Psychology & Marketing

WILEY

RESEARCH ARTICLE OPEN ACCESS

Natural Language Processing Algorithms to Improve Digital Marketing Data Quality and Its Ethical Implications

Sergi Pons[1] | Ruben Huertas-Garcia[1] | Jorge Lengler[2] | Daniel Luiz de Mattos Nascimento[1]

[1]Business Department, Universitat de Barcelona, Barcelona, Spain | [2]Durham University Business School, Durham University, Durham, UK

Correspondence: Jorge Lengler (jorge.lengler@durham.ac.uk)

Received: 29 November 2024 | **Revised:** 6 March 2025 | **Accepted:** 7 March 2025

Funding: Special thanks to Marc Vergés Santiago, a student of the double degree in Mathematics and Computer Engineering at the University of Barcelona, who has collaborated in the design and development of the algorithms presented in this paper.

Keywords: data quality | database cleaning | digital marketing | ethical implications | levenshtein algorithm | NLP algorithm | personalization | RapidFuzz

ABSTRACT

The ethical implications of personalization in digital marketing are significantly greater when companies adapt their marketing actions to individual consumer preferences. While this approach helps to reduce oversaturation and a sense of irrelevance among consumers, it also raises concerns about privacy and potential algorithmic bias. One form of personalization is self-referencing, where companies use the customer's name in all communications with that person. For this to be effective, customer data must be accurate and sourced from a high-quality database. This study presents a real case of data mining by a lead generation company, illustrating the sequential process of cleaning a database containing the names and surnames of 100,000 customers. In the final filtering step, we compared the performance of two Natural Language Processing (NLP) algorithms, Levenshtein and RapidFuzz, using ratio tests. The results demonstrate that the Levenshtein algorithm outperformed RapidFuzz, the former achieving a 93.43% clean data set compared to the latter's 92.93%. Finally, we discuss the ethical challenges posed by the privacy-personalization paradox, explore the theoretical and managerial implications, and propose future research directions that balance digital marketing interests with consumer privacy.

1 | Introduction

The widespread growth of digital marketing, and particularly via email, has led to a saturation of messages in the consumer's inbox, which makes them feel irrelevant (Zhang et al. 2017) and, moreover, goes against their natural desire to stand out from the crowd (Chandra et al. 2022). To preserve this sense of uniqueness, marketing practitioners have developed personalization strategies within the marketing mix (Surprenant and Solomon 1987). However, this strategy requires accurate personal information about the customer (name, address, email, telephone number) and their behavior (consumption habits, sources of information, topics of interest) to create better consumer experiences (Lim et al. 2022).

Unlike offline environments, the digital ecosystem, with its numerous data-generating sources, such as social media platforms, user-generated reviews, lead-generating activities, and online transaction records, offers ample possibilities for marketing personalization (Jürgensmeier and Skiera 2024). Due to its data-intensive nature, statistical analysis requires the use of machine learning (ML) algorithms, which generate extremely precise segmentation and forecasting results, enabling marketing decision-makers to create attractive and personalized

commercial offers, thus transforming the essence of customer engagement (Chandra et al. 2022; Sáez-Ortuño, Sanchez-Garcia, et al. 2023).

However, ML algorithms require accurate databases, which need be refined to eliminate errors and extraneous elements (Jürgensmeier and Skiera 2024). Data collected from the internet often contain many shortcomings, such as inconsistencies, missing information, typographical errors, invalid entries, and redundancies (Garcia et al. 2016; Rahm and Do 2000). This matters, because poor data quality not only affects the performance of ML algorithms but also impacts all multi-database integration processes (Rahm and Do 2000). Data filtering, or data cleaning, is a solution that involves detecting and removing errors and inconsistencies to improve quality (Garcia et al. 2016).

Email marketing, a specialized form of electronic marketing, consists of sending offers and commercial messages to potential customers via email. However, its success largely depends on access to accurate lists of names and email addresses (Hudák et al. 2017), which can pose a major challenge. A study of data provided by more than five million Spanish participants in internet-based contests and sweepstakes found errors in 5.87% of the records, 17.20% of which were "typos" and the remainder intentional (Sáez-Ortuño, Forgas-Coll, et al. 2023). Once data are captured, the natural step is to filter or clean it. However, since databases typically contain millions of records, manual cleaning is impractical, and automated solutions are hence required.

Although researchers and practitioners have made significant efforts to improve the performance of ML algorithms (e.g., by incorporating neural architectures), less attention has been paid to enhancing data quality, despite its critical role in generating accurate analyses and reliable recommendations (Jain et al. 2020). More effort should be dedicated to understanding both structured and unstructured data sets, designing metrics to detect potential errors, and developing operations that enable error correction to improve database quality and enhance the performance of ML models (Jain et al. 2020).

This study responds to this call by analyzing the quality of unstructured data, which are considered to be harder to correct than the structured variety (Collins et al. 2018). Specifically, we present a qualitative case study of a complex big data management experience within a lead generation company (Yin and Yin 1994), specifically the filtering of 100,000 data records in the form of a "string" (consumer names). This case is useful for highlighting the challenges involved in ensuring data quality. We also compare the performance of two ML algorithms, Levenshtein and RapidFuzz, in automating the final phase of the process by correcting misspelled words (Dumont et al. 2019).

Levenshtein is considered a classical ML algorithm, while RapidFuzz is more modern, but both pertain to the field of Natural Language Processing (NLP), a broad discipline aimed at enabling machines to understand, interpret, generate and respond to human language meaningfully, and which includes the identification and correction of strings (Chowdhury and Nath 2001).

In addition to proposing a filtering process for unstructured data, our research also investigates whether state-of-the-art algorithms consistently outperform classical algorithms or if performance depends on the type of data.

