REMOVING OF MULTIPLE VOTES BY USING DE-DUPLICATION ANALYSIS

Dr. R. Poorinma
Professor, Malla Reddy University
Hyderabad

Adithi Kulkarni
Student, Malla Reddy University
Hyderabad
2011cs020023@mallareddyuniversity.ac.in

G. Ashrutha
Student, Malla ReddyUniversity
Hyderabad
2011cs020026@mallareddyuniversity.ac.in

K.Vijaya
Student, Malla Reddy University
Hyderabad
2011cs020045@mallareddyuniversity.ac.in

B. Sai Neha
Student, Malla Reddy University
Hyderabad
2011cs020053@mallareddyuniversity.ac.in

Abstract: De-duplication analysis is crucial in maintaining the integrity of datasets, particularly in contexts like elections and surveys, where accurate representation is paramount. Ensuring that each eligible voter is represented only once in the voter register upholds the democratic principle of one person, one vote. The "Removing of Multiple Votes by using De-Duplication Analysis" project is dedicated to enhancing the accuracy and reliability of voting systems by leveraging advanced data analysis techniques and the 'DataVoter.xlsx' database. By identifying and eliminating duplicate or multiple entries for individual voters, this project significantly improves the accuracy and transparency of the voting process. Its success lies in its ability to address the critical need for precision and fairness within electoral systems, ultimately reinforcing trust in democratic processes.

Index terms *– Multiple votes, De-duplication Analysis.*

1. INTRODUCTION

In the realm of data collection, ensuring the integrity and accuracy of responses is paramount. Whether it be in the context of elections, surveys, or other data-gathering endeavors, the presence of multiple votes or responses from the same individual can distort results and comprise the credibility of the process. This is where de-duplication analysis comes into play.

De-duplication analysis is a meticulous and systematic approach aimed at identifying and eliminating duplicate or multiple entries within a dataset. By employing this technique, one can sift through large volumes of data and distill it into a more accurate and reliable representation of unique responses.

In the landscape of data-driven decision-making, the presence of duplicate or multiple entries within datasets poses a significant change.

This issue becomes particularly pronounced in scenarios such as elections, surveys, and opinion polls, where the accuracy of results hinges on the uniqueness of each participant's contribution.

It is important that a voter register should only include one current record for each eligible voter, to ensure the register facilitates the democratic principle of one person, one vote.

The problem statement centers on devising systematic and transparent methodologies to identify and remove duplicate entries with datasets, ensuring that the final data accurately reflects the distinct input of each participant.

This prompts exploration to de-duplication analysis.

Introducing our groundbreaking project focused on enhancing the integrity of voting systems: De-Duplication Analysis for the Removal of Multiple Votes. In a world where accurate representation is paramount, the challenge of eliminating duplicate votes within electoral processes has gained prominence.

Our innovative solution leverages advanced data analytics and to identify and eliminate multiple votes, ensuring that each individual's voice is counted justly. By mitigating the risk of fraudulent practices and enhancing the transparency of voting systems.

Our project aims to uphold the fundamental principles of democracy and reinforce public trust in the electoral process.

2. LITERATURE SURVEY

Data de-duplication, the process of identifying and removing duplicate or multiple entries from datasets, is crucial in various fields such as elections, surveys, and data management systems. This literature survey explores different techniques and approaches proposed by researchers to address data de-duplication challenges. The survey covers methods from hashing algorithms to data mining techniques, aiming to provide insights into the current landscape of data de-duplication research.

In their paper, Aishwarya et al. proposed a solution to data de-duplication issues on the cloud using hashing and MD5 techniques [8]. Hashing functions play a vital role in identifying duplicate records efficiently. By generating unique hash values for each record and comparing them, duplicate entries can be detected and removed. MD5, a widely used hashing algorithm, ensures data integrity and aids in the de-duplication process.

Selvi et al. conducted a survey on the removal of duplicate records using data mining techniques [9]. Data mining algorithms such as clustering, association rule mining, and decision trees have been applied to identify patterns and similarities among records, facilitating the detection of duplicates. These techniques offer robust solutions for de-duplication tasks, especially in large datasets where manual inspection is impractical.

