Communications of the Association for Information Systems

Volume 53 Paper 8

10-10-2023

Improving Information Systems Sustainability by Applying Machine Learning to Detect and Reduce Data Waste

Bastin Tony Roy Savarimuthu
Department of Information Science Otago Business School

Jacqueline Corbett
FSA ULaval Université Laval

Muhammad Yasir
Department of Information Science Otago Business School

Vijaya Lakshmi
FSA ULaval Université Laval

Recommended Citation

Savarimuthu, B., Corbett, J., Yasir, M., & Lakshmi, V. (2023). Improving Information Systems Sustainability by Applying Machine Learning to Detect and Reduce Data Waste. Communications of the Association for Information Systems, 53, 189-213. https://doi.org/10.17705/1CAIS.05308

Improving Information Systems Sustainability by Applying Machine Learning to Detect and Reduce Data Waste

Cover Page Footnote

This manuscript underwent peer review. It was received 10/27/2022 and was with the authors for six months for two revisions. The Associate Editor chose to remain anonymous. This special section was reviewed during the tenure of editor-in-chief Fred Niederman.

Communications of the Association for Information Systems

Special Section DOI: 10.17705/1CAIS.05308 ISSN: 1529-3181

Improving Information Systems Sustainability by Applying Machine Learning to Detect and Reduce Data Waste

Bastin Tony Roy Savarimuthu
Department of Information Science
Otago Business School
Dunedin, New Zealand

Jacqueline Corbett
FSA ULaval
Université Laval
Quebec, Canada

Muhammad Yasir
Department of Information Science
Otago Business School
Dunedin, New Zealand

Vijaya Lakshmi
FSA ULaval
Université Laval
Quebec, Canada

Abstract:

Big data are key building blocks for creating information value. However, information systems are increasingly plagued with useless, waste data that can impede their effective use and threaten sustainability objectives. Using a constructive design science approach, this work first, defines digital data waste. Then, it develops an ensemble artifact comprising two components. The first component comprises 13 machine learning models for detecting data waste. Applying these to 35,576 online reviews in two domains reveals data waste of 1.9% for restaurant reviews compared to 35.8% for app reviews. Machine learning can accurately identify 83% to 99.8% of data waste; deep learning models are particularly promising, with accuracy ranging from 96.4% to 99.8%. The second component comprises a sustainability cost calculator to quantify the social, economic, and environmental benefits of reducing data waste. Eliminating 5948 useless reviews in the sample would result in saving 6.9 person hours, $2.93 in server, middleware and client costs, and 9.52 kg of carbon emissions. Extrapolating these results to reviews on the internet shows substantially greater savings. This work contributes to design knowledge relating to sustainable information systems by highlighting the new class of problem of data waste and by designing approaches for addressing this problem.

Keywords: Data Waste, Information Systems, Information Management, Sustainability, Machine Learning, Deep Learning, Reviews.

This manuscript underwent peer review. It was received 10/27/2022 and was with the authors for six months for two revisions. The Associate Editor chose to remain anonymous. This special section was reviewed during the tenure of editor-in-chief Fred Niederman.

1 Introduction

Every day, over 2.5 quintillion bytes of data are produced (Marr, 2018) by myriad automated and human processes (Jones, 2019). While big data are a key building block for information value creation (Koch et al., 2021), the rise of big data creates new technological, environmental, social, and intellectual challenges (Ekbia et al., 2015; Lucivero, 2020). This has led to greater awareness of data-driven problems, such as information overload (Bawden & Robinson, 2009; Wilson, 1995), misinformation (Lee et al., 2022), information proliferation (Hills, 2019), information management waste (Hicks, 2007), and data waste (Corbett et al., 2020a, 2020b).

The diversity of terms used to describe the problem of data that become "*unused, chucked, disregarded or forgotten*" (Gildersleeve, 2020, p. 135), in effect, data waste, provides evidence of the fragmented nature of research and practice. When data stores were small, the adverse effects of data waste were negligible, if not manageable. However, the explosion of digital data generation, especially on the internet, amplifies existing challenges and creates new ones. A critical question becomes how to efficiently manage and make use of the data that are available (Smith, 2020). Data waste creates noise within data stores and contributes to overload that can lead to lost time due to searching and hesitation in decision making (Fan et al., 2021). Additionally, organizations incur substantial costs to develop the required infrastructures to store, transmit, and process data. The growth of internet services and data has increased demand for sophisticated data centers that carry substantial environmental costs (Lucivero, 2020; Zhang & Yang, 2021) in terms of the extraction of rare and hazardous materials, consumption of large amounts of energy and water, and climate-changing carbon emissions (Siddik et al., 2021).

In industrial settings, the challenges of information waste (Bevilacqua et al., 2015) and digital waste (Romero, 2018) have been highlighted. We argue that digital data waste is an important new class of problem for which innovative solutions are required (Niederman & March, 2012; Wagner et al., 2021). As a start, various remedial strategies have been proposed for information waste based on the principles of lean management (Hicks, 2007). By treating data similarly to other physical inputs into a production process, managers can identify and eliminate the sources of data and information waste to reduce costs and improve efficiencies (Baglee & Marttonen-Arola, 2018; Hicks, 2007; Hölttä et al., 2010). Meanwhile, information-centric fields, including information systems (IS) have focused on extracting relevant information from data through processes and techniques, such as data mining (Qi et al., 2016), analytics (Martens & Maalej, 2019), and machine learning (Meyer et al., 2014), while internet research has also examined how characteristics of online content affect its helpfulness in decision making (Ghasemaghaei et al., 2018; Lee & Park, 2022).

Reducing data waste has the dual benefits of improving information value and sustainability. As Watson et al. (2012) argue, a sustainable information strategy must consider the inherent resource demands of increasing data stores throughout their lifecycle. However, limited research on the topic is dispersed across disciplines and concrete solutions are scarce. Accordingly, we adopt a design science research (DSR) approach, focusing on the creation of 'how to' knowledge (Gregor, 2023) and the development of an artifact that can be used to reduce the real-world problem (Nagle et al., 2022; Peffers et al., 2018) of data waste. The aim of this paper is to extend and enhance design knowledge and practice by designing and developing a generalized (and extensible) approach for detecting data waste and determining the sustainability costs of data waste that could be saved if such data waste were removed *at source*, that is when the data enters the information processing lifecycle.

Building a viable solution for this class of problems is a multi-step process, requiring the development of an ensemble artifact with two main components. An *ensemble* is the "coming together of elements forming a whole, a unified or interrelated group"[1]. An ensemble artifact is a collection of artifacts that holistically addresses a particular problem. In our case, the ensemble artifact addresses the data waste problem by detecting data waste and measuring its impacts. As a part of the multi-step process, first, a definition of the problem space is required. Traditionally, kernel theories from reference disciplines were considered fundamental to DSR (Walls et al., 1992), however, justificatory knowledge and practical theories-in-use (Jones & Gregor, 2007) have become accepted for informing design ideas (Iivari, 2020). In this work, we combine three main streams of theory and empirical study to inform our ideation and initial artifact design.

[1] https://en.wiktionary.org/wiki/ensemble

Second, a mechanism for identifying data waste is required. In the digital era, manual processes are not sufficient for identifying data waste because humans can be easily overwhelmed by the volume and velocity of data being created. To address this problem, machine learning (ML) approaches offer a promising avenue for detecting data waste. For instance, Tun and Tun (2019) propose content outlier mining as a mechanism for improving the efficiency and effectiveness of web searches, while Amrit, Wijnhoven, and Beckers (2015) suggest learning algorithms can be used to detect internet waste and improve the relevance of the information provided to users. Other scholars have tackled the issue of extracting useful information from app reviews (Gao et al., 2018; Malgaonkar et al., 2022) using ML approaches. Despite these examples, there is little direction for identifying data waste using both traditional ML and deep learning approaches. Thus, our first research question (RQ1) asks: ***how can traditional ML and deep learning approaches be used to identify data waste?***

Recognizing that data and related data waste come in many different forms, we focus on text data. Automated text classification is a challenging task because text data has high dimensionality: the data contain many features that need to be identified and extracted. This process requires strong domain knowledge to identify features and determine appropriate computational techniques to extract them. Thus, we chose online reviews as the research setting and examined the occurrence and costs of data waste within two popular domains: app reviews and restaurant reviews.

The third step involves measuring the sustainability impacts associated with the identified data waste. This is an often-overlooked step in green IS research (Gholami et al., 2016) but it is essential for sensitizing people – individuals and organizations – to the problem, setting targets for improvement, and assessing progress (Corbett, 2018; Lucivero, 2020). Thus, building upon the results of RQ1, the second question (RQ2) aims to quantify data waste and its impacts by asking: ***how does data waste and its associated sustainability costs differ between domains?***

This research contributes to IS research and the advancement of sustainable information practices (Chowdhury & Koya, 2017; Watson et al., 2012) by highlighting the presence of digital data waste and the potential for using ML approaches to detect it. The development of an ensemble artifact for detecting and measuring the sustainability impacts of data waste situates this research within the construction quadrant of design knowledge contributions (Maedche et al., 2021). While certain components of the solution (e.g., ML and deep learning techniques) are not in themselves new, we identify features and assemble them in an innovative way to address a new class of problem, consistent with an exaptation contribution type (Wagner et al., 2021). The development of 13 models, including four deep learning approaches, is novel and the results suggest they could offer a promising avenue for tackling an emerging social and business challenge. In addition, the proposed sustainability cost calculator allows for the quantification of social, economic, and environmental costs associated with data waste. In terms of practical impact, this research provides a proof of concept (Nagle et al., 2022), showing how data waste can be identified and measured, which can lead to more informed and sustainable decisions and data management.

2 Conceptual Background

When engaging in DSR, researchers can draw inspiration from kernel theories as well as justificatory knowledge to inform the development of an innovative solution (Iivari, 2020). Walls et al. (1992) propose that kernel theories for IS design research come from the natural or social sciences and mathematics and direct the design requirements. Jones and Gregor (2007) extend the idea of kernel theory to justificatory knowledge that includes the underlying knowledge or theory used as the basis and explanation for the design. Justificatory knowledge is not limited to formal theories but may also include practitioner-in-use theories or evidence-based justification (Iivari, 2020; Jones & Gregor, 2007). Taking this latter perspective, we combine three main knowledge sets to inform the design of the artifact: we draw upon lean information management and data waste perspectives for the conceptualization of data waste; we examine common practice to understand the sustainability costs of data waste; and, we survey the use of ML approaches for detecting data waste.

