Khan, U. A., Kauttonen, J., Aunimo, L., & Alamäki, A. (2024). A system to ensure information trustworthiness in artificial intelligence enhanced higher education. *Journal of Information Technology Education: Research, 23*, Article 13. https://doi.org/10.28945/5295

Journal of Information Technology Education: Research

An Official Publication of the Informing Science Institute
InformingScience.org

JITEResearch.org

Volume 23, 2024

A System to Ensure Information Trustworthiness in Artificial Intelligence Enhanced Higher Education

Umair Ali Khan*	RDI & Competences, Haaga-Helia University of Applied Sciences, Helsinki, Finland	Umairali.khan@haaga-helia.fi
Janne Kauttonen	RDI & Competences, Haaga-Helia University of Applied Sciences, Helsinki, Finland	janne.kauttonen@haaga-helia.fi
Lili Aunimo	RDI & Competences, Haaga-Helia University of Applied Sciences, Helsinki, Finland	lili.aunimo@haaga-helia.fi
Ari Alamäki	RDI & Competences, Haaga-Helia University of Applied Sciences, Helsinki, Finland	ari.alamaki@haaga-helia.fi

* Corresponding author

Abstract

Aim/Purpose	The purpose of this paper is to address the challenges posed by disinformation in an educational context. The paper aims to review existing information assessment techniques, highlight their limitations, and propose a conceptual design for a multimodal, explainable information assessment system for higher education. The ultimate goal is to provide a roadmap for researchers that meets current requirements of information assessment in education.
Background	The background of this paper is rooted in the growing concern over disinformation, especially in higher education, where it can impact critical thinking and decision-making. The issue is exacerbated by the rise of AI-based analytics on social media and their use in educational settings. Existing information assessment techniques have limitations, requiring a more comprehensive AI-based approach that considers a wide range of data types and multiple dimensions of disinformation.

Accepting Editor Janice Whatley | Received: September 23, 2023 | Revised: January 17, February 27, March 11, April 8, April 14, April 22, May 10, 2024 | Accepted: May 16, 2024.
Cite as: Khan, U. A., Kauttonen, J., Aunimo, L., & Alamäki, A. (2024). A system to ensure information trustworthiness in artificial intelligence enhanced higher education. *Journal of Information Technology Education: Research, 23,* Article 13. https://doi.org/10.28945/5295

Methodology	Our approach involves an extensive literature review of current methods for information assessment, along with their limitations. We then establish theoretical foundations and design concepts for EMIAS based on AI techniques and knowledge graph theory.
Contribution	We introduce a comprehensive theoretical framework for an AI-based multimodal information assessment system specifically designed for the education sector. It not only provides a novel approach to assessing information credibility but also proposes the use of explainable AI and a three-pronged approach to information evaluation, addressing a critical gap in the current literature. This research also serves as a guide for educational institutions considering the deployment of advanced AI-based systems for information evaluation.
Findings	We uncover a critical need for robust information assessment systems in higher education to tackle disinformation. We propose an AI-based EMIAS system designed to evaluate the trustworthiness and quality of content while providing explanatory justifications. We underscore the challenges of integrating this system into educational infrastructures and emphasize its potential benefits, such as improved teaching quality and fostering critical thinking.
Recommendations for Practitioners	Implement the proposed EMIAS system to enhance the credibility of information in educational settings and foster critical thinking among students and teachers.
Recommendations for Researchers	Explore domain-specific adaptations of EMIAS, research on user feedback mechanisms, and investigate seamless integration techniques within existing academic infrastructure.
Impact on Society	This paper's findings could strengthen academic integrity and foster a more informed society by improving the quality of information in education.
Future Research	Further research should investigate the practical implementation, effectiveness, and adaptation of EMIAS across various educational contexts.
Keywords	information assessment, artificial intelligence, higher education

INTRODUCTION

The abundance of intentionally false, manipulated, misleading, or satirical content in the digital age presents significant challenges, such as fake news, disinformation, and propaganda (European Commission, 2018). Disinformation can have severe consequences, including deceiving or manipulating individuals for economic or political gain, negatively impacting institutions, propagating false narratives, and escalating geopolitical tensions (European Commission, 2018). Furthermore, during times of crisis, such as the COVID-19 pandemic, disinformation can pose an even greater threat to public health and safety (National Academy of Medicine, 2023).

In the context of education, the effects of disinformation can be particularly detrimental, as it can negatively impact students' critical thinking skills and decision-making abilities (Nygren et al., 2020). The trustworthiness of information has been a significant topic in educational discussions (Dame Adjin-Tettey, 2022). Students and teachers use different information sources and systems in writing and reading in the classrooms, doing assignments and homework at home, and searching information for authoring essays and learning tasks virtually in any place where they have access to the Internet. Access to accurate information in research is vital, as it affects the quality and reliability of research findings and conclusions (Eslake, 2006). Since students and teachers in a higher education institute

are involved in different levels of research, using inaccurate information can lead them down false avenues of investigation, wasting time and resources.

Disinformation can spread rapidly in educational environments, making it challenging for teachers, students, and administrators to counter its effects (European Commission, 2022). Disinformation affects both teachers and students in similar ways. It can result in students not receiving a comprehensive education and being inadequately prepared for their future careers (Weiss et al., 2020). In the same way, disinformation creates confusion and uncertainty among teachers regarding the accuracy of the information and can lead to the dissemination of incorrect knowledge and beliefs. At the same time, it also adds to their workload as they need to ascertain the trustworthiness of information before using it in their teaching. Overall, disinformation not only weakens teachers' ability to maintain a neutral, evidence-based approach to teaching, as they may be influenced by biased or false information, but it also limits students' access to accurate information and creates barriers for them to seek out the required information (Moyer, 2018). Hence, disinformation undermines the credibility of information and erodes trust in credible sources, thereby affecting the quality of education and research (European Commission, 2022).

The rapid adoption of artificial intelligence (AI) based analytics and personalization features in social media (Hermann, 2022), the gradually expanding use of AI-enhanced learning in education (Cardona et al., 2023), and the systematic production of intentional disinformation by several politically colored organizations (Muhammed T & Mathew, 2022) are now challenging traditional self-directed information searching in educational institutes.

To address these issues, we first review the existing techniques for information assessment in general, then education and learning. We argue why artificial intelligence-based techniques are most promising in this regard and what their limitations and challenges are. We propose the conceptual design of a multimodal, explainable information assessment system for education and learning that not only rates the quality of a piece of information but also provides more insights into its decision-making so that the user can develop trust in the system. This work is aimed at providing a roadmap to researchers that meets the contemporary requirements of information assessment. First, we investigate the existing research gaps in information assessment and the adoption of AI-in-Education (AIED) tools. We then highlight the requirements of an information assessment system in education. This leads us to propose a conceptual framework based on a repository composed of an evolving knowledge graph, which not only contains the trustworthiness metrics of each piece of information but also contains the relationships between the information sources and the influencing entities in the form of a multimodal knowledge graph. This approach transcends the traditional binary approach of dis(mis)information detection, which merely scratches the surface of this problem. Due to the ever-growing surge in information production in multiple formats, we also justify addressing multiple modalities and the potential ways to do this. In addition, we also discuss the potential machine learning techniques for training multimodal models for information assessment. We also highlight the challenges of the development of the proposed information assessment system.

Our formulated research questions are given as follows.

RQ1: What criteria define information assessment in academia, and which taxonomy serves this purpose?

RQ2: Which design foundations and methods can create a multimodal information assessment system, offering intuitive explanations or visuals that support trust and comprehension of information validity for educators and learners?

RQ3: What functions are required by the explainable, multimodal information assessment system for the objectives described in RQ2?

RQ4: What are the major challenges when integrating AI-powered information assessment in educational contexts, and how can these be ethically and effectively navigated?

The rest of the paper is structured as follows. We address RQ1 and RQ2 by presenting an overview of the related work, identifying research gaps, providing a literature review on AI's role in educational information assessment, outlining the requisites of the AI-enhanced education system, and introducing a pertinent taxonomy. Subsequently, we address RQ3 by describing the methodical development of EMIAS, discussing its individual layers and their functions. RQ4 is addressed by exploring challenges, potential risks, and solutions related to EMIAS development. Additionally, we present methodologies for information management utilized by educators and students. We provide an in-depth discussion and highlight the opportunities presented by EMIAS integration in tertiary education. Finally, we chart out potential avenues for future research and conclude the paper.

RELATED WORK AND RESEARCH GAPS

While some efforts have been made to combat disinformation, there are still significant research gaps in information assessment. Existing systems have limited ability for information assessment and focus on assessing credibility alone, neglecting other important aspects such as propaganda, hate speech, biasedness, intention, and manipulation using generative AI. There is also limited support for multilingual and multimodal information assessment, and existing fact-checking resources depend on a significant amount of human involvement. Additionally, the unique features of AI generative techniques pose new challenges for information assessment. Therefore, there is a need to develop new and more sophisticated methods for detecting disinformation and enhancing the accuracy and reliability of information assessment.

While traditional disinformation detection systems have focused on assessing the credibility of information (Al-Ahmad et al., 2021; Sahoo & Gupta, 2021; Xiang, 2022), it is crucial to consider other aspects of information assessment, including in-depth analysis of individual constituent parts of a piece of information, differentiating between disinformation and misinformation (Lecheler & Egelhofer, 2022), identifying the fake information sources and their relationships, detecting information with harmful intent (Sharma et al., 2022), and detecting manipulated information. The rapid growth of social media and digital platforms has made it easier to generate and spread fake, inaccurate, and harmful information in a variety of formats (Muhammed T & Mathew, 2022). As a result, information assessment is no longer limited to textual data alone but also involves analyzing other data modalities such as images, videos, and audio (Hameleers, 2023). Though the problem of fake news detection has been addressed by techniques using multimodal features (Giachanou et al., 2020; Kumari & Ekbal, 2021; Segura-Bedmar & Alonso-Bartolome, 2022; Song et al., 2021), these techniques address this issue as a binary classification problem (Kim et al., 2021) and have significant limitations in addressing the complexity of the problem (Thota et al., 2018). While binary classification assumes that all pieces of information are either true or false (e.g., Balshetwar et al., 2023; Jeyasudha et al., 2022), fake information can be much more nuanced and complex (Iceland, 2023). Fake information can be ambiguous and contain both true and false elements, making it difficult to classify it as simply true or false (Hoy & Koulouri, 2021). This complexity may not be captured using a binary classification model. Moreover, the meaning of fake information can change depending on the context and is challenging to address using a binary classification model.

As artificial intelligence (AI) has advanced, it has become easier to generate fake information that can be difficult to detect. In particular, the use of generative AI, such as deep learning models, has made it easier to manipulate digital media and create fake content that appears to be genuine (Almars, 2021). This includes image forgery, doctored videos and audio, and fake articles. The ability to generate convincing fake information using AI has further complicated the challenge of assessing information accuracy and factuality (Helmus, 2022). In response, there is a need to develop multimodal

systems for information assessment that can detect AI generative elements in addition to assessing the factuality of the information. Despite the importance of detecting AI generative elements in fake information, the existing techniques for information assessment have not yet addressed this dimension in combination with the dimension of factuality. Therefore, there is a need to develop new and more sophisticated methods for detecting AI generative elements and integrate these methods with existing techniques for assessing the factuality of the information. This will help to enhance the accuracy and reliability of information assessment and enable better detection of fake information that uses AI generative techniques.

Current techniques for detecting fake information do not integrate both human and technological knowledge (Lampridis et al., 2022). A human-centered information assessment detection system is important for several reasons. First, humans are inherently better at making sense of ambiguous, complex, and contextual information than machines. Information assessment is not just about classifying information as true or false but also about understanding the broader context, interpreting the meaning of the information, and identifying potential sources of bias. Human expertise is critical in making these nuanced judgments. For example, the trustworthiness of a text is strongly associated with such properties as information content, neutrality, clarity, sentiment, and logic (Kauttonen et al., 2020). Second, a human-centered approach helps ensure ethical considerations are considered in the detection and assessment of fake information (Lampridis et al., 2022). Machines can make decisions based purely on data without considering the broader ethical implications. Humans, however, can bring a more nuanced ethical perspective to the decision-making process, considering factors such as privacy, bias, and fairness. Finally, a human-centered approach helps to build trust in the information assessment detection. Users are more likely to trust a system that is transparent, accountable, and open to feedback (Schoenherr et al., 2023). A system that incorporates human expertise can help to provide these qualities, by ensuring that the decision-making process is transparent and that decisions are made based on sound ethical principles.

Transparency and explainability are two essential features that an information assessment system must have (Szczepański et al., 2021). First, transparency helps to build trust in the system by making the process of information assessment more accessible and understandable to end-users. Users are more likely to trust the results of the system if they can understand how it works and the criteria used to evaluate the information. Transparency can help to increase the accountability of the system and ensure that it operates ethically and responsibly (Memarian & Doleck, 2023). Second, explainability is important for enabling users to understand the rationale behind a decision made by the system (Szczepański et al., 2021). If users can see the steps taken by the system to arrive at a decision, they can make more informed judgments about the information's credibility. Explainability also helps to identify any potential biases or errors in the system, enabling them to be addressed and improved. At the same time, it allows the user to gauge the authenticity of the decision and provide feedback to the system for continuous improvement. Finally, transparency and explainability are essential for promoting the wider adoption of information assessment systems. If the system is opaque or difficult to understand, it is unlikely to gain widespread adoption or trust from end-users. In contrast, if the system is transparent and explainable, it is more likely to be adopted and used to support the wider goals of promoting accurate and reliable information online.

