Data Duplication Removal by MD5 Using Random Forest Algorithm

Mrs. Kavyashree H. N.
Assistant Professor
Computer Engineering
Nutan Maharashtra Institute of Engineering and
Technology,Talegaon,Pune.

Mr. Himanshu T Sarode.
Computer Engineering
Nutan Maharashtra Institute of Engineering and
Technology,Talegaon,Pune.

Mr. Mohanish K Sarode.
Computer Engineering
Nutan Maharashtra Institute of Engineering and
Technology,Talegaon,Pune.

Mr. Yash S Pawar .
Computer Engineering
Nutan Maharashtra Institute of Engineering and
Technology,Talegaon,Pune.

Abstract— In many domains, data duplication is a major problem that results in inefficient processing, storage, and analysis. In this work, we investigate the use of machine learning methods for tasks related to data duplication and deduplication, with a particular emphasis on textual and image data. This project provides preparation pipeline for textual data that includes tokenizing text, addressing missing values, and standardizing formats. Then, feature extraction techniques like word embeddings and TF-IDF are used to mathematically represent text. These attributes can be used to train machine learning models, such as Support Vector Machines (SVM) or clustering methods like K- means, which are used to efficiently identify and eliminate duplicate entries. These models are capable of identifying duplicate photos because they learn hierarchical representations of images. To successfully recognize duplicates, this entails creating hybrid models that integrate textual and visual information.

Keywords—Data security, Deduplication, Authorization, Authentication, Access control, Cloud Security, Cipher Technology, Data synchronization.

I. INTRODUCTION

In today's data-driven world, organizations across a range of industries collect vast amounts of data to support their operations, strategic goals, and decision-making processes. Nonetheless, duplicating data in these databases remains a persistent problem. Duplicate data not only costs money in storage space but also taints data analysis, introduces inconsistencies, and jeopardizes the dependability of electronic systems [1].

This study aims to tackle the issue of redundant data by utilizing a robust deduplication method. Deserialization is the process of discovering and removing duplicate entries from datasets. It is also commonly referred to as duplication discovery or duplicate removal. Eliminating duplicates can help organizations maintain data quality, improve the efficacy of data-driven operations, and streamline processes [2]. The primary objective of this project is to design and develop an efficient deduplication system that can find and remove duplicate items in large databases. This system will employ complex algorithms and approaches to efficiently identify duplicates while lowering false positives and preserving data integrity [3]. Furthermore, this study aims to explore several aspects of deduplication, including algorithm selection, scalability, performance improvement, and interaction with existing.

By using a comprehensive approach, the proposed solution aims to provide a comprehensive way to handle duplicate data [4]. With this project's assistance, hope help enhance data quality control methods and let companies make the most of their data assets [5]. By lessening the effects of data duplication, organizations can enhance decision-making, foster creativity, and promote operational efficiency.

II. LITERATURE SURVEY

While providing quality-of-service provisioning for the original private information, a secondary encryption secure transmission approach backup plan. The proposed method uses the secure secondary communications to encrypt the principal confidential messages, giving the secondary system access to specific spectrum opportunities. To be more specific, the primary system encrypts the primary information using the secure secondary messages; if this is not possible, the spectrum will be used for secondary transmission [6]. In situations where the primary system is secure but the secondary messages can be transmitted securely, the primary system can transmit the primary information directly. Project analyze the performance of the primary ergodic secrecy rate and the average secondary throughput for the proposed architecture.

Reversible data hiding has drawn a lot of interest recently because of its many applications in a range of industries, such as cloud computing and the transfer of medical imaging. In this study, we describe a novel method that allows reversible data concealment in encrypted images. In this method, the data hider can embed one additional data bit into a small block (B × 2 pixels) from the encrypted image [11]. The non-overlapping areas, or blocks, of the encrypted image will all be processed by accessing those blocks in a preset order. An encrypted image block can have the bit value 0 without requiring any pixel values to be changed.

