Creating Blockchain-Powered Tools Enhanced by Artificial Intelligence for Evaluating Semantic Concordance in Scientific Documents. (A research proposal)

Shamsollah Ghanbari

No Institute Given

1 Abstract

Intellectual ownership of scientific works is a primary concern for universities, institutes, and organizations when it comes to publishing research papers. Nowadays, web-based software tools like Turnitin are widely used to assess the similarity of textual content across documents. While these tools may have certain limitations, we would not focus the limitation here. It should be emphasized that semantic similarity or content similarity is an increasingly significant matter that has garnered attention from researchers in recent years. Additionally, prominent concepts such as deep learning and Large Language Models (LLM) have emerged as robust learning strategies, leading to remarkable advancements in the field of natural language processing (NLP). To explore content similarity, several recent studies have been taken into consideration. We have a block chain based approach to the problem. In this research we have a plan to propose a human computer interaction model for investigating the similarity of context i.e. ideas.

2 Introduction

The protection of intellectual property in scientific research is crucial for universities, institutes, and organizations involved in publishing research papers. To evaluate the similarity of textual content across documents, web-based software tools like Turnitin have gained widespread adoption. However, in this discussion, we will not delve into the limitations of these tools. Instead, we want to emphasize the growing importance of semantic similarity and content similarity, which have garnered significant attention from researchers in recent years.

The field of natural language processing (NLP) has undergone a revolution with the advancements in deep learning and the emergence of Large Language Models (LLM). These powerful learning strategies have led to remarkable progress in understanding and analyzing textual data.This project presents a novel approach that utilizes blockchain technology to enhance the identification of contextual resemblance within scientific text by considering various human factors and methods. Our research proposes a blockchain-based solution to tackle

this issue effectively. Our objective is to develop an interactive model for human-computer interaction that specifically concentrates on evaluating the similarity of ideas in a given context. By leveraging the capabilities of blockchain technology, we aim to establish a secure and transparent framework for assessing and comparing the conceptual similarity of scientific texts. Through this project, we aspire to contribute to ongoing endeavors in safeguarding intellectual property rights and promoting ethical research practices. By employing cutting-edge technologies such as blockchain, data mining, LLM, and NLP, we strive to offer a dependable and efficient solution for universities, institutions, and organizations to ensure the integrity and authenticity of scientific works. This study presents a novel approach that utilizes blockchain technology to enhance the identification of contextual resemblance within scientific text by considering various human factors and methods. Our research proposes a blockchain-based solution to tackle this issue effectively. Our objective is to develop an interactive model for human-computer interaction that specifically concentrates on evaluating the similarity of ideas in a given context. By leveraging the capabilities of blockchain technology, we aim to establish a secure and transparent framework for assessing and comparing the conceptual similarity of scientific texts. Through this project, we aspire to contribute to ongoing endeavors in safeguarding intellectual property rights and promoting ethical research practices. By employing cutting-edge technologies such as blockchain, data mining, LLM, and NLP, we strive to offer a dependable and efficient solution for universities, institutions, and organizations to ensure the integrity and authenticity of scientific works.By embracing these cutting-edge technologies, we envision a future where researchers can confidently protect their intellectual ownership and foster a culture of trust and integrity in the scientific community. Human-computer interaction (HCI) has evolved significantly with the advent of artificial intelligence and natural language processing technologies. This literature review provides an overview of recent advancements in HCI systems, personalized interactions, text mining, large language models, blockchain technology, knowledge extraction, and cognitive workload measurement. The growing interest in talking-head generation systems for HCI is evident [1]. These systems offer the potential for highly personalized interactions, aligning with the trend toward personalized HCI experiences [2]. Personalization in HCI is crucial for enhancing user experience and engagement, and advances in personalized HCI have been driven by the integration of AI and user modeling techniques [2]. Text mining has gained prominence in the context of services management, offering valuable insights from unstructured data [3]. Moreover, the emergence of large language models, such as ChatGPT, has revolutionized natural language understanding and generation tasks [1,2,4] present comprehensive surveys of large language models and their evaluation techniques, highlighting the significance of such models in various applications. Blockchain technology has also influenced HCI, particularly in ensuring trust and security in digital interactions [6,7]. Trust in blockchain-enabled exchanges has been a focal point, shaping the future landscape of blockchain marketing [8]. Additionally, OpenAI's ChatGPT has been leveraged for generating content and assessing similarity in-