2.1 Conceptualizations of Data Waste

Two streams of research have developed around the definition of data waste. First, within the industrial manufacturing context, lean management practices emphasize the reduction of waste of all types, where waste is defined as any nonvalue-adding activity (Romero et al., 2018). Here, data waste is measured in terms of processing activities that do not add value. Information management waste arises from additional

actions or inactivity that occur when the information consumer does not have immediate access to appropriate, accurate, and up-to-date information (Hicks, 2007). Waste from data and information management can take a variety of forms, including excess data collection and storage, redundant or erroneous processing, and ineffective communications (Bevilacqua et al., 2015; Cottyn et al., 2008; Hicks, 2007). Romero (2018) further highlights the possibility of passive digital waste that results from missed opportunities to leverage existing data and active digital waste that results from poor management that fails to deliver the right amount of information at the right time to the right actor.

Second, there has been some limited work in the information systems and related disciplines (e.g., information science, computer science) on the topic of data waste (Amrit et al., 2015; Gildersleeve, 2020; Hasan & Burns, 2011, 2013; Wijnhoven et al., 2012). These works tend to view waste data from an object perspective (Gildersleeve, 2020), rather than a process perspective. Hasan and Burns (2013) broadly define waste data as any data element that has no utility for a user in a given context. Data waste, which includes unused, disregarded, unwanted, or forgotten data (Gildersleeve, 2020), arises due to the creation of data that is collected, managed, transmitted or stored for no tactical, operational or strategic reasons (Loshin, 2013). Waste can also include once-meaningful data that has served its purpose (Tun & Tun, 2019). Data waste may also be considered as low-quality data that is practically useless to the owner or was never useful in the first place (Hasan & Burns, 2011). From the archival perspective, Wijnhoven et al. (2012, p. 135), define information waste as "information which is unnecessary (e.g., redundant) and unusable (e.g., not understandable) and which are the consequences of human limitations of knowing which data are of no use and could thus be removed or stored on a non-direct access medium." A commonality among these definitions is their reliance on the idea of usefulness, which is subjective depending on the context and the user (Amrit et al., 2015). Moreover, data can often be repurposed in unanticipated ways (Gildersleeve, 2020), implying it may never be possible to categorize a data element as waste with complete certainty.

In this work, we integrate the two main perspectives described above, adopting an object view of data waste while recognizing that the retention of data waste within data stores and information systems can lead to information processing and management wastes and create barriers to information value creation and sustainability. Although it may be difficult to determine objectively which bits of data are waste, we contend there is a continuum along which data elements exist, from highly useful to completely useless (Corbett et al., 2020b). In other words, there are some data that, due to their incompleteness, poor quality, incomprehensibility, or other characteristics, have no likelihood of creating information value (i.e., being useful or helpful to any stakeholder (Fan et al., 2021; Lee & Park, 2022)). Thus, we define data waste as ***data that are not, or are no longer, fit for purpose and thus have no value-adding potential***. This data waste, if not eliminated, creates important organizational and societal costs.

2.2 Sustainability Costs of Data Waste

The United Nation's Sustainable Development Goals (SDGs) for 2030 highlight the need for sustainable information practices that take into account social, economic, and environmental considerations (Chowdhury & Koya, 2017). The process perspective of information management waste is useful for structuring the discussion on the sustainability costs of data waste, where the data lifecycle consists of four main activities: data capture and storage, processing, transmission, and consumption.

The main social costs of data waste arise from wasted cognitive effort and time as humans try to make sense of data. Data waste can contribute to information overload (Wilson, 1995), which "occurs when information received becomes a hindrance rather than a help" (Bawden & Robinson, 2009), and obstruct individual and organizational decision making (Fan et al., 2021; Romero, 2018). For example, individuals must spend time scrolling or searching through web pages or online reviews to find relevant information, while organizations must process or filter this data waste, which consumes additional time and resources. Thus, eliminating data waste would improve search efficiency and effectiveness, and reduce social costs.

From an economic perspective, big data come with big costs. Such costs principally take the form of capital infrastructure investments and operating costs of data centers and cloud services. Global data center infrastructure spending could reach US $350 billion by 2026 (Bicheno, 2022). These investments include physical facilities, primary and backup power systems, environmental controls, as well as servers, networks, and communications hardware. The increasing costs of data centers relate to the increasing volume of data collected, processed, and transmitted. For example, between 2010 and 2018, global data center storage capacity increased by a factor of 25 and the volume of file transfers increased more than six-fold (Masanet et al., 2020). Running a large data center can cost between $10 million and $25 million

per year (Stream, 2022). Considering that the typical transmission of 1GB of data over the Internet consumes 5 kWh of energy (Costenaro & Duer, 2012), inefficiencies caused by data waste create substantial economic costs at each stage of the data lifecycle.

Data waste can also create threats to environmental sustainability (Lykou et al., 2018). Environmental costs from big data waste include the demand for natural resources (rare materials, water, energy) in the construction and operation of physical installations and equipment (Siddik et al., 2021). Thus, data waste contributes to carbon emissions, natural resource extraction, waste production, and other harmful environmental impacts directly or indirectly attributable to data-driven infrastructures (Bietti & Vatanparast, 2019). By 2025, data centers could globally account for 20% of electricity consumption and 3.2% of carbon emissions. Given continued data growth by 2040, storing digital data could account for 14% of worldwide carbon emissions (Trueman, 2019). Recognition of the environmental threats caused by energy use in data centers has inspired significant work related to improving energy efficiency, resource utilization, and green data centers (Zhang & Yang, 2021), however, the issue of data waste has largely been ignored.

2.3 Machine Learning Approaches to Data Waste Reduction

Identification of data waste is a classification problem (Li et al., 2020) where the data is classified as waste or useful. To this end, various automated approaches have been proposed. Wijnhoven et al. (2014) develop a file waste indicator to classify a file as information waste or non-waste. Kim et al. (2020) develop an algorithm to reduce video data waste (downloaded-but-unwatchable video data) which could reduce waste by 10-70%. Tun and Tun (2019) apply web content outlier mining, with 94% precision, to enhance the quality and effectiveness of web searches by detecting irrelevant and redundant documents. Studies in the app reviews domain have focused on extracting new features requested by users and existing features requiring improvements (Chen et al., 2014; Gao et al., 2018). Prior works have developed classification routines to separate informative reviews containing helpful information (Fan et al., 2021) from non-informative reviews based on the frequency of words, readability, emotional expressions, vague descriptions, or unclear and irrelevant questions (Genc-Nayebi & Abran, 2017; Mudambi & Schuff, 2010). These works confirm that ML techniques are useful in extracting helpful information and accessing hidden knowledge within big data.

ML is an exploratory process where the accuracy and performance of models vary, based on the characteristics of variables and observations in a study (Austin et al., 2013). Two ML approaches can be used to classify text: traditional ML approaches and deep learning approaches. Common traditional ML algorithms used for building classification models are Logistic Regression, Naïve Bayes Classifier, Support Vector Machine (SVM), Decision Tree, Random Forest (RF), Neural networks (Ayodele, 2010), and boosting algorithms such as XGBoost (Chen & Guestrin, 2016) and AdaBoost (Freund & Schapire, 1997). Feature selection is an integral part of ML algorithm training as it increases algorithmic performance by reducing the dimensionality of text data, improves the accuracy of the model, and reduces overfitting. Features, the key to any ML approach, are measurable properties of the phenomenon being studied, which form independent variables of the models developed to predict the outcome variables. In the domain of online reviews, features such as review length, unigram, and ratings are important to predict the helpfulness of reviews (Kim et al., 2006). Review length, review age, richness, sentiment, and readability are also important features in predicting the helpfulness of online customer reviews (Akbarabadi & Hosseini, 2018). Upon building a traditional ML model using selected features, the model needs to be validated by subjecting it to a test data set. K-Fold cross-validation is a common approach employed for this purpose (Hastie et al., 2009).

While traditional ML algorithms demonstrate good performance, deep learning models powered by pre-trained language models recently have been shown to be more promising (Adoma et al., 2020; Minaee et al., 2021). BERT, RoBERTa, XL-Net and GPT-3, in particular, have shown improvements in accurately classifying textual information in various contexts (Devlin et al., 2018; Floridi & Chiriatti, 2020; Yang et al., 2019). However, these algorithms have not been tested in the context of detecting data waste. The above-mentioned algorithms are promising because they do not need features to be identified by the modeler (i.e., they are model free) and can compute optimal values for a large number of parameters. However, this means the model building is more time-consuming and resource-intensive, so this trade-off must be considered.

3 Methodology

A significant body of knowledge has developed around IS design research (Baskerville et al., 2018; Gregor, 2023; Jones & Gregor, 2007; Maedche et al., 2021; Nagle et al., 2022; Niederman & March, 2012; Wagner et al., 2021; Walls et al., 1992). Seven main elements of DSR have been identified: problem identification, presence of an artifact, adherence to a specified DSR process, iterative design, evaluation, practical impact, and knowledge contribution (Nagle et al. 2022). We integrated these seven elements into our methodological processes. In section 2.1, we provided a conceptualization of data waste as the target *problem* for the research. Our data collection and coding processes are described in section 3.1. Then, as we elaborate in section 3.2, we built an *ensemble artifact* using a constructive DSR approach appropriate for a problem-solving paradigm (Hevner & Chatterjee, 2010). With respect to *DSR process*, we followed the Design Science Research Methodology (DSRM) (Peffers et al., 2007). The steps involved the following: a) defining the objective of the solution (or the artifact), which is to quantify data waste, and develop ML models; and, b) evaluating the goodness of the models using various performance metrics, quantifying the amount of waste that can be identified, and determining economic, social and environmental costs. We also compared the amount of waste captured in two different domains (app reviews and restaurant reviews). These steps were operationalized as a part of answering the two research questions. The ensemble artifact is the result of a six-year, *iterative* process including empirical and peer *evaluation*. An artifact using a rules-based approach to detecting data waste was initially developed (Corbett et al., 2020a). Based on testing and peer feedback, we then developed a second artifact using ML, which showed improved effectiveness at identifying data waste as compared to the rules-based approach (Savarimuthu et al., 2020). The third iteration, which is the subject of this work, integrates deep learning approaches. It also extends the application of the ensemble artifact to a second domain to provide more generalized design knowledge on how to identify data waste as well as empirical insights into the scope of the data waste problem in different contexts. We used real, historical data sets to evaluate the effectiveness of the new artifact and report the results in section 4.1. Sustainability calculations showing the savings achieved by removing the data waste at source are presented in section 4.3. The *practical impact* and *knowledge contributions* of this work are discussed in section 5.