The rest of the article is organized as follows. The following section provides the conceptual background, introducing email marketing and the importance of message personalization. It also discusses the NLP and key string-matching algorithms used to analyze and compare first names. We then present a data mining case study that illustrates the process of cleaning name data and compares the performance of the Levenshtein and RapidFuzz ML algorithms. We conclude with reflections on the ethical implications of over-personalization and the use of NLP algorithms, which can often discriminate against consumers with less common first names or surnames, and offer avenues for future research to advance the fields of data mining and personalized marketing.

2 | Conceptual Backgrounds

Email marketing is one of the most widely used communication tools for both B2B and B2C interactions (Zhang et al. 2017) as it is such a cost-effective method for sending commercial offers to large volumes of potential customers and receiving responses (Chittenden and Rettie 2003). According to a 2020 global survey of more than 2000 marketers, the average return on investment (ROI) generated by email is $36 for every $1 spent, significantly higher than all other channels (Moller 2020).

Email is also considered an effective tool for fostering customer loyalty and encouraging repeat purchases (Kumar et al. 2014). Sahni et al. (2017), after conducting 70 randomized field experiments, found that email promotions not only have the immediate effect of increasing average spending after recipients receive the message, but that this effect can be sustained for up to a week. Additionally, customer familiarity and loyalty influence engagement with email marketing. For example, new recruits tend to be more active in opening emails and visiting promoters' websites than long-time recipients, who tend to respond less enthusiastically. However, this does not necessarily mean that their purchase rates decline (Zhang et al. 2017). In other words, email marketing generates short-, medium-, and long-term effects, and marketers must tailor their communication strategies to different stages of the customer journey.

However, one of the most complex challenges in email marketing is excessive mailing and inbox saturation, which leads consumers to take measures such as using firewalls to block access or unsubscribing (Zhang et al. 2017). A 2024 survey conducted by the GetApp agency found that 4 out of 10 US consumers (496 respondents) unsubscribed from at least one brand's emails each week due to the overwhelming volume of commercial messages (four or more per month from the same brand) (Jani 2024).

Previous research has shown that message oversaturation creates a sense of irrelevance among customers (Zhang et al. 2017). One way to mitigate this issue is the use of personalized content and tailored communication strategies (Jani 2024).

2.1 | Customer Personalization in Email Marketing

The *Cambridge Dictionary* defines personalization as 'the act of making something suitable for the needs of a particular person'. In other words, it is about enhancing the uniqueness of each consumer (Blom 2020). Email marketing employs targeted messaging using the consumer's name and an offer tailored to their needs (Murthi and Sarkar 2003), with the goal of enhancing the customer experience and building relational bonds that increase their value (Peppers and Rogers 1997).

However, the literature has not reached a consensus on the definition of personalization. Several meanings have been proposed (Table 1 provides a summary), including "individualization" (Riemer and Totz 2003), "segmentation" (Smith 1956), "one-to-one marketing" (Peppers and Rogers 1997) and "customization" (Davis 1987). For the purposes of this study, we focus on distinguishing between personalization and customization. Chandra et al. (2022) explain that personalization is a business initiative aimed at tailoring the marketing mix to each potential consumer, whereas customization is a business response to consumer demands—for example, when customers request variations on a standard offer, such as specific features for their laptop (Montgomery and Smith 2009). In other words, personalization is a proactive strategy designed to enhance customer relationships, whereas customization is reactive. According to Boudet et al. (2019), the effective implementation of personalization can lead to an increase in revenue by 5%–15% and an improvement in marketing efficiency of 10%–30% within a single channel.

So, how are personalization policies implemented in e-commerce marketing? Peppers and Rogers (1997) proposed a four-stage process: (1) identify the customer; (2) determine their needs; (3) establish communication; and (4) adapt the product to meet those needs. The most critical stages are the first two, when the precision achieved determines the effectiveness of communication and the degree to which the offer aligns with the customer's preferences. This study focuses on customer identification, which involves collecting and filtering information (Jürgensmeier and Skiera 2024). Personal data is typically gathered by website applications or lead generation companies—intermediary businesses that connect retailers with potential customers—while customer preferences are often collected from reviews, social media posts, and similar sources (Lim et al. 2022). However, because customer-provided data is not always reliable, it must be filtered before being passed on to retailers for message personalization (Sáez-Ortuño, Forgas-Coll, et al. 2023).

Regarding the use of personalization in online communication, Cavdar-Aksoy et al. (2021) identified three key approaches: (1) Self-reference, which entails addressing customers by name, whether in commercial offers, greetings, or congratulatory messages; (2) Anthropomorphism, which involves replicating human-like behaviors via tools such as chatbots, often using voices, gestures, and emotions; and (3) System features, referring to various mechanisms—such as artificial intelligence (AI) algorithms—that are employed to generate personalized recommendations (Cavdar-Aksoy et al. 2021). In our research, we focus on self-reference, and the need to ensure that messages correctly address consumers by name to enhance personalization.

2.2 | Ethical Considerations

The immense analytical power of digital marketing, particularly through the use of ML algorithms to process big data and generate highly precise consumer profiles, raises ethical concerns among both consumers and academia (Tarbit et al. 2023). Aguirre et al. (2016) introduced the personalization-privacy paradox, which describes a situation in which consumers appreciate personalized marketing initiatives tailored to their preferences, but express concerns about the lack of transparency in how their personal data is collected, analyzed, and used (Acquisti et al. 2016).

This growing concern over privacy is worrisome for the industry due to its potential consequences, such as consumers' reduced willingness to share personal information on digital platforms, the declining effectiveness of personalization strategies, and increased regulatory scrutiny (Martin et al. 2017). Indeed, the European Commission (2012) introduced the General Data Protection Regulation (GDPR), which established directives

TABLE 1 | Summary of discrepancies in personalization.