dices in the domain of Library & Information Science [9]. Knowledge extraction from text has been a subject of interest in various domains, with AI playing a pivotal role [10]. This is echoed by the work of Mol & Kumar [**?**], which focuses on knowledge extraction in the agriculture domain, and Yu et al. [11], who present a survey of knowledge-enhanced text generation. Furthermore, the application of large language models in extracting financial information from text has been explored in FinBERT [13]. The potential of latent knowledge discovery in language models without supervision has also been investigated [15]. Text mining, particularly in the context of big data analytics, holds promise for discovering valuable insights from vast unstructured data sets [15]. In HCI, measuring cognitive workload is crucial for understanding user engagement and task performance [16, 18] Additionally, knowledge graph acquisition and applications have been the focus of research, highlighting the importance of knowledge representation in various domains [16]. The discovery of knowledge graph schema from short natural language text via dialog has also been explored, emphasizing the potential of extracting structured knowledge from unstructured data [15]. The paper holds significant scientific novelty and importance for several compelling reasons. Firstly, it addresses a notable research gap by conducting a comprehensive study that aims to deliver definitive and practical outcomes. Given the crucial role of defining content and semantic similarity in scientific fields and educational institutions, this project offers a valuable solution. It is expected that the project will lead to substantial scientific advancements and economic benefits.

Secondly, the outcomes of this study will culminate in the development of a software tool specifically designed to support organizations involved in handling scientific texts, such as universities, journal editors, and conference organizers. Beyond its economic impact, this research also contributes to the advancement of fundamental sciences. The software being developed will play a pivotal role in accurately determining the intellectual property of scientific works, which holds utmost significance.

Moreover, this research aims to enhance the confidence of idea owners and producers in publishing their scientific achievements. Additionally, the project exhibits potential for international application. We intend to disseminate the results through articles published in reputable ISI journals and conferences, ensuring widespread access and recognition. The objective of this research is to develop an intelligent application system that utilizes artificial intelligence and business intelligence to identify similarities in content among scientific texts. This project holds significant value for various institutions including universities, scientific journals and magazines, conference organizers, and business idea owners.

3 Importance of the research

In this project, we introduce a cutting-edge commercial product based on advanced theoretical concepts. Given the growing demand for intellectual property

within scientific and industrial communities, the resulting product holds great value for both researchers and business owners. It is evident that this project will pave the way for a revenue-generating software product with significant market potential.

4 Objective

The objective of this research project is to develop an intelligent application system that utilizes artificial intelligence, natural language processing (NLP), large language models (LLM), and block chain technology to identify similarities in scientific texts. If a new and innovative concept is introduced in a specific article, and a similar idea is expressed in another language or with different wording by a different author in a new article, the proposed project will have the capability to detect this similarity. This project holds significant importance for the purpose of publishing and can provide substantial benefits to various institutions, including universities, scientific journals and magazines, conference organizers, as well as entrepreneurs and business owners.

5 Methodology

This section explores the methodology and project environment. The project's methodology comprises two distinct components: the front-end and back-end processes. The front-end consists of three integral elements, namely, the User Interface, Knowledge Discovery from Text, and Knowledge Clustering. The project's back-end encompasses four key components, which are the Pre-trained Transformer, the development of trained decision blocks, the assignment of knowledge clusters to these blocks, and the refinement of results through expert input. In due to this section, we will explain the operational and non-operational environments. A general framework of the project's methodology has been illustrated in Figure 1.

5.1 User Interface:

As part of this project, we have developed a user interface that is both intuitive and user-friendly. This interface allows users to easily input scientific documents in formats such as Word or PDF, and receive similarity scores or rankings based on contextual similarities. The system employs algorithms that calculate the percentage of semantic and content similarities between the provided text and other texts, presenting the results in an output report. It's important to note that the output report can be customized to meet the specific needs of each user, making it a truly personalized experience. Knowledge Discovery from Text (KDT): Through a dynamic process of human-computer interaction, text mining algorithms are employed to acquire knowledge, which then undergoes meticulous refinement. In this phase, the user-provided text is skillfully transformed into

