Application of Blockchain and the SimHash Algorithm to Detect Plagiarism

Roman Dorosh

Department of Information Technology
Vasyl Stefanyk Precarpathian National University
Ivano-Frankivsk, Ukraine

***Abstract*—In this paper, a novel, unified approach for detecting plagiarism was proposed. It is built on the Ethereum blockchain environment and uses the SimHash algorithm for similarity detection, scalability and plagiarism detection in education field as main area of application. Through a network of social and smart contracts, Blockchain technology will function as an open technology that protects against unwanted access, but at the same time will save timestamp and current state of the analyzed documents. Also, the unified algorithm was compared to widely used TF-IDF algorithm on the same dataset and showed improvement by nearly 7%.**

Keywords— Similarity Detection; SimHash; Blockchain; Smart-contracts; Education; Natural Language Processing; Ethereum Virtual Machine

I. INTRODUCTION

Blockchain also provides the main characteristics: autonomy, decentralization, fault-tolerance, distribution, openness, immutability to previous version of the network and transparency [1], which led to interest of researchers into this area and current state of research is poor. Blockchain as a foundational technology is useless without being built upon. Building apps that interact with data on distributed nodes rather than just appending to databases is made possible by smart contracts, which enable programmatic support for blockchain. Everybody may access the content and data of the public blockchain since smart contracts are executed there. However, unless another version is deployed to a public network of connected nodes, they are immutable.

This article also offers a high-level description of the system, a plagiarism detection algorithm with support for different languages using word embeddings, and a smart contract that was implemented in a test EVM (Ethereum Virtual Machine). Also included in this study are real-world results, a detailed description of the method used to compare techniques using blockchain and SimHash, comparisons to other approaches, and a future outlook for this initial effort.

II. RELATED WORK

Following the release of Satoshi Nakamoto's initial paper [2], which was primarily intended to demonstrate how to obtain consensus through the decentralization of servers, blockchain began to acquire prominence. Thus, "miners" that serve the blockchain on their own hardware or use cloud-based services are rewarded for sustaining the infrastructure. The PoW (Proof of Work) algorithm is a type of algorithm that creates transactions that cannot be modified without repeating the PoW.

Following the popularity of Bitcoin and the realization of the limitations of this method of creating blockchains, Ethereum [3], a second widely used system based on PoS (Proof of Stake), emerges. Ethereum is a unique environment that enables the development of decentralized applications utilizing smart-contracts as rule-based systems and an API. Developers now have the ability to create web applications and publish them on the Ethereum blockchain.

Natural language processing is a common method for extracting relevant data from documents and normalizing it. Then, the simple cosine similarity, a slightly modified version of the cosine similarity [4], or the TF-IDF algorithm may be used to determine the degree of similarity between the texts. Neural networks, in particular, and machine learning may be used in fields like software defined demodulation of weak radio signals [5] , effective identification document recognition [7], and weak radio signal demodulation. Research [8] demonstrates the significance of using novel educational strategies, particularly for online engineering education.

A Blockchain-Based Non-Fungible Tokens strategy with various similarity metrics was suggested by Pungilă et al. [9]. They come to the conclusion that their method outperforms conventional similarity assessments in terms of speed and is applicable to real-world circumstances. However, there are other downsides, including lots of created NFT (which translates into high prices) and the inability to recognize minimally changed versions of the same data that are presented in multiple ways (using the SHA256 hash).

III. METHODOLOGY

Scalability and speed are the SimHash algorithm's main benefits. The original study [10] demonstrates its practical applicability for locating near-duplicate documents in huge file storage (multi-billion repository). SimHash was therefore selected as the platform for plagiarism detection based on its implementation areas and properties. According to this method, we determine the SimHash of each element in the set of documents S = "D1, D2, D3,... Dn" by looking through the documents (D). Then, each entry's hash SHd is compared to each document S individually based on the distance between them. This procedure will produce a number in the range [0, 1], which may be converted to the relevant percentage. Additionally, the Sh set of hashes is saved on the Ethereum blockchain via a smart contract and cannot be removed by a centralized authority.