Term	Definition	Proactive/Reactive	References
Personalization	Tailoring the marketing mix to each consumer based on the information collected.	Proactive	Chandra et al. (2022)
Individualization	Creating unique experiences or products for each consumer.	Proactive	Riemer and Totz (2003)
Segmentation	Dividing consumers into groups based on common characteristics to target specific strategies to each segment.	Proactive	Smith (1956)
One-to-One Marketing	Develop direct communication with personalized content or offers.	Proactive	Peppers and Rogers (1997)
Customization	Adapt an offer to the consumer demand (e.g., allowing customers to choose or modify product features).	Reactive	Davis (1987)

requiring companies to obtain informed consent before collecting personal data, to provide users with options to delete their data, and to ensure the ethical use of AI and ML algorithms in data processing (AEPD 2018). However, ethical concerns persist regarding transparency and control over personal data.

The ability of NLP algorithms to personalize email messages enhances consumer satisfaction and generates profits for marketers. However, it also raises ethical concerns related to algorithmic bias and consumer autonomy (Karami et al. 2024). Algorithmic bias occurs when consumers are misclassified or underrepresented based on their language or cultural background. For instance, the exclusion of certain names from databases can result in the elimination of specific immigrant groups from marketing campaigns, reinforcing existing social inequalities (Barocas et al. 2019).

Concerns about consumer autonomy arise due to the informational asymmetry between consumers and marketers. The vast computational power of ML algorithms, combined with access to extensive databases, allows digital marketers to monitor and track consumer behavior on an unprecedented scale. This enables them to collect personal information, track purchases (surveillance), design highly personalized messages, and implement persuasive techniques (such as hooks) that make offers hard to resist (Yeung 2017).

However, some scholars, such as Matzner (2019), argue that despite the sophistication of ML algorithms, their analytical power does not necessarily undermine consumer autonomy. Instead, he suggests that these algorithms reshape and adapt consumer behavior to modern times. In other words, just as traditional advertising has lost some of its persuasive power over time as consumers have become more familiar with its tactics, the same may eventually happen with personalized digital marketing stimuli (Perloff 1993).

2.3 | NLP Algorithms

NLP is a field of AI that focuses on learning, analyzing, and reproducing human language to facilitate human-machine interactions. It is an interdisciplinary domain that bridges linguistics and data science to help machines understand and generate language as naturally and flexibly as humans do (Chowdhury and Nath 2001). Although NLP has been studied for over a century, its development has accelerated significantly in the last decade due to advances in computational power and access to big data (Evans and Aceves 2016).

Advancements in ML and deep learning (DL) have enabled NLP to perform tasks such as sentiment analysis (Moon et al. 2021), machine translation (Kumar and Rathore 2016), and text classification (Evans and Aceves 2016). Furthermore, given the importance of personalization in email marketing, ML tools for text classification and error detection are essential.

In a recent literature review on AI in marketing, Mariani et al. (2022) emphasized linguistic analytics as a key research area, highlighting its role in data filtering. García et al. (2016) argue that to ensure the optimal performance of ML algorithms in big data contexts, data preprocessing is required, which includes tasks such as data cleaning, integration, and transformation (encoding). For instance, effective personalization and customization strategies require customer names in databases to be error-free and standardized. Otherwise, errors such as misspellings or duplicate messages can create a negative impression on recipients (Anshari et al. 2019).

2.4 | String Matching Algorithms

String matching algorithms are used to determine the degree of similarity between two texts (Navarro 2001). Although string comparisons have been used for over a century to correct spelling errors, it was not until the 1950s that Shannon (1951) revolutionized language analysis by introducing concepts such as entropy and predictability. Based on these concepts, approximate similarity algorithms were developed to overcome the limitations of strict word matching (Navarro 2001). Damerau (1964) introduced the classification of errors, proposing that if a word does not match the correct dictionary entry, it is due to one of the following four possible mistakes: (1) inclusion of an incorrect letter, (2) omission of a correct letter, (3) addition of extra letters, or (4) transposition of letters. Using this classification, he was able to identify 95% of spelling mistakes. Shortly afterwards, Levenshtein (1966) proposed a metric, now known as edit distance, to measure the difference between two sequences of strings (words). This metric represents the minimum number of single-letter changes (insertions, deletions, or substitutions) needed to transform one word into another. For example, converting “cat” to “car” requires only changing “t” to “r,” resulting in an edit distance of 1. This method quantifies similarity by calculating the minimum required changes, allowing for the correction of common misspellings, such as transposed or omitted letters.

In parallel, Fellegi and Sunter (1969) addressed the issue of record linkage (or data linkage), which involves searching for and identifying information about the same item (e.g., a person) across different databases that may or may not share common identifiers (files, books, websites, etc.). They proposed a probabilistic approach using vector matching for various fields, such as name, birthdates, or addresses, to determine whether two records belonged to the same person,—even in the presence of typographical errors and data variations. Building on this probabilistic framework, Jaro (1989) and later, Jaro-Winkler (Winkler 1990) introduced an improved string similarity metric, which assigns a greater weight to matches occurring at the beginning of strings. However, it is important to note that although this method is referred to as a distance metric, it does not strictly adhere to the mathematical definition, as it does not satisfy triangular inequality.

In today's big data environment, new algorithms have emerged, such as RapidFuzz, an open-source Python library licensed from MIT and developed in 2020 by Bachmann as an improvement on its predecessor, FuzzyWuzzy. RapidFuzz combines multiple string-matching algorithms, including

Levenshtein distance, Jaro distance, and Jaro-Winkler distance (Bachmann 2024). One of its key outputs is the calculation of similarity thresholds between two strings, which helps determine whether two names are similar enough to be considered a match. This capability enhances the detection and correction of typos or variations in name spelling.