Bit value 1 embedding will map all the pixels in the first column of the selected picture block. This project's primary objective is to design and develop an efficient deduplication system that can find and remove duplicate items in large databases. Duplicate data not only costs money in storage space but also contaminates data analysis, introduces errors, and jeopardizes the dependability of electronic systems.

They provide a secondary encryption secure transmission strategy to protect the original privacy information and offer quality-of-service provisioning for the secondary system. According to the suggested plan, the secondary system can gain certain spectrum opportunities while the primary system encrypts the primary confidential messages using the secure secondary messages [9]. To be more precise, in cases where the primary system is secure, the primary information can be transmitted directly; in cases where the primary system is not secure but the secondary messages can be transmitted securely, the primary system encrypts the primary information using the secure secondary messages; in other cases, the spectrum will be used for secondary transmission. They study the performances of the average secondary throughput and the primary ergodic secrecy rate for the suggested scheme. Quantitative findings have indicated that the primary privacy messages can be protected and the secondary transmission throughput can be increased by using the secondary encryption secure transmission scheme.

Information transfer is far more convenient in the realm of information development. But there's always a chance that the transmission mechanism will be breached, stolen, or altered, raising the possibility that the data source is inaccurate. For this reason, some academics suggested using passwords to secure sensitive data [7]. The 3D-Playfair Cipher with Message Integrity Using MD5 was proposed by Alok et al. 3D-Playfair encryption is used in this paper. Nevertheless, the author suggests that since basic 3D-playfair encryption cannot ensure the integrity of data while it is being transmitted, combined with MD5 to guarantee the data's integrity, but because the data source's legitimacy is in question, this study employs XOR calculation techniques to confirm the data's veracity one more time. In the event of a man-in-the-middle attack, the attacker intercepts the packet. Furthermore, it remains possible to reliably ascertain whether the data's original sender is its source even while changing with its content. This approach increases the data's credibility while ensuring its integrity.

Due to its numerous uses in a variety of fields, including cloud computing and the transmission of medical images, reversible data concealing has received a lot of attention in recent years. In this paper, we present a brand-new method for reversibly hiding data in encrypted pictures. Under this approach, a small block (B × 2 pixels) from the encrypted image can have one bit of additional data embedded by the data hider [12]. The encrypted image's non-overlapping portions, or blocks, will all be processed by accessing those blocks in a predetermined order. There is no need to change any pixel values in order to incorporate the bit value 0 in an encrypted image block. All of the pixels in the first column of the chosen image block will be mapped if bit value 1 is to be embedded. According to a predetermined function, into a new pixel value. Data extraction and picture recovery are performed at the receiver side by comparing the proximity of pixels in the neighboring columns of the decrypted image's pixels in every block.

III. METHODOLOGY

The machine learning approach to data duplication elimination is a methodical process that is customized for text and visual data. Text normalization, stop word removal, tokenization, and punctuation removal are the first steps in the text duplication elimination procedure. Sentence embeddings and word embeddings are two feature extraction approaches that turn text into numerical representations. The likeness of text samples is then evaluated by measuring similarity using measures such as cosine similarity. Similar texts are grouped by clustering techniques like k-means, and duplicates are classified based on a threshold. After that, duplicates that have been found are eliminated to preserve original data [8].

On the other hand, preparation for picture duplication removal include scaling images and utilizing pretrained CNNs to extract features. Distance measurements are used to measure similarity between photos, and then clustering is used to group visually comparable images. Determining the threshold and then removing duplicates are done, with a focus on maintaining a variety of image content. Iterative improvement is driven by continual monitoring and feedback, with evaluation criteria such as precision and recall serving as validation of system performance [10]. This approach guarantees effective duplicate elimination, improving data quality, and supporting well-informed decision-making through smooth integration into current pipelines and vigilant monitoring.

The process of extracting meaningful representations from preprocessed data is known as feature extraction, and it comes after data preparation. Text data is frequently converted into numerical representations using feature extraction approaches like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings like Word2Vec or GloVe. By capturing the semantic meaning of words and phrases, these strategies help the model comprehend the text's context [13]. Utilizing pre-trained CNN (Convolutional Neural Network) models like VGG and ResNet, as well as methods like SIFT (Scale-Invariant Feature Transform) and perceptual hashing (e.g., pHash), features can be retrieved from image data. These features help identify duplicates by capturing significant visual patterns and characteristics found in the photos.