SimHash technique provides ability to arrange similar papers into groups and identify many instances of plagiarism with a single database query. The biggest drawback in this situation is that the algorithm cannot accurately display the locations where the cheater copied from the original document. To facilitate multi-language plagiarism detection, the text processing consists of the following steps: tokenization, stop-word removal, stemming, lemmatization, and transfer to word embeddings. In this research, I suggest combining the decentralized blockchain technology and SimHash, two techniques of plagiarism detection, into a single algorithm (Figure 1)

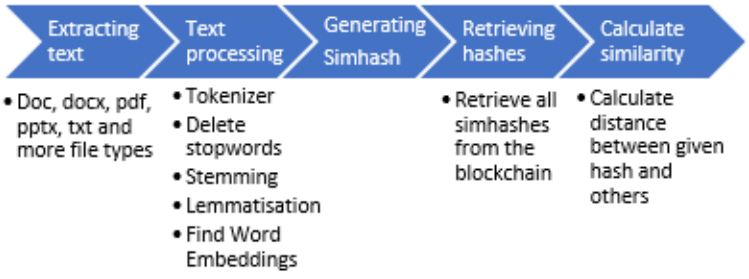

Figure 1 – Unified Algorithm

Finding plagiarism and calculate the similarity is not the easy task, especially when it comes to the multi-language comparison. For evaluation of the algorithm performance and accuracy I've implemented the next pseudo-algorithm (Figure 2)

Algorithm 2 Evaluation similarity pseudo algorithm

for `each added` *Document* **do**
  `process` *Document*
  `generate` *SimHash* `for processed` *Document*
  if *generatedSimHash* can be found on the Blockchain **then**
    added *Document* totally plagiarized ($S \leftarrow 1$)
  else
    calculate similarity S
  end if
return 10 most common *SimhHashes* and their S (similarity) score in range [0, 1]
end for

Figure 2 – Evaluation similarity pseudo algorithm

The outer loop makes request to the server with appropriate document for future analysis and processing. If the generated SimHash for given document is equal 1, it means that this document is totally plagiarized. If not, the proposed algorithm calculated similarity between current document and the all documents that available on the blockchain and return the 10 most "similar" with their scores. Also, the threshold for the document identified as plagiarized is equal to 25% and can be set programmatically.

IV. Results

Precision and recall are the primary performance indicators for information retrieval, pattern recognition, and classification activities. While recall (sensitivity) is the percentage of relevant documents that were recovered, or alternatively, the question of how many relevant documents were obtained, precision (positive predictive value) is the fraction of positive detect plagiarism papers to all the retrieved documents. Other metrics and techniques exist for measuring the effectiveness of an information system, such as area under the ROC curve (AUC), which is also taken into account in this study. The dataset [11] for evaluation proposed in this paper consists of 143,000 news articles from 15 major publications but used only 10,000. The language of the articles is English, because multi-language support is not fully implemented. Also, I can formalize this task, as a classification with two classes (plagiarism - 0 and not plagiarism - 1).

I have performed testing on Processor 11th Gen Intel(R) Core(TM) i7-1185G7 @ 3.00GHz, 2995 Mhz, 4 Core(s), 8 Logical Processor(s) with 32 GB RAM and 512 GB SSD drive, OS is Windows 10. The Python library for visualizing metric results *sklearn* was used. So, the results of testing my algorithm is given below on the Figure 3 and Figure 4 (ROC AUC score).

	precision	recall	f1-score	support
0	0.90	0.90	0.90	4988
1	0.90	0.91	0.90	5012
accuracy			0.90	10000
macro avg	0.90	0.90	0.90	10000
weighted avg	0.90	0.90	0.90	10000

Figure 3 – Metrics of proposed algorithm

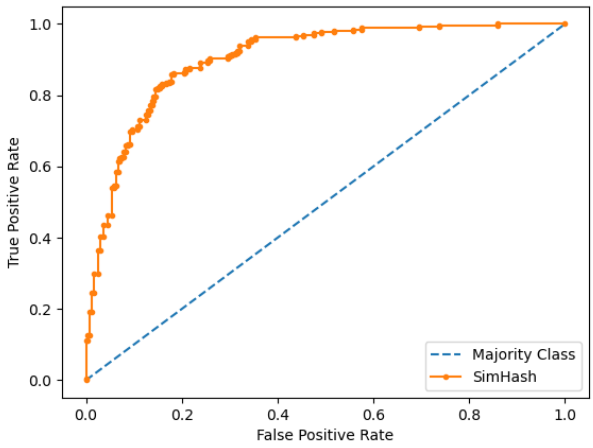

Figure 4 – ROC AUC score of proposed algorithm

V. Discussion

In this work, I suggest a brand-new technique for detecting plagiarism that makes use of the Ethereum blockchain and SimHash. My algorithm has decent precision, recall, and f1-score based on the findings; however, it is not suitable for usage in production. Overall, it performs 7% better than the well-known algorithm TF-IDF.