Liu et al. proposed a novel optimization method to improve the performance of de-duplication storage systems [10]. Optimization techniques enhance the efficiency of de-duplication processes by minimizing computational overhead and storage space requirements. By optimizing data storage and retrieval mechanisms, these methods contribute to faster and more accurate de-duplication outcomes, ensuring timely processing of datasets.

Luciv et al. introduced the Duplicate Finder Toolkit, a comprehensive framework for data de-duplication tasks [11]. This toolkit provides researchers and practitioners with a suite of tools and algorithms designed specifically for duplicate detection and removal. By offering a unified platform for de-duplication activities, the toolkit streamlines the process and improves the reproducibility and scalability of de-duplication tasks.

Data de-duplication is a critical aspect of data management and analysis, with applications in various domains including elections, surveys, and cloud computing. This literature survey highlights the diverse range of techniques and approaches proposed by researchers to tackle de-duplication challenges. From hashing and MD5 techniques to data mining algorithms and optimization methods, each approach contributes to the advancement of de-duplication research. By understanding the strengths and limitations of these techniques, researchers can develop more effective solutions for ensuring data integrity and accuracy in diverse datasets.

3. METHODOLOGY

i) Proposed Work:

The proposed system seeks to address the challenges associated with duplicate record removal in systematic reviews by introducing a comprehensive and automated approach to de-duplication analysis. Leveraging five different de-duplication strategies, the system aims to evaluate their efficacy in terms of time efficiency and accuracy. By tracking the time taken for de-duplication in each option, as well as quantifying false positives and false negatives, the system provides a holistic assessment of the performance of each method. Key features of the proposed system include its ability to streamline the de-duplication process, reducing the manual effort required and minimizing the likelihood of errors. Automation ensures consistency and reliability in duplicate detection, enhancing the overall efficiency of systematic review procedures. Additionally, the system's comprehensive approach considers various factors such as time efficiency and accuracy, enabling researchers to make informed decisions regarding the selection of de-duplication strategies. Overall, the proposed system offers a robust solution to the challenges posed by duplicate record removal in systematic reviews, ultimately improving the quality and reliability of research outcomes.

ii) System Architecture:

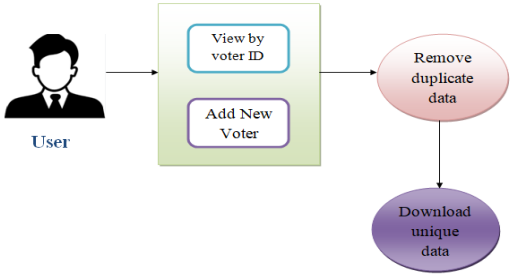


Fig 1 Proposed Architecture

iii) Modules:

To implement this project we have designed the following modules: They are User signup, User login, Search voter ID, Add / register new voter, remove duplicates, download unique data

a) User signup:

The User Signup module enables new users to register for the system by providing essential information such as username, email, and password. Through a simple

and intuitive interface, users input their details to create a new account, facilitating their access to the system's functionalities. This module streamlines the signup process, ensuring a seamless and user-friendly experience for individuals seeking to join the platform. By collecting necessary information upfront, the module sets the foundation for user authentication and personalized user interactions within the system.

b) User login:

The User Login module enables registered users to access the system by entering their credentials, including username and password. Through this module, users authenticate their identity to gain entry into their accounts and utilize the system's functionalities. By providing a secure and straightforward login process, the module ensures that authorized users can access their personalized settings, data, and features within the platform. This facilitates a seamless user experience and promotes efficient interaction with the system's resources and services. Additionally, the module enhances system security by verifying user credentials before granting access to sensitive information.

c) Search voter ID:

The Search Voter by ID module allows users to query the system for voter records using a specific voter ID. Users input the unique voter ID they wish to search for, and the system retrieves the corresponding voter information from the database. This module streamlines the process of accessing individual voter details, providing users with quick and efficient access to specific voter records. By offering a targeted search functionality based on voter IDs, the module enhances user productivity and facilitates seamless access to voter information within the system.

d) Add/Register New Voter to Existing 'DataVoter.xlsx' File:

