Family Studies*, *28*, 2358–2364.

Banker, S., & Khetani, S. (2019). Algorithm overdependence: How the use of algorithmic recommendation systems can increase risks to consumer well-being. *Journal of Public Policy & Marketing*, *38*(4), 500–515.

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good News About Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News. *Journal of Cognition*, *3*(1), Article 2. https://doi.org/10.5334/joc.91

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion Shapes the Diffusion of Moralized Content in Social Networks. *Proceedings of the National Academy of Sciences, USA*, *114*, 7313–7318. https://doi.org/10.1073/pnas.1618923114

Brandt, M. (2017). *So (un)höflich geht es im Netz zu [This is how (un)polite things are on the web]*. Statista. https://de.statista.com/infografik/8067/abschneiden-beim-digitalen-hoeflichkeits-index/

Breakstone, J., McGrew, S., Smith, M., Ortega, T., & Wineburg, S. (2018). Teaching students to navigate the online landscape. *Social Education*, *82*(4), 219–221.

Brusilovsky, P. (2003). A distributed architecture for adaptive and intelligent learning management systems. *11th International Conference on Artificial Intelligence in Education. Workshop Towards Intelligent Learning Management Systems*, *4*.

Bucher, T. (2016). The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Https://Doi.Org/10.1080/1369118X.2016.1154086*, *20*(1), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

Caulfield, M. (2018). For Online Media Literacy That Works, Speed and Ease Matters. *Medium*. https://medium.com/trust-media-and-democracy/for-online-media-literacy-that-works-speed-and-ease-matters-896dba85b54c

Chan, J., Ghose, A., & Seamans, R. (2016). The internet and racial hate crime. *Mis Quarterly*, *40*(2), 381–404.

Chan, T.-W., & Baskin, A. B. (1990). *Learning Companion Systems*. *6-33*.

Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). News in an online world: The need for an “automatic crap detector.” *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–4. https://doi.org/10.1002/PRA2.2015.145052010081

Chesney, R., & Citron, D. K. (2018). *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*. https://doi.org/10.2139/ssrn.3213954

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. https://doi.org/10.1016/0364-0213(89)90002-5

Chou, C. Y., Chan, T. W., & Lin, C. J. (2003). Redefining the learning companion: the past, present, and future of educational agents. *Computers & Education*, *40*(3), 255–269. https://doi.org/10.1016/S0360-1315(02)00130-6

Clarebout, G., Elen, J., Johnson, W. L., & Shaw, E. (2002). Animated pedagogical agents: An opportunity to be grasped? *Journal of Educational Multimedia and Hypermedia*, *11*(3), 267–286.

Combs, B. H. (2018). Everyday racism is still racism: The role of place in theorizing continuing racism in modern US society. *Phylon (1960-)*, *55*(1 & 2), 38–59.

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS One*, *12*(5), e0175799.

Costello, M., Hawdon, J., Bernatzky, C., & Mendes, K. (2019). Social group identity and perceptions of online hate. *Sociological Inquiry*, *89*(3), 427–452.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559.

der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2), 1600008.

Diaz Eleanor. (2017). *New resolution addresses accurate information and media manipulation | News and Press Center*. https://www.ala.org/news/member-news/2017/02/new-resolution-addresses-accurate-information-and-media-manipulation

Diener, E., Lusk, R., DeFour, D., & Flax, R. (1980). Deindividuation: Effects of group size, density, number of observers, and group member similarity on self-consciousness and disinhibited behavior. *Journal of Personality and Social Psychology*, *39*(3), 449.

Donabauer, G., Ognibene, D., Kruschwitz, U., Hernández-Leo, D., Fulantelli, G., & Hoppe, H. U. (2023). The "Courage Companion"–An AI-Supported Environment for Training Teenagers in Handling Social Media Critically and Responsibly. *Higher Education Learning Methodologies and Technologies Online: 4th International Conference, HELMeTO 2022, Palermo, Italy, September 21–23, 2022, Revised Selected Papers*, 395.

EU parliament. (2016). *REGULATION (EU)*. https://gdpr-info.eu/

Eurobarometer. (2018). *Fake news and disinformation online - March 2018 - - Eurobarometer survey*. https://europa.eu/eurobarometer/surveys/detail/2183

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *The Harvard Kennedy School Misinformation Review*, *1*(2). https://doi.org/10.37016/mr-2020-009

Fedorov, A. (2015). Media Literacy Education. *Fedorov, Alexander. Media Literacy Education. Moscow: ICO "Information for All*.