After the features have been retrieved, the next step is model training, in which machine learning techniques are used to find patterns and connections in the data. Classification algorithms like logistic regression, random forest, or neural networks can be trained to predict whether or not pairs of texts are duplicates in order to remove text duplication [14]. These models generate a probability score that indicates the possibility of duplication from the feature vectors of two texts as input. Classification models, in particular CNNs, are trained to identify pairs of images as duplicates or non-duplicates based on their derived features in a similar manner for the purpose of image duplication removal. During the training phase, the model is fed labeled data, its parameters are iteratively adjusted to reduce prediction errors, and its performance is assessed using metrics like accuracy, precision, recall and F1 score[19].

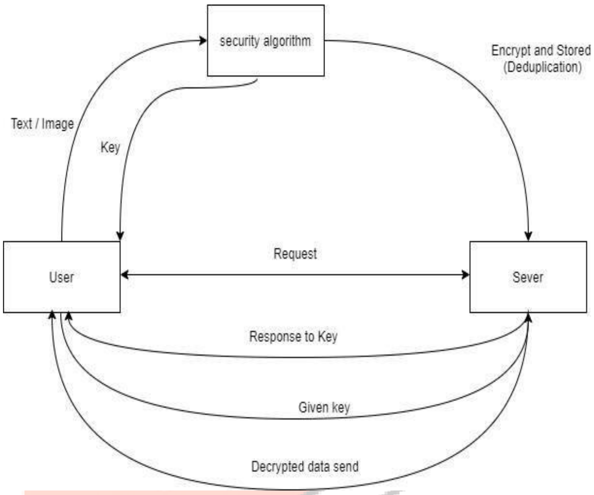

Figure 1- System Architecture

The last step, known as duplicate detection, involves comparing pairs of data samples to find duplicates once the model has been trained. Duplicate pairings in text data are found by calculating similarity scores using metrics such as cosine similarity between their feature representations. When two pairs of similarity scores go above a certain threshold, they are deemed duplicates and are deleted from the dataset. Similarity thresholds are used to identify duplicates in picture data, and similarity scores are calculated using the retrieved features [20]. This procedure ensures the quality and integrity of the dataset by efficiently removing redundant data.

To improve the duplication elimination system's resilience and performance, a few more factors must be taken into account during the process. These include ensemble approaches (combining many models for increased accuracy), data augmentation techniques (generating more training data), and active learning strategies (incorporating human feedback and iteratively improving the model's performance) [17]. Sustained assessment and system improvement are also necessary to properly handle new problems or adjustments to the data distribution.

In conclusion, data preprocessing, feature extraction, model training, and duplicate detection are all steps in the methodical and iterative process of eliminating data duplication using machine learning. A strong and efficient duplication removal system may be created to guarantee the integrity and quality of datasets in a variety of applications by combining these steps and taking into account extra elements like data augmentation and model validation[18].

IV. ADVANTAGES AND APPLICATIONS

A. Advantages *:*

1) Efficiency: The amount of manual labor and time needed to find and remove duplicate data instances is greatly decreased by automated duplicate removal techniques [15]. vast datasets can be handled by machine learning algorithms, which can process vast amounts of data efficiently. This makes the duplicate elimination procedure scalable.
2) Accuracy: By recognizing intricate links and patterns in the data, machine learning models are able to accurately identify duplicate occurrences [16]. The system minimizes the possibility of false positives and false negatives by achieving high precision and recall rates through the use of sophisticated algorithms and feature representations.
3) Scalability: The approach can handle a wide range of data kinds, including text, photos, audio, and more. It is very flexible and scalable. It is appropriate for a broad range of applications across several industries, including e-commerce, healthcare, banking, and more, due to its ability to handle a variety of data sources and formats.