The Add/Register New Voter to Existing 'DataVoter.xlsx' File module enables users to input and register new voter information directly into the existing 'DataVoter.xlsx' Excel file. Users provide details such as voter ID, name, address, and other relevant information for the new voter. Upon submission, the system processes the input and seamlessly integrates the new voter's details into the 'DataVoter.xlsx' file. This module simplifies the process of updating voter records, ensuring that the Excel file remains up-to-date with the latest voter information. By offering a user-friendly interface for data entry, the module enhances efficiency and accuracy in voter registration procedures.

e) Remove Duplicates and Display Chart with All Records and Unique Records Length:

The Remove Duplicates and Display Chart module is dedicated to data deduplication within the 'DataVoter.xlsx' file. It systematically identifies and eliminates duplicate records from the dataset. Following the deduplication process, the module generates a visual chart that illustrates the count of records before and after deduplication. This chart provides users with a clear comparison, showcasing the reduction in duplicate entries and the resulting unique record count. By offering visual insights into the deduplication outcome, the module enhances user understanding and ensures the integrity and accuracy of voter data within the system.

f) Download Unique Data:

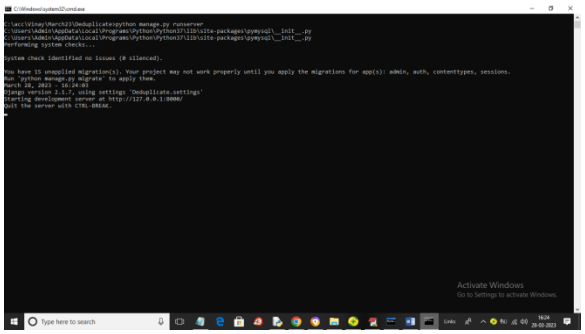
The Download Unique Data module facilitates users in obtaining a file comprising solely unique records

post-deduplication. Users are presented with a download link, granting them access to the filtered dataset devoid of any duplicates. This downloadable file serves as a refined version of the original dataset, ensuring that users can access and utilize accurate and streamlined voter information. By providing easy access to the unique data subset, this module enhances user convenience and promotes efficient utilization of deduplicated records for further analysis or integration into other systems.

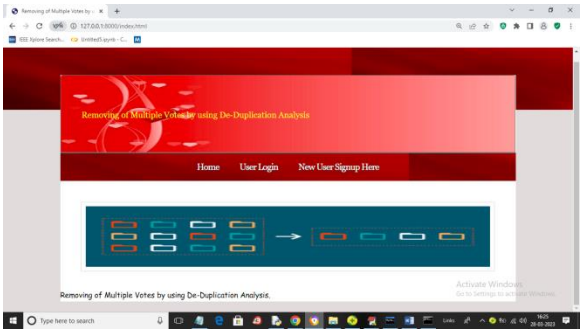
4. EXPERIMENTAL RESULTS

To run project install python 3.7.0 and then install all packages given in requirement file and then install MYSQL database and give password as 'root' and after installation open MYSQL and copy content from DB.txt file and paste in MYSQL to create database.

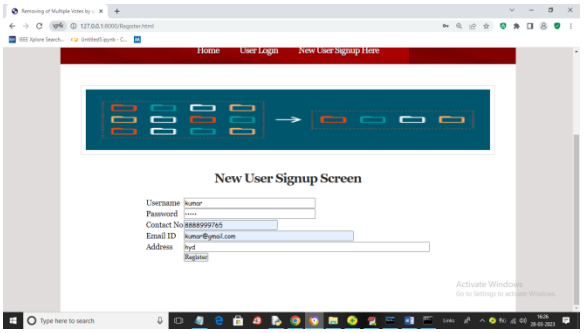
Now double click on 'run.bat' to start python web server and get below page



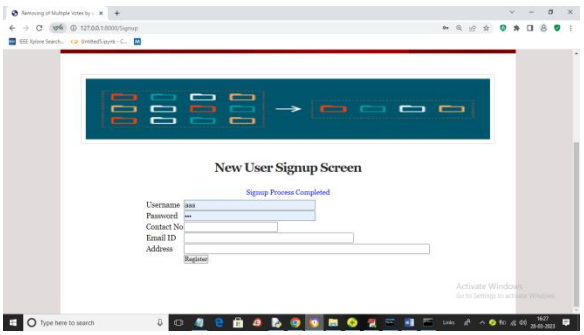
In above screen server started and now open browser and enter URL as http://127.0.0.1:8000/index.html and press enter key to get below page