String-matching algorithms have been widely applied across various fields, including text retrieval, signal processing, and computational biology (Yujian and Bo 2007). Table 2 presents some of these applications. Despite being developed in 1966, Levenshtein's algorithm is still used today in ML applications. In some cases, it even outperforms more recent methods, such as the Jaro-Winkler algorithm, when applied to specific databases (Kiawkaew et al. 2023). Medhat et al. (2015) adapted the Levenshtein algorithm for record linkage in multilingual data sets, comparing it with eleven other algorithms, including Arabic Soundex and Editex. Their findings showed that Levenshtein achieved an accuracy of 91.6%, significantly outperforming the other methods, with Editex being the closest competitor at 85% accuracy. In a related study, Syarafina and Palandi (2021) tackled typographical errors in article manuscripts, using Levenshtein's algorithm for misspelling correction. Their experiment, conducted on texts ranging from 100 to 500 words, yielded correction accuracies between 95% and 90%, respectively. Kiawkaew et al. (2023) compared Thai and English personal names by first converting Thai names into the Roman alphabet and then applying Levenshtein distance and Jaro-Winkler distance. Their results showed that Levenshtein was significantly more effective, achieving an average accuracy of 99.98%, outperforming Jaro-Winkler.

String-matching algorithms have also been applied in digital marketing. For instance, Levenshtein's algorithm has been used to: (1) improve Search Engine Optimization (SEO) by helping to match consumer search queries with retail product listings by identifying typos and close matches (Zelenetska et al. 2023); and (2) personalize email marketing campaigns by introducing minor variations in email subjects and headers, thus increasing opening rates by preventing messages from being flagged as redundant (Balakrishnan and Parekh 2014). However, Levenshtein's algorithm has notable limitations: (1) it only analyzes character-level changes without considering the semantic meaning of words; (2) it does not account for contextual usage, which could further improve spelling correction; and (3) it requires a significant amount of memory as string size increases, making it computationally expensive (Syarafina and Palandi 2021).

As it is so new, research on the use of RapidFuzz in text comparison applications is still limited, although some early studies have explored its potential. Day (2022) applied it to the analysis of peer review comments in academic journals, using it to measure overlap and to detect duplicate or partially duplicated feedback that could indicate misconduct. Cabrera-Diego and Gheewala (2024) proposed the use of RapidFuzz for pseudonymization,—replacing real names with pseudonyms in legal documents to ensure compliance with GDPR. As a result, they developed Psilence, a tool that applies this technique to international arbitration documents in English.

To our knowledge, no studies to date have compared the performance of Levenshtein's algorithm and RapidFuzz in detecting errors in name spelling.

3 | Methodologies

This study employs a case study approach to examine the data filtering process used in digital marketing personalization. It describes the different phases of information standardization and compares the performance of two error correction algorithms—Levenshtein and RapidFuzz.

TABLE 2 | Summary of comparative studies.

Approach	Brief explanation	Results	Authors (year)
Levenshtein distance algorithm adapted for cross-language comparison	Used in a data mining experiment to link records written in different languages or by people from different cultures	Achieved 91.6% precision, compared to other algorithms that scored above 90%.	Medhat et al. (2015)
Levenshtein distance algorithm based on approximate string matching	Used to correct misspelled words in texts ranging from 100 to 500 words	Obtained average correction accuracies between 95% and 90% respectively	Syarafina and Palandi (2021)
Levenshtein and Jaro-Winkler distance	Used to compare the similarity between Romanized Thai and English personal names	Levenshtein proved more effective, with an average accuracy of 99.98% compared to Jaro-Winkler	Kiawkaew et al. (2023)
RapidFuzz	Used to estimate the degrees of overlap between peer review comments to detect duplication and partial duplication	Helped identify potential candidates for misconduct for further investigation	Day (2022)
RapidFuzz	Used to support the pseudonymization of names in legal documents for GDPR compliance	Proposed as a tool called Psilence for consistent pseudonymization across public documents	Cabrera-Diego and Gheewala (2024)

A case study is a qualitative research method that involves in-depth analysis of a certain problem, managerial experience, or proposed theory (Ghauri 2004). Case studies are particularly useful in complex or relatively unexplored research areas, and when the goal is to propose hypotheses. In other words, they attempt to answer "how" and "why" questions (Yin and Yin 1994).

This case study replicates the complex process of standardizing and cleaning a sample of 100,000 customer records, randomly selected from a database of over three million records collected by a lead-generation company. The data set was compiled between January and September 2024, which was verified using the Home Location Register (HLR). That is, the information provided on the phone number was checked against the Global System for Mobile Communications (GSM), which is standard in current digital marketing practices to prevent registration fraud. A sample of 100,000 records was chosen as a representative subset of the full database, allowing for a more manageable analysis in terms of computing time and storage cost, while ensuring that the results can be generalized to the complete data set (Cochran 1977).

The database was provided under a confidentiality agreement with Coregistros, a lead-gathering company that attracts customers by advertising online sweepstakes and quizzes (on history or geography, e.g.,). To receive sweepstakes prizes or to know the number of correct answers, participants must register on the sponsor's website by providing their personal data. However, the collected data is not always accurate, and therefore needs to be cleaned by removing false information and correcting errors (Sáez-Ortuño, Forgas-Coll, et al. 2023). Before utilizing the data, we ensured that the company was compliant with the GDPR (European Commission 2012) and the Spanish data protection act, LOPD-GDPR ("Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y Garantía de los Derechos Digitales") (AEPD 2018).

We specifically focus on the name filtering phase, applying Boolean algebra functions (if the name exists, then 1; otherwise, 0) and, in the final phase, comparing the performance of two NLP-based error correction algorithms. Misspelled names can be either unintentional ("typos") or intentional (to conceal identity or impersonate) (Sáez-Ortuño, Forgas-Coll, et al. 2023). Typographical errors are identified, corrected, and verified using filters and algorithms, while records containing intentional errors are deleted and blacklisted.