Due to the growing importance of explainability and transparency in AI systems, techniques for detecting fake information have attempted to address this issue (Chien et al., 2022; Kou et al., 2022; Shang et al., 2022; Yang et al., 2019). However, the current methods have certain limitations, such as applying explainable solutions for unimodal data and binary classifiers or only applying them to specific sources of information. Additionally, most multimodal techniques are designed to assess only text and image modalities, disregarding the increasing prevalence of videos and audio as a means of disseminating information online. With the advancement of deep fake technologies, videos and audio have become more advanced and are increasingly used to spread false information (Helmus, 2022).

Disinformation is a problem that exists globally and is not limited to one particular language or nation (Colomina et al., 2021). Therefore, any information assessment system that aims to detect fake news must be language-agnostic. This means that the system must be capable of handling multiple languages, and its methods must be generalizable to other languages as well (Dementieva et al., 2023). Currently, there is limited support for multilingual fake news detection in existing techniques. However, some techniques (Ahuja & Kumar, 2023; Dementieva et al., 2023; Hammouchi & Ghogho, 2022) attempt to address this problem using a cross-lingual evidence approach. The basic idea behind this approach is that if a news item is true, it should be widespread in different languages and across media with different biases. Furthermore, the facts mentioned in the news should be identical. On the other hand, if the news is fake, it will receive less attention in the foreign press than true news. However, the effectiveness of these techniques relies on an unproven hypothesis. Other techniques, such as those presented by Chu et al. (2021) and Dementieva and Panchenko (2021) attempt to address the multilingual fake news problem by utilizing cross-language and cross-domain feature transfer. However, the accuracy of these techniques is currently limited.

A myriad of online resources for fact-checking exists (e.g., Snopes [https://www.snopes.com], PolitiFact [https://www.politifact.com], FactCheck [https://www.factcheck.org]). These resources typically classify articles as true or false or provide some other form of classification (X. Zhang & Ghorbani, 2020). However, most of these resources are either partially automated or depend entirely on manual detection by professional experts and organizations. This process is both time-consuming and expensive, as it requires a significant amount of human involvement to maintain such detection systems (Dale, 2017). Some of these sources can either only check claims or statements which contain statistical information or require high-quality training datasets and labels.

This section has explored the current methodologies for detecting disinformation and the associated challenges. The next section explores how current AI-in-Education (AIED) systems are reshaping educational landscapes through AI-based methods and platforms for trustworthy information.

EXISTING SYSTEMS USING OPEN LEARNING RESOURCES REPOSITORIES OR LINKED DATA

As AIED systems are helping students and teachers more and more in their activities, they must be equipped with techniques and methods for ensuring the trustworthiness of information. In this paper, a framework for detecting mis- and dis-information from any data, especially data found on the Internet, is presented. This new framework is called the EMIAS framework.

There is also another approach to ensuring the trustworthiness of information. It uses trustworthy repositories of learning resources and/or linked data and employs AI in tasks such as personalized recommendations and the ingestion of novel learning resources into the repositories. The preceding section presented previous work and research gaps concerning the detection of mis- and dis-information from any data, concentrating especially on data available online. This section presents previous research on AI-based methods and platforms for trustworthy information that are based, at least to some extent, on trustworthy repositories for learning resources, linked data, or ontologies.

Trustworthy and safe search engines for children are examples of AIEDs that may be used to provide the teacher with a set of suitable and trustworthy teaching materials. Along the same lines, a search engine for pupils should ensure that the information provided is reliable. There are some techniques for trustworthy search (Ramachandran et al., 2009) and search engines for children (Gossen et al., 2013) as well as for safe search (Patel & Singh, 2016). Along with search engines for education, there are several AI-based systems for scraping the web for teaching and learning resources. These include tools such as X5GON [https://www.x5gon.org/] and two commercial tools, Teacher Advi-

sor IBM [https://www.ibm.com/remotelearning/ibmresources.html] and Clever Owl [https://cleverowl.education/#About]. The tools under X5GON are freely available and have been developed as a part of the Horizon 2020 program for research and innovation funded by the European Union.

In addition to specially tailored search engines, linked data and learning object repositories offer an alternative resource for web scraping to find trustworthy learning materials. Pereira et al. (2018) present an educational recommender system that is based on social network interactions and linked data. Their system provides the learner with educational materials based on his educational context. The recommendation system is multimodal as it recommends text, video, and audio materials. Instead of freely scraping the web, the recommendation system is based on repositories of educational resources. The sheer number of open educational resources with varying subjects and learning goals distributed in different locations on the web calls for a recommendation system so that the user can find the relevant materials. The AI-based X5Learn (Perez-Ortiz et al., 2021) platform is another AI-based platform that is based on a repository of open learning resources. It contains multimodal data, such as video and text, and is aimed at teachers and students wishing to make the most of trustworthy open educational resources.

One solution for ensuring the trustworthiness of educational materials has been to resort only to educational content provided by trusted and well-known organizations. To increase findability and interoperability, trustworthy educational content has typically been provided using widely spread standards such as the Learning Object Metadata (LOM), originating from the IEEE Learning Technology Standards Committee (Duval & Hodgins, 2003). Many universities and educational institutions publish resources as LOMs. There are hundreds of repositories that are freely available on the web. Some of them are open repositories founded by consortia, and some are repositories managed by a single educational institution. The Merlot learning object repository [http://www.merlot.org] and MIT open courseware [https://ocw.mit.edu/] (Abelson, 2008) repositories are examples of these, respectively.

Metadata alone does not solve the problem of interoperability, and existing learning object repositories often suffer from interoperability issues that hinder their full exploitation. The adoption of linked data principles is one solution to alleviate the interoperability problem (Kawase et al., 2013). The OpenScout portal provides an example of its usage along with metadata. Linked data principles and the metadata scheme Dublin Core [https://www.dublincore.org/specifications/dublin-core/dces/] DC - LOM (Sutton & Mason, 2001) were used when building the OpenScout portal, which is the top repository of open educational data in the field of Business and Management (Kawase et al., 2013). OpenScout integrates over 23 repositories of learning objects.

Another example of using linked data to provide recommendations for people is the BROAD-RSI educational recommender system. It uses social network interactions and linked data to create recommendations for learning materials and for people with similar interests (Pereira et al., 2018). In their system, people and their interests are modeled using the FOAF (Friend-of-a-Friend) (Brickley & Miller, 2014) and SIOC (Semantically-Interlinked Online Communities) (Breslin et al., 2006) ontologies. These ontologies are commonly used to model people and their relationships and can also be used when assessing the trustworthiness of an author.

Having examined the current resources, repositories, and linked data employed by existing AI in Education (AIED) systems, we proceed to explore the specific requirements and information taxonomy necessary for information assessment by these systems in the next section. This section lays the foundation for the development of the targeted information assessment system, outlining its key requirements, functionalities, and the specific information taxonomy on which it is intended to operate.

AIED SYSTEMS FOR INFORMATION ASSESSMENT - REQUIREMENTS AND INFORMATION TAXONOMY

In addition to assessing the credibility of information, another vital aspect of information assessment is the harmfulness (Parker & Ruths, 2023). Harmful content can contain elements such as hate speech, propaganda, graphic content, vulgarity, and bullying (S.-Y. Lin, Kung, et al., 2022). The detection of harmful information is particularly important in educational institutions as it can significantly impact the learning process and the educational environment. When harmful content is present, it can hinder students' mental faculties and spread uncertainty, fear, and unrest within the institution. Therefore, detecting harmful information is essential to maintain a safe and conducive learning environment.

Another important aspect worth addressing in information assessment is building teachers' and students' trust and confidence in the assessment system. Users are usually skeptical about the decisions of such systems, which could hamper their adoption. To address this issue, teachers and students must be provided with the rationale for a decision. The explanation of a decision might include the justification of assigning a specific score to the piece of information by highlighting its individual parts that contributed to the scoring. Explanations can provide transparency in the assessment process and help educate teachers and students on the criteria and factors that are considered in determining the trustworthiness of the information. Providing explanations can also serve as a form of feedback to teachers and students. For example, if the information is rated as unreliable, the explanation can highlight why it was rated that way, providing valuable insights into potential biases, inaccuracies, or other issues with the information. More importantly, the explanations can provide a basis for discussion and debate, encouraging teachers and students to engage in meaningful dialogue about the credibility of information.

Due to the constant growth in social media platforms, information generation is not limited to only text. It is observed that visual modalities such as images and videos tend to attract more audiences on social media and spread faster than text-based information (Uppada et al., 2022). Hence, an information assessment system must be able to work with multiple modalities. Multimodal information assessment is still an emerging idea. There is no complete solution that considers a wide range of modalities (text, images, videos, audio, network features, spatio-temporal context, user and source context, community context, etc.). Also, the aspects of assessing credibility and harmfulness (such as hate speech, propaganda, graphic content, vulgarity, bullying, etc.) have not been studied in combination (Alam et al., 2021). Other limitations of the existing techniques include limited amount and diversity of data, susceptibility of performance degradation for manipulated or adversarial examples, limited ability to effectively combine features from different modalities with different dimensions, limited language scope, and limited interpretability. Hence, there is a need for a complete ecosystem through which the multimodal information must pass before reaching the user. The system must be able to distinguish between information authentication and harmfulness. At the same time, the user must be given insights into the factors that drive trustworthiness and quality of information. This is indispensable to gaining users' trust in the information assessment system and providing a human-centric approach to preserve human autonomy and oversight.

Information assessment by trustworthiness scoring could be more useful than a binary decision of fake information detection because it provides a more nuanced and comprehensive evaluation of the credibility of information. A binary decision of fake or genuine for a given piece of information oversimplifies the complexity of information and its evaluation. In reality, information can be misleading, incomplete, or have different degrees of accuracy and credibility. A binary decision reduces this to a simple true or false judgment, ignoring important nuances and context. This can lead to incorrect or misleading conclusions and decisions. Additionally, the determination of whether the information is

fake or genuine can be subjective and influenced by various biases and perspectives. A binary decision fails to take into account these subjective factors, which can result in inconsistent or unreliable evaluations.

To address these issues, the concept presented in this article goes beyond traditional disinformation detection techniques. Instead of treating credibility as a binary decision, we consider a multidimensional approach that includes assessing factuality, intention, and manipulation. This approach enables us to evaluate the information more comprehensively and to ensure a more accurate assessment of the information's impact on the educational environment. We emphasize that the problem of information assessment in education must be treated as a multidimensional issue that can be analyzed across three dimensions: factuality, intention, and manipulation. Factual assessment aims to determine the accuracy of information by identifying whether it is true, false, misleading, or fabricated. Intention assessment aims to uncover the motives or purpose behind the dissemination of information and the content tone, while manipulation assessment seeks to detect any manipulated, engineered, or AI-generated content. To achieve this, we developed a hierarchical taxonomy of disinformation that considers multiple dimensions of information.

The taxonomy outlined in Table 1 is based on several studies (Chong & Choy, 2020; Kapantai et al., 2021; Kumar & Shah, 2018; Wardle & Derakhshan, 2017) and shapes our system's design. It particularly influences the analytics operations necessary for assessing information, as detailed in the next section. Moreover, it contributes to determining the annotations process, which is vital for training AI algorithms to evaluate information across various dimensions effectively.

Table 1. Taxonomy for information assessment in education

Type of Information		Definition	Classes	Definition
Factuality		Determining the factual accuracy of the information.	Accurate	Contents based on verifiable facts.
			Misleading	Misleading, out-of-context content
			Fabricated	Wholly invented, devoid of fact content
Intention	Spread Intention	Determining motives behind the dissemination of information.	Informative	Contents shared for genuine information dissemination without hidden motives.
			Misinformation	Information spread without intent to deceive
			Propaganda	Content aimed to persuade or agitate, may or may not be factual.
			Hate Speech	Contents attacking individuals or groups based on their characteristics.
			Biased	Content that represents a one-sided view or omits key information.
			Neutral Intent	Aimed to be unbiased in presentation, though the content may still have biases.
	Content Tone	Describes the nature of language and presentation.	Indecent	Content that might be deemed crude, offensive, or vulgar based on societal standards.

Type of Information		Definition	Classes	Definition
			Biased Tone	Content that carries a tone favoring a particular stance or viewpoint, regardless of factual accuracy.
			Neutral Tone	Unbiased and balanced in presentation.
Manipulation		Nature of how content is displayed or modified.	Original	Unaltered, genuine content
			Doctored	Manipulated images, videos, and audio with the intent to deceive.
			Generative AI	Contents generated by AI systems.

For several reasons, analyzing information in the three dimensions of factuality, intention, and manipulation is crucial in the context of education and research. Factuality assessment ensures that accurate and reliable information is used for teaching and research purposes, avoiding the dissemination of false or misleading information that can lead to misinformed decisions or conclusions. Intention assessment helps to understand the motives or purpose behind the dissemination of information, which can affect how the information is interpreted and used. Propaganda or hate speech may have a different impact than accurate and objective information. Finally, manipulation assessment helps to detect any engineered or AI-generated content that may be used to influence opinions or perceptions. Manipulation analysis in the context of generative AI content is crucial because these contents are becoming more common in education and research. Students may use AI generative tools to create assignments, and fake images and videos related to education and research data may be circulated. These manipulations can lead to the spread of false or misleading information, which can harm the reputation of individuals or institutions or even affect the validity of research results.

