Kadali Devi Sindhuja, 2025, 13:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

SMARTCLEAN: AI–POWERED DIGITAL HOARDING CLEAN

Professor K Deepa Shree, Sanjana S, Shreya R, Sinchana Adiga, Sharanya

Department of Computer Science Engineering, Dayananda Sagar Academy of Technology and Management, Karnataka, India

Abstract- In today's data-driven world, individuals and organizations generate vast amounts of digital data, often leading to inefficient storage and digital clutter—commonly known as digital hoarding. Traditional storage cleaners primarily focus on file size or duplication and lack the intelligence to make context-aware decisions. This paper proposes an AI-powered Digital Hoarding Cleaner that leverages advanced Natural Language Processing (NLP) and Machine Learning (ML) to analyze file content and user behavior. The system integrates models such as BERT, GPT, and BART for file summarization and semantic understanding. It offers personalized recommendations to retain, delete, or reorganize files based on relevance, using techniques like semantic similarity detection, fuzzy matching, and behavioral analytics. Cloud support for services such as Google Drive and OneDrive ensures seamless integration, while a user-friendly web dashboard provides insights into storage patterns and suggested actions. Emphasizing data privacy, the tool enables local processing and secure communication. Overall, the system aims to optimize storage usage, reduce cognitive overload, and deliver intelligent file management beyond traditional cleaning methods.

Keywords: Digital Hoarding, NLP, Machine Learning, File Summarization, Behavioral Analytics, Semantic Analysis, Cloud Integration, Data Privacy.

I. INTRODUCTION

With the proliferation of smartphones, cloud storage[1], and collaborative work platforms, individuals often store files across multiple devices and services, making it harder to manage and organize them efficiently. The problem compounds when files with similar names, formats, or vague titles are saved repeatedly, making manual filtering tedious and error-prone. Users often encounter scenarios where they search endlessly for a specific document, only to find multiple irrelevant or outdated versions. This disorganization impacts personal productivity and increases the cognitive burden of digital life management.

Another dimension of digital hoarding is its psychological aspect[2]. Research in digital behaviour shows that users often form emotional attachments to digital files, such as photos, emails, or work documents. This attachment makes the decision to delete or archive files more difficult, even when they no longer hold practical value. Furthermore, the fear of losing potentially important files due to accidental deletion adds to the reluctance to engage in regular cleanup

activities. As a result, users end up holding onto files "just in case," leading to further accumulation.

Businesses and organizations are not exempt from this issue. In enterprise settings, unmanaged digital clutter can result in wasted storage costs, reduced operational efficiency, and security vulnerabilities due to redundant or outdated data being improperly stored. Compliance regulations like GDPR and data retention laws also necessitate that files be managed more rigorously. Traditional IT practices are insufficient to cope with the scale and complexity of modern data storage, calling for more intelligent solutions that go beyond simple rule-based systems.

AI and data science technologies offer a promising path forward. By analyzing the actual content and context of files, AI models can determine their relevance, importance, and usage patterns far more effectively than conventional tools. Natural Language Processing enables machines to understand and summarize documents, while machine learning can detect behavioral trends in file access or modification. Combining these capabilities allows for a dynamic system that adapts to user needs and preferences over time[3].

This paper aims to fill the gap left by existing storage cleaners by proposing an intelligent system that not only categorizes and summarizes content but also learns from user behavior to make context-aware recommendations. By integrating these advanced techniques into a user-friendly interface with cloud support, the system empowers users to regain control of their digital environments. In doing so, it fosters better digital hygiene, reduces stress associated with disorganization, and contributes to more sustainable digital practices.

II. RELATED WORKS

- **Digital Hoarding:** The Rising Environmental and Personal Costs of Information Overload – Dr. A. Shaji George:

It explores how digital hoarding affects both personal productivity and environmental sustainability. It links data accumulation to stress and the carbon footprint of data centers. The author advocates for AI tools that guide users in sustainable file management. It emphasizes the need for policy interventions and user awareness. The work highlights digital hoarding as a broader social and ecological issue.