B. Applications *:*

1) E-commerce Platforms: A lot of product data is frequently handled by e-commerce websites. Eliminating redundant product listings makes search results more relevant, improves user experience, and keeps buyers from becoming confused.
2) Content Management Systems (CMS): Textual content is managed by Content Management Systems (CMS) platforms. By identifying and removing duplicate blog posts, articles, or user- generated information, text deduplication preserves content quality and enhances search engine optimization (SEO).
3) Document Management Systems: To store and arrange massive amounts of papers, organizations rely on document management systems. By ensuring that duplicate documents are found and eliminated, text deduplication maximizes storage capacity and expedites document retrieval.

V. ALGORITHMS

1)Shingling:
Make each page into a series of shingles, which are fixed-length, brief, overlapping subsequences. Three shingles, for instance, would be {"The quick brown", "quick brown fox", and "brown fox jumps"} for a document titled "The quick brown fox jumps."
2)Hashing:
on produce a hash value, apply hash functions on every shingle. Usually, several hash functions are employed for each shingle. Hash values should be distributed uniformly and quickly via hash functions.
3)Signature MinHash:
Make a MinHash signature a brief collection of hash values that represents the document for every document. The minimum hash value for each hash function over all of the document's shingles is chosen to create the signature.
4)Compute Similarity:
Utilizing the MinHash signatures of each image, determine the Jaccard similarity between pairs of photographs. It is possible to modify similarity levels in accordance with the particular features of the image data.
5)Identifying Duplicates:
Determine which image pairs are likely duplicates if their Jaccard similarity is higher than the selected threshold. You can also choose to carry out other verifications, including analyzing the information of the images or refining duplicate detection with more sophisticated similarity metrics.

VI. RESULT AND ANALYSES

Data duplication removal and image text de- duplication are successfully implemented in this project. This project text and Image data. It has to keep secure in a cloud server. Digital images have to be protected over the communication, however generally personal identification details like copies of pan card, Passport, ATM, etc., to store on one's own pc. So, we are protecting the text file and image data for avoiding the duplication in our proposed system. Figure gives us the result and output of the project:

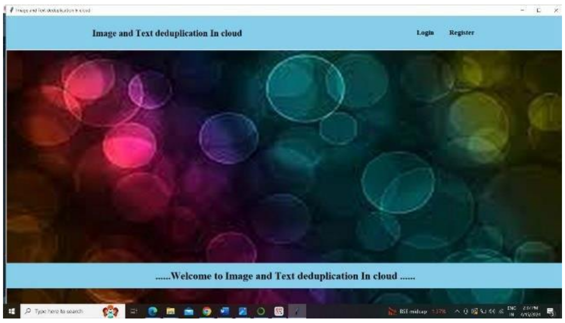
Image and Text deduplication In cloud
Login
Register
......Welcome to Image and Text deduplication In cloud ......

Figure 2- Image Text Duplication

This is a homepage which can perform registration and login activities for user.

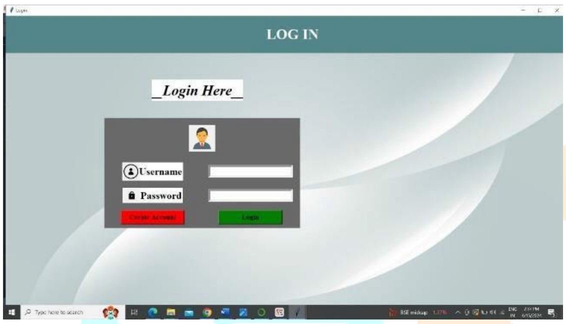

Figure 3- Login Page

Login page is for user to login in the system to perform individual actions.

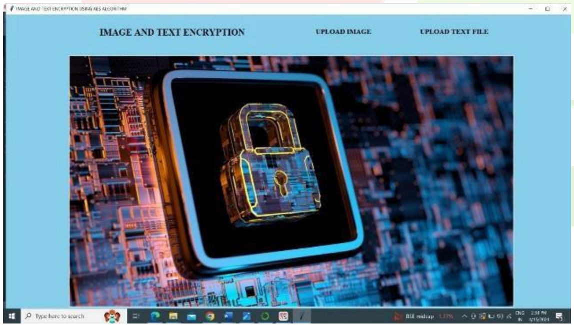

Figure 4- Image, Text Encryption

This page is for image uploading and text uploading to the local server and get encrytion and decryption of specific image or text.