Following our detailed exploration of the requirements, functionalities, and information taxonomy needed for an information assessment system, the next section will present the architecture of the proposed explainable, multimodal information assessment system. This includes a thorough discussion of each component's functions within the system, along with their specific requirements. This section is designed to provide a clear and detailed framework of the system, offering insights into how each component contributes to the overall performance of the information assessment process.

FRAMEWORK OF EXPLAINABLE, MULTIMODAL INFORMATION ASSESSMENT SYSTEM (EMIAS)

EMIAS conceptual framework targets the education sector to enhance students' ability to evaluate the information they come across critically. By taking input from the student, such as their answers to questions about the information they want to evaluate, the evaluation model can be tailored to the student's specific needs and understanding. This personalized approach can help students develop a better understanding of how to assess the accuracy and reliability of the information. Additionally, by involving the student in the evaluation process, the system can act as a tool to teach critical thinking skills rather than just a black box that works by itself. This approach can also help students understand the impact of factors such as tone/sentiment, source/author, and sharing behavior on the accuracy of information. Additionally, the systems built on the EMIAS framework can serve as recommendation engines to provide reliable and authentic sources for a required piece of information. Ultimately, the implementation of EMIAS in education can help students become more informed and responsible consumers of information in a world where disinformation and misinformation are increasingly prevalent. In our exploration of RQ3, we present the EMIAS framework, comprised of three interrelated layers as depicted in Figure 1. This section not only continues to elaborate on RQ3

but also intersects with RQ4. In the next sections, we explain the specific functions of these layers and how they collectively work towards accomplishing the objectives set forth in RQ3. This dual approach ensures a comprehensive understanding of the framework's operational dynamics that are in line with the research questions.

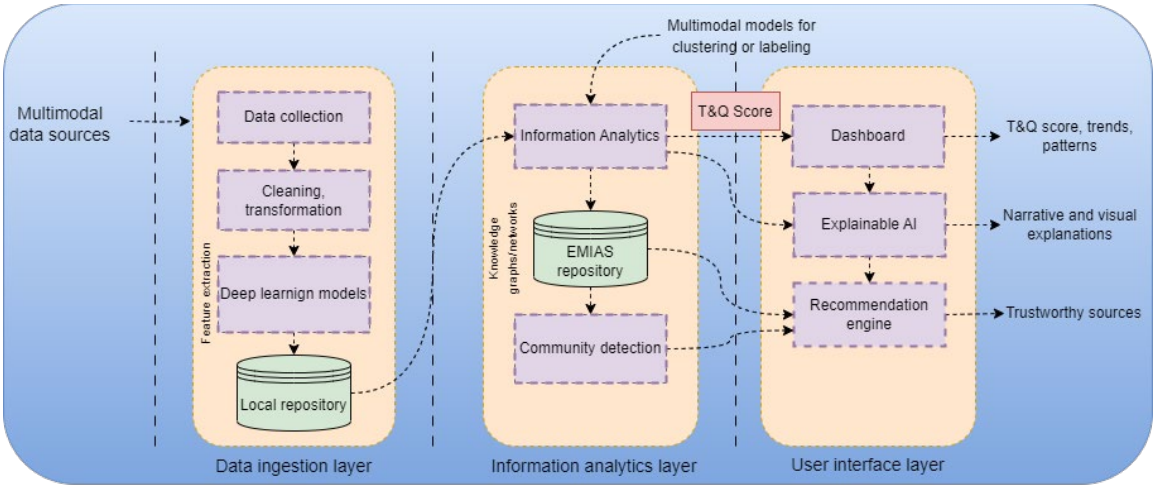

Figure 1. High-level depiction of the EMIAS framework

Layer 1: Data Ingestion

Data ingestion is the first layer of EMIAS, which plays a crucial role in the education use-case of the information assessment system by collecting, importing, and processing multimodal data from various sources. Relevant multimodal data sources could include digital textbooks, e-learning platforms, academic journals, research papers, online discussion forums, and social media platforms used by students and educators. To build a robust system for the education use case, data is collected from both trustworthy and untrustworthy sources to ensure the system can accurately identify the T&Q level of a wide range of information. For example, reputable academic journals, textbooks, and e-learning platforms, as well as online forums, social media posts, and other unverified sources, could be used.

The data collection component of the data ingestion layer shows several data collectors, as depicted in Figure 2. Web crawlers browse web pages to extract relevant data. Specific domains can be targeted, such as educational blogs or online learning platforms, to gather information related to education. Crawlers can extract various types of information, such as text, images, videos, and audio. Popular open-source tools used for web crawling include Scrapy [https://scrapy.org/], Pyspider [https://github.com/binux/pyspider], Webmagic [https://webmagic.io/en/], and Googlebot [https://developers.google.com/search/docs/crawling-indexing/googlebot]. Unlike web crawlers, which generally index the content, web scrapers target specific information from web pages for collection and analysis. They parse the web page's content and extract the required data. Popular open-source web scraping tools include BeautifulSoup [https://pypi.org/project/beautifulsoup4/] and ParseHub [https://www.parsehub.com/]. Similar to web scrapers, data harvesters are used to gather specific data from various sources, including the web. They are often more sophisticated, potentially combining data from multiple sources or different types of data. Data harvesters can use a combination of techniques, including APIs (Application Programming Interfaces) and direct scraping. Data collectors are broader in scope compared to the other tools. Data collectors gather information from various sources, not limited to web sources. This could include data from sensors, databases, file systems, and other repositories. Tools used for data collection vary widely depending on the data source and can range from simple scripts written in programming languages like Python or JavaScript to more complex data integration platforms.

Once the data is collected, it undergoes the preprocessing stage to ensure it is in a suitable format for analysis. In the education context, the preprocessing stage may involve cleaning the data by removing irrelevant or outdated information, handling missing or incomplete data, and standardizing text, images, and videos to a consistent format. The goal of this stage is to prepare the data for analysis by removing inconsistencies and irrelevant information.

After preprocessing, the data needs to be annotated with appropriate labels to train the AI model. In the education context, this involves assigning T&Q levels to each piece of information based on its trustworthiness and quality. Domain experts or fact-checkers could evaluate the information and provide the necessary annotations. For example, academic papers could be evaluated based on the reputation of the journal or the author, the quality of the research methodology, and the accuracy of the results. The annotators may also use a standardized set of criteria or guidelines to ensure consistency in the annotation process.

Once the data is preprocessed and annotated, it is stored in the ingestion repository. In the education context, the ingestion repository could store diverse forms of data, including text, images, videos, and audio, in the format in which they are originally or processed if it suits the specific use case. The repository holds structured and unstructured data at any scale and is a general-purpose repository for storing diverse forms of data.

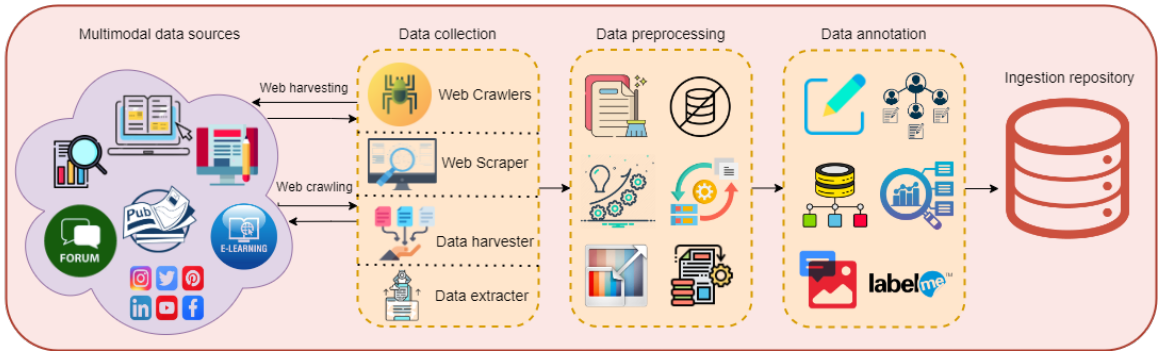

Figure 2. Multimodal data ingestion layer

LAYER 2: INFORMATION ANALYTICS

The information analytics layer is responsible for running inferences on the data collected in the first layer and generating the information assessment results. These results are stored in a repository for further analysis. The following sections describe in detail the function of the information analytics layer.

Data analytics module

The Analytics Layer is a crucial part of the EMIAS system that uses content analysis services to extract useful information from the multimodal data collected in Layer 1. The T&Q score is a critical aspect of the EMIAS system as it helps identify the trustworthiness and quality of a piece of information based on factuality, intent, and manipulation. EMIAS computes the T&Q score using a weighted algorithm. The weight assigned to each factor may vary based on the specific use case or context. For example, in an education context, the weight assigned to factuality may be higher than that assigned to intent or manipulation. The T&Q score is computed on a scale of 0 to 1, with 1 representing the highest level of trustworthiness and quality and 0 representing the lowest. The score can be used to classify the information into different categories, such as highly trustworthy, moderately trustworthy, or untrustworthy.

The data analytics module extracts useful information from the multimodal data collected in Layer 1. The extracted features can be used to detect clusters with consistent information features. The features can be extracted using deep learning models, which encode the information as dense vectors. Various deep learning models can be used for feature extraction, such as classical non-contextual word embeddings like Word2Vec and contextual paragraph embeddings like BERT and more recent GPT-family models. Recently, embeddings have also become available for images, such as convolutional neural networks (CNNs), and videos, such as 3D convolutional neural networks. These models can be used separately for each modality or combined to create multimodal embeddings, such as text and images (Rombach et al., 2022) or videos. Some recent approaches, such as ImageBind (Girdhar et al., 2023), can learn joint representations across multiple modalities and could be more effectively employed for extracting multimodal embeddings. Multimodal embedding models such as Vertex AI's multimodal model [https://cloud.google.com/vertex-ai/docs/generative-ai/embeddings/get-multimodal-embeddings] and Amazon's Titan [https://docs.aws.amazon.com/bedrock/latest/userguide/titan-multiemb-models.html] can also generate embeddings of multiple modalities in the same semantic space with the same dimensionality.

Once the embeddings are obtained, clustering can be applied. Clustering helps to group similar pieces of information based on their T&Q level. A cluster will have a specific T&Q level based on the characteristics of fake or harmful elements contained in it. The AI model trained to identify the cluster for a specific piece of information can also learn to identify key influencers within the cluster who are particularly effective at spreading such information. A common approach is to integrate data modalities through learning and embedding that encodes multiple data modalities into a shared latent space, from which any clustering algorithm that accepts the embedding as input could be used for clustering (X. Lin, Tian, et al., 2022).

The factuality, intent, and manipulation of the information can be detected based on the characteristics of the cluster. For instance, if a cluster contains mostly factual and unbiased information, it will have a high T&Q score. On the other hand, if a cluster contains a large number of fake or manipulated pieces of information, it will have a low T&Q score. Similarly, if a cluster contains a significant number of pieces of information with malicious intent, such as spreading rumors or misinformation, it will have a low T&Q score. A recent ensemble learning framework called SnapCCESS (Yu et al., 2023) employs variational autoencoders to create multimodal embeddings and learn the multiple embeddings to generate consensus clusters. SnapCCESS is implemented as an open-source Python package [https://github.com/PYangLab/SnapCCESS]. Though originally purposed for clustering cells, this framework is useful for clustering pieces of information.

Zero-shot learning (Xian et al., 2019) can also be used as an alternative approach to computing T&Q scores. Zero-shot learning has been recently used for multi-modal, multi-label classification (He et al., 2022). The primary advantage of zero-shot learning is that it can be trained on a smaller corpus, and it can efficiently predict unseen examples. This approach can be applied as a multi-label classification problem, where the labels are the three types of disinformation in the taxonomy - factuality, intent, and manipulation.

Figure 3 shows the information assessment using supervised and (semi-)supervised learning. To define the label space, we use our taxonomy (Table 1), which defines the semantic attribute or classes for this task. The training dataset for the system also includes multimodal examples that cover the hierarchical taxonomy. The data ingestion layer collects the relevant dataset. The deep learning models extract the multimodal features of the dataset. Then, a sufficient number of examples covering each taxonomy can be shown to train the system. For instance, the system can be trained to recognize factual statements by showing it a set of factual and non-factual statements. Similarly, to train the system to recognize intent and manipulation, a set of statements with different levels of intent and manipulation can be shown.

Once the system is trained to recognize these factors, it can compute the T&Q score of a piece of information. For instance, consider an article about a recent education policy. The system can analyze the language used in the article and determine the factuality of the claims made, the intent behind the article, and whether it contains any elements of manipulation or bias. In this multi-label classification problem, the individual probabilities of each class in taxonomy can be predicted, and the weight of each class can be found. If the article has a high probability of being factual but a low probability of having malicious intent, the system can assign a higher weight to factuality and a lower weight to intent in the computation of the T&Q score. This can lead to a more accurate assessment of the trustworthiness and quality of the information. This computation can be represented by the following equation:

$$T\&Q = W_1F + W_2I + W_3M \quad (1)$$

where F represents the factuality score of the information, I represents the intent score of the information, M represents the manipulation score of the information, and W_1, W_2, W_3 are percentages assigned to each factor, respectively. The weights can be adjusted based on the specific use case, e.g., more weight to W_1 than W_2 and least to W_3.