In the following subsections, we detail the steps involved in the construction and evaluation of the artifact and its two main components.

3.1 Data Collection and Coding

We collected online data from app reviews from the Android store and restaurant reviews available on Google Maps. Specific contexts were selected in both domains because a targeted context is important for the accurate detection of data waste. Such contextualization facilitates the creation of appropriate ML approaches without introducing too much complexity in the design.

The app reviews domain was chosen because it has attracted significant interest from researchers (Fan et al., 2021; Ghasemaghaei et al., 2018; Lee & Park, 2022). The Android store was chosen over other app stores because it hosts the largest number of apps. Within the broad domain of apps, we chose to focus on the entertainment category because it is one of the top five app categories (Statista, 2020). Moreover, apps in this category are diverse (e.g., games, social media) and have the potential to attract rich and varied sets of comments.

We used a custom-developed software program to extract reviews from the top 19 entertainment apps available in the New Zealand Android App Store between 1 December 2016 and 15 January 2017. New Zealand was chosen because the two authors who collected the data (including the first author) were from New Zealand and they were able to better understand the context mentioned in the reviews. Sample apps include Netflix, YouTube Kids, and Talking Ben the Dog. For each review, we captured the app name, rating, title of review, and description. A total of 21,921 unique reviews were extracted from multiple apps. The data set was then loaded into the R programming environment and was cleaned to remove the non-ASCII characters (e.g., emoticons). This corresponds to the Extract-Load-Transform (ELT) approach widely employed in data management (Singhal & Aggarwal, 2022). The cleaned data was stored in a spreadsheet. Then, a sample of 15,576 reviews (71% of all reviews in the dataset) was manually coded. The coding was split across four coders. Each coder read the review and specified whether it was data waste (no informational value) or potentially useful to either the app developer or user/potential users. We adopted a conservative approach: if there was any possibility of inferring value (i.e., information) from the review, we categorized it as useful. Of the sample, 5740 reviews were coded by at least two coders with

an average agreement of 98.53%. Given this high level of consensus, we felt confident that the categorization of the reviews by a sole coder was reliable.

The second domain, restaurant reviews, was chosen because of its interest to, and uptake by, individuals in large and small cities: restaurant reviews are often read by many hundreds even thousands of people to determine whether a restaurant can enhance their culinary experience (Sparks et al., 2003). The Google Maps platform was chosen because of its global popularity for both reviewers and review users. We collected reviews from restaurants in Auckland and Wellington, as the cities are amongst the largest in New Zealand, thus the restaurants are able to attract a large number of reviews, including those that may represent data waste. In a manner similar to app reviews, we created a software program to extract reviews. We collected the data between 1 December 2021 and 5 February 2022[2]. We only included reviews for restaurants with more than 500 reviews. We retrieved a total of 96,541 reviews, spanning restaurants with different cuisines (e.g., Chinese, Indian, Italian, and Kiwi). We cleaned the dataset by removing duplicates and empty reviews (i.e., reviews without any textual comments). From the resulting 48,712 reviews, we randomly selected 20,000 reviews to create a sample comparable to the size of the app review dataset. This sample comprised reviews from 94 restaurants (65 from Auckland and 29 from Wellington). A manual evaluation of 200 reviews (out of 48,712) showed that reviews having seven words or less had the potential to be data waste. Of the 20,000 reviews, 7226 met that criteria, suggesting they could be data waste. The remaining 12,774 reviews in the sample were deemed to be useful data. Two coders read and coded 10% (i.e., 723) of the 7226 reviews as either data waste or not waste; 95% consensus was achieved. Discussions were held between the two coders to resolve differences, and the remaining data (i.e., 6503 reviews) were coded by a single coder. Subsequently, another 650 reviews were coded by the second coder, this time with a consensus reaching 99%. Further, they discussed the 1% and reached full consensus.

Samples of data waste from both domains are shown in Table 1. These reviews are deemed data waste because they do not have value-adding potential (i.e., the potential to generate actionable insights for stakeholders). These reviews neither reveal what the user appreciates in a product (an app) or service (in a restaurant), nor what specific aspects they expect to change. Also, some of these reviews reveal information that is already known (i.e., whether the review is positive or negative) captured through the star rating the user provides (e.g., out of a scale of 1 to 5, from worst to best), and as such do not contain value-adding potential. All the comments in the two domains considered in this study can be found at this link: https://github.com/muhammad-yasir/data_waste_project/tree/main/Data.

Table 1. Examples of Data Waste in Reviews

App Reviews Domain	Restaurant Reviews Domain
“Dumb app” “Too bad, cannot do much with this” “Lovin it”	“Looks ok” “Not great” “It’s ok nothing special”

3.2 Model Development

In this work, the problem of identifying data waste is addressed by designing and developing 13 machine learning models that can detect data waste in online reviews. The effectiveness of these artifacts is evaluated to identify the best performing model. Below, we outline the details of traditional ML and deep learning models developed.

3.2.1 Traditional ML Models

These models were developed by first identifying features in the dataset, then by employing appropriate algorithms to fit the data. We built a set of nine models: Logistic Regression (Wright, 1995), Naïve Bayes (John & Langley, 1995), Decision Trees (Safavian & Landgrebe, 1991), Random Forest (Breiman, 2001), SVM (Vapnik, 1995), K Nearest Neighbor (KNN) (Dasarathy, 1991) and Artificial Neural Network (ANN) (Hagan et al., 1996), XGBoost (Chen & Guestrin, 2016) and AdaBoost (Freund & Schapire, 1997). These models were selected based on prior work in the arena of text classification (Abdel-Karim et al., 2021; Minaee et al., 2021).

[2] The second domain was also selected to demonstrate the generalizability of our work (i.e., to show the presence of data waste across domains, and across different time periods).

Feature selection: Based on the literature, we identified 17 textual features which were extracted from the reviews and used to develop the models. These features comprise five feature categories (see column 4 of Table 2). The first category, *review length,* covers both character count and word count in reviews. These were obtained using Python's built-in functions for counts. The second category is *Parts of Speech* (POS) *tags*. Prior work shows that nouns, verbs, adverb and adjectives signify key information within the text and the presence of these keywords suggest useful information is being provided (Keertipati et al., 2016). We used the NLTK library in Python to obtain these counts. The third category, *sentiments,* captures positive, negative, and neutral sentiments in reviews. We employed the SentimentIntensityAnalyzer library in Python for this purpose. The fourth category captures *emotions* expressed in reviews. We considered four specific emotions – joy, anger, fear, and sadness, based on prior work (Corbett & Savarimuthu, 2022), which were identified using IBM Watson's Tone Analyser. The last category captures features related to *language style,* capturing the extent to which a review contains analytical, confident, and tentative keywords. These were obtained using IBM Watson's Tone Analyser.

Table 2. Features in the Traditional Models and their Description

Feature number	Feature name	Feature description	Feature Category and (references employing the feature)
1	Review length	Character count in a review	Review length (Mudambi & Schuff, 2010)
2	Word count	Word count in a review	Review length (Mudambi & Schuff, 2010)
3	Proper noun count	Counts of proper nouns	POS tags (Keertipati et al., 2016)
4	Common noun count	Counts of common nouns	POS tags (Keertipati et al., 2016)
5	Verb count	Counts of verbs	POS tags (Dalpiaz & Parente, 2019)
6	Adverb count	Counts of adverbs in a review	POS tags (Kurtanovic & Maalej, 2017)
7	Adjectives count	Counts of adjectives in a review	POS tags (Dalpiaz & Parente, 2019)
8	Positive sentiment count	Score for positive sentiment	Sentiment (Hutto & Gilbert, 2014)
9	Negative sentiment count	Score for negative sentiment	Sentiment (Hutto & Gilbert, 2014)
10	Neutral sentiment count	Score for negative sentiment	Sentiment (Hutto & Gilbert, 2014)
11	Joy count	Counts of words indicating joy	Emotion (Corbett & Savarimuthu, 2022; Ren & Hong, 2019)
12	Anger count	Counts of words indicating anger	Emotion (Corbett & Savarimuthu, 2022; Ren & Hong, 2019)
13	Fear count	Counts of words indicating fear	Emotion (Corbett & Savarimuthu, 2022; Ren & Hong, 2019)
14	Sadness count	Counts of words indicating Sadness	Emotion (Corbett & Savarimuthu, 2022; Ren & Hong, 2019)
15	'Analytical' word count	Score for analytics words based on IBM Tone Analyser	Language style (Al Marouf et al., 2019)
16	'Confident' word count	Score for confident words based on IBM Tone Analyser	Language style (Al Marouf et al., 2019)
17	'Tentative' word count	Score for tentative words based on IBM Tone Analyser	Language style (Al Marouf et al., 2019)

After extracting features from the text, the next step was to select features having the most impact on the outcome variable. Feature selection involves reducing the number of input variables when developing a model by selecting the most prominent features that contribute to the prediction results. Apart from allowing for better predictability, feature selection helps in reducing the risk of over fitting a model and also reduces computational effort of the model to predict the dependent variable (Shilaskar & Ghatol, 2013). Two well-known techniques for feature selection are forward selection and backward elimination. In forward selection, variables are added progressively into a larger dataset and a model is built and tested incorporating each added variable. The set of variables that produce the best prediction results is chosen as the feature set for model development. Alternatively, backward elimination starts with all variables and proceeds with a stepwise deletion of the least promising variables. The elimination of variables is halted when no further improvement in the result is obtained (Guyon & Elisseeff, 2003). We employed the

forward selection technique because it is computationally more efficient than backward elimination. Through this process, fifteen features (out of seventeen) were selected for developing the ML models. The two removed features were proper nouns from POS and neutral sentiments from the sentiment category. All ML models are binary classifiers where the models classify whether a review is data waste or not.