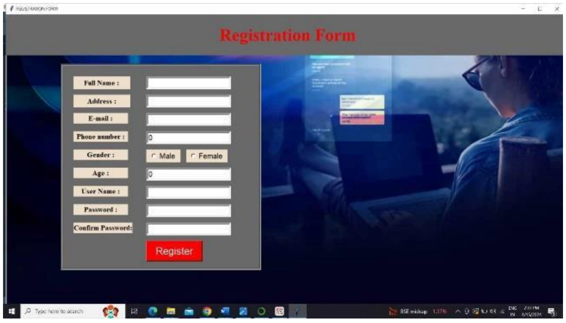

Figure 5- Registration Page

Registration page for the first time user to save the information of user like full name, email, username and password, etc.

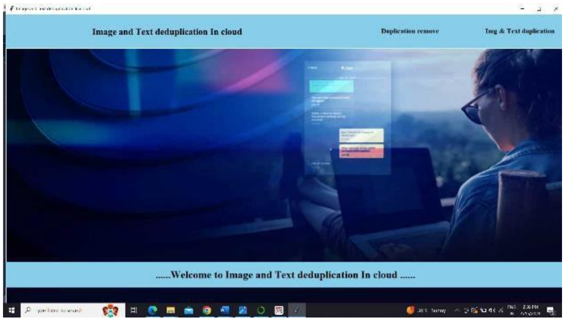
Image and Text deduplication In cloud
Duplication remove
Img & Text duplication
......Welcome to Image and Text deduplication In cloud ......

Figure 6- Duplication Removal Page

This page checks whether the uploaded file already exists or not if not then encrypt it.

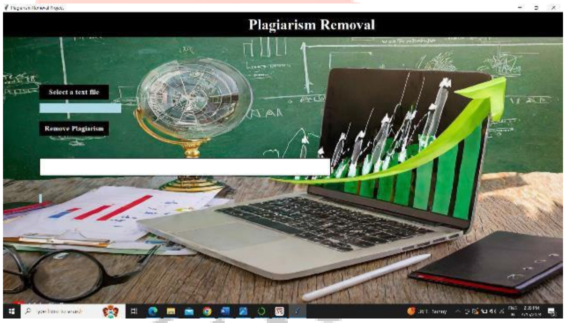

Figure 7- Plagiarism Removal Page

This page accepts text file only and remove the duplicates.

VII. FUTURE SCOPE

The future potential of machine learning and related techniques for data duplication reduction holds great promise for improving efficiency, accuracy, and adaptability in a variety of fields. In the future, research will probably concentrate on improving the performance and sophistication of algorithms designed especially for duplicate removal jobs as machine learning models continue to develop . This could entail creating deep learning architectures, attention mechanisms, and reinforcement learning strategies to discover and remove duplicate data instances with greater efficiency and accuracy.

Furthermore, there will be an increasing demand for duplication removal systems that can handle a variety of data formats as multi-modal data sources, including text, photos, audio, and video, proliferate. In addition, it is anticipated that privacy-preserving methods like federated learning and differential privacy would be crucial in guaranteeing the security and integrity of sensitive data while permitting efficient duplication elimination.

It will be necessary to have real-time duplicate detection capabilities in order to facilitate prompt decision-making and response, especially in industries like cybersecurity and healthcare. Additionally, applications for future systems may extend outside traditional fields, such as e-commerce, cybersecurity, and healthcare, improving productivity and data quality in a variety of industries.

VIII. CONCLUSION

A strong framework for improving data quality, integrity, and efficiency across numerous domains is presented by the methodology of eliminating data duplication utilizing machine learning and related techniques. Duplication removal systems are scalable and precise approaches for finding and removing duplicate data instances by combining sophisticated algorithms, feature extraction strategies, and model training procedures. Even if there are still obstacles to be solved, such as computational resources, data quality dependencies, and algorithm complexity, continued research and innovation are well-positioned to resolve these problems and open up new avenues for future developments.