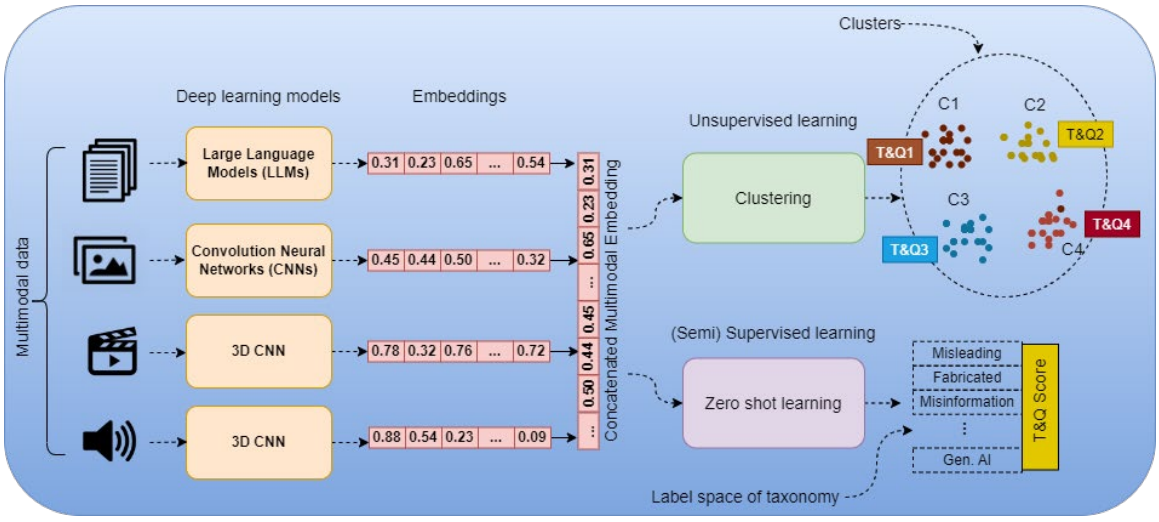

Figure 3. Information assessment using unsupervised and (semi) supervised learning

EMIAS repository

The EMIAS Repository is a database that stores the output of the information analytics module in the form of knowledge graphs. This is useful for identifying communities of densely interconnected data points and understanding the underlying structure of the data. For instance, knowledge graphs built on information analysis can be used to identify groups of researchers or organizations that are working together in a particular field. This can provide insights into the underlying structure of the data and help to identify any biases or conflicts of interest that may exist within the community.

The EMIAS Knowledge Graph (KG) is a formal mathematical and computational structure, denoted as **G = {V, E, A}**. It's a directed, attributed, multi-relational graph aimed at representing a multitude of elements and their relationships within the educational context of the EMIAS framework.

- **V** is the set of nodes. Each node **v ∈ V** in the graph represents an entity within the EMIAS system. Entities fall into two main categories: information sources such as lecture videos, research papers, textbooks, educational podcasts, and case studies, and influential entities such as teachers, educational institutes, publishers, and researchers.

- **E** is the set of edges in which every edge $\mathbf{e} \in \mathbf{E}$ is an ordered pair $(\mathbf{v_1}, \mathbf{v_2})$ that represents a relationship from node $\mathbf{v_1}$ to node $\mathbf{v_2}$. The edges are multi-relational, representing different types of relationships like similarity, influence, or contribution. For example, an edge between two information sources could represent an information exchange such as 'research papers cite educational material' (Fig. 3). An edge from an influential entity to an information source might represent influence or contribution.
- **A** is the set of attributes in which each node **v** has an associated attribute vector $\mathbf{a} \in \mathbf{A}$, which contains specific properties of the entity. The attributes are multi-dimensional, capturing different aspects of the entity like T&Q scores, publication date, authorship, reputation score, or influence score.

In Figure 4, we present a visual representation of the EMIAS Educational Repository. The diagram depicts the complex relationships between different types of information sources and influential entities in the higher education ecosystem. The influential entities contribute to the creation, dissemination, evaluation, and promotion of information sources within EMIAS. The EMIAS Educational Repository node serves as a central hub, connecting to different clusters of information sources based on their T&Q scores and relationships. Each cluster contains a diverse mix of information source types. These types include lecture videos, research papers, textbooks, educational podcasts, and case studies, each represented by a unique shape and color.

The EMIAS repository node links to two primary nodes: information sources and influential entities. Each information source can branch into sub-categories. For instance, *education material* splits into textbooks, video lectures, and podcasts, and these can further branch out. Similarly, influential entities can have their sub-categories. These nodes can interrelate or have connections within their categories. An example of a relationship within influential entities is *teachers collaborating with researchers*, as depicted in Figure 3. Another instance is the link between an influential entity and an information source, such as *teachers using textbooks*.

A number of open-source graph databases can be used to build EMIAS knowledge graphs and repositories. Some notable open-source, massively scalable graph databases include Neo4j [https://neo4j.com/], JanusGraph [https://janusgraph.org/], ArangoDB [https://arangodb.com/], and OrientDB [https://orientdb.org/]. A GitHub repository [https://github.com/totogo/awesome-knowledge-graph] provides a curated list of multimodal knowledge graphs, including datasets and research papers, which could serve as a valuable resource for building multimodal knowledge graphs and implementing its operations.

Community detection

Community detection and trustworthiness assessment are interrelated in the sense that community detection can be used as a tool to identify connections between pieces of information that can be used for identifying trustworthy and harmful communities in the field of education. This can provide insights into the underlying structure of the data and help to identify any biases or conflicts of interest that may exist within the community. Community detection algorithms, such as the Louvain algorithm or the Girvan-Newman algorithm, can be applied to the EMIAS's knowledge graphs to identify clusters of highly interconnected nodes or entities.

Furthermore, community detection can also be used to identify clusters of sources with similar characteristics, which can help in assessing the trustworthiness of the sources. For example, if a community of sources is found to be closely interconnected and consistently publishing high-quality academic data, then it can be considered a highly trustworthy community. In addition, community detection can be used to identify groups of sources that are spreading disinformation or misinformation. For instance, if a community of sources is found to be interconnected and consistently publishing false or misleading information, then it can be considered a harmful community. This information can be used to inform the T&Q score of the information associated with these sources, allowing for a more accurate assessment of its trustworthiness and quality.

Neo4j graph database implements several community detection algorithms in knowledge graphs, which are explained in detail in Neo4j's official manual [https://neo4j.com/docs/graph-data-science/current/algorithms/community/].

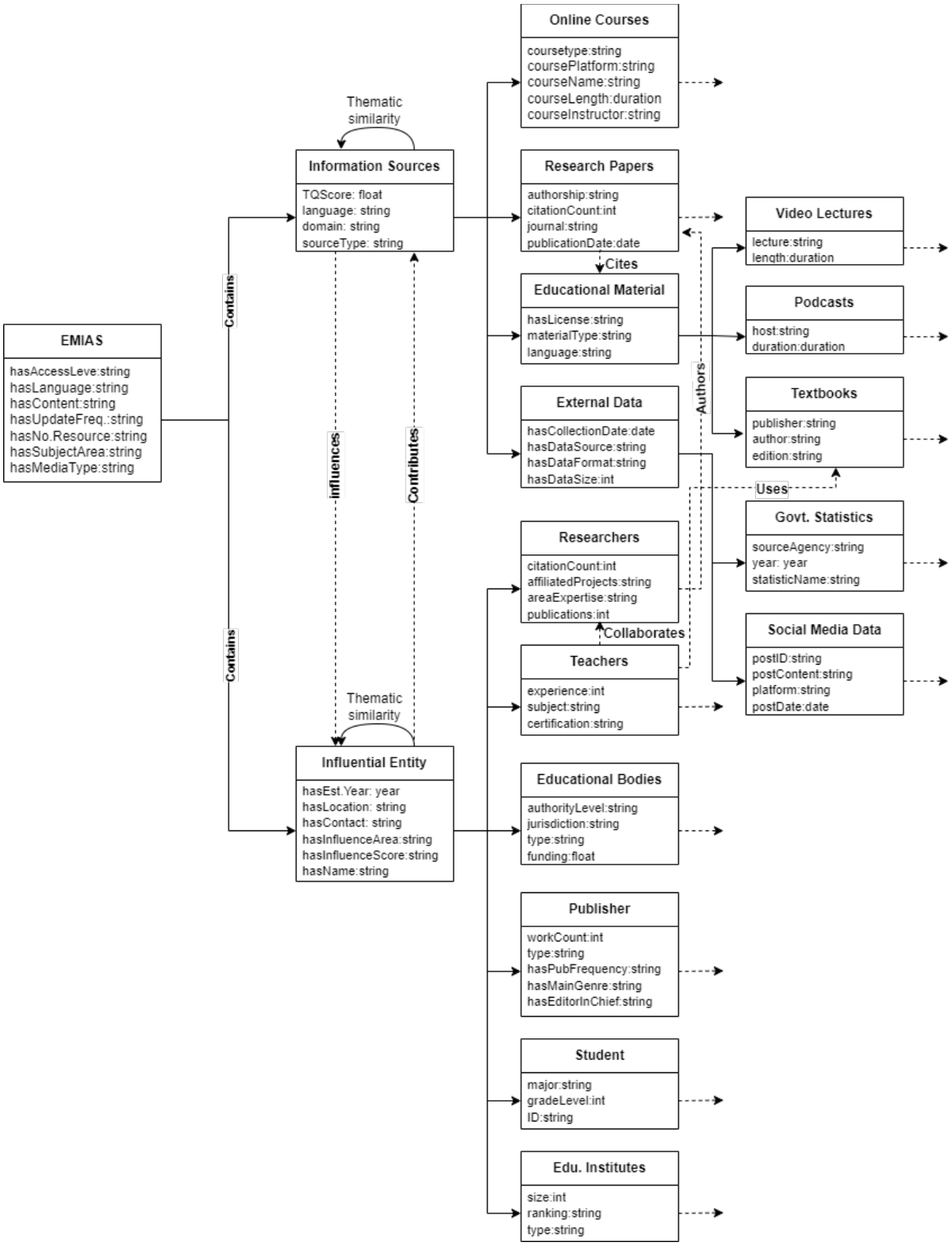

Figure 4. Ontological representation of the EMIAS repository

Layer 3: User Interface (UI) Layer

This section focuses on one of RQ3's key objectives: providing explanations for AI decisions to foster trust and understanding. The third layer of EMIAS, the UI layer, is responsible for providing users with access to the Explainable AI module, which allows them to understand how the system makes decisions and recommendations. The Explainable AI module in Layer 3 provides a user-friendly interface for interacting with the AI models and accessing explanations of their decisions. The module can provide users with information about the data used in the analysis, the algorithms and techniques employed, and the outcomes of the analysis. This transparency and interpretability can help build trust in the system and allow users to verify the validity of the recommendations made.

The User Dashboard is another important component of the UI layer. This is where users can access and interact with the data and results generated by the system. The dashboard can display information such as the T&Q scores of individual pieces of information, trends, and patterns in the data, as well as alerts for potentially harmful or suspicious content. The dashboard can be customized to meet the needs of different users, such as educators, researchers, or administrators. For instance, the dashboard can provide a teacher with a comprehensive view of their student's assignments and their compliance with the T&Q taxonomy.

Figure 5 depicts the EMIAS user-interface layer, which addresses the 'interpretability' and 'explainability' aspects of RQ3. The Explainable AI module can analyze the content of each assignment and assign a T&Q score, which is then displayed in the dashboard. To check students' assignments based on the taxonomy, the teacher can filter the assignments according to the T&Q score and the specific factors of the taxonomy that are relevant to the assignment. For example, if the assignment is a research paper on a particular topic, the teacher can filter the assignments according to the "factuality" factor to assess the accuracy of the information presented. The teacher can also use the dashboard to identify patterns or trends in the students' assignments. For instance, if a particular student consistently receives low scores for the "intent" factor, the teacher may need to provide additional guidance to help the student understand how to avoid biased or misleading language in their writing. Furthermore, the reports generated by the EMIAS system can help the teacher to identify any potential issues with the assignments. For example, if the reports show that a particular source cited by multiple students has a low T&Q score, the teacher can investigate whether this source is trustworthy and make appropriate adjustments to the assignments if necessary.

The Explainable AI module in the UI layer can help students find relevant sources for their research projects, assignments, and examinations. They can input their research topic or question, and the system can provide a list of sources that are relevant to the topic. The sources can be filtered based on the taxonomy mentioned earlier, which can help students in finding credible sources. Similarly, the dashboard can be used by students to check their assignments for quality and credibility. For instance, the T&Q score can be used to assess the trustworthiness of the sources used in the assignment, and the report generated can be used to identify areas for improvement.

Several explainable AI frameworks can be employed for interpreting AI decisions. Notable open-source explainable AI frameworks for this purpose include SHAP [https://shap.readthedocs.io/en/latest/], LIME [https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/], What-if Tool [https://pair-code.github.io/what-if-tool/], and AIX360 [https://github.com/Trusted-AI/AIX360].

This section provided an in-depth exploration of the design elements of an explainable, multimodal information assessment system, detailing the functionalities of its individual components. Moving forward, the next section will address the various challenges inherent in the development of this system.