Custom tools for information standardization often need to be developed before correction algorithms can be applied. Common challenges include: (1) abbreviations—consumers may write shortened versions of their names, especially for compound names (e.g., "J. Carlos" instead of "Juan Carlos"), which need to be replaced with the full forms; (2) Inversion between first and last names—(e.g., "Rodriguez Angel" instead of "Angel Rodriguez"), which is detected and corrected using a Boolean algorithm; (3) Invalid characters—users may mistakenly enter numbers instead of letters, such as typing "0" (zero) instead of "O" (uppercase O).

To evaluate the incremental improvements at each step, the registered and corrected names are compared against the Spanish National Institute of Statistics' (INE) official database of names, which contains 58,423 unique names (28,644 male and 29,779 female), with name frequencies ranging from 20 to 614,853 for males and 20 to 630,253 for females. In this study, following established practices in the literature, we apply the Threshold Occurrence Level (TOL), excluding names that appear fewer than 20 times in Spain (Sáez-Ortuño, Forgas-Coll, et al. 2023).

All standardization and correction steps not only enhance computational efficiency but also ensure traceability for further analysis. However, despite these efforts, some name records remain uncorrected, with a conditional formula result of False (0). For these irreducible cases, we apply two error correction algorithms—Levenshtein and RapidFuzz—and compare their performance in detecting and correcting name errors.

3.1 | Filtering

The data cleaning process is based on Damerau (1964) and followed five phases: (1) elimination of extraneous elements and standardization; (2) correction of abbreviations; (3) rectification of first and second name order; (4) amendment of details; and (5) name correction. Meanwhile, the names in the INE database were divided into five sets: male names (28,644), female names (29,579), second names or surnames (82,134), the union of male and female names (57,330), and the intersection of male and female names (names used for both genders) (893).

The filtering and standardization process is illustrated in Figure 1 and Table 3 presents the improvements achieved at each step. The phases are described below:

Phase 1. *Removal of Extraneous Elements and Normalization*. The process began with the detection of irregular text, including numbers, excessively short names, repetitive letter sequences, or unusual combinations of vowels and consonants. Boolean conditional functions were applied to eliminate non-alphanumeric characters, duplicate letters within a word, double spaces, and blank cells, returning a value of 1 (true) or 0 (false). A filter was also applied to detect very short names (two letters or fewer), identifying 19 names. Finally, all names were normalized by conversion to lowercase.

Phase 2. *Substitution of Abbreviations*. This phase is also part of standardization, as it involves detecting the most common abbreviations of compound names and replacing them with their full forms (e.g., José Antonio, Francisco Javier, Ana Belen). Two databases were compiled containing the most common abbreviations and their corresponding full names: one for the first part of compound names (722 records) and another for the second part (14 records). For example, common abbreviations of José Antonio included José A., Pepe, Joséan, and J.A. A conditional function replaced these with their full forms by cross-referencing the INE name database. After completing Phases 1 and 2, 87,085 of the 100,000 records matched the INE database, while 12,915 did not.

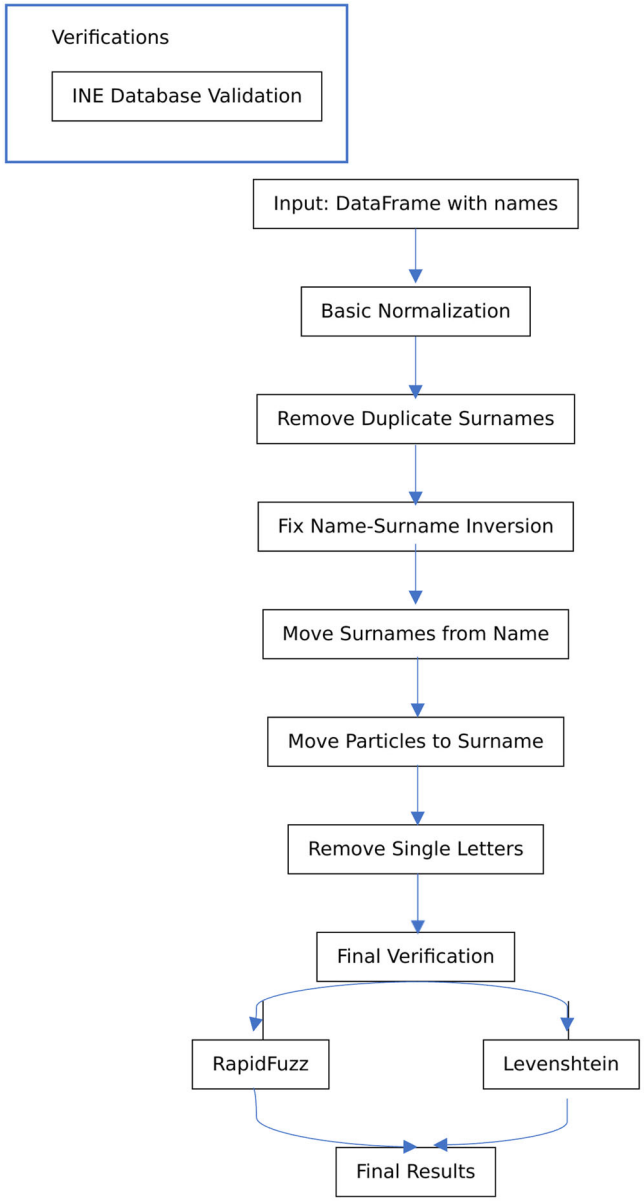

FIGURE 1 | Sequential data filtering process diagram.

Phase 3. *Checking and Correcting the Order of First and Second Names*. This filter was developed in three stages: (1) Detecting cases of two identical names appearing in the second name field and removing the duplicate; (2) Detecting cases of the same word appearing in the first and second name fields, and deleting the entry from the first; and (3) Detecting names that did not match the INE list of first names but did match the second name list (or vice versa), and correcting the order. These three steps increased the number of matches with the INE database from 87,085 to 92,075, representing an improvement of 5.7%.