Additionally, the system displays flawless results when an exact replica of copied content is submitted, but this is a very simple example. The algorithm detects some percentage of plagiarism when genuine and original work is contributed, but it does not provide false-positive results. The most difficult test instance is when a cheater plagiarizes original material and attempts to hide it by paraphrasing the entire line. As a result, the suggested approach outperforms the TF-IDF model while having the lowest success rate in detecting it among the three test instances. As a result, I must update my corpora with related and synonym words in order to detect paraphrasing and incorporate more intricate multilingual support.

VI. Future research

Since this algorithm was tested on the dataset presented only in English, detection between different languages is also important in further development. Also, the algorithm was compared only with the TF-IDF algorithm. Further research is possible for more constructive pre-processing and application of text models of machine learning.

VII. Conclusion

The proposed unified algorithm performs well on detecting exact or partial copies of the documents; additionally, the performance does not degrade with scalability and size of the documents; changing words to their synonyms or paraphrasing can also be detected; identification of multi-language plagiarism performs on the basic level; complex language analysis must be performed to achieve a more precise solution. Furthermore, the SimHash algorithm performs well on enormous numbers of documents (billions), therefore the pace of plagiarism detection is consistent between low and high document numbers.

References

[1] M. S. Ali, M. Vecchio, M. Pincheira, K. Dolui, F. Antonelli and M. H. Rehmani, "Applications of Blockchains in the Internet of Things: A Comprehensive Survey," in *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1676-1717, Secondquarter 2019, doi: 10.1109/COMST.2018.2886932.

[2] S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System”, Aug 2022, [online] Available: https://bitcoin.org/bitcoin.pdf.

[3] V.Buterin, “Ethereum: A Next-Generation Smart Contract and Decentralized Application Platform”, Aug 2022, [online] Available: https://ethereum.org/669c9e2e2027310b6b3cdce6e1c52962/Ethereum_Whitepaper_-_Buterin_2014.pdf.

[4] Anzelmi, Daniele, Domenico Carlone, Fabio Rizzello, Robert Thomsen and Dil Muhammad Akbar Hussain. “Plagiarism Detection Based on SCAM Algorithm”. In *Proceedings of the International MultiConference on Engineers and Computer Scientists 2011* (Vol. Volume I, pp. 272-277). Newswood Limited, International Association of Engineers, IAENG

[5] M. Kozlenko, I. Lazarovych, V. Tkachuk and V. Vialkova, "Software Demodulation of Weak Radio Signals using Convolutional Neural Network," *2020 IEEE 7th International Conference on Energy Smart Systems (ESS)*, 2020, pp. 339-342, doi: 10.1109/ESS50319.2020.9160035.

[6] M. Kozlenko and V. Vialkova, "Software Defined Demodulation of Multiple Frequency Shift Keying with Dense Neural Network for Weak Signal Communications," *2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*, 2020, pp. 590-595, doi: 10.1109/TCSET49122.2020.235501.

[7] M. Kozlenko,V. Sendetskyi, O. Simkiv, N. Savchenko, A. Bosyi, “Identity Documents Recognition and Detection using Semantic Segmentation with Convolutional Neural Network (short paper)”. *Cybersecurity Providing in Information and Telecommunication Systems 2021* (CPITS), 2021, pp. 234-242.

[8] Dutchak, M., Kozlenko, M., Lazarovych, I., Lazarovych, N., Pikuliak, M., Savka, I. “Methods and Software Tools for Automated Synthesis of Adaptive Learning Trajectory in Intelligent Online Learning Management Systems”. *Innovations in Smart Cities Applications Volume 4. SCA 2020. Lecture Notes in Networks and Systems*, vol 183. Springer, Cham. https://doi.org/10.1007/978-3-030-66840-2_16.

[9] C. Pungilă, D. Galis, V. Negru, “A New High-Performance Approach to Approximate Pattern-Matching for Plagiarism Detection in Blockchain-Based Non-Fungible Tokens (NFTs)”, arXiv:2205.14492 [cs.CR], May 2022.

[10] G.S. Manku, A. Jain, A.D. Sarma, “Detecting Near-Duplicates for Web Crawling”, research.google.com, https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/33026.pdf (accessed Aug. 23, 2022).

[11] “All the news dataset”. Kaggle.com https://www.kaggle.com/datasets/snapcrack/all-the-news (accessed Aug. 27, 2022).