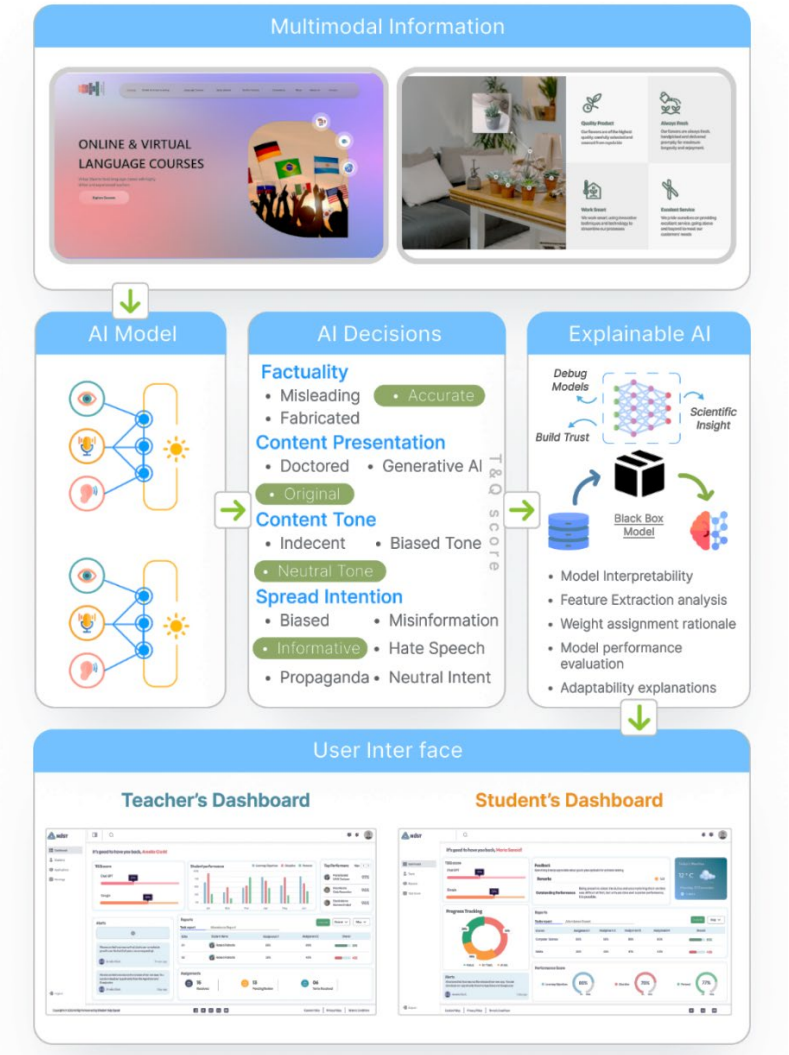

Figure 5. Depiction of EMIAS user-interface layer

CHALLENGES IN DEVELOPING EMIAS

Having addressed RQ1-RQ3, which guided us in developing the EMIAS framework, we now shift our focus to RQ4. This section is dedicated to discussing the potential challenges associated with the development of EMIAS.

The endeavor to integrate AI into the academic landscape is marked by both promise and complexity. While the potential benefits of this union are vast, the road to achieving a seamlessly integrated Explainable Multimodal Information Assessment System (EMIAS) also has potential challenges. One of the most significant issues is the overpowering influence of major technology corporations in shaping the AI ecosystem. Their dominance often dictates the direction, development, and accessibility of AI tools, which might not always align with the unique and diverse needs of the academic community. Additionally, the intrinsic nature of AI, which is a reflection of its human developers and the data it is trained on, means that biases – whether overt or subtle – are often embedded within its algorithms. Such biases, when left unchecked, can introduce distortions and inaccuracies in information assessment. This is especially concerning for the academic world, where the accuracy and fairness of information are paramount. The complex process of AI-enabled information assessment in

higher education is not only about integrating technology but also about navigating these broader socio-technical issues, ensuring ethical considerations are at the forefront, and consistently refining the approach based on evolving needs and discoveries.

Table 2 provides an overview of these challenges. It not only pinpoints the roadblocks that educators and technologists might encounter but also emphasizes the depth and breadth of considerations required when fusing AI and academia.

Following the identification of challenges in developing EMIAS, the next section will discuss how this system can effectively support both students and teachers in their roles concerning information assessment. Additionally, this section will also discuss the ethical considerations integral to the EMIAS framework.

Table 2. The challenges in developing ethical and trustworthy AI in education

Challenge	Description	Potential risks	Potential solutions
Diverse information sources	Addressing various types of media, including AI-generated content.	Academic misinformation due to misleading sources.	Implement stringent content vetting processes; promote cross-referencing.
AI & media literacy	Educating users on AI's role in shaping media content.	Over-reliance or blind trust in AI outputs.	Develop comprehensive educational programs detailing AI's media influence.
Language & cultural diversity	Providing equitable resources for a global academic audience.	Knowledge gaps for underrepresented languages.	Integrate multilingual AI modules; foster partnerships with diverse academic entities.
Data dynamics & integration	Continuously updating content and integrating AI within the broader educational framework.	Data inconsistencies and high integration costs.	Adopt continuous integration and deployment frameworks.
Educator & student training	Enabling educators and students to effectively utilize AI systems.	Misapplication of AI and compromised educational standards.	Offer immersive AI training modules emphasizing hands-on experience.
Content bias & appropriateness	Addressing inherent AI biases and ensuring pedagogical relevance.	Distorted academic perspectives and loss of learning objectives.	Deploy AI audit mechanisms; align AI evaluations with pedagogical goals.
Privacy & data governance	Protecting student data while employing AI for content trustworthiness.	Ethical concerns and potential privacy breaches.	Implement robust encryption standards; clarify data handling protocols.
Data Acquisition and Annotation	Collecting and accurately annotating diverse datasets.	Quality issues leading to incorrect AI insights.	Collaborate with stakeholders for dataset collection; employ expert annotators.
Biased annotations	Addressing potential biases in AI data labeling.	AI predictions that perpetuate existing biases.	Use bias-detection algorithms; conduct manual audits periodically.

Challenge	Description	Potential risks	Potential solutions
Modalities fusion	Integrating various types of media for comprehensive analysis.	Incomplete or inaccurate content representation.	Apply advanced data integration techniques; ensure cohesive content analysis.
Generalizability, Interpretability and Robustness	Ensuring AI systems work across various scenarios and are understandable.	Reduced AI reliability and utility in diverse academic settings.	Design modular AI architectures; emphasize user-friendly interfaces.
Tech company dominance and proprietary models	Addressing the control of large tech companies over AI advancements.	Reduced openness and innovation, increase dependence on external tech entities.	Promote open-source AI initiatives; enforce regulation; reduce reliance on tech giants.
Data availability	Ensuring sufficient and varied data for AI system training.	Inadequate AI model training leading to skewed outputs.	Forge diverse data partnerships; encourage varied and representative datasets.
Dynamic data	Addressing the ever-evolving nature of online academic content.	Outdated content leading to academic obsolescence.	Create feedback loops with the academic community; regularly update content.
AI integration challenges	Ensuring seamless fusion of AI systems with other educational tools.	Disruptions in the learning experience and system inefficiencies.	Prioritize user experience (UX) design; ensure seamless AI integration.
AI isolation and legacy systems	Addressing standalone AI apps lacking broader network integrations	Reduced system versatility and over-reliance on singular AI tools	Encourage interconnected AI ecosystems; discourage standalone applications.

STUDENTS AND TEACHERS MANAGING INFORMATION

Building on our response to RQ4, this section explores how EMIAS can significantly aid both students and educators in effectively managing and assessing online information, ensuring a reliable and informed educational experience. In addition, it also discusses the ethical considerations vital for the effective and responsible incorporation of the EMIAS framework in educational contexts.

SUPPORTING STUDENTS AND TEACHERS FOR FACT-BASED DECISION MAKING

AI-based solutions in education should be considered as a support tool for students and teachers. The previously presented taxonomy (Table 1) and the EMIAS framework (Figure 5) can potentially support students and teachers when they need to evaluate the accuracy, originality, tone neutrality, and intention of information. The EMIAS framework supports students and teachers while they are making decisions to trust and use online information. Figure 5 shows that the EMIAS aims to add value by creating AI-based decisions regarding the trustworthiness of information. In practice, the AI models of EMIAS try to find cues and indications about misleading, fabricated, doctored, or biased information, hate speech, or propaganda. Its explainable AI features provide adaptability explanations, interpretations of used models, and performance evaluation for students and teachers. This helps them to understand how and why the EMIAS ends up making such decisions.

Students use social media and other internet information sources, but they do not consider them the most reliable source for searching for information (Vranias & Kyridis, 2022). The conceptual approach of the EMIAS framework could support students in making decisions about the trustworthiness of information by using various Internet information sources such as social media. The level of trust in information has also decreased during the last year (Latkin et al., 2020), which creates more demand for solutions like EMIAS. For example, the surveys of Statista (2024a) and Statista (2024b) show that a significant number of adults do not trust the news media or Internet-based information. However, the level of trust differs around the world, which makes national variations between countries and sets some customization requirements for solutions like EMIAS. On average, 59% of adults trusted search engines, 59% trusted traditional media, and social media is the least trusted media (Statista, 2024a). According to Statista's news media research (Statista, 2024b), the highest number of adults (69%) trusted news media in Finland, whereas only 26% of people in the United States, 29% of French, 41% of Australians, or 44% of Japanese trusted news media. In addition to the news media itself, other actors on the Internet may also affect people's trust in information. Sterrett et al. (2019) found that people trust social media information more if it has been shared by a public figure that they trust than by someone they do not trust, regardless of the source of news. The AI could also model and evaluate the trustworthiness of opinions among famous people and politics.

Balancing AI and Media Literacy in Education

Adopting AIED systems in higher education changes the pedagogical requirements of media literacy education. While students and teachers begin to use EMIAS-typed tools for detecting dis- or misinformation in their studies and teaching, they should learn to collaborate with AI-based tools in their decision-making. The traditional way to detect trustworthy information from fake news and other biased information has been to develop the media literacy skills of students in higher education. The development of students' media literacy is an educational goal where teachers aim to teach skills to detect trustworthy information and reliable information sources. Thus, media literacy education improves students' media use, practices, cultures, media literacies, and teachers' instructional methods and pedagogy of media literacy (Rasi et al., 2019). The EMIAS tools digitize a part of media literacy practices by setting new learning and teaching goals regarding human-AI interaction. The goals of AI literacy provide elements from which media literacy education could benefit. Namely, skills to understand, use, and evaluate artificial intelligence in general have begun to be emphasized in higher education.

As artificial intelligence is integrated into almost every sector of our daily lives, researchers have begun to conceptualize artificial intelligence literacy as a generic learning skill in education (Laupichler et al., 2022). One of the key goals of AI literacy is to improve individuals' abilities to evaluate the fairness, accountability, transparency, ethics, and safety of AI (Ng et al., 2021). As data plays a significant role in AI-based analytics, the trustworthiness of information is dependent on the quality of data used in AI solutions (Alamäki et al., 2019). For example, UNESCO's report (Pedró et al., 2019) highlights ethics and transparency in data collection, use, and dissemination as one of the educational challenges regarding AI. Hence, students and teachers need to be able to evaluate the trustworthiness of information, whether it is generated and shared through traditional digital media or AI-enhanced digital media. Researchers have framed the concept of AI readiness as a contextualized way of helping people understand data-driven AI (Luckin et al., 2022). AI readiness focuses on the basic understanding of how AI is used and how it creates value in a specific context. Students and teachers need to understand the basics of AI, namely, they should have sufficient AI readiness as a part of their media literacy skills to successfully collaborate with the EMIAS-typed tools while interacting with EMIAS. In practice, this means skills to critically evaluate some suggestions that AI generates and understand that tools like EMIAS are continuously developing while they are used. Students and teachers can also misinterpret the analysis although the source information is correct. In addition, students and teachers should make the final decision to trust the information that they are using and

processing since AI does not have accountability, but it should be seen rather as a support tool than a decision maker.

Ethical Considerations in the EMIAS Framework

The presented EMIAS framework potentially improves the ethical use of online information, but simultaneously, it might create ethical concerns if misused. Ethical guidelines in education are needed as students and teachers can misuse AI by analyzing, using, or monitoring content, behavior, or performance that they should not need to do or know. AI tools, as rapidly developing technologies or with poorly trained and tested algorithms and data, may also produce biased analyses that teachers or administration can use. Ethical concerns arise in situations where AI-generated analyses cannot be verified transparently or credibly. These issues happen in situations when teachers or administration do not know why something happens or why AI recommends or draws such conclusions. The explainability of AI features in the EMIAS aims to decrease those visibility risks.

The European Commission, UNESCO, and other organizations have released their ethical guidelines for artificial intelligence (European Union, 2022; UNESCO, 2021). The ethical guidelines form a comprehensive framework to review all aspects of AI and its adoption that influences the ethical adoption and use of AI. The ethics of AI means artificial intelligence that works correctly from the perspective of trustworthiness, morals, human life, and social norms. Trustworthy AI plays a significant role in policies (e.g., European Union, 2022; OECD Legal Instruments, 2024) and research papers. The European Commission's ethical guidelines for education (European Union, 2022) list the following key requirements for trustworthy AI: human agency and oversight, transparency, diversity, non-discrimination, and fairness, societal and environmental well-being, privacy and data governance, technical robustness, and safety and accountability. The requirements are aligned with the European Commission's High-Level Expert Group's (European Commission, 2020) assessment list for trustworthy artificial intelligence. According to Adams et al. (2023), AI-enhanced learning and teaching ethics guidelines, both K-12 and others, employ the following core principles: transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, and freedom & autonomy. While developing and implementing solutions like the EMIAS framework proposes, those ethical guidelines provide comprehensive checklists to ensure the ethical use of AI.

After discussing how EMIAS supports students and teachers in their respective roles and what the ethical considerations in the EMIAS framework are, we will ultimately proceed to a discussion on the proposed EMIAS framework, highlighting the potential opportunities unraveled by EMIAS in higher education and its transformative role in higher education.