Phase 4. *Amendment of Details*. This phase ensured that in the exchange process between the parts of compound names (e.g., María del Carmen, Juan de Dios) there were no loose letters (which are eliminated) and that the particles "de," "la," "del" were correctly positioned in names. A function was implemented to move these particles from the end of the first part of a compound name to the beginning of the second part when necessary. As a result, the number of matches with the INE database increased to 92,457, leaving 7543 non-matching records.

Phase 5. *Correction of Misspelled Names*. The final phase compared the performance of the two NLP-based error correction algorithms—RapidFuzz and Levenshtein. The efficiency estimator used was the degree of similarity between the corrected names and the official INE database.

3.2 | Comparative Performance Analysis Between RapidFuzz and Levenshtein

To compare the efficiency of two NLP algorithms, we selected RapidFuzz, a modern algorithm developed by Bachmann in 2020, and the classic Levenshtein (1966). The latter has demonstrated robust performance in multiple comparative analyses (Kiawkaew et al. 2023; Medhat et al. 2015) but, to the best of our knowledge, has never been compared to RapidFuzz.

- RapidFuzz is one of the latest additions to NLP algorithms for correcting names with typos or spelling variations. It functions as a hybrid, synthesizing several other techniques, including Levenshtein and Jaro-Winkler distance, to calculate a similarity score ranging from 0 to 100.

  In this case, the RapidFuzz algorithm processed names discarded during the filtering sequence and searched the INE database for the most similar matches. It assigned a percentage of similarity to each of them and replaced the original name with the highest-scoring match. For the initial stage, RapidFuzz used the hybrid WRatio algorithm, which combines an estimate of the minimum number of operations required to transform one string into another with the Jaro-Winkler distance. This approach attributes greater weight to matches at the beginning of strings (e.g., the first few letters) and yields a similarity percentage estimate. The process usually generates between 1 and 5 candidate names, each with an associated probability of being the best replacement for the misspelled name. Names achieving a match rate above 95% are usually selected (Bachmann 2024).

  To further refine the selection process, additional information was incorporated to determine the most suitable candidate. In this study, the results were cross-referenced with the complementary data on declared sex (male or female), concurrence of which is added to the probability of string matching. Finally, the names are replaced accordingly.

  For example, if a discarded name is "raqel" and the declared sex is female, the algorithm narrows the search to the list of female names. The WRatio algorithm then compares the name 'raqel' with the names in the INE list of female names and selects the string of letters that has the highest percentage match. In this case, the candidate "raquel" received the highest score of 96, which is above the threshold of 95. The name was therefore replaced in the

TABLE 3 | Filtering process.

Stage	INE matches	%	INE non-matches	%_non_matches	% Improvement
Initial data after applying the abbreviation filter	87,085	87.09%	12,915	12.91%	—
Remove duplicate surnames	89,212	89.21%	10,788	10.79%	2.13%
Fix name-surname inversion	90,003	90.00%	9997	10.00%	0.79%
Move surnames from name	92,075	92.08%	7925	7.92%	2.07%
Move particles to surname	92,227	92.23%	7773	7.77%	0.15%
Remove single letters	92363	92.36%	7637	7.64%	0.14%
Results of the filtering process	92,457	92.46%	7543	7.54%	0.09%

TABLE 4 | Comparative results of the algorithms.

Algorithm applied	Names not matched with INE	% of total sample	Names corrected	% improvement	% improvement over total sample	Names matched with INE	% of total sample
RapidFuzz	7543	7.54	476	6.31	0.48	92,933	92.93
Levenshtein	7543	7.54	1969	26.10	1.97	94,426	94.43

analyzed data, and the process continued with the next word.

- The Levenshtein algorithm follows a similar process. It compares the list of discarded names with the INE database, selects candidates, assigns a score, and replaces the incorrect word with the highest-scoring candidate. As described above, this algorithm calculates the minimum number of operations required to transform one string (a name) into another, and is particularly useful for detecting misspellings and correcting small differences between names. The distance is derived from the cost matrix $(m+1) \times (n+1)$, where m is the length of the original string and n is the length of the string being compared. This matrix records the partial transformation costs for all substrings.

For example, to calculate the distance between "andrs" (a string of 5 letters +1) and "andres" (a string of 6 letters +1), the algorithm constructs a 6×7 matrix. The first row corresponds to indices from 0 to 5 ("andrs"), and the first column to indices from 0 to 6 ("andres"). The matrix is then completed by estimating the cost of transforming one word into another, which depends on the number of letters that need to be changed to match the two words. When the letters in one word match those in the other (as in the case of "a," "n," "d," "r," and "s"), the cost is 0. However, when an insertion is required, the cost is 1. In this case, where only the letter "e" needs to be inserted between "r" and "s," the cost is 1, for only one insertion is required to transform "andrs" into "andres."

Candidate names are selected if their distance is less than or equal to a predetermined tolerance threshold. In our case, the acceptable tolerance is 1. If multiple candidates meet this criterion, the name with the highest frequency is selected.

3.3 | Results

To assess the predictive capacity of the RapidFuzz and Levenshtein algorithms, both were applied to the 7534 names remaining after the filtering process. In the next phase, the model's predictive capacity was validated. The results, summarized in Table 4, show that the more modern RapidFuzz algorithm achieved a lower improvement rate (6.31% [476/7543]) than the older Levenshtein algorithm (26.10% [1969/7743]). A Chi-square test for proportions rejected the null hypothesis of equality of proportions, indicating that the differences between the two algorithms are significant ($X^2 = 911.67$, $p < 0.000$). To complement the analysis, a Z-test was performed to compare the conversion rates of the two algorithms, yielding a Z-statistic of -32.99, thus confirming the extremely significant difference between both ratios ($p < 0.001$), with Levenshtein demonstrating a significantly higher conversion rate than RapidFuzz.