Fetters Maloy, A., & Branigin, A. (2023). *An AI-generated puffy-coat pope fooled us all. How much does it matter? - The Washington Post*. https://www.washingtonpost.com/lifestyle/2023/03/27/pope-francis-coat-puffy-white-ai-fake/

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378.

Gabriel, S. (2021). Can Digital Games Improve Critical Information Literacy? *ECGBL 2021 15th European Conference on Game-Based Learning (p. 244). Academic Conferences Limited.*, 244–246. https://doi.org/10.34190/GBL.21.056

Gerstenfeld, P. B., Grant, D. R., & Chiang, C.-P. (2003). Hate online: A content analysis of extremist Internet sites. *Analyses of Social Issues and Public Policy*, *3*(1), 29–44.

Geschke, D., Lorenz, J., & Holtz, P. (2019). The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology*, *58*(1), 129–149.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53–96. https://doi.org/https://doi.org/10.1111/j.1539-6053.2008.00033.x

Gillani, N., Yuan, A., Saveski, M., Vosoughi, S., & Roy, D. (2018). Me, my echo chamber, and I: introspection on social media polarization. *Proceedings of the 2018 World Wide Web Conference*, 823–831.

Glynn, Carroll J, Ostman, Ronald E, McDonald, & Daniel G. (1995). THEORETICAL FOUNDATIONS. *Public Opinion and the Communication of Consent*, 249.

Gnewuch, U., Morana, S., Adam, M., & Maedche, A. (2018). Faster is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. *Research Papers*. https://aisel.aisnet.org/ecis2018_rp/113

Goodman, B., Linton, F., & Gaimari, R. (2016). Encouraging Student Reflection and Articulation Using a Learning Companion: A Commentary. *International Journal of Artificial Intelligence in Education*, *26*(1), 474–488. https://doi.org/10.1007/s40593-015-0041-4

Guardian, T. (2018). The cambridge analytica files: A year-long investigation into facebook, data, and influencing elections in the digital age. *The Guardian*.

Hafenbrädl, S., Waeger, D., Marewski, J. N., & Gigerenzer, G. (2016). Applied decision making with fast-and-frugal heuristics. *Journal of Applied Research in Memory and Cognition, 5*(2), 215–231.

Hietala, P., & Niemirepo, T. (1998). The Competence of Learning Companion Agents 1. *International Journal of Artificial Intelligence in Education*, *9,* 178–192.

Hsu, S.-H., Chou, C.-Y., Chen, F.-C., Wang, Y.-K., & Chan, T.-W. (2007). An investigation of the differences between robot and virtual learning companions' influences on students' engagement. *2007 First IEEE International Workshop on Digital Game and Intelligent Toy Enhanced Learning (DIGITEL'07)*, 41–48.

Jaursch, J. (2019). *Regulatory reactions to disinformation: How Germany and the EU are trying to tackle opinion manipulation on digital platforms*. https://www.stiftung-nv.de/sites/default/files/regulatory_reactions_to_disinformation_in_germany_and_the_eu.pdf

Jin, Y., & Liu, B. (2010). The blog-mediated crisis communication model: Recommendations for responding to influential external blogs. *Journal of Public Relations Research*, *22*(4), 429–455. https://doi.org/10.1080/10627261003801420

Johnson, W. L., Rickel, J. W., Lester, J. C., & others. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, *11*(1), 47–78.

Jones, L. M., & Mitchell, K. J. (2016). Defining and measuring youth digital citizenship. *New Media & Society*, *18*(9), 2063–2079.

Kirkinis, K., Pieterse, A. L., Martin, C., Agiliga, A., & Brownell, A. (2018). Racism, racial discrimination, and trauma: A systematic review of the social science literature. *Taylor & FrancisK Kirkinis, AL Pieterse, C Martin, A Agiliga, A BrownellEthnicity & Health, 2021•Taylor & Francis*, *26*(3), 392–412. https://doi.org/10.1080/13557858.2018.1514453

Komaç, G., & Ça˘gıltay, K. (2021). Raising Awareness Through Games: The Influence of a Trolling Game on Perception of Toxic Behavior. In Ö. Cordan (Ed.), *Game + Design Education* (Vol. 13, pp. 143–149). Springer Nature Switzerland AG. https://doi.org/https://doi.org/10.1007/978-3-030-65060-5_12

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, *21*(3), 103–156. https://doi.org/10.1177/1529100620946707

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(24), 8788.