Model construction and validation: Of the two datasets, the restaurant reviews dataset was imbalanced. So, we addressed this issue using the oversampling method (Haixiang et al., 2017). We employed K-Fold cross validation (Hastie et al., 2009), a common resampling approach used to evaluate the performance of the models developed. This approach judges how well the models perform on unseen data ("test data"). The models are fitted with the training data and the performance is evaluated based on the test data. In a K-Fold cross-validation, the full data set is divided into K equal subsets. Each time, one of the K subsets is used as the test data set, and the other K-1 subsets are put together to form a training set. This holdout method is repeated K times. The value of K in our experiment was 10 as 10-fold cross validation is the most common form of K-Fold validation (Berrar, 2019).

Comparison of ML models: The performances of the developed classifier models were compared using the following five metrics: accuracy, precision, recall, F1-measure, and Mathew's Correlation Coefficient (MCC). Accuracy is the most commonly used metric for assessing a classification model's performance (Tharwat, 2018). It is the ratio of the number of correct predictions (sum of the number of True Positives (TP) and True Negatives (TN)) to the total number of input samples (i.e.TP, TN, False Positives (FP), and False Negatives (FN)).

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Precision is the number of true positive results divided by the sum of true and false positives predicted by the classifier. Precision is expressed as follows:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Recall is the number of true positive results divided by the number of all relevant samples. Recall is expressed as follows:

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

F1-Score is the harmonic mean between precision and recall. This is a better metric than accuracy for datasets with class imbalance, where the number of data items belonging to each class are unequal (Lipton et al., 2014). F1-measure aims to find the balance between precision and recall. Thus, most research work uses both accuracy and F1-Score. F1-measure is formally expressed as:

$$F1 - Score = 2 * \frac{\mathrm{Precison} * \mathrm{Recall}}{\mathrm{Precison} + \mathrm{Recall}}$$

Mathew's Correlation Coefficient (MCC) quantifies the relationship between actual and predicted values. Researchers have recently argued that MCC is a better metric than accuracy and F1-Score since it produces a high score only if the prediction produces good results for all the four confusion matrix categories (TP, FP, TN, FN) and it does not inflate results, especially on imbalanced datasets unlike the other two metrics, accuracy and F1-Score (Chicco & Jurman, 2020). Mathematically, MCC is represented as:

$$MCC = \frac{\mathrm{TP\ x\ TN} - \mathrm{FP\ x\ FN}}{\sqrt{(\mathrm{TP\ x\ FP})(\mathrm{TP\ x\ FN})(\mathrm{TN\ x\ FP})(\mathrm{TN\ x\ FN})}}$$

The value of accuracy, precision, recall, and F1-Score lies between 0 and 1. Models with values closer to 1 for both accuracy and F1-Score metrics imply that the models fit the data better. MCC score ranges between -1 and +1. Scores closer to 1 imply a perfect prediction, 0 represents an average random prediction, and -1 implies an inverse prediction.

3.2.2 Deep Learning Models

We employed four transformer-based deep learning algorithms — BERT, XL-Net, RoBERTa and GPT-3 — to detect data waste. These transformer-based deep learning models were selected as they have shown the most promise in text classification tasks in other domains (Adoma et al., 2020; Minaee et al., 2021). These algorithms are model-free: the features do not have to be specified by the model developer, unlike the traditional ML models discussed above. BERT (Devlin et al., 2018) employs self-attention mechanism that pays specific attention to word tokens that contribute significantly to the desired outcomes (e.g., class labels in a classification problem). RoBERTa (Liu et al., 2019) improves on BERT by using more training data and by using dynamic masking instead of random masking employed by BERT. It is also better suited to dealing with longer sentences. XLNet (Yang et al., 2019) also improves on BERT by allowing all tokens to be predicted rather than BERT's approach to predicting the 15% masked tokens. Finally, GPT-3 (Floridi & Chiriatti, 2020) uses 175 billion parameters when compared to 110 million parameters for BERT. GPT-3 is trained on 499 billion words: as a result of the large training space, GPT-3 has the ability to employ few-shot learning. In other words, GPT-3 requires fewer examples from which to learn. GPT-3's disruptive and transformative potential is expected to revolutionize natural language processing tasks, such as answer generation, creative writing, translating text into different languages, and generating software code (Gruetzemacher, 2022; Gruetzemacher & Paradice, 2021).

We used the following implementations of the deep learning algorithms. For BERT, we used *bertbase-uncased* model, for XLnet we used *xlnet-base-cased* model, and for RoBERTa we used *roberta-base* model. These implementations are available from HuggingFace transformer library of Python. We fine-tuned those models on our training datasets. These three models and the nine traditional ML models in the previous section were run on a high-end desktop that had AMD Ryzen 9 5900x processor with 32 GB RAM, using Windows 11 OS. For GPT-3, we used OpenAI library in Python and the algorithm was run on a cloud server. Four epochs of training were conducted for the four deep learning algorithms. The two datasets from the two domains (with features) and the Python code for all 13 algorithms can be found at https://github.com/muhammad-yasir/data_waste_project.

After constructing the 13 models (9 traditional and 4 deep learning), we evaluated their performance by calculating five metrics (accuracy, precision, recall, f1-score, and MCC coefficient) as described in section 3.2.1.

3.2.3 Quantifying Data Waste Costs

The results of the best performing model were used to quantify the amount of data waste and its resulting social, economic, and environmental costs. The costs were derived from values found in the literature, as summarized in Table 2, and are split into server and client-middleware costs (columns) and the four different phases of the data management lifecycle, namely, storage, processing, transmission, and consumption (rows). The server costs are incurred for persistent storage of reviews and CPU usage cost is incurred when a review is processed at the server in response to a user request. The middleware used between client and server (e.g., routers) consumes power for various activities, such as temporary storage of queried data, processing of data packets, and transmission of data. *Economic cost* is measured as the sum of persistence cost, CPU usage cost and power used costs (shown in columns 2-4). *Social cost* is measured as the extra time taken for processing reviews that have no informational value and loading them on a browser, as well as additional transmission time, and increased consumption (reading) time for the user (shown in column 5). *Environmental costs* in the form of CO_2 emissions are computed based on the power used (1kWh = 0.429 kg of CO_2, https://carbonfund.org/calculation-methods/). In Table 3, the shaded cells represent costs assumed to be negligible, although this assumption is conservative because if data is stored in cloud storage there will be costs associated with data replication (i.e., additional storage cost, transmission cost and power cost).

Table 3. Unit Costs for Phases of Data Lifecycle

	Server costs		Middleware and client costs	
	Persistence cost	**CPU usage cost**	**Power used**	**Time taken (seconds)**
Storage	$276 per TB (Nhen, 2016)		1700 KWh per TB (Posani, 2018)	
Processing		$0.389 per CPU hour (Fusaro et al., 2011)	520 KWh per TB (Baliga et al., 2011)	2031.25 seconds per TB (Prakash et al., 2013)
Transmission			5000 KWh per TB (Costenaro & Duer, 2012)	750 seconds per TB (Xiong et al., 2012)
Consumption				2.8 seconds per review (Guzman et al., 2015)

4 Results

4.1 Effectiveness of ML Approaches for Detecting Data Waste

For construction-type DSR, researchers must evaluate the design artifact to demonstrate its appropriateness and capability to resolve the problem (Maedche et al., 2021). The effectiveness of the nine traditional ML models and the four deep learning models are presented in Tables 4 and 5 corresponding to app reviews and restaurant reviews respectively; the best results for each metric category have been bolded for traditional ML and deep learning models. Among traditional ML models, random forest outperforms the others with 93.2% accuracy in the app reviews domain and 99.6% accuracy in the restaurant reviews domain. The other eight models yield an accuracy between 82.8% and 90.8% for the app reviews domain and between 83.9% and 99.3% for restaurant reviews. Naïve Bayes shows the lowest accuracy in both domains. All the transformer-based deep learning models performed well, with accuracy at or greater than 96.4% for app reviews and 99.5% for restaurant reviews. RoBERTa and GPT-3 show the most promise in both domains.

Table 4. Results of Traditional ML and Deep Learning Models - App Reviews

Model type	ML Models	Accuracy	Precision	Recall	F1-Score	MCC
Traditional ML	Logistic regression	0.868	0.905	0.823	0.862	0.740
	Naïve Bayes	0.828	0.933	0.707	0.804	0.677
	Decision Tree	0.908	0.936	0.876	0.905	0.818
	Random Forest	**0.932**	**0.948**	**0.915**	**0.931**	**0.866**
	Support Vector Machine (SVM)	0.871	0.874	0.868	0.871	0.743
	K Nearest Neighbor (KNN)	0.875	0.881	0.867	0.874	0.751
	Artificial Neural Network (ANN)	0.892	0.890	0.895	0.892	0.785
	XGBoost	0.906	0.910	0.900	0.905	0.809
	AdaBoost	0.874	0.871	0.879	0.875	0.749
Deep Learning	BERT	0.964	0.968	0.961	0.964	0.929
	RoBERTa	**0.968**	0.970	0.967	**0.969**	**0.937**
	XLNet	0.964	0.960	**0.968**	0.964	0.928
	GPT-3	0.964	**0.971**	0.957	0.964	0.928