- User Behavior Analysis and Prediction Model Construction in Higher Education Management Information Systems – Shanshan Yu, Lihua Zhong, Yuwang Liu :

Though focused on education systems, the paper demonstrates how user behavior modeling can predict system interactions. It applies machine learning to analyze data usage and trends. The methods support adaptive interfaces and intelligent recommendations. Its techniques can inform behavior-based digital file relevance scoring. The paper supports integrating predictive analytics into file management systems.

- **DeclutterAR:** Mobile Diminished Reality and Augmented Reality to Address Hoarding by Motivating Decluttering and Selling on Online Marketplace - Samantha W.T Chan, Bektur Ryskeldiev, Suranga Nayanakkara:

DeclutterAR[4] uses AR and gamification to encourage physical and digital decluttering. It visualizes clutter and integrates with online selling platforms for motivation. While focused on physical items, its user engagement strategies are applicable to digital tools. It emphasizes interactive design to change hoarding behaviors. The work inspires UI/UX features for digital decluttering platforms.

- **The Psychology of Digital Hoarding:** How Unlimited Cloud Storage is Changing Consumer Behaviour – Naga Vaishak S K:

It explores how unlimited cloud storage influences hoarding behaviors. It reveals that users keep unnecessary files due to loss aversion and emotional attachment. The study recommends behaviorally-informed file management systems. It supports using summaries and nudges to encourage cleanup. The work emphasizes the need for psychological understanding in AI systems[6].

Digital Hoarding and Personal Use Digital Data – Elizabeth Sillence, Jordan A. Dawson, Richard D. Brown, Kerry McKellar & Nick Neave :
The study explores user motivations for retaining digital files, including sentiment and uncertainty. It classifies users into types such as sentimental or pragmatic hoarders. These insights help personalize AI file cleaning tools. The paper recommends reflective prompts and behavior-aware suggestions. It's valuable for designing emotionally intelligent file assistants[7].

- **Modern-day hoarding:** A model for understanding and measuring digital hoarding – Darshana Sedera , Sachithra Lokuge , Varun Grover c:

It presents a model with measurable dimensions of digital hoarding like volume, attachment, and redundancy. It introduces a digital hoarding index and categorizes user types. The model helps train adaptive AI systems for decluttering. It highlights mental health links and advocates for diagnostic alerts. The framework is useful for behavior-aware file management[8].

- Automatic Assessment of Hoarding Clutter from Images Using Convolutional Neural Networks – M. Ozan Tezcan, Janusz Konrad, Jordana Muroff:

Using CNNs, this study assesses physical clutter levels through images. While focused on physical spaces, its techniques apply to digital file organization. It proves AI can evaluate abstract clutter concepts. Its scoring system can inspire digital clutter metrics. The research supports visual analytics in file management tools.

- **Time for De-cluttering:** Digital Clutter Scaling for Individuals and Enterprises – Naciye Güliz Uğur, Kübra Çalışkan Akbıyık:

It introduces a digital clutter index for individuals and organizations. It uses metrics like file age, duplication, and usage to assess clutter. The model supports automated, benchmark-based cleanup recommendations. It advocates dashboards and visual feedback. The approach aids in tracking storage hygiene and cleanup progress.

- Please Delete That! Why Should I? – Ute Schmid, Christoph Niessen, and Ute Schmid:

It explores user reluctance to delete files and proposes explanation-based deletion prompts. It uses usage history to justify AI recommendations. Transparent suggestions increase user compliance. The work emphasizes explainable AI for trust in file management tools. It validates the importance of interpretable AI in decluttering systems[5].

III. PROPOSED SYSTEM ARCHITECTURE

The proposed architecture for the AI-Powered Digital Hoarding Cleaner is designed to intelligently analyze digital files, understand user interaction patterns, and recommend actions with justification. It integrates several key components that work collaboratively to automate and personalize the decluttering process.