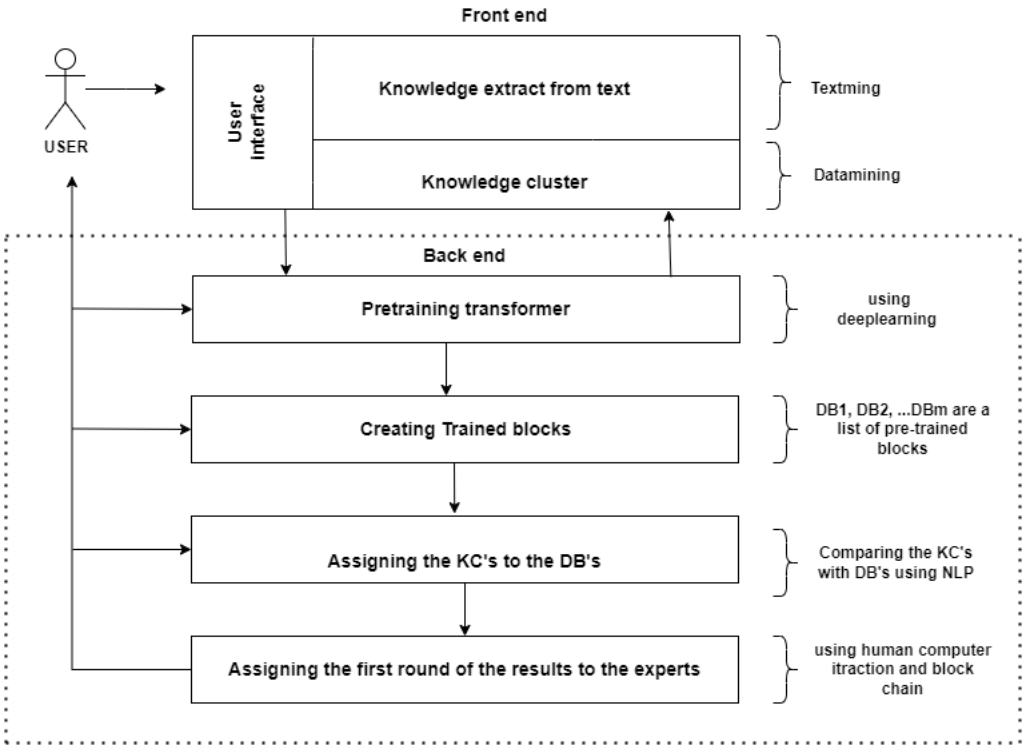

Fig. 1: **A general framework of the proposed project.**

coherent and meaningful units of knowledge. To ensure accuracy and reliability, the project incorporates block chain technology to consolidate expert opinions pertaining to the generated knowledge. Additionally, state-of-the-art techniques such as Natural Language Processing (NLP) and Large Language Models (LLM) are utilized to develop robust programs within this domain, thereby enhancing the overall effectiveness of our systems.

5.2 Knowledge Discovery from Text (KDT):

Through a dynamic process of human-computer interaction, text mining algorithms are employed to acquire knowledge, which then undergoes meticulous refinement. In this phase, the user-provided text is skillfully transformed into coherent and meaningful units of knowledge. To ensure accuracy and reliability, the project incorporates block chain technology to consolidate expert opinions pertaining to the generated knowledge. Additionally, state-of-the-art techniques such as Natural Language Processing (NLP) and Large Language Models (LLM) are utilized to develop robust programs within this domain, thereby enhancing the overall effectiveness of our systems.

5.3 Knowledge Clustering (KC):

Utilizing advanced data mining techniques, the knowledge generated in the preceding step is distributed into specialized clusters. To ensure utmost accuracy in this process, subject matter experts are engaged. These experts have already been registered and their expertise is documented within our databases. By incorporating their specialized domain knowledge, we strive to enhance the precision and quality of the knowledge distribution phase.

5.4 Back End:

The project's back-end encompasses four key components, which are the Pre-trained Transformer, the development of trained decision blocks, the assignment of knowledge clusters to these blocks, and the refinement of results through expert input.

5.5 Pre-trained Transformer:

A pre-trained transformer plays a crucial role in the project. It involves collecting and preparing a large dataset of text using techniques like self-supervised or supervised learning. The model is then fine-tuned on downstream tasks, and its performance is evaluated based on data quality and size. The training process greatly affects the model's performance, while the architecture determines how input text is processed to generate meaningful outputs. By fine-tuning the model on specific tasks using smaller datasets and continuously evaluating and refining it, we can enhance its performance. Designing a pre-trained transformer requires careful consideration of various factors, including architecture, data quality, training processes, and performance evaluation. This subsystem serves as the

core component of the project, leveraging existing knowledge and determining the need for new knowledge.

5.6 Creating trained Decision blocks:

The trained decision blocks, denoted as Db1, Db2, ..., Dbk, serve as pre-trained modules responsible for acquiring and processing published knowledge. Think of these decision blocks as akin to miners in a blockchain system.

5.7 Assignment the KC's to the train decision block:

In this section, the clusters obtained from the higher layer will be allocated to decision blocks. The objective is to assess the similarity of the knowledge stored within these clusters using a set of pre-trained decision blocks. The outcome of this process determines the degree of similarity of the idea, expressed as an error value (which may not reach 100 percent).

5.8 Assigning the first round of the results to the experts:

The outcomes derived from the preceding section may exhibit inaccuracies attributed to the specialized nature of the subject matter. Consequently, a preliminary review and refinement process is imperative. These results are then forwarded to subject matter experts within blockchain groups for validation and enhancement. Upon achieving completeness, the final results are conveyed to the user. However, should issues persist, the results are further routed to subsequent layers for resolution.

6 Research Environment

This section provides an in-depth analysis of the project's environments, focusing on both operational and non-operational aspects.