Table 5. Results of Traditional ML and Deep Learning Models - Restaurant Reviews

	ML Models	Accuracy	Precision	Recall	F1-Score	MCC
Traditional ML	Logistic regression	0.874	0.889	0.854	0.871	0.749
	Naïve Bayes	0.839	0.909	0.754	0.824	0.689
	Decision Tree	0.993	1	0.987	0.993	0.987
	Random Forest	**0.996**	**1**	**0.992**	**0.996**	**0.992**
	Support Vector Machine (SVM)	0.891	0.893	0.889	0.891	0.783
	K Nearest Neighbor (KNN)	0.985	1	0.970	0.985	0.970
	Artificial Neural Network (ANN)	0.989	0.999	0.979	0.989	0.978
	XGBoost	0.990	1	0.980	0.990	0.981
	AdaBoost	0.919	0.932	0.904	0.917	0.838
Deep Learning	BERT	0.995	1	0.989	0.992	0.98
	RoBERTa	0.997	1	0.995	0.997	0.995
	XL-Net	0.997	1	0.994	0.997	0.994
	GPT-3	**0.998**	**1**	**0.997**	**0.998**	**0.997**

Comparing the results in Tables 4 and 5, it can be observed that the average accuracy for traditional and deep learning algorithms across the two domains were 91% and 98% respectively. The distributions of accuracy scores in the two algorithmic groups differed significantly (Mann–Whitney U = 21, n1 = 18, n2 = 8, P < 0.05 two-tailed). The accuracy of deep learning models is on average 6.8% better than traditional ML algorithms. These results are in agreement with the generalized conclusion of previous studies (conducted in different domains) that have shown deep learning algorithms, such as BERT, RoBERTa and XL-Net, perform better than traditional ML algorithms in classification tasks (Kamath et al., 2018; Minaee et al., 2021). GPT-3 implementations have only recently been available for researchers and our results show superior performance of the algorithm when compared to traditional ML models. Results using GPT-3 have also started to emerge in other domains showing better performance than other deep learning algorithms such as BERT (Liu et al., 2021).

We also compared the effectiveness of the ML models between the two domains. The average accuracy scores for app reviews and restaurant reviews domains were 91% and 96% respectively. The distributions of accuracy scores in the domains differed significantly (Mann–Whitney U = 34.5, n1 = n2 = 13, P < 0.05 two-tailed). These results suggest that ML approaches are slightly better (across all models) in detecting data waste in restaurant reviews than app reviews. The lower performance in the app review domain can be attributed to the fact that each app is designed for a specific purpose and audience containing specific functionalities. Hence, the nature of reviews (e.g., vocabulary used) will be different. The model needs to learn these nuances across different apps, which makes it somewhat difficult to create a generalized model. On the other hand, there is more uniformity amongst restaurant reviews independent of cuisines (e.g., use of terms such as food, service, quality, and cleanliness), enabling algorithms to generalize better.

In summary, both traditional ML and deep learning models demonstrate high effectiveness at identifying data waste across the two domains, with the average accuracy of the two types of models being 91.3% and 98.1% respectively. Transformer-based deep learning models show near-perfect accuracy. With this high-level of accuracy, these models can be deployed effectively to identify data waste in online reviews. These results provide a proof-of-concept, validating the proposed approach for identifying data waste.

4.2 Savings from Reducing Data Waste

Table 6 summarizes the results with respect to data waste and sustainability savings that could be achieved from reducing data waste at the source in the two domains. Detailed calculations are shown in Tables 7 and 8, and detailed computations for costs can be viewed in the spreadsheets at https://github.com/muhammad-yasir/data_waste_project/tree/main/cost_calculator.

Table 6. Comparison of Estimated Savings Across Two Domains

	App reviews domain	Restaurant reviews domain
Percentage of data waste (number of reviews)	35.8% (5576 out of 15576)	1.9% (372 out of 20000)
Maximum length of the review (in characters)	350	4000
Social costs: Time taken (seconds)	20055	4742
Economic costs: Server, middleware, client costs (in US$)	$1.83	$1.10
Environmental costs: carbon-equivalent emissions (kg)	5.95	3.57

In the app review domain, 5567 reviews were identified as data waste (35.8%) by the best performing model, RoBERTa. By eliminating these reviews before they enter the information processing lifecycle, 20,055 seconds of user time can be saved (i.e., by preventing users from reading useless data) and $1.83 in total cost can be saved (in USD) by reducing storage and processing costs at the server, middleware and the client. In addition, 13.86 kWh less energy would be consumed, resulting in a savings of 5.95 kg of carbon emissions.

In the restaurant review domain, the amount of data waste was substantially lower: the best performing model, GPT-3 categorized 372 of 20,000 reviews (1.9%) as data waste. Reviewers in this domain appear to genuinely appreciate their dining experience or make efforts to indicate areas for improvement instead of providing vague comments that do not provide actionable information. By reducing data waste in this domain, 4742 seconds of user time can be saved and $1.10 can be saved in server costs (server, middleware and client costs). In addition, power usage could be reduced by 8.32 kWh, resulting in 3.57 kg of carbon emission reductions. Although the percentage of data waste was lower in this domain, the cost savings are comparable to the app reviews domain because the maximum allowed length in reviews was 4000 characters for Google reviews compared with 350 characters for app reviews in the Android store. This means a more than 10-fold increase in storage and processing costs for restaurant reviews on Google. In both domains, it can be observed that data processing and transmission costs are much more than that of storage cost (see Tables 7 and 8).

Table 7. Estimated Savings from all Reviews Identified as Data Waste (Based on RoBERTa Model Results in the App Reviews Domain)

	Server costs		Middleware and client costs			Carbon emissions
	Persistence cost (US$)	CPU usage cost (US$)	Power used (US$ / KWh)		Time taken (seconds)	(Based on power use, in kg)
Storage	0.0007		0.0005	0.004		
Processing		0.001	0.1722	1.305	5.0999	0.0018
Transmission			1.6558	12.553	1.8830	0.5600
Consumption					20048	5.3854
Total	0.0007	0.001	1.8285	13.863	20054.98	5.9474
Total (with units)	$ 1.83			13.863 KWh	20054.98 seconds	5.9474 kg

Table 8. Estimated Savings from all Reviews Identified as Data Waste (Based on GPT-3 Model Results in the Restaurant Reviews Domain)

	Server costs		Middleware and client costs			Carbon emissions
	Persistence cost (US$)	CPU usage cost (US$)	Power used (US$ / KWh)		Time taken (seconds)	(Based on power use, in kg)
Storage	0.0004		0.0003	0.0025		
Processing		0.0006	0.1033	0.7836	3.0610	0.0011
Transmission			0.9938	7.5348	1.1302	0.3361
Consumption					4737.60	3.2324
Total	0.0004	0.0006	1.0985	8.3210	4741.79	3.5697
Total (with units)	$ 1.10			8.3210 KWh	4741.79 seconds	3.57 kg

5 Discussion

Data waste is emerging as a new class of problem associated with the exponential growth of big data stored in information systems. Although big data has the potential to yield deeper insights than traditional data stores, collecting, storing, transmitting, and processing useless data creates substantial costs. Although data waste can occur at multiple stages in the data life cycle, including data that are no longer fit for purpose (i.e., data that were at one time useful), this work focuses on reducing data waste *at source*. The artifact we developed can help to prevent useless data generated in the future rather than eliminating now-useless data that may have been generated in the past, which would have been stored for archival purposes. Other techniques and innovative solutions may be required for dealing with data waste at other stages in the data lifecycle.

We suggest the data waste problem must be tackled through the redesign of work practices and systems at the collective, rather than individual level (Wilson, 1995). A main challenge is the human limitation of knowing what data are useful (Wijnhoven et al., 2012). The speed at which digital data is being created makes it impossible to place human intervention at the center of data waste identification and reduction. Instead, automated methods are required. This research takes an important step toward more sustainable information systems and data management by providing a generalized approach for using ML techniques to detect waste data and a sustainability cost calculator for quantifying savings that can be obtained from eliminating that waste. In the following sub-sections, we discuss the research contributions, implications for practice, limitations, and avenues for future research.

5.1 Contributions and Implications for Research

In considering the contributions of this work, we refer to Baskerville et al. (2018), who note that the foci of DSR research can be on producing artifacts, theories, or both and that most research in DSR produces artifacts in the initial stages before contributing to theory development. Adopting a constructive approach (Jones & Gregor, 2007), we have given priority to the former. The contributions of this paper are aligned with two research objectives that a DSR paper may pursue: 1) the development of design artifacts and 2) demonstrating the impacts of the resultant artifacts (Baskerville et al., 2018).

Firstly, this study contributes to design knowledge through the development of an ensemble artifact comprised of two main components to effectively demonstrate how innovative IS solutions can be built to address data waste. We constructed and compared the performance of 13 different ML models, including nine traditional ML and four deep learning models. Consistent with exaptation knowledge contributions (Wagner et al., 2021), we applied the nine traditional ML models in a novel context. We identified 15 features that saliently contribute to obtaining good performance. These features can be used in building reliable models by future design research pursuing a similar goal. Further, none of the prior work in the app review domain, which comprised the part of the justificatory knowledge informing the work, combined the five feature categories considered in this work. Particularly, the use of tone with other categories is novel. This work also demonstrates that proper nouns and neutral sentiment features offer little value for building traditional models. Thus, this research contributes back to the initial corpus of knowledge by suggesting that researchers exercise caution when using these two features while building models at the app domain level.

While traditional ML approaches have been used in a variety of contexts to extract meaningful information, the application of deep learning techniques is relatively recent and represents a *second contribution*. Notably, the deep learning models performed better than traditional ML approaches, suggesting that they should be given preference in the design of data waste detection systems. While accuracy is a chief concern, our interest in sustainability led us to examine factors besides performance as suggested by Watson et al. (2012). While their performance is superior, deep learning algorithms take a longer time to create models and, as a result, consume more power during the training phase than the traditional algorithms (see Table A1 in Appendix for a comparison of model development times). Also, deep learning models have a higher memory footprint than traditional ML models (Fu & Menzies, 2017); hence these models should be run on servers with efficient, high-end hardware specifications. The trade-off between accuracy and computational cost of deep learning algorithms is a topic of debate (Fu & Menzies, 2017; Janiesch, et al., 2021). Still, based on our results in the text-based domain of online reviews, the costs associated with one-off model development and ongoing higher memory requirements should be more than offset by the savings that accrue over time by avoiding the storage, processing, and transmission costs associated with an improved ability to identify data waste. In the future, the burden of manual label

assignment of traditional ML models could be reduced by few-shot deep learning algorithms that need fewer examples from which to learn (e.g., GPT-3). Furthermore, the computational costs associated with deep learning approaches are likely to decrease with advancements in high-performance computing architectures, such as network compression and acceleration (Thompson et al., 2020), making deep learning models a viable solution.