Krämer, N. C., & Bente, G. (2010). Personalizing e-learning. The social effects of pedagogical agents. *Educational Psychology Review, 22*, 71–87.

Kuss, D. J., & Griffiths, M. D. (2011). Online social networking and addiction—a review of the psychological literature. *International Journal of Environmental Research and Public Health*, *8*(9), 3528–3552.

Kyza, E. A., Varda, C., Karageorgiou, M., Perfumi, S. C., Iftikhar, S., Shah, H., & Hosseini, A. S. (2020). *Combating misinformation online: re-imagining social media for policy-making Nadejda Komendantova International Institute for Applied Systems Analysis*. *Internet Policy Rev…*. https://doi.org/10.14763/2020.4.1514

Kyza, E., Varda, C., Konstantinou, L., Karapanos, E., Perfumi, S. C., Svahn, M., & Georgiou, Y. (2021). SOCIAL MEDIA USE, TRUST AND TECHNOLOGY ACCEPTANCE: INVESTIGATING THE EFFECTIVENESS OF A CO-CREATED BROWSER PLUGIN IN MITIGATING THE SPREAD OF MISINFORMATION ON SOCIAL MEDIA. *AoIR Selected Papers of Internet Research*. https://doi.org/10.5210/SPIR.V2021I0.12197

Lavenia, G. (2012). Internet e le sue dipendenze. *Dal Coinvolgimento Alla Psicopatologia. Milano: Franco Angeli*.

Lawson, R. (2006). *The science of cycology: Failures to understand how everyday objects work*. https://www.liverpool.ac.uk/~rlawson/PDF_Files/L-M&C-2006.pdf

Layug, A., Krishnamurthy, S., McKenzie, R., & Feng, B. (2022). The Impacts of Social Media Use and Online Racial Discrimination on Asian American Mental Health: Cross-sectional Survey in the United States During COVID-19. *JMIR Formative Research*, *6*(9), e38589.

Lee, R. S. C., Hoppenbrouwers, S., & Franken, I. (2019). A systematic meta-review of impulsivity and compulsivity in addictive behaviors. *Neuropsychology Review, 29*, 14–26.

Lester, J. C., Voerman, J. L., Towns, S. G., & Callaway, C. B. (1999). Deictic believability: Coordinated gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence*, *13*(4–5), 383–414.

Levy, S. R., Lytle, A., Shin, J. E., & Hughes, J. M. (2016). Understanding and reducing racial and ethnic prejudice among children and adolescents. *Social Issues and Policy Review, 10*(1), 201–234.

Lewandowsky, S. (2020). The “post-truth” world, misinformation, and information literacy: A perspective from cognitive science. In S. Goldstein (Ed.), *Informed societies—Why information literacy matters for citizenship, participation and democracy* (pp. 69–88). Facet Publishing.

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the “post-truth” era. *Journal of Applied Research in Memory and Cognition*, *6,* 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

Linke, L. H. (2012). Social closeness and decision making: Moral, attributive and emotional reactions to third party transgressions. *Current Psychology, 31*, 291–312.

Lomonaco, F., Taibi, D., Trianni, V., & Ognibene, D. (2022, September). A game-based educational experience to increase awareness about the threats of social media filter bubbles and echo chambers inspired by “wisdom of the crowd”: preliminary results. *4th International Conference on Higher Education Learning Methodologies and Technologies Online HELMeTO2022*.

Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C., & Hertwig, R. (2020). How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behavior*. https://doi.org/10.1038/s41562-020-0889-7

Malzahn, N., Aprin, F., Hoppe, H. U., Eimler, S. C., & Moder, S. (2021). Measures of disagreement in learning groups as a basis for identifying and discussing controversial judgements. *Proceedings of the 14th International Conference on Computer-Supported Collaborative Learning-CSCL 2021*.