In summary, of the 100,000 names in the sample, 87% were matched to a name in the official INE list, leaving 12,915 names categorized as unknown. Next, a sequence of filters was applied, summarized in Table 3, which reduced the number of unknown names to 7543. Finally, two NLP algorithms were applied to detect and correct possible transcription errors. A comparison of the results shows that the Levenshtein algorithm performed better, correcting 1969 names and reducing the number of names requiring manual review to 5574.

4 | Conclusions

E-mail marketing is faced with a dilemma. Excessive letterboxing is turning it into junk mail, making it more likely to be

ignored by consumers (Chandra et al. 2022). Moreover, growing privacy concerns among consumers may be further reducing its effectiveness (Martin et al. 2017). To address these challenges, marketers are advised to space out communications, make smart use of personalization strategies (Jani 2024), and address ethical concerns by improving transparency (Barocas et al. 2019).

Our study has focused on personalization, a complex strategy that involves multiple phases including self-reference, anthropomorphism and system characteristics (Cavdar-Aksoy et al. 2021). This study has particularly focused on the self-reference process, which involves using the customer's name in all company communications. However, to be effective, the information must be correct, accurate, and reliable, and this requires a data cleaning process. Our case study illustrates the complexity of this process in data mining, by illustrating the steps involved in cleaning a database containing 100,000 first names and surnames of potential customers. Boolean conditional functions were used to standardize and clean the data, while two string similarity algorithms detected and corrected misspelled names. After completing the data preparation, filtering, and error correction stages, the number of records matching the INE database rose from 87,086 to 94,426, leaving just over 5500 records requiring human review.

To automate the final phase of data cleaning, two NLP algorithms were compared, the findings showing that the Levenshtein algorithm outperformed RapidFuzz, in line with previous results reported in the literature (Kiawkaew et al. 2023; Syarafina and Palandi 2021). While RapidFuzz, which integrates algorithms such as Jaro-Winkler, typically presents high accuracy when analyzing short strings, such as exact name matching, it is not so accurate when working with 7500 items of data. In addition, Levenshtein proved effective at handling multilingual data, as in our case, which included numerous Moroccan and Romanian names, for example, and also for text normalization applications (Medhat et al. 2015; Kiawkaew et al. 2023). Analysis of the unmatched names highlights two challenges: (1) a strong suspicion that the vast majority are false entries; and (2) the presence of names that are strange for the Spanish language (e.g., formed by two letters like Yu, the inclusion of repeated vowels like Mariia, or more than two consonants in a row like Jingwen) that are usually detected as errors even though they belong to names of citizens of foreign origin. These findings underscore the sensitivity of both the filtering processes and NLP algorithms in rejecting names with complex variations for the base language.

The sequential filtering process by stages not only favors the computational analysis of algorithms, whether for error detection or ML in message personalization, but also ensures the traceability of all corrections made. This allows for corrective actions at the appropriate stage if false positives or negatives are identified during manual review. However, NLP algorithms are not a panacea. They cannot fully resolve all problems related to error detection in texts. Therefore, a combination of traditional filtering methods and advanced data recognition techniques is necessary for effective database cleaning, particularly when registering first and second names.

In short, this study contributes to the debate on the quality of the output generated by ML algorithms, emphasizing the need for reliable data to enhance their effectiveness. Indeed, database errors can introduce bias, reduce the accuracy of model estimates and, ultimately, compromise decision-making in digital marketing (Jain et al. 2020).

5 | Ethical Implications

This study expands knowledge about the principles and practices of ML quality theory by addressing aspects such as data quality, model performance, robustness and overall reliability of ML applications (Rahm and Do 2000). It highlights often overlooked issues around big data, such as data cleaning and standardization, which are essential for effective application of ML algorithms. Specifically, the study describes the various stages of the filtering process for a certain data point: the consumer's name. To achieve this, selection criteria such as TOL were proposed, which excludes names that appear fewer than 20 times in the country (Sáez-Ortuño, Forgas-Coll, et al. 2023). While this serves as a control mechanism against false name registrants, it also raises ethical dilemmas around the exclusion of names that are foreign to the native language of the country (Karami et al. 2024).

Although the personalization-privacy paradox has been discussed in the literature (Aguirre et al. 2016), the ethical implications of data mining in the NLP context have not been fully addressed. Three key concerns have typically been identified: the appropriate use of data, algorithmic bias, and the impact on consumer autonomy.

Consumer concerns about the appropriate use of data stems from a lack of knowledge about what companies do with the collected information, and its potential consequences (Acquisti et al. 2016). This uncertainty heightens privacy concerns and makes consumers more reluctant to share personal data with marketers and lead collectors (Ponte et al. 2024). To address this information asymmetry and safeguard consumer privacy, the GDPR outlines two primary objectives: minimizing collection and anonymizing personal data (European Commission 2012). To rebuild consumer trust, marketers need to be more transparent by clearly explaining how data is used, limiting collection to what is strictly necessary, and implementing anonymization mechanisms (Ponte et al. 2024).

Another ethical issue is algorithm bias, which can lead to discrimination due to ML algorithms and databases reflecting, for example, historical inequalities that disproportionately exclude certain consumer groups, as algorithms tend to attribute greater value to higher frequencies (Mariani et al. 2022). In this study, the TOL selection criterion excluded names with frequencies lower than 20 to minimize noise. However, a review of the discarded names revealed that most were difficult to spell in Spanish. The RapidFuzz and Levenshtein algorithms also failed to process many names of immigrant origin, such as “Josibel” and “Umalt.” This issue aligns with ML equity theory, which advocates for algorithmic solutions that do not disadvantage specific groups based on gender, ethnicity, disability, or linguistic origin (Barocas et al. 2019). To address this, marketers should conduct regular audits of their algorithms to ensure fair outcomes. If biases are detected, they should be addressed by

refining both the data collection methods and algorithmic processes (Mehrabi et al. 2021).