The second component of the artifact, the sustainability cost calculator, is a *third contribution* of this research. In developing the calculator, we show how IS researchers can identify and measure the different types of costs that can be incurred during the data lifecycle if data waste is not removed. While IS research has a long-standing interest in economic costs and benefits of IS, less attention has been given to measuring the environmental and social costs (Gholami et al., 2016). Conceptualizing lost time due to data waste is particularly novel and to our knowledge has not been captured in previous research. Measurement of these costs is important because making meaningful changes in IS design and usage requires an understanding of the negative impacts inherent in these systems (Chen et al., 2008; Gholami et al., 2016). Additionally, by applying the models in two different online review domains, the research highlights that different data contexts can generate different amounts of waste. Specifically, we found the restaurant review domain was relatively waste-free compared to the app review domain. These results suggest some domains will benefit more from adopting waste reduction strategies and systems redesigns than others.

Beyond its contributions to design knowledge, this research extends the literature on IS and sustainable development – a *fourth contribution*. Data have been identified as a key lever for the realization of the SDGs (Hassani et al., 2021), but they come with their own inherent demands for resources (Watson et al., 2012). Through this research, we provide a quantification of certain sustainability costs associated with data waste. Admittedly, the sustainability savings that could be obtained from our sample is small due to limited sample size (15,576 app reviews and 20,000 restaurant reviews). However, more substantial benefits could be realized when the universe of reviews is considered. For example, we extrapolated our results for Google restaurant reviews to other well-known platforms (Google, Yelp, Facebook, TripAdvisor, Zomato, Foursquare, OpenTable, Zagat), and food delivery platforms that also have review functionality (Grubhub, DoorDash, UberEats, and Seamless). With a conservative estimate of one billion reviews available from these 12 platforms[3] and 1000 read reviews per app per year, 15.98 person years and $53,155 would be saved due to removing data waste. The environmental savings would be 180,499 kg of carbon emissions, equivalent to about 448,036 miles driven by an average passenger vehicle[4]. Also, our estimations show that the cost of developing and running the deep learning algorithm (XLNet which takes the most time) is $82.05 (see Table A2 in the Appendix). This cost is a very small fraction (0.15% of $53,155) of the savings that can be achieved by reducing data waste in the restaurant rating domain. Further, extrapolating the results from the restaurant reviews domain to include other popular online review segments such as products (e.g., Amazon), movies (e.g., IMDB), books, general businesses, and reviews posted on social networking websites (e.g., Facebook, Twitter, Instagram, Whatsapp, Snapchat) and the associated data replication costs in cloud-based data centers, the savings due to waste reduction is likely to be at least 10 times higher than the results reported above (assuming just 10 domains). The sensibilization of the IS community to these costs is an important step towards engaging in more impactful green IS research.

5.2 Implications for Practice

Beyond the contribution this work makes to research, it also has tangible implications for practice. The research presents the proof-of-concept for an artifact capable of detecting the presence of data waste in incoming data streams and quantifying their sustainability costs. Practitioners can use the models developed to identify data waste and quantify it in the app review and restaurant reviews domains. They can also extend these models and apply them in other text-based domains (e.g., movie reviews). Furthermore, cost savings from avoiding data waste can be quantified using the sustainability cost calculator to help identify where there is potential to reduce social, environmental or economic impacts. Specific values within the calculator (e.g., cost of time or energy use, carbon emission conversions) can be easily adjusted by practitioners to accommodate specific geographic or organizational contexts, thus making the tool flexible for use.

[3] TripAdvisor alone has 760 million reviews of five million restaurants around the world.
[4] See https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator

As our extrapolation results suggest (see Section 5.1), significant cost savings could be realized in a text-based domain with limited data waste (i.e., restaurant reviews). Still, the savings are likely to be manifold for resource-intensive domains that use videos and image data such as YouTube and Instagram that store large swathes of data. In addition, by highlighting the potential costs associated with data waste, we hope to inspire the industry and review platform providers to prioritize the quality of reviews over the quantity. Our findings can inform the development of an information system for detecting data waste *at source*. When a review is posted, the system can assess its potential information value and suggest improvements to the information content by providing domain-specific suggestions, such as including information about food quality, service, and portion sizes. These prompts can nudge users to improve the quality of posts, thus enhancing the sustainability of the review ecosystem by reducing social, economic, and environmental costs. Such systems are being considered in other domains to address other problems, such as reducing hate speech in online communities (Cheriyan et al., 2021).

5.3 Limitations and Future Research

One limitation of this study stems from the timeframe during which the data was gathered. The data set used in the entertainment app review domain was collected in 2016-2017 and the data from the restaurant review domain was obtained in 2021-2022. We acknowledge that the amount of data waste, particularly in the entertainment domain could be different now as compared to when we created our sample. Thus, there is an opportunity to undertake studies with more recent data and to replicate the analysis at given intervals over time to monitor data waste trends and progress (we hope) toward more sustainable platforms. That said, our main goal in this work is to demonstrate an approach for detecting and quantifying data waste, to demonstrate scalability across domains, and not to determine the exact percentage of data waste. We encourage researchers to pursue data waste studies spanning multiple domains with recent data.

A second limitation of this work is that we do not provide explanations for the differences in data waste between domains. This question was outside of the scope of our research. However, we encourage future research to delve into this question, using a variety of different approaches and practices that include ML. For instance, research could investigate how platform owners and operators encourage or discourage the creation of data waste. Alternatively, social scientists can continue to explore individual motivations and behaviors when it comes to submitting reviews and examine whether 'green' nudges that flag a review as data waste may help to reduce the problem. Engaging a ML-driven chatbot to inform and educate individuals on how to provide meaningful reviews could be a beneficial avenue for future research and development (Adamopoulou & Moussiades, 2020).

A third limitation is that, in the sustainability cost calculator, we considered linear growth of savings in the simple text-based domain where processing is a function of text length. Despite this limitation, researchers can use the proposed sustainability cost calculator to quantify data waste and investigate the nature of the data waste problem across different domains and platforms and then develop IS solutions to reduce this issue. Furthermore, we emphasize that reducing data waste in other domains, such as videos, may have non-linear savings growth (e.g., exponential) due to a reduction in complex processing (e.g., extracting different features such as objects, audio and text with different granularity levels such as edges, lines and shapes), and this is a fertile avenue for future investigation.

In the future, developing a general framework for detecting and reducing data waste across different data formats would be of immense value to the community. Such a framework is likely to involve five steps: 1) *feature compilation* (identifying features in different data formats such as audio, images and video that are salient in the classification), 2) *model building* (i.e., different models will need to be built based on data formats), 3) *classification* (i.e., identifying whether an image or video is data waste or not), 4) *explanation* (i.e., explaining to the user why an image/video is data waste), and 5) *recommendation* (i.e., suggesting a course of action to reduce data waste such as not posting or posting after modifying). Such a framework can also be used to handle individual articles such as new items, blogs and tutorials that contain multiple data/media types embedded in them, to identify specific objects containing data waste.

Finally, we suggest researchers consider the adoption of deep learning approaches for waste detection, given the promise of these approaches (accuracy of 96.4% to 99.8% across these two domains). The ability of these models to detect waste represents the *interior impact* of our designed artifacts (Baskerville et al., 2018). Researchers can extend these models by supplementing new textual features or improving the performance of the models by employing ensemble learning approaches (although this may increase the computational cost due to running multiple models in parallel). To this end, our research can inform

the adoption of a well-rounded perspective on the nature and utility of collected data, encompassing both the advantages of useful data and also the negative sustainability impacts of data waste.

6 Conclusion

This paper proposes approaches to detect and quantify useless data collected in the realm of reviews. Using data from the app review and restaurant review domains, it develops and compares the performance of 13 ML algorithms, which can detect up to 90% of data waste, with the deep learning approaches showing the most promise. It also quantifies the amount of social, economic, and environmental savings that can be obtained by eliminating identified data waste before it enters the information processing lifecycle. The approach can be extended and tested in data-heavy domains to enhance the informational value and sustainability of information systems.

Acknowledgments

Jacqueline Corbett acknowledges the two grants received from the Natural Sciences and Engineering Research Council of Canada (grants RGPGP-2013-328942 and RGPIN-2019-05599).

References

Abdel-Karim, B. M., Pfeuffer, N., & Hinz, O. (2021). Machine learning in information systems-a bibliographic review and open research issues. *Electronic Markets, 31*(3), 643-670.

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications, 2*, 100006.

Adoma, A. F., Henry, N.-M., & Chen, W. (2020). Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition [presentation]. In *17th International Computer Conference on Wavelet Active Media Technology and Information Processing* (ICCWAMTIP).

Akbarabadi, M., & Hosseini, M. (2018). Predicting the helpfulness of online customer reviews: The role of title features. *International Journal of Market Research*, 1-16.

Al Marouf, A., Hossain, R., Sarker, M. R. K. R., Pandey, B., & Siddiquee, S. M. T. (2019). Recognizing language and emotional tone from music lyrics using IBM Watson Tone Analyzer [presentation]. In *IEEE International Conference on Electrical, Computer and Communication Technologies* (ICECCT).

Amrit, C., Wijnhoven, F., & Beckers, D. (2015). Information waste on the world wide web and combating the clutter [presentation]. In *European Conference on Information Systems*.

Austin, P. C., Tu, J. V., Ho, J. E., Levy, D., & Lee, D. S. (2013). Using methods from the data-mining and machine-learning literature for disease classification of heart failure subtypes. *Journal of Clinical Epidemiology, 66*(4), 398-407.

Ayodele, T. O. (2010). Types of machine learning algorithms. *New Advances in Machine Learning*, 19-48.

Baglee, D., & Marttonen-Arola, S. (2018). *An approach to identifying waste in data management processes* [presentation]. In *14th International Conference on Data Science* (ICDATA 2018).