Malzahn, N., Schwarze, V., Eimler, S. C., Aprin, F., Moder, S., & Hoppe, H. U. (2023). How to measure disagreement as a premise for learning from controversy in a social media context. *Research and Practice in Technology Enhanced Learning*, *18,* 012–012. https://doi.org/10.58459/RPTEL.2023.18012

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, *22*(2), 205–224.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, *USA*, *114*, 12714–12719.
https://doi.org/10.1073/pnas.1710966114

McCright, A. M., & Dunlap, R. E. (2017). Combatting misinformation requires recognizing its types and the factors that facilitate its spread and resonance. *Journal of Applied Research in Memory and Cognition*, *6*, 389–396. https://doi.org/10.1016/j.jarmac.2017.09.005

McDonald, J. D. (John D., & Levine-Clark, M. (2017). *American Library Association (ALA)*. 67–84. https://doi.org/10.1081/E-ELIS4

McGrew, S., Smith, M., Breakstone, J., Ortega, T., & Wineburg, S. (2019). Improving university students' web savvy: An intervention study. *British Journal of Educational Psychology*, *89*(3), 485–500.

Meyers, E. M., Erickson, I., & Small, R. V. (2013). Digital literacy and informal learning environments: an introduction. *Learning, Media and Technology*, *38*(4), 355–367.

Meyers, Z. (2022). *Will the Digital Services Act save Europe from disinformation?* https://www.cer.eu/

Milano, S., Taddeo, M., & Floridi, L. (2021). Ethical aspects of multi-stakeholder recommendation systems. *The Information Society*, *37*(1), 35–45.

Miville, M. L., Romans, J. S. C., Johnson, D., & Lone, R. (2004). Universal-diverse orientation: Linking social attitudes with wellness. *Journal of College Student Psychotherapy*, *19*(2), 61–79.

Mladenović, M., Ošmjanski, V., & Stanković, S. V. (2021). Cyber-aggression, cyberbullying, and cyber-grooming: a survey and research challenges. *ACM Computing Surveys (CSUR)*, *54*(1), 1–42.

Moreno, R. (2001). Software agents in multimedia: An experimental study of their contributions to students' learning. *Human-Computer Interaction Proceedings*, 275–277.

Musetti, A., & Corsano, P. (2018). The internet is not a tool: reappraising the model for internet-addiction disorder based on the constraints and opportunities of the digital environment. *Frontiers in Psychology*, *9*, 558.

Nakayama, H., & Higuchi, S. (2015). Internet addiction. *Nihon Rinsho. Japanese Journal of Clinical Medicine*, *73*(9), 1559–1566.

NATO Strategic Communications Centre of Excellence. (2017). *Digital hydra: Security implications of false information online*. https://www.stratcomcoe.org/digital-hydra-security-implications-false-information-online

Neubaum, G., & Krämer, N. C. (2017). Opinion climates in social media: Blending mass and interpersonal communication. *Human Communication Research*, *43*(4), 464–476.

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, *223*, 103525.

Nicol, A. A. M., & De France, K. (2016). The big five's relation with the facets of right-wing authoritarianism and social dominance orientation. *Personality and Individual Differences*, *98*, 320–323.

Nikolov, D., Oliveira, D. F. M., Flammini, A., & Menczer, F. (2015). Measuring online social bubbles. *PeerJ Computer Science*, *1*, e38.

NRW, S. M. A. (2022). *Hate Speech forsa-Studie 2022. Zentrale Untersuchungsergebnisse [Hate Speech forsa Study 2022: Key Results]*. https://www.medienanstalt-nrw.de/themen/hass/forsa-befragung-zur-wahrnehmung-von-hassrede.html

Ognibene, D., Wilkens, R., Taibi, D., Hernández-Leo, D., Kruschwitz, U., Donabauer, G., Theophilou, E., Lomonaco, F., Bursic, S., Lobo, R. A., Sánchez-Reina, J. R., Scifo, L., Schwarze, V., Börsting, J., Hoppe, U., Aprin, F., Malzahn, N., & Eimler, S. (2023). Challenging social media threats using collective well-being-aware recommendation algorithms and an educational virtual companion. *Frontiers in Artificial Intelligence*, *5*, 654930. https://doi.org/10.3389/FRAI.2022.654930/BIBTEX