The future of duplication removal looks very promising with the development of machine learning models, the growth of multi-modal data sources, and the growing need for real-time and privacy-preserving solutions. Duplication removal systems will continue to be essential to guaranteeing data integrity, dependability, and usability across a wide range of sectors and applications by seizing these opportunities and utilizing cutting-edge technologies. The ongoing development and application of duplicate removal techniques will be crucial for enabling actionable insights, spurring innovation, and building confidence in the data-driven decision-making process as we negotiate the difficulties of data management in the digital era.

ACKNOWLEDGEMENT

We wish to thank our parents and associates for their valuable support and encouragement throughout the development of the project work and we would also like to thank our guide Prof. Kavyashree H. N. for guiding us throughout the project work.

REFERENCES

[1] Schütze, H., Raghavan, P., and Manning, C.D. (2008). Information Retrieval Overview. Cambridge University Press.

[2] In 2014, Leskovec, J., Rajaraman, A., and Ullman, J.D. Extensive Data Mining. Cambridge University Press.

[3] Zaki, M.J., Ganesan, K., and Hacid, M.S. (2008). finding duplicate sets of websites. Systems of Knowledge and Information, 17(1), 1–28.

[4] Wiederhold, G., Wang, J.Z., and Li, J. (2002). Semantics-Sensitive Integrated Matching for Picture Libraries: SIMPLIcity. 23(9), 947-963, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[5] T. Mishra, A. Kumar, and S. Satapathy, "Designing and Implementing Mobile Applications with Flutter," in Proc. of 2021 International Conference on Computing and Communication Systems (IC3S), 2021

[6] Schütze, H., and C.D. Manning (1999). Statistical Natural Language Processing's foundations. Press of MIT. In 2016,

[7] He, K., Zhang, X., Ren, S., and Sun, J. Image Recognition with Deep Residual Learning. IEEE Conference on Computer Vision and Pattern Recognition Proceedings, 770-778.

[8] Jordn , M. I., Ng, A.Y., and Blei, D. M. (2003). Allocation of Latent Dirichlet. 3, 993_1022, Journal of Machine Learning Reasearch.

[9] Chen, K., Corrado. G., Mikolov, T., & Dean, J. (2013)Efficient Vector Space Word Representation Estimation preprint arXiv.

[10] Ba, J., and D. P. Kingma (2014). Adam: A Stochastic Optimization Method Method. arXiv preprint 1412.6980 arXiv:1412.

[11] R. Steinberger, B. Pouliquen and J. Hagman, "Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC", in Proceedings of the 3rd Inter. Conf. of Computational Linguistics

and Intelligent Text Processing (Mexico City, Mexico, 2002, pp. 415-424.

[12] S. Brin, J. Davis and H. Garcia-Molina, “Copy Detection Mechanisms for Digital Documents”, in the ACM SIGMOD International Conference on Management of Data (San Jose, California, USA, May 22-25 1995), 1995, pp. 398-409.

[13] A. Si, H.V. Leong and R.W.H. Lau, “CHECK: A document plagiarism detection system”, in Proceedings of ACM Symposium for Applied Computing, ACM (San Jose, California, USA, February 28 - March 1 1997), 1997, pp. 70-77.

[14] M.C. Burl, M. Weber, and P. Perona, ªA Probabilistic Approach to Object Recognition Using Local Photometry and Global Geome- try,º Proc. European Conf. Computer Vision, pp. 628-641, June 1998.

[15] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom et al. ªQuery by Image and Video Content: The QBIC System,º IEEE Computer, vol. 28, no. 9, 1995.

[16] Y. Rubner, L.J. Guibas, and C. Tomasi, ªThe Earth Mover's Distance, Multi-Dimensional Scaling, and Color-Based Image Retrieval,º Proc. DARPA Image Understanding Workshop, pp. 661-668, May 1997.

[17] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157–166, 1994.

[18] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580, 2012.

[19] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to hand- written zip code recognition. Neural computation, 1989.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.