Baliga, J., Ayre, R. W., Hinton, K., & Tucker, R. S. (2011). Green cloud computing: Balancing energy in processing, storage and transport. *Proceedings of the IEEE, 99*(1), 149-167.

Baskerville, R., Baiyere, A., Gregor, S., Hevner, A., & Rossi, M. (2018). Design science research contributions: Finding a balance between artifact and theory. *Journal of the Association for Information Systems, 19*(5), 358-376.

Bawden, D., & Robinson, L. (2009). The dark side of information: Overload, anxiety and other paradoxes and pathologies. *Journal of Information Science, 35*(2), 180-191.

Berrar, D. (2019). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology, 1*, 542-545.

Bevilacqua, M., Ciarapica, F. E., & Paciarotti, C. (2015). Implementing lean information management: The case study of an automotive company. *Production Planning & Control, 26*(10), 753-768.

Bicheno, S. (2022). *Data centre spending to increase 61% by 2026 – Dell'Oro*. Retrieved from https://telecoms.com/513294/data-centre-spending-to-increase-61-by-2026-delloro/

Bietti, E., & Vatanparast, R. (2019). Data waste. *Harvard International Law Journal, 61*. Retrieved from https://ssrn.com/abstract=3584251

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32.

Chen, A. J., Boudreau, M. C., & Watson, R. T. (2008). Information systems and ecological sustainability. *Journal of systems and Information technology, 10*(3), 186-201.

Chen, N., Lin, J., Hoi, S. C., Xiao, X., & Zhang, B. (2014). AR-miner: Mining informative reviews for developers from mobile app marketplace [presentation]. In *36th International Conference on Software Engineering*.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system [presentation]. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD'16).

Cheriyan, J., Savarimuthu, B. T. R., & Cranefield, S. (2021). *Towards offensive language detection and reduction in four software engineering communities* [presentation]. In *Proceedings of Evaluation and Assessment in Software Engineering*.

Chicco, D., & Jurman, G. (2020). The advantages of Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification. *BMC Genomics, 21*(1), 1-13.

Chowdhury, G., & Koya, K. (2017). Information practices for sustainability: Role of iSchools in achieving the UN sustainable development goals (SDGs). *Journal of the Association for Information Science and Technology, 68*(9), 2128-2138.

Corbett, C. J. (2018). How sustainable is big data? *Production and Operations Management, 27*(9), 1685-1695.

Corbett, J., & Savarimuthu, B. T. R. (2022). From tweets to insights: A social media analysis of the emotion discourse of sustainable energy in the United States. *Energy Research & Social Science, 89*, 102515.

Corbett, J., Savarimuthu, B. T. R., & Lakshmi, V. (2020a). Separating treasure from trash: Quantifying data waste in App reviews [presentation]. In *Americas Conference on Information Systems* (AMCIS).

Corbett, J., Savarimuthu, B. T. R., & Lakshmi, V. (2020b). Separating treasure from trash: Quantifying data waste in app reviews. In *Americas Conference on Information Systems, Virtual Conference*.

Costenaro, D., & Duer, A. (2012). The megawatts behind your megabytes: Going from data-center to desktop [presentation]. In 2012 *ACEEE Summer Study on Energy Efficiency in Buildings*.

Cottyn, J., Stockman, K., & Van Landegnhem, H. (2008). The complementarity of lean thinking and the ISA 95 standard [presentation]. In *WBF 2008 European Conference*.

Dalpiaz, F., & Parente, M. (2019). *RE-SWOT: From user feedback to requirements via competitor analysis* [presentation]. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*.

Dasarathy, B. V. (1991). *Nerest Neighbor (NN) norms: NN pattern classification techniques*. IEEE Computer Society Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. Retrieved from https://arxiv.org/abs/1810.04805

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V. R., Tsou, A., Weingart, S., & Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology, 66*(8), 1523-1545.

Fan, W., Liu, Y., Li, H., Tuunainen, V. K., & Lin, Y. (2021). Quantifying the effects of online review content structures on hotel review helpfulness. *Internet Research, 32*(7).

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*(4), 681-694.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119-139.

Fu, W., & Menzies, T. (2017). Easy over hard: A case study on deep learning [presentation]. In *Proceedings of the 11th Joint Meeting on Foundations of Software Engineering*.

Fusaro, V. A., Patil, P., Gafni, E., Wall, D. P., & Tonellato, P. J. (2011). Biomedical cloud computing with Amazon Web Services. *PLoS Computational Biology, 7*(8), 1-6.

Gao, C., Zeng, J., Lyu, M. R., & King, I. (2018). Online app review analysis for identifying emerging issues [presentation]. In *Proceedings of the 40th International Conference on Software Engineering*.

Genc-Nayebi, N., & Abran, A. (2017). A systematic literature review: Opinion mining studies from mobile app store user reviews. *Journal of Systems and Software, 125*, 207-219.

Ghasemaghaei, M., Eslami, S. P., Deal, K., & Hassanein, K. (2018). Reviews' length and sentiment as correlates of online reviews' ratings. *Internet Research*, *28*(3), 544-563.

Gholami, R., Watson, R. T., Hasan, H., Molla, A., & Bjorn-Andersen, N. (2016). Information systems solutions for environmental sustainability: How can we do more? *Journal of the Association for Information Systems, 17*(8), 521-536.

Gildersleeve, R. E. (2020). The knowledge imperative in academic waste(lands). *Taboo: The Journal of Culture and Education, 19*(3).

Gregor, S. (2023). A design research journey. *Communications of the Association for Information Systems, 52*(1), 15.

Gruetzemacher, R. (2022). The power of natural language processing. Retrieved from https://hbr.org/2022/04/the-power-of-natural-language-processing

Gruetzemacher, R., & Paradice, D. (2021). Deep transfer learning & beyond: Transformer language models in information systems research. *ACM Computing Surveys (CSUR), 54*(10S), 1-35.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157-1182.

Guzman, E., Aly, O., & Bruegge, B. (2015). Retrieving diverse opinions from app reviews [presentation]. In *2015 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (ESEM).

Hagan, M. T., Demuth, H. B., & Beale, H. D. (1996). *Neural network design*. PWS Publishing.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications, 73*, 220-239.

Hasan, R., & Burns, R. (2011). *Where have you been? Secure location provenance for mobile devices*. arXiv. Retrieved from https://arxiv.org/abs/1107.1821

Hasan, R., & Burns, R. (2013). The life and death of unwanted bits: Towards proactive waste data management in digital ecosystems [presentation]. In *Third International Conference on Innovative Computing Technology* (INTECH 3013).

Hassani, H., Huang, X., MacFeely, S., & Entezarian, M. R. (2021). Big data and the united nations sustainable development goals (UN SDGs) at a glance. *Big Data and Cognitive Computing, 5*(3), 28.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Hevner, A., & Chatterjee, S. (2010). Design science research in information systems. In *Design Science Research in Information Systems* (pp. 9-22). Springer.

Hicks, B. J. (2007). Lean information management: Understanding and eliminating waste. *International Journal of Information Management, 27*, 233-249.

Hills, T. T. (2019). The dark side of information proliferation. *Perspectives on Psychological Science, 43*(3), 323-330.

Hölttä, V., Mahlamäki, K., Eisto, T., & Ström, M. (2010). Lean information management model for engineering changes. *World Academy of Science, Engineering and Technology, 4*, 1213-1220.

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text [presentation]. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Iivari, J. (2020). A critical look at theories in design science research. *Journal of the Association for Information Systems, 21*(3), 10.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets, 31*(3), 685-695.

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers [presentation]. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*.

Jones, D., & Gregor, S. (2007). The anatomy of a design theory. *Journal of the Association for Information Systems*, *8*(5).

Jones, M. (2019). What we talk about when we talk about (big) data. *The Journal of Strategic Information Systems, 28*(1), 3-16.

Kamath, C. N., Bukhari, S. S., & Dengel, A. (2018). Comparative study between traditional machine learning and deep learning approaches for text classification [presentation]. In *Proceedings of the ACM Symposium on Document Engineering 2018*.

Keertipati, S., Savarimuthu, B. T. R., & Licorish, S. A. (2016). Approaches for prioritizing feature improvements extracted from app reviews. In *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*.

Kim, S. M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). Automatically assessing review helpfulness [presentation]. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*.

Koch, H., Chipidza, W., & Kayworth, T. R. (2021). Realizing value from shadow analytics: A case study. *Journal of Strategic Information Systems, 30*(2).

Kurtanovic, Z., & Maalej, W. (2017). Mining user rationale from software reviews [presentation]. In *IEEE 25th International Requirements Engineering Conference* (RE).

Lee, J., Britt, B. C., & Kanthawala, S. (2022). Taking the lead in misinformation-related conversations in social media networks during a mass shooting crisis. *Internet Research*, *33*(6).

Lee, J., & Park, C. (2022). The effects of corporate, review and reviewer characteristics on the helpfulness of online reviews: The moderating role of culture. *Internet Research, 32*(5), 1562-1594.

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2020). *A survey on text classification: From shallow to deep learning*. arXiv. Retrieved from https://arxiv.org/abs/2008.00364

Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). Thresholding classifiers to maximize F1 score. *Machine Learning and Knowledge Discovery in Databases, 8725*, 225-239.

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & Tang, J. (2021). *GPT understands, too*. arXiv. Retrieved from https://arxiv.org/abs/2103.10385

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach*. arXiv. Retrieved from https://arxiv.org/abs/1907.11692

Loshin, D. (2013). *Big data analytics: From strategic planning to enterprise integration with tools, techniques, NoSQL, and graph*. Elsevier.

Lucivero, F. (2020). Big data, big waste? A reflection on the environmental sustainability of big data initiatives. *Science and Engineering Ethics, 26*(2), 1009-1030.

Lykou, G., Mentzelioti, D., & Gritzalis, D. (2018). A new methodology toward effectively assessing data center sustainability. *Computers & Security, 76*, 327-340.

Maedche, A., Gregor, S., & Parsons, J. (2021). Mapping design contributions in information systems research: The design research activity framework. *Communications of the Association for Information Systems, 49*(1), 12.