Discussion

Transitioning from traditional literacy, students and teachers need to be well-versed in navigating the digital landscape. They must understand the subtleties of online information dissemination, its platforms, and the potential pitfalls it conceals (Khan et al., 2021). As we enter an era dominated by technology, proficiency in utilizing AI tools becomes indispensable, particularly in the fields of education and research (Cardona et al., 2023). These tools have transformed various aspects of academia, from expediting literature reviews to facilitating global collaborations among researchers and enhancing scientific writing (UNESCO, 2023). Introducing students to these tools early on familiarizes them with their capabilities and limitations. However, as we embrace this AI-focused future, both students and educators must possess the ability to critically evaluate and validate the output of generative AI (Cornell University, Center for Teaching Innovation, n.d.). This skill set is crucial for ensuring information authenticity, guarding against algorithmic biases, and nurturing a comprehensive understanding that combines technological expertise with critical human inquiry.

How Does EMIAS Address Existing Issues?

EMIAS addresses the research gaps outlined in the Related Work and Research Gaps section. It employs a multi-modal approach that extends beyond textual data analysis to include images, videos, and audio content. This is especially relevant given the advancements in generative AI technologies that are increasingly used to manipulate digital media and generate convincing fake content across different data modalities (Almars, 2021). The design philosophy guiding EMIAS addresses the gaps pertaining to simple binary classification, as identified by Kim et al. (2021) and Thota et al. (2018).

Incorporating both human and technological knowledge (Lampridis et al., 2022) makes EMIAS a robust solution that leverages humans' ability to interpret complex and contextual information. The integration of transparency and explainability features into its operation ensures users have clear insight into its decision-making processes, fostering trust and promoting wider adoption, the issues mentioned by Scharowski et al. (2023) and Szczepański et al. (2021). In this way, EMIAS ensures that while AI plays an instrumental role in processing vast amounts of data rapidly and accurately, final decision-making power remains with the human users who can leverage their understanding to evaluate AI-generated outcomes.

Addressing multilingual capabilities is another critical aspect where EMIAS strives to meet user needs effectively. Given that disinformation is not restrained by language or geographical boundaries (Colomina et al., 2021), it is imperative for an effective solution like EMIAS to be capable of handling multiple languages while maintaining its efficacy across them, an issue highlighted by Dementieva et al. (2023).

By reducing dependency on manual fact-checking resources, an issue highlighted by Dale (2017), EMIAS eliminates potential bottlenecks associated with time-consuming human involvement without compromising accuracy or reliability. It proposes to use advanced machine learning algorithms capable of rapidly sifting through substantial amounts of online data from various sources and identifying false information.

Why Does the Need for a Human-in-the-Loop Approach Persist?

While automated information assessment systems like EMIAS can substantially alleviate the information assessment load for stakeholders in an educational environment, the competency to validate AI system outcomes remains a critical human task. This skill can be effectively imparted through hands-on workshops, integrating real-world case studies and interactive simulations that focus on understanding the underlying mechanics and potential biases of AI.

While EMIAS is intended to offer an advanced approach to filtering out disinformation, educators, as primary knowledge sources for students, still face the challenge of the subtleties of the information they present. Their role goes beyond mere content delivery; they interpret, contextualize, and enrich knowledge. Thus, a holistic strategy is crucial even in the age of AI-driven solutions. Professional development that emphasizes digital literacy and critical thinking is indispensable for teachers (Widana, 2020). EMIAS might identify potential issues, but teachers need the skills to discern and contextualize this feedback, distinguishing between genuine content and potential misinformation. Platforms where educators collaboratively verify AI-flagged information leverage the power of collective expertise, ensuring a human touch remains in the interpretation process. Frequent dialogues with IT and AI experts can equip teachers to understand the limitations and capabilities of systems like EMIAS, enabling them to use such tools more effectively. In essence, by fostering continuous learning and adaptability, we can ensure that educators are not only relying on AI systems but are active participants in the fight against disinformation.

What Opportunities Arise from the Integration of EMIAS in Higher Education?

The integration of an advanced system like EMIAS promises a multitude of opportunities. Not only does it cater to enhancing the technological framework, but it also extends benefits to students, teachers, organizations, and society at large. As institutions of higher education consider the implications of integrating such a system, it is essential to understand the comprehensive spectrum of possibilities it unfolds. Table 3 presents the myriad of opportunities across various sectors, illustrating the transformative potential of EMIAS in contemporary education.

Table 3. Multifaceted opportunities enabled by the integration of EMIAS in higher education

Category	Opportunities
Technological Opportunities	• Seamless integration for real-time assessment and feedback • Advanced algorithms for multi-modal data analysis • Creation of student/teacher-specific dashboards • Secure data storage and transmission methods
Data-Driven Opportunities	• Access to diverse data sources for enhanced accuracy • Collaborative data acquisition and annotation for more effective system training • Data validation checkpoints to ensure reliability • Optimized data processing methods for rapid feedback
Student-Centric Opportunities	• User-friendly interface aiding in research and content validation tasks • Integration into learning management systems for real-time corrections • Training modules on system functionalities • Encouraging student feedback for continuous improvement
Teacher-Led Opportunities	• Customized functionalities assisting in the course material vetting • Professional development modules for educators on system use • Adaptability to set system parameters based on class-specific needs • Collaborative content validation among educators
Organizational Opportunities	• Investment in robust IT infrastructure for optimal system performance • Regular audits and quality checks to ensure effectiveness • Engagement with system developers for institutional-specific solutions • Embracing AI-driven decision-making
Societal Opportunities	• Public demos and seminars to raise awareness • Gathering feedback to align the system with societal values • Collaboration with community stakeholders for ethical use • Encouraging debates on AI's role in education and potential biases
Policy and Regulatory Opportunities	• Advocacy for policies supporting the system's integration in institutions • Compliance with data protection regulations • Engagement with policymakers on the system's advantages and concerns • Guidelines for the ethical use of such systems in education

The technological opportunities stemming from EMIAS' user interface layer allow real-time assessment and feedback for students and teachers, addressing the need for quick response times in the

fast-paced digital era. The application of advanced algorithms for multi-modal data analysis provides improved accuracy and reliability in information assessment. Customized dashboards cater to specific user needs, while secure data storage ensures the confidentiality and integrity of user information.

The user interface layer further offers student-centric opportunities geared towards enriching learning experiences. An intuitive user interface aids navigation and simplifies system interaction, promoting active use among students. Real-time correction capabilities integrated into learning management systems facilitate immediate remediation, enhancing student learning outcomes. System-specific training modules increase student competency in system operation, while continuous improvement is encouraged through active solicitation of student feedback.

The user interface layer also offers teacher-led opportunities focused on empowering educators through customized functionalities assisting in course material vetting. This ensures content alignment with learning objectives and suitability for class level. Professional development modules improve educator competency regarding system utilization. Flexibility to set parameters based on classroom requirements addresses teaching variability, while collaborative content validation provides collective expertise to validate AI outcomes.

EMIAS's data ingestion layer produces data-driven opportunities that arise from accessing diverse data sources, enhancing accuracy through variety and volume, along with collaborative data acquisition that aids in training more effective AI systems. Ensuring data validation at several checkpoints upholds the reliability of the assessment system while optimized processing methods expedite feedback provision.

Organizational opportunities stemming from EMIAS' data analytics layer emphasize investment in IT infrastructure for optimal performance. Regular audits ensure effectiveness while engaging with system developers and allow customization to meet institutional needs. It also encourages embracing AI-powered decision-making capabilities to leverage its predictive power for improving academic outcomes.

EMIAS's aspects of ethics and information management, as mentioned in the previous section on ethics, provide social opportunities that underscore fostering societal engagement through public demos raising awareness about EMIAS's utility. Gathering feedback enables alignment with social values, while collaboration with community stakeholders supports ethical use. Encouraging debates around AI's role in education helps inform potential bias issues arising from deploying such technologies within educational contexts. Similarly, policy and regulatory opportunities advocate policies supporting EMIAS's integration into institutions, ensuring compliance with necessary data protection regulations. Engaging policymakers facilitates constructive dialogue regarding advantages and concerns related to implementing such technology-based solutions on a broader scale while outlining clear guidelines ensuring ethical usage within an educational context.

What Are the Salient Takeaways of this Study?

Our research highlights the pivotal role of AI in enhancing the trustworthiness of content accessed by educators and students. The theoretical foundations suggest the importance of knowledge graph structures, marking a shift from traditional isolated document analysis and emphasizing the interconnectedness of digital information. We argue that knowledge graph structure is necessary for accurately determining trustfulness and identifying disinformation in documents. This is because the knowledge graph can model relationships between documents to bring information that cannot be obtained from the document itself (Verma et al., 2023). Several recent studies have strongly supported the integration of AI models with domain-specific knowledge graphs for better performance (Holzinger et al., 2023; Pan et al., 2023; Trajanoska et al., 2023). At the same time, knowledge graphs enhance AI models such as Large Language Models (LLMs) by providing external knowledge for inference and interpretability (Pan et al., 2023), which improves the accuracy of LLMs and makes them suitable for domain-specific applications.

While AI can be a powerful ally in combating misinformation, it cannot completely replace human input. A comprehensive education plan should seamlessly incorporate lessons that develop skepticism and encourage analytical thinking. While technology aids in evaluating information, the key factor remains students' and teachers' proficiency in distinguishing between different types of information and employing appropriate assessment methods. Despite the promises of AI-centered solutions for automation and efficiency, the irreplaceable human element, characterized by subtlety and subjective judgment, is essential in verifying information accuracy and contextual relevance (Ninaus & Sailer, 2022). This human oversight is crucial for in-depth analysis and as a means to improve and fine-tune AI systems through feedback, ensuring their continuous improvement (Laux, 2023). The theoretical framework of our study serves as a foundation for future research, emphasizing the multidimensional approach to content verification, combining technology, education, and human discernment.

For a practical contribution, our study offers a comprehensive blueprint for institutions considering the development and deployment of such advanced systems. This blueprint details the pathway for potential technological enhancements, outlining clear strategies for system integration, user interface design, and multi-modal analysis capabilities for students and educators. Emphasizing modularity and adaptability also ensures future scalability in line with technological advancements. This research thus serves as a tangible guide, equipping educational institutions with the tools and insights to harness AI's full potential in maintaining academic integrity and advancing the educational journey. It does this by helping educators and students discern trustworthy content. Its user-friendly interface simplifies interaction while its integration into learning management systems enhances learning outcomes. It also fosters critical thinking, encouraging users to validate AI-generated outputs, thereby advancing the educational journey in a technologically advanced era. Moreover, it supports professional development among educators and engages institutional stakeholders for continuous improvement and adaptive use.

Following our examination of how EMIAS can revolutionize higher education, particularly in its capacity to combat disinformation and foster new opportunities, we will next transition to exploring potential future directions and potential advancements for EMIAS.

RESEARCH DIRECTIONS

This study encourages researchers in both computer science and education fields to examine human-AI interaction from the viewpoint of media and AI literacy. Students need skills to use AI-based solutions, but they also need a basic understanding of the principles of detecting trustworthy learning content. More research is needed on the micro-level activities of student-AI interaction while using AI-based solutions in analyzing information trustworthiness, as well as pedagogical practices to train students and teachers to benefit from AI-based solutions in analyzing information trustworthiness. Future studies should also be conducted on the micro-, meso-, and macro-level impacts of improved information trustworthiness and how they are interrelated.

The expanding landscape of AI offers ample opportunities for developing the EMIAS system. For instance, harnessing the capabilities of recent advancements in Multimodal LLMs (MLLMs) offers a promising avenue, especially within the second layer of the EMIAS framework. MLLMs, with their adeptness at interpreting a spectrum of data types ranging from text and images to audio, convert varied data into a standardized encoding realm. Notably, recent innovations have led to the emergence of several multimodal learning frameworks, catering to a wide array of tasks from fundamental perception to specialized applications, multimodal document understanding, and intricate data mining (Ye et al., 2023; Y. Zhang et al., 2023).

Recent breakthroughs in the construction and utilization of multimodal knowledge graphs (Chen et al., 2023; Ektefaie et al., 2023) lay the foundation for the EMIAS repository, which acts as a comprehensive system for evaluating information, bridging the connections between various information

sources and key influential entities. EMIAS knowledge graph is dynamic by nature as it is aimed to be updated regularly by streaming information. Hence, the advanced techniques for dynamic knowledge graph learning (e.g., (Barry et al., 2022; Wu et al., 2022)) could be leveraged to get insights from continuously updated data. Similarly, advanced, automated data annotation, preparation, and cleaning techniques can be further investigated in the data ingestion layer (layer 1) (Demrozi et al., 2023; Goyle et al., 2023).

Domain-specific adaptations of EMIAS need to be explored, focusing on how it can be optimized for distinct academic fields with diverse requirements. Such specialization will ensure its relevance and efficacy across varied disciplines. Similarly, methods for collecting and incorporating user feedback can be explored for continual improvement, allowing developers to identify pain points, gauge user satisfaction, and adapt to evolving user needs.

Another promising avenue for research involves the exploration of seamless integration techniques for EMIAS within the existing educational infrastructure. This would entail understanding how EMIAS can harmoniously coalesce with current educational platforms, databases, and other academic tools to ensure streamlined operations and augment the existing capabilities without causing disruptions or redundancies.