- **File Ingestion Module:** This component acts as the entry point for the system, scanning both local directories and cloud platforms such as Google Drive and OneDrive. It collects metadata (e.g., file size, format, creation/modification dates) and extracts the actual file content for further semantic analysis. Its goal is to unify access across platforms while maintaining data integrity and speed.

- **Content Analysis (NLP Engine):** Once files are ingested, this module uses sophisticated natural language models such as BERT, GPT, and BART to process textual content. It performs tasks such as summarization, keyword extraction, and named entity recognition. This helps the system understand the document's subject matter and importance, forming the basis for relevance scoring and recommendation[9].

- **Categorization Module:** Based on the semantic content and metadata, files are automatically grouped into logical categories like personal, work, entertainment, financial, or archival. ML classifiers are trained to perform this task, and unsupervised clustering algorithms are used to handle unfamiliar file

types. This structure improves file discoverability and organization.

- **Behavioral Analytics Engine:** This engine monitors how frequently and recently a file has been accessed or modified. It also considers user-specific interaction patterns to evaluate a file's ongoing relevance. It calculates a dynamic relevance score that evolves with user behavior, supporting personalized decluttering recommendations.

- **Recommendation Engine:** Leveraging inputs from both the content and behavioral modules, this engine suggests whether a file should be retained, deleted, or archived. It uses a hybrid scoring algorithm that weighs semantic importance and usage data. Each recommendation includes a justification based on summaries and access history to enhance user confidence and promote clarity.

- **User Interface (UI):** The interface offers an intuitive and visually engaging dashboard displaying categorized files, relevance scores, and actionable suggestions. Users can preview summaries, confirm or override AI suggestions, and customize rules for automation. The UI is developed to support both web and mobile platforms for broader accessibility.

- **Cloud Integration Layer:** To support seamless storage management across services, this component provides secure synchronization with cloud platforms like Google Drive and OneDrive. It ensures changes made within the system reflect across all user platforms, while also handling version control and conflict resolution to prevent data loss.

IV. FUTURE WORK

As the project transitions into the implementation phase, one of the key focus areas will be the creation and annotation of a rich and diverse dataset. This dataset will include various file formats, content types, and user behavior patterns. High-quality data is essential to effectively train the NLP and ML models that will drive intelligent content analysis and categorization. Special attention will be paid to ensuring data diversity, which allows the system to generalize well across different user contexts and digital ecosystems.

Another major development goal is the integration of explainable AI techniques. These will help make the system's recommendations transparent and understandable to users. By providing clear justifications for file suggestions (retain, delete, archive), user trust and adoption are expected to improve significantly[11]. Methods like attention visualization, summary previews, and relevance scoring explanations will play a central role in ensuring that users remain in control and informed[10].

Additionally, the system's functionality will be expanded to support multi-language documents. As users store content in different languages, enabling NLP pipelines for multilingual processing becomes crucial. This will involve leveraging or fine-tuning existing language models and evaluating their accuracy across linguistic contexts. Moreover, platform compatibility will be prioritized, with versions developed for desktop and mobile environments to ensure a seamless cross-platform user experience[12-13].

Lastly, performance evaluation will be conducted using standard metrics such as precision, recall, F1-score, and user satisfaction ratings. These metrics will be used to iteratively fine-tune the system and identify areas of improvement. Real-world testing and feedback collection will form a core part of the validation process. The insights gained will be crucial in refining the tool to meet user expectations while maintaining its efficiency and accuracy in managing digital hoarding.

V. CONCLUSION

Digital hoarding has emerged as a significant challenge in today's data-driven world, affecting both individual users and organizations. The ease of generating and storing digital information has led to uncontrolled data accumulation, resulting in cluttered systems and inefficient workflows. Existing

tools designed to manage this issue are primarily reactive and lack the intelligence to evaluate file relevance based on content and usage patterns. This often leads to either overly aggressive deletion or passive data buildup. As a result, users experience cognitive overload, difficulty in finding important files, and unnecessary storage costs.