Paris, B., & Donovan, J. (2019). *Deepfakes and cheap fakes. The manipulation of audio and visual evidence*. https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1-1.pdf

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2019). Shifting attention to accuracy can reduce misinformation online. *PsyArXiv*. https://doi.org/10.31234/osf.io/3n9u8

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science, 31*, 770–780. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188,* 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin and Review, 24*(6), 1774–1784. https://doi.org/10.3758/S13423-017-1242-7

Pezzullo, L. G., Wiggins, J. B., Frankosky, M. H., Min, W., Boyer, K. E., Mott, B. W., Wiebe, E. N., & Lester, J. C. (2017). "Thanks Alisha, keep in touch": Gender effects and engagement with virtual learning companions. *Artificial Intelligence in Education: 18th International Conference, AIED 2017, Wuhan, China, June 28–July 1, 2017, Proceedings 18,* 299–310.

Postmes, T., & Spears, R. (1998). Deindividuation and antinormative behavior: A meta-analysis. *Psychological Bulletin, 123*(3), 238.

Rideout, V., Fox, S., Peebles, A., & Robb, M. B. (2021). Coping with COVID-19: How young people use digital media to manage their mental health. *San Francisco, CA: Common Sense and Hopelab*.

Roberts, M. (2018). *Censored: distraction and diversion inside China's Great Firewall*. Princeton University Press.

Rojecki, A., & Meraz, S. (2016). Rumors and factitious informational blends: The role of the web in speculative politics. *New Media & Society, 18,* 25–43. https://doi.org/10.1177/1461444814535724

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications, 5*. https://doi.org/10.1057/s41599-019-0279-9

Rubin, V., Brogly, C., Conroy, N., Chen, Y., Cornwell, S. E., & Asubiaro, T. V. (2019). *litrl/litrl_code: Litrl Browser Experimental 0.14.0.0 Public*. https://doi.org/10.5281/ZENODO.2588566

Rubin, V. L. (2019). Disinformation and misinformation triangle: A conceptual model for "fake news" epidemic, causal factors and interventions. *Journal of Documentation, 75*(5), 1013–1034. https://doi.org/10.1108/JD-12-2018-0209/FULL/XML

Schafer, J. A. (2002). Spinning the web of hate: Web-based hate propagation by extremist organizations. *Journal of Criminal Justice and Popular Culture*.

Schultze-Krumbholz, A., Jäkel, A., Schultze, M., & Scheithauer, H. (2012). Emotional and behavioural problems in the context of cyberbullying: A longitudinal study among German adolescents. *Emotional and Behavioural Difficulties*, *17*(3–4), 329–345. https://doi.org/10.1080/13632752.2012.704317

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shearer, E. (2018, December 10). *Social media outpaces print newspapers in the U.S. as news source | Pew Research Center*. https://www.pewresearch.org/short-reads/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/

Skilbred-Fjeld, S., Reme, S., of, S. M.-C. J., & 2020, undefined. (2020). Cyberbullying involvement and mental health problems among late adolescents. *Duo.Uio.No*, *14*(1). https://doi.org/10.5817/CP2020-1-5

Sparks, J. R., Katz, I. R., & Beile, P. M. (2016). Assessing digital information literacy in higher education: A review of existing frameworks and assessments with recommendations for next-generation assessment. *ETS Research Report Series*, *2016*(2), 1–33.

Stewart, A. J., Mosleh, M., Diakonova, M., Arechar, A. A., Rand, D. G., & Plotkin, J. B. (2019). Information gerrymandering and undemocratic decisions. *Nature*, *573*(7772), 117–121.

Sunstein, C. R. (2016). People prefer system 2 nudges (kind of). *Duke LJ*, *66*, 121.
Taibi, D., Börsting, J., Hoppe, U., Ognibene, D., Hernández-Leo, D., Eimler, S. C., & Kruschwitz, U. (2022). The role of educational interventions in facing social media threats: overarching principles of the courage project. *International Workshop on Higher Education Learning Methodologies and Technologies Online*, 315–329.