Future research directions could also explore the broader societal and economic outcomes of enhanced information trustworthiness, especially as systems like EMIAS become more prevalent. Understanding how increased accuracy and credibility in information can foster public confidence and support decision-making processes is crucial. Moreover, the potential of AI technologies to be tailored to enhance the media literacy skills of both students and teachers warrants attention. Investigating the development of intelligent, adaptive learning environments for students and professional development programs for teachers could reveal ways to promote critical thinking and discernment of reliable information sources. Equally important is the incorporation of information trustworthiness skills into AI literacy education, necessitating the creation of new curricula, assessment tools, and teaching strategies. Moreover, as this domain continues to evolve, there may be emerging commercial prospects, presenting opportunities for startups, tech companies, and educational institutions to innovate and collaborate in delivering solutions that cater to the demands of trustworthy information dissemination and consumption.

Building upon the groundwork established by the EMIAS framework proposed in this study, our objective is to systematically execute its application by developing an information assessment system for educational institutions operating within selected domains.

This section has explored future research directions in the realm of explainable, multimodal emotion recognition. Our proposed framework for an explainable, multimodal information assessment system marks the beginning of new research pathways, offering a fertile ground for inquiry and innovation for researchers in the field. The following section will summarize the key insights and overall contribution of this paper.

CONCLUSION

This paper has presented a systematic blueprint for developing an Explainable Multimodal Information Assessment System (EMIAS) uniquely designed to cater to the educational sector's needs. We have addressed five critical research questions throughout this study.

For the first research question, we undertook a comprehensive survey of existing information assessment techniques utilizing machine learning, simultaneously highlighting their inherent limitations. We identified the primary characteristics of an efficient information assessment system that include multi-modality, human-centric design, explainability, and the capability to evaluate various dimensions of information. Our perspective emphasized that information assessment extends beyond mere binary decision-making into a complex, multifaceted process. Further exploration involved analyzing

existing AI-enhanced Education systems and resources currently in use by educational entities to gauge their information assessment level.

In answering the second research question, we specified the requirements of an efficient information assessment system in educational institutions and proceeded to formulate a taxonomy based on our literature review. This taxonomy guided and informed the functions within the EMIAS framework.

The solutions identified for both RQ1 and RQ2 provided us with foundational groundwork, leading us to develop the EMIAS framework and a detailed examination of its functions, addressing RQ3.

Finally, in addressing RQ4, we shed light on key challenges faced in implementing EMIAS effectively within educational environments. Alongside this, we explored how EMIAS could serve students and educators in several aspects of information assessment while staying mindful of ethical considerations. Future advancements in the EMIAS framework are also hinted at through potential research directions discussed towards the end.

As the educational landscape continues to evolve in the digital age, the integration of such systems becomes paramount to ensure robust, transparent, and accurate information assessment. This study serves as a comprehensive guide for educators, technologists, policymakers, and academic institutions, aiming to navigate the complexities of AI integration in education. We envision a future where the insights and methodologies discussed herein act as foundational pillars, supporting the pursuit of academic excellence in an age of artificial intelligence.

REFERENCES

Abelson, H. (2008). The creation of OpenCourseWare at MIT. *Journal of Science Education and Technology*, *17*, 164–174. https://doi.org/10.1007/s10956-007-9060-8

Adams, C., Pente, P., Lemermeyer, G., & Rockwell, G. (2023). Ethical principles for artificial intelligence in K-12 education. *Computers and Education: Artificial Intelligence*, *4*, 100131. https://doi.org/10.1016/j.caeai.2023.100131

Ahuja, N., & Kumar, S. (2023). Mul-FaD: Attention based detection of multilingual fake news. *Journal of Ambient Intelligence and Humanized Computing*, *14*, 2481–2491 https://doi.org/10.1007/s12652-022-04499-0

Al-Ahmad, B., Al-Zoubi, A. M., Abu Khurma, R., & Aljarah, I. (2021). An evolutionary fake news detection method for COVID-19 pandemic information. *Symmetry*, *13*(6), 1091. https://doi.org/10.3390/sym13061091

Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., Dimitrov, D., Da San Martino, G., Shaar, S., Firooz, H., & Nakov, P. (2021). *A survey on multimodal disinformation detection.* https://arxiv.org/abs/2103.12541

Alamäki, A., Mäki, M., & Ratnayake, R. (2019). Privacy concern, data quality and trustworthiness of AI analytics. *Proceedings of Fake Intelligence Online Summit*, 37–42. https://issuu.com/satakunnan_ammattikorkeakoulu/docs/2019_d_1_samk_proceedings_fakeintel/37 and https://urn.fi/URN:NBN:fi-fe2019061420477

Almars, A. M. (2021). Deepfakes detection techniques using deep learning: A survey. *Journal of Computer and Communications*, *9*(5), 20–35. https://doi.org/10.4236/jcc.2021.95003

Balshetwar, S. V., RS, A., & R, D. J. (2023). Fake news detection in social media based on sentiment analysis using classifier techniques. *Multimedia Tools and Applications*, *82*, 35781–35811. https://doi.org/10.1007/s11042-023-14883-3

Barry, M., Bifet, A., Chiky, R., El Jaouhari, S., Montiel, J., El Ouafi, A., & Guerizec, E. (2022, December). Stream2Graph: Dynamic knowledge graph for online learning applied in large-scale network. *Proceedings of the IEEE International Conference on Big Data*, *Osaka, Japan,* 2190–2197. https://doi.org/10.1109/BigData55660.2022.10020885

Breslin, J., Decker, S., Harth, A., & Bojars, U. (2006). SIOC: An approach to connect web-based communities. *International Journal of Web Based Communities*, *2*(2), 133–142. https://doi.org/10.1504/IJWBC.2006.010305

Brickley, D., & Miller, L. (2014). *FOAF vocabulary specification.* http://xmlns.com/foaf/spec/20140114.html

Cardona, M. A., Rodríguez, R. J., & Ishmael, K. (2023). *Artificial intelligence and the future of teaching and learning: Insights and recommendations.* U.S. Department of Education. https://www2.ed.gov/documents/ai-report/ai-report.pdf

Chen, X., Zhang, J., Wang, X., Wu, T., Deng, S., Wang, Y., Si, L., Chen, H., & Zhang, N. (2023). *Continual multimodal knowledge graph construction.* ArXiv:2305.08698.

Chien, S.-Y., Yang, C.-J., & Yu, F. (2022). XFlag: Explainable fake news detection model on social media. *International Journal of Human-Computer Interaction*, *38*(18–20), 1808–1827. https://doi.org/10.1080/10447318.2022.2062113

Chong, M., & Choy, M. (2020). An empirically supported taxonomy of misinformation. In K. Dalkir, & R. Katz (Eds.), *Navigating fake news, alternative facts, and misinformation in a post-truth world* (pp. 117–138). IGI Global. https://doi.org/10.4018/978-1-7998-2543-2.ch005

Chu, S. K. W., Xie, R., & Wang, Y. (2021). Cross-language fake news detection. *Data and Information Management*, *5*(1), 100–109. https://doi.org/10.2478/dim-2020-0025

Colomina, C., Margalef, H. S., Youngs, R., & Jones, K. (2021). *The impact of disinformation on democratic processes and human rights in the world.* European Parliament, Brussels.

Cornell University, Center for Teaching Innovation. (n.d.). *Ethical AI for teaching and learning.* https://teaching.cornell.edu/generative-artificial-intelligence/ethical-ai-teaching-and-learning

Dale, R. (2017). NLP in a post-truth world. *Natural Language Engineering*, *23*(2), 319–324. https://doi.org/10.1017/S1351324917000018

Dame Adjin-Tettey, T. (2022). Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education. *Cogent Arts & Humanities*, *9*(1), Article 2037229. https://doi.org/10.1080/23311983.2022.2037229

Dementieva, D., Kuimov, M., & Panchenko, A. (2023). Multiverse: Multilingual evidence for fake news detection. *Journal of Imaging*, *9*(4), 77 https://doi.org/10.3390/jimaging9040077

Dementieva, D., & Panchenko, A. (2021). Cross-lingual evidence improves monolingual fake news detection. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop* (pp. 310–320). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-srw.32

Demrozi, F., Turetta, C., Al Machot, F., Pravadelli, G., & Kindt, P. H. (2023). *A comprehensive review of automated data annotation techniques in human activity recognition.* https://arxiv.org/pdf/2307.05988

Duval, E., & Hodgins, W. (2003, May). A LOM research agenda. *Proceedings of the 12th International World Wide Web Conference. Budapest, Hungary.*

Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M., & Zitnik, M. (2023). Multimodal learning with graphs. *Nature Machine Intelligence*, *5*(4), 340–350. https://doi.org/10.1038/s42256-023-00624-6

Eslake, S. (2006). The importance of accurate, reliable and timely data. https://www.anz.com/documents/economics/the_importance_of_data.pdf

European Commission. (2018, April 26). *Communication - Tackling online disinformation: A European approach.* https://digital-strategy.ec.europa.eu/en/library/communication-tackling-online-disinformation-european-approach

European Commission. (2020, July 17). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment.* https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

European Commission. (2022, October 11). *Guidelines for teachers and educators on tackling disinformation and promoting digital literacy through education and training.* https://education.ec.europa.eu/news/guidelines-for-teachers-and-educators-on-tackling-disinformation-and-promoting-digital-literacy-through-education-and-training

European Union. (2022). *Ethical guidelines on the use of Artificial Intelligence (AI) and data in teaching and learning for educators*. https://op.europa.eu/en/publication-detail/-/publication/d81a0d54-5348-11ed-92ed-01aa75ed71a1/language-en

Giachanou, A., Zhang, G., & Rosso, P. (2020). Multimodal fake news detection with textual, visual, and semantic information. In P. Sojka, I. Kopeček, K. Pala, & A. Horák (Eds.), *Text, speech, and dialogue* (pp. 30-38). Springer. https://doi.org/10.1007/978-3-030-58323-1_3

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023, June). Imagebind: One embedding space to bind them all. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *Vancouver, BC, Canada,* 15180–15190. https://doi.org/10.1109/CVPR52729.2023.01457

Gossen, T., Hempel, J., & Nürnberger, A. (2013). Find it if you can: Usability case study of search engines for young users. *Personal and Ubiquitous Computing*, *17*(8), 1593–1603. https://doi.org/10.1007/s00779-012-0523-4

Goyle, K., Xie, Q., & Goyle, V. (2023). *DataAssist: A machine learning approach to data cleaning and preparation.* https://arxiv.org/abs/2307.07119

Hameleers, M. (2023). Disinformation as a context-bound phenomenon: Toward a conceptual clarification integrating actors, intentions and techniques of creation and dissemination. *Communication Theory*, *33*(1), 1–10. https://doi.org/10.1093/ct/qtac021

Hammouchi, H., & Ghogho, M. (2022). Evidence-aware multilingual fake news detection. *IEEE Access*, *10*, 116808–116818. https://doi.org/10.1109/ACCESS.2022.3220690

He, S., Guo, T., Dai, T., Qiao, R., Ren, B., & Xia, S.-T. (2022). *Open-vocabulary multi-label classification via multi-modal knowledge transfer.* The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23). https://arxiv.org/abs/2207.01887

Helmus, T. C. (2022). *Artificial intelligence, deepfakes, and disinformation.* Rand Corporation. https://www.rand.org/pubs/perspectives/PEA1043-1.html

Hermann, E. (2022). Artificial intelligence and mass personalization of communication content – An ethical and literacy perspective. *New Media & Society*, *24*(5), 1258–1277. https://doi.org/10.1177/14614448211022702

Holzinger, A., Saranti, A., Hauschild, A.-C., Beinecke, J., Heider, D., Roettger, R., Mueller, H., Baumbach, J., & Pfeifer, B. (2023). Human-in-the-loop integration with domain-knowledge graphs for explainable federated deep learning. In A. Holzinger, P. Kieseberg, F. Cabitza, A. Campagner, A. M. Tjoa, & E. Weippl (Eds.), *Machine learning and knowledge extraction* (pp. 45–64). Springer. https://doi.org/10.1007/978-3-031-40837-3_4

Hoy, N., & Koulouri, T. (2021). *A systematic review on the detection of fake news articles.* https://arxiv.org/abs/2308.02727

Iceland, M. (2023). *How good are SOTA fake news detectors.* https://arxiv.org/abs/2308.02727

Jeyasudha, J., Seth, P., Usha, G., & Tanna, P. (2022). Fake information analysis and detection on pandemic in Twitter. *SN Computer Science*, *3*, Article 456. https://doi.org/10.1007/s42979-022-01363-y

Kapantai, E., Christopoulou, A., Berberidis, C., & Peristeras, V. (2021). A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society*, *23*(5), 1301–1326. https://doi.org/10.1177/1461444820959296

Kauttonen, J., Hannukainen, J., Tikka, P., & Suomala, J. (2020). Predictive modeling for trustworthiness and other subjective text properties in online nutrition and health communication. *PLoS ONE*, *15*(8), e0237144. https://doi.org/10.1371/journal.pone.0237144

Kawase, R., Fisichella, M., Niemann, K., Pitsilis, V., Vidalis, A., Holtkamp, P., & Nunes, B. (2013). Openscout: Harvesting business and management learning objects from the web of data. *Proceedings of the 22nd International Conference on World Wide Web* (pp. 445–450). Association for Computing Machinery https://doi.org/10.1145/2487788.2487962

Khan, M. N., Ashraf, M. A., Seinen, D., Khan, K. U., & Laar, R. A. (2021). Social media for knowledge acquisition and dissemination: The impact of the COVID-19 pandemic on collaborative learning driven social media adoption. *Frontiers in Psychology*, *12*, 648253. https://doi.org/10.3389/fpsyg.2021.648253

Kim, B., Xiong, A., Lee, D., & Han, K. (2021). A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions. *PLoS ONE*, *16*(12), e0260080. https://doi.org/10.1371/journal.pone.0260080