6.1 Operational Environment:

In this subsection, we delve into the operational environment of the project. We explore key components such as artificial intelligence, natural language processing, large language model programming, and data mining algorithms. Additionally, we employ essential tools like distributed systems implementation using containers and blockchain, as well as utilizing Go Lang environments. Furthermore, we emphasize the significance of project-related data and thoroughly discuss the data flow to ensure the validity and reliability of the proposed model. In this sub-section, we delineate the operational environment of the project, encompassing a suite of critical components:

AI and Data Mining Algorithms: 1 The project revolves around the pivotal utilization of artificial intelligence, data mining, and text mining algorithms. These algorithms are the cornerstone of our approach, facilitating the enhancement of raw data through learning processes. Additionally, human expertise is integrated into certain aspects of the project.

Essential Tools for Implementing Distributed Systems: Given the geographically dispersed nature of the project's operational environment, the utilization of distributed system algorithms and tools is imperative. This includes containerization technology, such as Docker, and orchestration platforms like Kubernetes, which are instrumental in ensuring seamless operation across diverse locations.

GO Programming Language: The selection of the Go programming language for module implementation is deliberate, as it offers a combination of simplicity and the capability to accommodate both artificial intelligence and distributed system requirements, ensuring the project's efficiency and effectiveness.

Blockchain Integration: A pivotal facet of our project involves harnessing collective intelligence. To realize this objective, we employ blockchain technology, which serves as a fundamental building block for collaborative efforts.

6.2 Non-operational Environment:

The non-operational environment encompasses the physical locations employed, the human resources engaged, and the pertinent constraints entailed by the project.

References

1. Zhao, Wayne Xin, et al. "A survey of large language models." arXiv preprint arXiv:2303.18223 (2023).
2. Chang, Yupeng, et al. "A survey on evaluation of large language models." arXiv preprint arXiv:2307.03109 (2023).
3. Zhen, Rui, et al. "Human-computer interaction system: A survey of talking-head generation." Electronics 12.1 (2023): 218.
4. Augstein, Mirjam, Eelco Herder, and Wolfgang Wörndl, eds. Personalized human-computer in-teraction. Walter de Gruyter GmbH & Co KG, 2023.
5. Malik, Shaily, and Dr Shashi Kant Gupta. "The Importance of Text Mining for Services Manage-ment." TTIDMKD 2.4 (2022): 28-33.
6. Liu, Yiheng, et al. "Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models." Meta-Radiology (2023): 100017.
7. WHIG, PAWAN. "Blockchain Revolution: Innovations, Challenges, and Future Directions." In-ternational Journal of Machine Learning for Sustainable Development 5.3 (2023): 16-25.

8. Tan, Teck Ming, and Saila Saraniemi. "Trust in blockchain-enabled exchanges: Future directions in blockchain marketing." Journal of the Academy of marketing Science 51.4 (2023): 914-939.
9. Kirtania, Deep Kumar, and Swapan Kumar Patra. "OpenAI ChatGPT Generated Content and Similarity Index: A study of selected terms from the Library & Information Science (LIS)." Qeios (2023).
10. Ma, Yongqiang, et al. "AI vs. human–differentiation analysis of scientific content generation." arXiv preprint arXiv 2301 (2023): 10416.
11. Nismi Mol, E. A., and M. B. Santosh Kumar. "Review on knowledge extraction from text and scope in agriculture domain." Artificial Intelligence Review 56.5 (2023): 4403-4445.
12. .Yu, Wenhao, et al. "A survey of knowledge-enhanced text generation." ACM Computing Sur-veys 54.11s (2022): 1-38.
13. Huang, Allen H., Hui Wang, and Yi Yang. "FinBERT: A large language model for extracting infor-mation from financial text." Contemporary Accounting Research 40.2 (2023): 806-841.
14. Burns, Collin, et al. "Discovering latent knowledge in language models without supervision." arXiv preprint arXiv:2212.03827 (2022).
15. Hassani, Hossein, et al. "Text mining in big data analytics." Big Data and Cognitive Computing 4.1 (2020): 1.
16. Ghosh, Subhasis, et al. "Discovering Knowledge Graph Schema from Short Natural Language Text via Dialog." Proceedings of the 21th Annual Meeting of the Special Interest Group on Dis-course and Dialogue. 2020.
17. Ji, Shaoxiong, et al. "A survey on knowledge graphs: Representation, acquisition, and applica-tions." IEEE transactions on neural networks and learning systems 33.2 (2021): 494-514.
18. Thomas, et al. "A survey on measuring cognitive workload in human-computer interaction." ACM Computing Surveys (2023).
19. Ghanbari, Amir Mohammad, Shamsollah Ghanbari, and Yaghoub Norouzi. "A new approach to architecture of human-computer interaction." 2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA). IEEE, 2017.