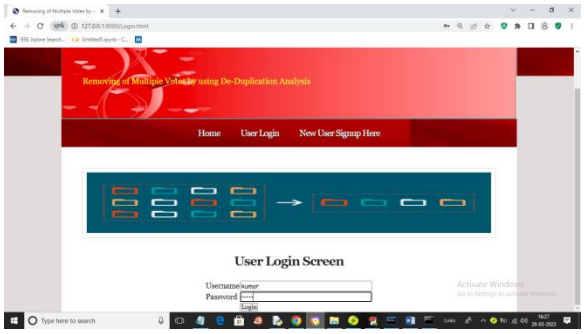
In above screen click on 'New User Signup Here' link to get below page



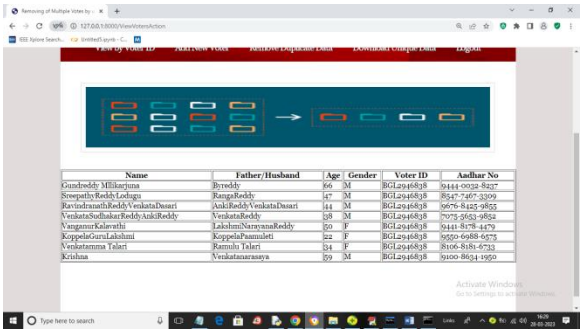
In above screen user is signing up and then press button to get below page



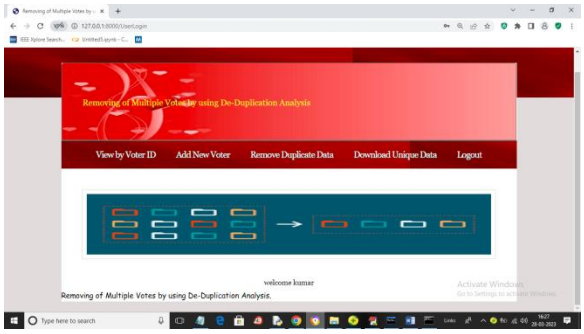
In above screen signup process completed and now click on 'User Login' link to get below page



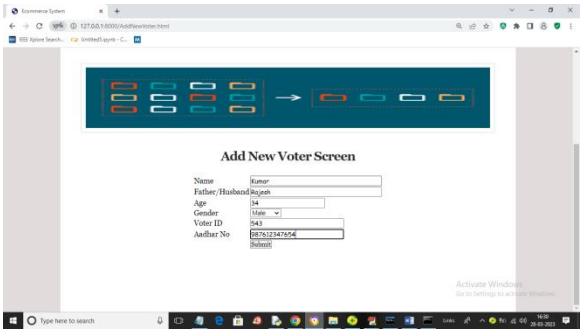
In above screen user is login and after login will get below page



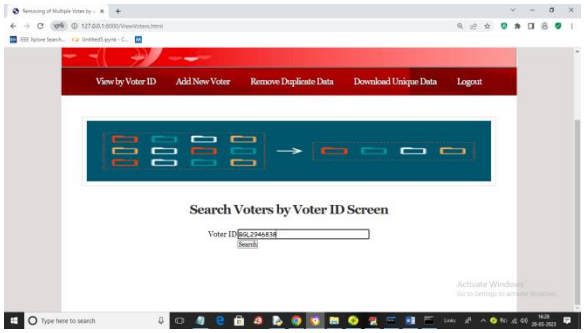
In above screen we got search result from given voter ID but there are multiple records exists on same ID and now click on ‘Add New Voter’ link to get below page



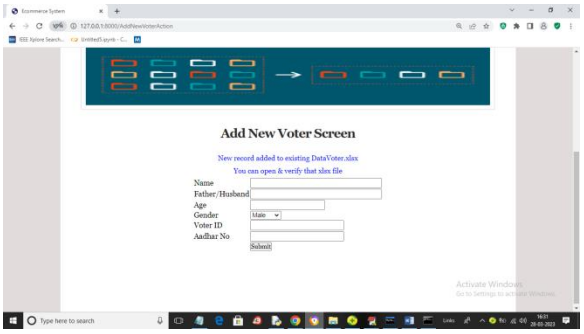
In above screen user can click on ‘View by Voter ID’ link to get below page