Tao, X., & Fisher, C. B. (2022). Exposure to social media racial discrimination and mental health among adolescents of color. *Journal of Youth and Adolescence*, 1–15. https://doi.org/https://doi.org/10.1007/s10964-021-01514-z

Taymur, I., Budak, E., Demirci, H., Akdağ, H. A., Güngör, B. B., & Özdel, K. (2016). A study of the relationship between internet addiction, psychopathology and dysfunctional beliefs. *Computers in Human Behavior*, *61*, 532–536.

Thaler, R. H., & Sunstein, C. R. (2008). Nudge: improving decisions about health. *Wealth, and Happiness*, *6*, 14–38.

Thaler, R. H. (2018). Nudge, not sludge. *Science*, *361*(6401), 431. https://doi.org/10.1126/science.aau9241

Theophilou, E., Schwarze, V., Börsting, J., Sánchez-Reina, R., Scifo, L., Lomonaco, F., Aprin, F., Ognibene, D., Taibi, D., Hernández-Leo, D., & others. (2022). Empirically investigating virtual learning companions to enhance social media literacy. *International Workshop on Higher Education Learning Methodologies and Technologies Online*, 345–360.

Vasilopoulos, P., & Lachat, R. (2018). Authoritarianism and political choice in France. *Acta Politica*, *53*, 612–634.

Walker, J. T. (1999). Statistics in Criminal Justice: Analysis and Interpretation. In *AN ASPEN*. Google Books: https://bit.ly/3rwy3H7

Walker, K. L. (2016). Surrendering information through the looking glass: Transparency, trust, and protection. *Journal of Public Policy & Marketing*, *35*(1), 144–158.

Wallisch, P. (2017). Illumination assumptions account for individual differences in the perceptual interpretation of a profoundly ambiguous stimulus in the color domain: “The dress.” *Journal of Vision*, *17*(4), 5–5. https://doi.org/10.1167/17.4.5

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking* (Vol. 27). Council of Europe Strasbourg.

Webb, H., Burnap, P., Procter, R., Rana, O., Stahl, B. C., Williams, M., Housley, W., Edwards, A., & Jirotka, M. (2016). Digital wildfires: Propagation, verification, regulation, and responsible innovation. *ACM Transactions on Information Systems (TOIS)*, *34*(3), 1–23.

Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific Reports*, *2*(1), 335.

Whitworth, B. (2007). Measuring disagreement. In *Handbook of Research on Electronic Surveys and Measurements* (pp. 174–187). IGI Global.

Wineburg, S., & McGrew, S. (2019). Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teachers College Record*, *121*(11), 1–40.

Woolf, B. P., Arroyo, I., Muldner, K., Burleson, W., Cooper, D. G., Dolan, R., & Christopherson, R. M. (2010). The effect of motivational learning companions on low achieving students and students with disabilities. *Intelligent Tutoring Systems: 10th International Conference, ITS 2010, Pittsburgh, PA, USA, June 14-18, 2010, Proceedings, Part I 10*, 327–337.

Wu, Q., Miao, C., & Shen, Z. (2012). A curious learning companion in virtual learning environment. *2012 IEEE International Conference on Fuzzy Systems*, 1–8.

Xu, S., Yang, H. H., MacLeod, J., & Zhu, S. (2019). Social media competence and digital citizenship among college students. *Convergence*, *25*(4), 735–752.

Yang, C., Chen, C., Lin, X., & Chan, M.-K. (2021). School-wide social emotional learning and cyberbullying victimization among middle and high school students: Moderating role of school climate. *School Psychology, 36*(2), 75.

Ybarra, M. L., Mitchell, K. J., & Korchmaros, J. D. (2011). National trends in exposure to and experiences of violence on the Internet among children. *Pediatrics, 128*(6), e1376–e1386. https://doi.org/https://doi.org/10.1542/peds.2011-0118

Zannettou, S., Sirivianos, M., Blackburn, J., & Kourtellis, N. (2019). *The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans*.

Zschache, U. (2022). *Extract from the Integrated Report on Trust and the Media, Part 3 "Fake news and Counter-strategies-Country reports" (September 2022) EnTrust research project: Enlightened Trust: An Examination of Trust and Distrust in Governance-Conditions, Effects and Remedies Disinformation and Counter-strategies in Challenging Times-The German Case*.
Entrust-project.eu : https://bit.ly/3RPdOz8

DOI: 10.17185/duepublico/81555
URN: urn:nbn:de:hbz:465-20240216-073248-2