The proposed AI-Powered Digital Hoarding Cleaner offers a forward-thinking solution by incorporating natural language processing and machine learning to enhance digital storage management[14-15]. Unlike conventional storage tools, this system takes into account semantic content, file context, and behavioral data to deliver tailored recommendations. It helps users better understand their digital assets through summarization and classification, promoting informed decisions on file retention or deletion. Furthermore, its cloud integration and cross-platform design ensure that digital hygiene is maintained seamlessly across devices and services.

The project also introduces transparency into automated file management through explainable AI, where every recommendation is backed by justification. This is essential to build user trust and foster wider adoption, especially among users who are wary of automated systems. By providing previews, relevance scores, and user feedback loops, the system empowers individuals to remain in control while receiving smart assistance. The incorporation of multilingual support and mobile compatibility broadens its applicability, making it suitable for diverse environments and user demographics.

Looking ahead, this system lays a strong foundation for the development of smarter, more adaptive digital organization tools[16-17]. With continuous improvement through real-world feedback and performance evaluations, the platform has the potential to significantly reduce digital clutter. By addressing both the technical and psychological dimensions of digital hoarding, it can help users reclaim control over their data, enhance productivity, and contribute to more sustainable digital practices. The AI-powered Digital Hoarding Cleaner thus stands as a promising innovation in the field of intelligent data management.

REFERENCES

1. T. Grosse-Puppendahl, K. Kunze, A. Schmidt, "Green Cloud," CHI Extended Abstracts, 2024.
2. J. Smith, R. Zhang, "Digital Hoarding: The Rising Environmental and Personal Costs," PMRJ, 2024.
3. X. Li, M. Chen, W. Huang, "An Exploratory Study of Digital Hoarding," LIS, 2023.
4. S. Kumar, P. Patel, "DeclutterAR," IEEE Conference Proceedings, 2023.
5. A. Gupta, S. Rao, "Literature Survey on Automated Storage Cleaner," JAPC, 2023.
6. N. Vaishak S K, "The Psychology of Digital Hoarding," SSRN, 2023.
7. E. Sillence, J.A. Dawson, R.D. Brown, K. McKellar, N. Neave, "Digital Hoarding and Personal Use Digital Data," Behaviour & Information Technology, 2023.
8. D. Sedera, S. Lokuge, V. Grover, "Modern-Day Hoarding: A Model for Understanding and Measuring Digital Hoarding," Information & Management, 2022.
9. M.O. Tezcan, J. Konrad, J. Muroff, "Automatic Assessment of Hoarding Clutter Using CNNs," IEEE Access, 2022.
10. N.G. Uğur, K.Ç. Akbıyık, "Time for De-cluttering: Digital Clutter Scaling for Individuals and Enterprises," IJIS, 2023.
11. U. Schmid, C. Niessen, "Please Delete That! Why Should I?," Cognitive Assistance Workshop, 2022.
12. S. Yu, L. Zhong, Y. Liu, "User Behavior Analysis in Education Management Systems," ACM ICIET, 2023.
13. C.W.T. Chan, B. Ryskeldiev, S. Nayanakkara, "DeclutterAR," ACM UbiComp Adjunct, 2023.
14. T. Tan, Y. Kim, "AI-Personalized File Organizer," IEEE Transactions on Emerging Topics in Computing, 2022.
15. S. Sharma, M. Rao, "Context-Aware Document Cleaner Using NLP," IJCAI, 2022.

16. M. Hossain, A. Ahmed, "Data Detox: Redundant File Cleaner for Mobile Devices," IEEE Mobile Data Management Conference, 2021.
17. J. Johnson, L. Lee, "Email Triage Using Behavioral Analytics," ACM CHI Extended Abstracts, 2021.