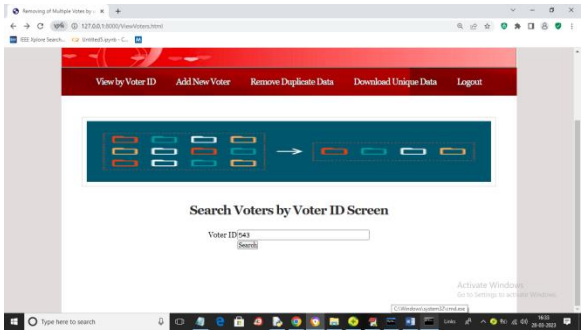
In above screen I am adding new voter details and then press button to add voter to excel file and get below output and I gave voter id as 543 and by giving this ID we can search record



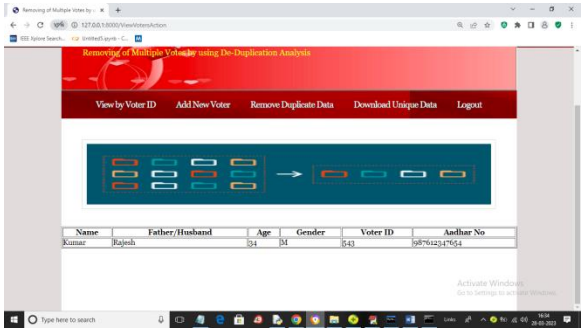
In above screen I entered some ‘Voter ID’ and then press button to get below page



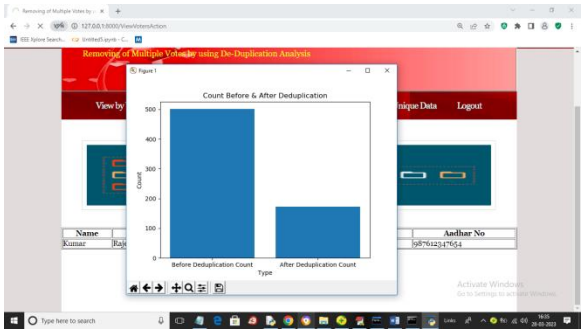
In above screen in blue colour text we can see the response as record added and now click on 'View by Voter ID' to search record again



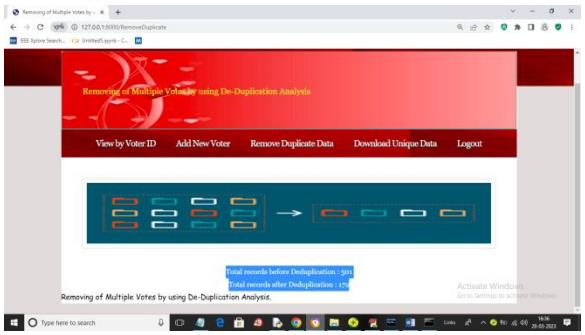
In above screen I am giving newly added Voter ID and press button to get below page



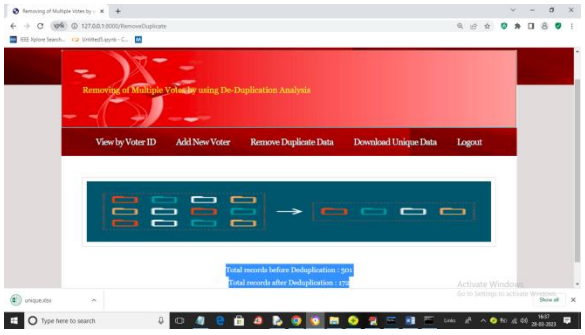
In above screen we can see newly added member record and now click on 'Remove Duplicate Data' link to get below output



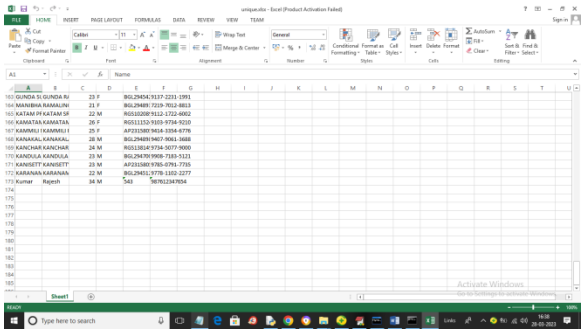
In above graph x-axis represents before and after duplication type and y-axis represents count and in above graph after removing duplicates we got unique records closer to 200 and now close above graph to get below output