Kou, Z., Zhang, Y., Zhang, D., & Wang, D. (2022). CrowdGraph: A crowdsourcing multi-modal knowledge graph approach to explainable fauxtography detection. *Proceedings of the ACM on Human-Computer Interaction*, *6*, Article 287. https://doi.org/10.1145/3555178

Kumar, S., & Shah, N. (2018). *False information on web and social media: A survey.* https://arxiv.org/abs/1804.08559

Kumari, R., & Ekbal, A. (2021). AMFB: Attention based multimodal factorized bilinear pooling for multimodal fake news detection. *Expert Systems with Applications*, *184*, 115412. https://doi.org/10.1016/j.eswa.2021.115412

Lampridis, O., Karanatsiou, D., & Vakali, A. (2022). MANIFESTO: A huMAN-centric explaInable approach for FakE news spreaders deTectiOn. *Computing*, *104*, 717–739. https://doi.org/10.1007/s00607-021-01013-w

Latkin, C. A., Dayton, L., Strickland, J. C., Colon, B., Rimal, R., & Boodram, B. (2020). An assessment of the rapid decline of trust in US sources of public information about COVID-19. *Journal of Health Communication*, *25*(10), 764–773. https://doi.org/10.1080/10810730.2020.1865487

Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence*, *3*, 100101. https://doi.org/10.1016/j.caeai.2022.100101

Laux, J. (2023). *Institutionalised distrust and human oversight of artificial intelligence: Toward a democratic design of AI governance under the European Union AI Act.* https://doi.org/10.2139/ssrn.4377481

Lecheler, S., & Egelhofer, J. L. (2022). Disinformation, misinformation, and fake news: Understanding the supply side. In J. Strömbäck, Å. Wikforss, K. Glüer, T. Lindholm, & H. Oscarsson (Eds.), *Knowledge resistance in high-choice information environments* (pp. 69–87). Routledge. https://doi.org/10.4324/9781003111474-4

Lin, S.-Y., Kung, Y.-C., & Leu, F.-Y. (2022). Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis. *Information Processing & Management*, *59*(2), 102872. https://doi.org/10.1016/j.ipm.2022.102872

Lin, X., Tian, T., Wei, Z., & Hakonarson, H. (2022). Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nature Communications*, *13*, Article 7705. https://doi.org/10.1038/s41467-022-35031-9

Luckin, R., Cukurova, M., Kent, C., & du Boulay, B. (2022). Empowering educators to be AI-ready. *Computers and Education: Artificial Intelligence*, *3*, 100076. https://doi.org/10.1016/j.caeai.2022.100076

Memarian, B., & Doleck, T. (2023). Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI), and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, *5*, 100152. https://doi.org/10.1016/j.caeai.2023.100152

Moyer, M. W. (2018, February 1). Schoolkids are falling victim to disinformation and conspiracy fantasies. *Scientific American.* https://www.scientificamerican.com/article/schoolkids-are-falling-victim-to-disinformation-and-conspiracy-fantasies/

Muhammed T, S., & Mathew, S. K. (2022). The disaster of misinformation: A review of research in social media. *International Journal of Data Science and Analytics*, *13*(4), 271–285. https://doi.org/10.1007/s41060-022-00311-6

National Academy of Medicine. (2023, July 21). *Discussion paper offers guidance on identifying credible sources of health information in social media.* https://nam.edu/discussion-paper-offers-guidance-on-identifying-credible-sources-of-health-information-in-social-media/

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI Literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*, 100041. https://doi.org/10.1016/j.caeai.2021.100041

Ninaus, M., & Sailer, M. (2022). Closing the loop – The human role in artificial intelligence for education. *Frontiers in Psychology*, *13*, 956798. https://doi.org/10.3389/fpsyg.2022.956798

Nygren, T., Wiksten Folkeryd, J., Liberg, C., & Guath, M. (2020). Students assessing digital news and misinformation. In M. van Duijn, M. Preuss, V. Spaiser, F. Takes, & S. Verberne (Eds.), *Disinformation in open online media* (pp. 63–79). Springer. https://doi.org/10.1007/978-3-030-61841-4_5

OECD Legal Instruments. (2024). *Recommendation of the Council on Artificial Intelligence.* https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2023). *Unifying large language models and knowledge graphs: A roadmap.* https://arxiv.org/abs/2306.08302

Parker, S., & Ruths, D. (2023). Is hate speech detection the solution the world wants? *Proceedings of the National Academy of Sciences*, *120*(10), e2209384120. https://doi.org/10.1073/pnas.2209384120

Patel, D., & Singh, P. K. (2016, October). Kids safe search classification model. *Proceedings of the International Conference on Communication and Electronics Systems*, *Coimbatore, India.* https://doi.org/10.1109/CESYS.2016.7914186

Pedró, F., Subosa, M., Rivas, A., & Valverde, P. (2019). *Artificial intelligence in education: Challenges and opportunities for sustainable development*. UNESCO.

Pereira, C. K., Campos, F., Ströele, V., David, J. M. N., & Braga, R. (2018). BROAD-RSI – Educational recommender system using social networks interactions and linked data. *Journal of Internet Services and Applications*, *9*, Article 7. https://doi.org/10.1186/s13174-018-0076-5

Perez-Ortiz, M., Dormann, C., Rogers, Y., Bulathwela, I., Kreitmayer, S., Yilmaz, E., Noss, R., & Shawe-Taylor, J. (2021). X5learn: A personalized learning companion at the intersection of AI and HCI. *Companion Proceedings of the 26th International Conference on Intelligent User Interfaces* (pp. 70–74). Association for Computing Machinery. https://doi.org/10.1145/3397482.3450721

Ramachandran, S., Paulraj, S., Joseph, S., & Ramaraj, V. (2009). Enhanced trustworthy and high-quality information retrieval system for web search engines. International Journal of Computer Science Issues, 5. https://web-archive.southampton.ac.uk/cogprints.org/6724/1/IJCSI-5-38-42.pdf

Rasi, P., Vuojärvi, H., & Ruokamo, H. (2019). Media literacy education for all ages. *Journal of Media Literacy Education*, *11*(2), 1–19. https://doi.org/10.23860/JMLE-2019-11-2-1

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022, June). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA*, 10684–10695. https://doi.org/10.1109/CVPR52688.2022.01042

Sahoo, S. R., & Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, *100*, 106983. https://doi.org/10.1016/j.asoc.2020.106983

Scharowski, N., Perrig, S. A. C., Svab, M., Opwis, K., & Brühlmann, F. (2023). Exploring the effects of human-centered AI explanations on trust and reliance. *Frontiers in Computer Science*, *5*, 1151150. https://doi.org/10.3389/fcomp.2023.1151150

Schoenherr, J. R., Abbas, R., Michael, K., Rivas, P., & Anderson, T. D. (2023). Designing AI using a human-centered approach: Explainability and accuracy toward trustworthiness. *IEEE Transactions on Technology and Society*, *4*(1), 9–23. https://doi.org/10.1109/TTS.2023.3257627

Segura-Bedmar, I., & Alonso-Bartolome, S. (2022). Multimodal fake news detection. *Information*, *13*(6), 284. https://doi.org/10.3390/info13060284

Shang, L., Kou, Z., Zhang, Y., & Wang, D. (2022). A duo-generative approach to explainable multimodal COVID-19 misinformation detection. *Proceedings of the ACM Web Conference* (pp. 3623-3631). Association for Computing Machinery. https://doi.org/10.1145/3485447.3512257

Sharma, S., Alam, F., Akhtar, M. S., Dimitrov, D., Martino, G. D. S., Firooz, H., Halevy, A., Silvestri, F., Nakov, P., & Chakraborty, T. (2022). Detecting and understanding harmful memes: A survey. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence* (pp. 5597-5606). https://doi.org/10.24963/ijcai.2022/781

Song, C., Ning, N., Zhang, Y., & Wu, B. (2021). A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management*, *58*(1), 102437. https://doi.org/10.1016/j.ipm.2020.102437

Statista. (2024a). *Most trusted sources in the world 2023*. https://www.statista.com/statistics/381455/most-trusted-sources-of-news-and-info-worldwide/

Statista. (2024b). *Share of adults who trust news media most of the time in selected countries worldwide as of Fabruary 2023*. https://www.statista.com/statistics/308468/importance-brand-journalist-creating-trust-news/

Sterrett, D., Malato, D., Benz, J., Kantor, L., Tompson, T., Rosenstiel, T., & Loker, K. (2019). Who shared it?: Deciding what news to trust on social media. *Digital Journalism*, *7*(6), 783–801. https://doi.org/10.1080/21670811.2019.1623702

Sutton, S. A., & Mason, J. (2001, October). The Dublin core and metadata for educational resources. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, *Tokyo, Japan*, 25–31.

Szczepański, M., Pawlicki, M., & Kozik, R., & Choraś, M. (2021). New explainability method for BERT-based model in fake news detection. *Scientific Reports*, *11*, Article 23705. https://doi.org/10.1038/s41598-021-03100-6

Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. (2018). Fake news detection: A deep learning approach. *SMU Data Science Review*, *1*(3), Article 10. https://scholar.smu.edu/datasciencereview/vol1/iss3/10

Trajanoska, M., Stojanov, R., & Trajanov, D. (2023). *Enhancing knowledge graph construction using large language models*. https://arxiv.org/abs/2305.04676

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000380455

UNESCO. (2023, September). *Guidance for generative AI in education and research*. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Uppada, S. K., Patel, P., & B., S. (2022). An image and text-based multimodal model for detecting fake news in OSN's. *Journal of Intelligent Information Systems*, *61*, 367–393. https://doi.org/10.1007/s10844-022-00764-y

Verma, S., Bhatia, R., Harit, S., & Batish, S. (2023). Scholarly knowledge graphs through structuring scholarly communication: A review. *Complex & Intelligent Systems*, *9*, 1059–1095. https://doi.org/10.1007/s40747-022-00806-6

Vranias, K., & Kyridis, A. (2022). Do Greek university students trust social media regarding science related issues? The Covid-19 case. *International Journal of Social Science Research*, *11*(1). https://doi.org/10.5296/ijssr.v11i1.20259

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Council of Europe. https://rm.coe.int/information-disorder-report-november-2017/1680764666

Weiss, A. P., Alwan, A., Garcia, E. P., & Garcia, J. (2020). Surveying fake news: Assessing university faculty's fragmented definition of fake news and its impact on teaching critical thinking. *International Journal for Educational Integrity*, *16*, Article 1. https://doi.org/10.1007/s40979-019-0049-x

Widana, I. W. (2020). The effect of digital literacy on the ability of teachers to develop HOTS-based assessment. *Journal of Physics: Conference Series*, *1503*, 12045. https://doi.org/10.1088/1742-6596/1503/1/012045

Wu, T., Khan, A., Yong, M., Qi, G., & Wang, M. (2022). Efficiently embedding dynamic knowledge graphs. *Knowledge-Based Systems*, *250*, 109124. https://doi.org/10.1016/j.knosys.2022.109124

Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2019). Zero-shot learning – A comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(9), 2251–2265. https://doi.org/10.1109/TPAMI.2018.2857768

Xiang, N. (2022). Deep learning-based fake information detection and influence evaluation. *Computational Intelligence and Neuroscience*, *2022*, Article 8514430. https://doi.org/10.1155/2022/8514430

Yang, F., Pentyala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E. D., Ji, S., & Hu, X. (2019). XFake: Explainable fake news detector with visualizations. *The World Wide Web Conference* (pp. 3600–3604). Association for Computing Machinery. https://doi.org/10.1145/3308558.3314119

Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Dan, Y., Zhao, C., Xu, G., Li, C., Tian, J., Qi, Q., Zhang, J., & Huang, F. (2023). *mPLUG-DocOwl: Modularized multimodal large language model for document understanding.* https://arxiv.org/abs/2307.02499

Yu, L., Liu, C., Yang, J. Y. H., & Yang, P. (2023). Ensemble deep learning of embeddings for clustering multimodal single-cell omics data. *Bioinformatics*, *39*(6), btad382. https://doi.org/10.1093/bioinformatics/btad382

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2), 102025. https://doi.org/10.1016/j.ipm.2019.03.004

Zhang, Y., Gong, K., Zhang, K., Li, H., Qiao, Y., Ouyang, W., & Yue, X. (2023). *Meta-transformer: A unified framework for multimodal learning.* https://arxiv.org/abs/2307.10802

AUTHORS

Umair Al Khan holds a Ph.D. degree in information technology and a Master's degree in intelligent transportation systems from the University of Klagenfurt, Austria. He is a senior researcher at Haaga-Helia University of Applied Sciences, Helsinki, focusing mainly on AI applications, digitalization, and data analysis using machine learning.

Janne Kauttonen works as a senior researcher at Haaga-Helia University of Applied Sciences. He received his PhD in statistical physics from the University of Jyväskylä and, after that, worked as a postdoctoral researcher at Aalto and Carnegie Mellon Universities. His work is mainly focused on AI applications and advanced data analytics.

Lili Aunimo is a Principal Lecturer at Haaga-Helia University of Applied Sciences. She received her Ph.D. in natural language technologies from the University of Helsinki. Her research focuses on applying data analysis, machine learning, and other AI techniques in several fields, such as human-computer interaction and the development of digital services.

Ari Alamäki is a Principal Lecturer at HHUAS and Adjunct Professor (technology education) at the University of Turku, Finland. His current research focuses on the applications of artificial intelligence in education and business services. He also worked in management positions in the ICT industry from 2000 to 2011 and was a visiting scholar at two US universities from 1997 to 1998.