Malgaonkar, S., Licorish, S. A., & Savarimuthu, B. T. R. (2022). Prioritizing user concerns in app reviews–A study of requests for new features, enhancements and bug fixes. *Information and Software Technology, 144*, 106798.

Marr, B. (2018). *How much data do we create every day? The mind-blowing stats everyone should read*. Forbes.com. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=47b2b1e160ba

Martens, D., & Maalej, W. (2019). Towards understanding and detecting fake reviews in app stores. *Empirical Software Engineering, 24*(6).

Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. *Science, 367*(6481), 984-986.

Meyer, G., Adomavicius, G., Johnson, P. E., Elidrisi, M., Rush, W. A., Sperl-Hillen, J. M., & O'Connor, P. J. (2014). A machine learning approach to improving dynamic decision making. *Information Systems Research, 25*(2), 239-263.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning--based text classification: A comprehensive review. *ACM Computing Surveys (CSUR), 54*(3), 1-40.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly, 34*(1), 185-200.

Nagle, T., Doyle, C., Alhassan, I. M., & Sammon, D. (2022). The research method we need or deserve? A literature review of the design science research landscape. *Communications of the Association for Information Systems, 50*, 358-395.

Nhen, A. (2016). *AWS S3: Understanding cloud storage costs and how to save*. Retrieved from https://www.apptio.com/emerge/aws-s3-understanding-cloud-storage-costs-to-save/

Niederman, F., & March, S. T. (2012). Design science and the accumulation of knowledge in the information systems discipline. *ACM Transactions on Management Information Systems (TMIS), 3*(1), 1-15.

Peffers, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: Introduction to the special issue on exemplars and criteria for applicable design science research. *European Journal of Information Systems*, *27*(2), 129-139.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45-77.

Posani, L. (2018). *The environmental footprint of a distributed cloud storage*. arXiv. Retrieved from https://arxiv.org/abs/1803.06973

Prakash, V. S., Wen, Y., & Shi, W. (2013). Tape cloud: Scalable and cost efficient big data infrastructure for cloud computing. In *Proceedings of Sixth International Conference on Cloud Computing*.

Qi, J., Zhang, Z., Jeon, S., & Zhou, Y. (2016). Mining customer requirements from online reviews: A product improvement perspective. *Information & Management, 53*(8), 951-963.

Ren, G., & Hong, T. (2019). Examining the relationship between specific negative emotions and the perceived helpfulness of online reviews. *Information Processing & Management, 56*(4), 1425-1438.

Romero, D., Gaiardelli, P., Powell, D., Wuest, T., & Thürer, M. (2018). Digital lean cyber-physical production systems: The emergence of digital lean manufacturing and the significance of digital waste. In *Proceedings of the International Conference on Advances in Production Management Systems*.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifer methodology. *IEEE Transactions on Systems, Man, and Cybernetics, 21*(3), 660-674.

Savarimuthu, B. T. R., Corbett, J., Yasir, M., & Lakshmi, V. (2020). Using machine learning to improve the sustainability of the online review market. In *Proceedings of International Conference on Information Systems* (ICIS).

Shilaskar, S., & Ghatol, A. (2013). Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. *Expert Systems with Applications, 40*(10), 4146-4153.

Siddik, M. A. B., Shehabi, A., & Marston, L. (2021). The environmental footprint of data centers in the United States. *Environmental Research Letters, 16*, 1-11.

Singhal, B., & Aggarwal, A. (2022). ETL, ELT and reverse ETL: A business case study. In *The 2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering* (ICATIECE).

Smith, G. (2020). Data mining fool's gold. *Journal of Information Technology, 35*(3), 182-194.

Sparks, B., Bowen, J., & Klag, S. (2003). Restaurants and the tourist market. *International Journal of Contemporary Hospitality Management, 15*(1), 6-13.

Statista. (2020). *Most popular Google Play app categories as of 4th quarter 2019, by share of available apps*. Retrieved from https://www.statista.com/statistics/279286/google-play-android-app-categories/

Stream. (2022). *Data center cost*. Retrieved from https://www.streamdatacenters.com/glossary/data-center-cost/#

Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*, 1-13.

Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). *The computational limits of deep learning*. arXiv. Retrieved from https://arxiv.org/abs/2007.05558

Trueman, C. (2019). *Why data centres are the new frontier in the fight against climate change*. Retrieved from https://www.computerworld.com/article/3431148/why-data-centres-are-the-new-frontier-in-the-fight-against-climate-change.html

Tun, T., & Tun, K. M. M. (2019). Web content outlier framework for enhancing web search results through mathematical approaches. In *Proceedings of the14th International Conference on Intelligent Systems and Knowledge Engineering* (ISKE).

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag.

Wagner, G., Prester, J., & Schryen, G. (2021). Exploring the scientific impact of information systems design science research. *Communications of the Association for Information Systems, 48*(1), 37.

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an information system design theory for vigilant EIS. *Information Systems Research, 3*(1), 36-59.

Watson, R. T., Corbett, J., Boudreau, M.-C., & Webster, J. (2012). An information strategy for environmental sustainability. *Communications of the ACM, 55*(7).

Wijnhoven, F., Dietz, P., & Amrit, C. (2012). Information waste, the environment and human action: Concepts and research. In *Proceedings of the IFIP International Conference on Human Choice and Computers*.

Wilson, P. (1995). Unused relevant information in research and development. *Journal of the American Society for Information Science, 46*(1), 45-51.

Wright, R. E. (1995). Logistic regression. In G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 217-244). American Psychological Association.

Xiong, H., Zhang, X., Yao, D., Wu, X., & Wen, Y. (2012). Towards end-to-end secure content storage and delivery with public cloud. In *Proceedings of the Second ACM Conference on Data and Application Security and Privacy*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems, 32*.

Zhang, Q., & Yang, S. (2021). Evaluating the sustainability of big data centers using the analytic network process and fuzzy TOPSIS. *Environmental Science and Pollution Research, 28*(14), 17913-17927.

Appendix

Table A1. Time Taken to Train Algorithms

ML Model		Time taken to train algorithms	
		App Reviews	**Restaurant Reviews**
Traditional ML	Logistic regression	4.75 seconds	9.17 seconds
	Naïve Bayes	236 milli seconds	373 milli seconds
	Decision Tree	1.09 seconds	14.3 seconds
	Random Forest	19.1 seconds	14.1 seconds
	Support Vector Machine (SVM)	1 minute, 58 seconds	6 minutes, 27 seconds
	K Nearest Neighbor (KNN)	2.26 seconds	6.69 seconds
	Artificial Neural Network (ANN)	31 minutes, 1 second	22 minutes, 57 seconds
	XGBoost	42.9 seconds	1 minute, 17 seconds
	AdaBoost	8.3 seconds	11.4 seconds
Deep Learning	BERT	10 hours, 28 minutes	20 hours, 43 minutes
	RoBERTa	10 hours, 30 minutes	20 hours, 48 minutes
	XLNet	14 hours, 15 minutes	28 hours, 14 minutes
	GPT-3	53 minutes	59 minutes

Table A2. Training and Running Costs of Traditional vs. Deep Learning Algorithms

Model type	One-off training costs	Running cost for a year	Total cost per year
Traditional ML model (ANN)	$0.31	$7.32	$7.63
Deep Learning model (XLNet)	$8.85	$73.20	$82.05

About the Authors

Bastin Tony Roy Savarimuthu is Professor of Information Science at the University of Otago, in Dunedin, New Zealand. He received his PhD from the University of Otago. His research interests are in the fields of Software Engineering, Multi-Agent Systems (a branch of distributed Artificial Intelligence) and Information Systems. His work focuses on social aspects of computing including designing socially aware software and studying social aspects in software development such as social norms, personality types and decision-making. His works in the Information Systems domain focus on developing artefacts (e.g., models) and systems that enable sustainable practices. His papers appear in the top-conferences in the three fields his research including International Conference on Information Systems (ICIS), International Conference on Software Engineering (ICSE) and International Joint Conference on Artificial Intelligence (IJCAI). He has been a Senior Programme Committee (SPC) member for the top conferences in these three fields. He has been a program chair and general chair of several international conferences.

Jacqueline Corbett is Professor of Management Information Systems in the Faculty of Business Administration at Université Laval, in Quebec City, Canada. She received her Ph.D. at Queen's University. Jacqueline's research focuses on the design and use of information systems (IS) to support sustainable development, with a particular interest on the use of data and digital innovations in the private and public sectors. Her papers have been published in top IS and management journals, including *Journal of the Association for Information Systems, Information Systems Journal, Strategic Entrepreneurship Journal*, *Journal of Business Ethics*, the *International Journal of Information Management*, and *Communications of the Association for Information Systems*. A Distinguished Member–Cum Laude of the Association for Information Systems (AIS), Jacqueline has served in leadership roles for the AIS Women's Network College, the AIS Special Interest Group on Green IS (SIGGreen), and as Associate Editor for various IS conferences and journals.

Muhammad Yasir has a PhD in Information Science from the University of Otago, Dunedin, New Zealand, where his research focused on leveraging agent-based modeling techniques in the energy domain to enhance the three pillars of sustainability. Currently serving as a Senior Data Intelligence Analyst at Wellington City Council, Yasir combines his expertise in software engineering and information systems to drive data-driven insights and strategic decision-making. He has published 20 research papers in renowned journals and conferences. His work encompasses a range of disciplines, including agent-based modeling, simulation, software engineering, and information systems. Yasir continues to explore innovative approaches to address complex challenges and promote sustainable practices.

Vijaya Lakshmi is a PhD candidate in Management Information Systems at Université Laval. Her research interests include artificial intelligence (AI), sustainable agriculture, green IS and sustainability. Her research has been published in conference proceedings of major IS conferences such as International Conference on Information Systems, Americas Conference on Information Systems, Pacific Asia Conference on Information Systems and Hawaii International Conference on System Sciences. Her research on AI and sustainable agriculture received an honourable mention at the Administrative Sciences Association of Canada Conference 2021 and was also nominated for Best Paper at the Hawaii International Conference on Systems Sciences (HICSS) 2022.