In above screen in blue colour text we can see number of records before and after duplication and now click on 'Download Unique Data' link to download file and get below output



In above screen in browser status bar we can see unique file is downloading and you can open and see that file like below screen



In above screen in unique file after removing duplicates we got 173 records. Similarly by following above screens you can run code

5. CONCLUSION

The proposed system aims to revolutionize de-duplication analysis by introducing a comprehensive solution that is not only more efficient and accurate but also ethically sound. By streamlining the process of maintaining data integrity, the system seeks to enhance the credibility of data-driven decision-making. Through its advanced methodologies and streamlined workflows, the system ensures that duplicate records are efficiently identified and removed, resulting in cleaner and more reliable datasets. Moreover, by upholding ethical principles such as privacy and fairness in data handling, the system fosters trust among users and stakeholders. As organizations increasingly rely on data to drive their decisions, the role of the de-duplication system becomes pivotal in shaping a future where data integrity is not merely a goal but a standard. By providing a robust framework for ensuring data accuracy and reliability, the system contributes to building a foundation of trust in data-driven processes. Ultimately, the proposed system emerges as a key player in promoting transparency, efficiency, and credibility in the realm of data management and analysis.

6. FUTURE SCOPE

The proposed de-duplication system serves as the cornerstone for ongoing improvement and adaptability. By incorporating enhanced data matching techniques, the system can further refine its ability to accurately identify and remove duplicate records. Additionally, increasing scalability ensures that the system can effectively handle larger datasets and growing user demands. These features enhance the system's effectiveness, enabling it to evolve in response to changing requirements and technological advancements, ultimately ensuring its continued relevance and impact in maintaining data integrity.

REFERENCES

[1] Chuanyi Liu; Yibo Xue; Dapeng Ju; Dongsheng Wang, et. al., “A Novel Optimization Method to Improve De-duplication Storage System Performance” published in ieee open Access, available at https://ieeexplore.ieee.org/document/5395260.

[2] C. I. Ezeife and T. E. Ohanekwu, “The use of smart tokens in cleaning integrated warehouse data,” International Journal of Data Warehousing and Mining (IJDWM), vol. 1, no. 2, pp. 1–22, 2005

[3] Dmitry Luciv; Dmitrij Koznov; George Chernishev; Hamid Abdul Basit; Konstantin Romanovsky , et. al., “Poster: Duplicate Finder Toolkit” published in ieee open Access, available at https://ieeexplore.ieee.org/document/8449485.

[4] E. Manogar and S. Abirami, “A study on data deduplication techniques for optimized storage,” in 2014 Sixth International Conference on Advanced Computing (ICoAC). IEEE, 2014, pp. 161–166.

[5] M. V. Maruti and M. K. Nighot, “Authorized data deduplication using hybrid cloud technique,” in 2015 International Conference on Energy Systems and Applications. IEEE, 2015, pp. 695–699.

[6] M. Maragatharajan and L. Prequiet, “Removal of duplicate data from encrypted cloud storage,” in 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS). IEEE, 2017, pp. 1–5.

[7] Nikita Medidar; Manik Chavan, et. al., “Data Duplicate Detection” published in ieee open Access, available at https://ieeexplore.ieee.org/document/8494135.

[8] O. A. FESTUS, “Data finding, sharing and duplication removal in the cloud using file checksum algorithm.”

[9] P. Puzio, R. Molva, M. Onen, and S. Loureiro, “Cloudedup: secure ¨ deduplication with encrypted data for cloud storage,” in 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, vol. 1. IEEE, 2013, pp. 363–370

[10] R. Aishwarya; K Sumanth Singh; S Mahesh Varma; Yogitha. R; G. Mathivanan , et. al., “Solving Data De-Duplication Issues on Cloud using Hashing and MD5 Techniques” published in ieee open Access, available at https://ieeexplore.ieee.org/document/9753902.

[11] Selvi, D.Shanmuga Priyaa, et. al., “A Perspective Analysis on Removal of Duplicate Records using Data Mining Techniques: A Survey” published in research gate open Access, available at https://www.researchgate.net/publication/316643996_36_P.