There are also ethical concerns around the impact of personalized marketing on consumer autonomy. The use of big data analysis and ML algorithms to design personalized offers can undermine consumer free will, making it harder for them to resist targeted promotions (Yeung 2017).

To prevent digital companies from overwhelming consumers—and thereby address saturation and consequent disengagement issues—marketers must balance short-term commercial interests with their ethical responsibility to protect consumer autonomy. Strategies such as spacing marketing messages and improving data accuracy can help ensure that personalized marketing provides real value to consumers (Jani 2024). Additionally, more transparent data collection is recommended, including the offer of explicit opt-in mechanisms, allowing users to decide whether they want to receive personalized content.

Ultimately, to address these ethical concerns, both marketers and data scientists should adopt a set of guiding principles based on transparency, bias detection and mitigation, as well as developing mechanisms that safeguard privacy (Dwork and Roth 2014).

6 | Limitations and Future Research Directions

This case study has certain limitations that also imply avenues for future research. First, it presents a five-stage consumer name filtering process based on the Levenshtein algorithm, achieving an estimated accuracy of 94.4%. Consequently, this procedure can be applied to other short text strings where users are prone to make typographical errors, such as email addresses, postal addresses, or user-generated reviews. However, further refinements are needed to improve accuracy, particularly when dealing with foreign names that are difficult to spell in national language or names that may have different transliterations into Latin script. Therefore, additional filtering criteria should be considered. NLP algorithms need to be able to analyze words from multiple languages in the same text. Therefore, future research should explore the development of hybrid approaches that combine traditional data cleaning techniques with advanced NLP models trained on multilingual data sets to minimize exclusion biases. Another future line of research is the exploration of the integration of Explainable AI (XAI) methods to make ML-based data processing more transparent and interpretable by addressing the "black box" nature of many models. This would give marketers a better understanding of how algorithms influence personalization outcomes (Barredo Arrieta et al. 2020).

Additionally, while this study estimates the internal validity of its findings by comparing the filtered data with the official INE database, a name matching official records does not necessarily confirm its authenticity. External validity also needs to be assessed. To improve accuracy, future research should consider cross-referencing names with additional sources, such as email addresses and phone numbers, to verify their authenticity.

Regarding future research on the use of big data with NLP algorithms for personalized digital marketing, some key questions are proposed: How should NLP models be designed to analyze more diverse data sets and reduce algorithmic bias in recognizing names with complex spellings? Should government watchdogs regulate the personalization strategies driven by big data and ML algorithms? And finally, what email marketing strategies should companies develop to balance personalization and privacy while maintaining consumer trust?

Acknowledgments

Special thanks to Marc Vergés Santiago, a student of the double degree in Mathematics and Computer Engineering at the University of Barcelona, who has collaborated in the design and development of the algorithms presented in this paper.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The authors have nothing to report.

References

Acquisti, A., C. Taylor, and L. Wagman. 2016. "The Economics of Privacy." *Journal of Economic Literature* 54, no. 2: 442–492.

AEPD. 2018. Ley Orgánica 3/2018, De 5 De Diciembre, De Protección De Datos Personales Y Garantía De Los Derechos Digitales. *Agencia Española de Protección de Datos*. https://www.aepd.es/.

Aguirre, E., A. L. Roggeveen, D. Grewal, and M. Wetzels. 2016. "The Personalization-Privacy Paradox: Implications for New Media." *Journal of Consumer Marketing* 33, no. 2: 98–110.

Anshari, M., M. N. Almunawar, S. A. Lim, and A. Al-Mudimigh. 2019. "Customer Relationship Management and Big Data Enabled: Personalization & Customization of Services." *Applied Computing and Informatics* 15, no. 2: 94–101.

Bachmann, M. 2024. RapidFuzz: Rapid Fuzzy String Matching in Python Using Various String Metrics. Github. https://github.com/maxbachmann/RapidFuzz.

Balakrishnan, R., and R. Parekh. 2014. "Learning to Predict Subject-Line Opens for Large-Scale Email Marketing." In *In 2014 IEEE International Conference on Big Data (Big Data)*, 579–584. IEEE.

Barocas, S., M. Hardt, and A. Narayanan. 2019. Fairness and Machine Learning, Accessed 1 February, 2025. http://www.fairmlbook.org.

Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." *Information Fusion* 58: 82–115.

Blom, J. 2020. "Personalization: A Taxonomy." In *CHI'00 Extended Abstracts on Human Factors in Computing Systems*, 313–314.

Boudet, J., B. Gregg, K. Rathje, E. Stein, and K. Vollhardt. 2019. *The Future of Personalization—And How to Get Ready for it*. McKinsey & Company.

Cabrera-Diego, L. A., and A. Gheewala. 2024. "PSILENCE: A Pseudonymization Tool for International Law." In *In Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo-2024)*, 25–36.

Cavdar-Aksoy, N., E. Tumer Kabadayi, C. Yilmaz, and A. Kocak Alan. 2021. "A Typology of Personalisation Practices In Marketing in the

Digital Age." *Journal of Marketing Management* 37, no. 11–12: 1091–1122.

Chandra, S., S. Verma, W. M. Lim, S. Kumar, and N. Donthu. 2022. "Personalization in Personalized Marketing: Trends and Ways Forward." *Psychology & Marketing* 39, no. 8: 1529–1562.

Chittenden, L., and R. Rettie. 2003. "An Evaluation of E-Mail Marketing and Factors Affecting Response." *Journal of Targeting, Measurement and Analysis for Marketing* 11: 203–217.

Chowdhury, S., and A. Nath. 2001. "Trends In Natural Language Processing: Scope and Challenges." *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 7, no. 6: 393–401.

Cochran, W. G. 1977. *Sampling Techniques*. Johan Wiley & Sons Inc.

Collins, E., N. Rozanov, and B. Zhang. 2018. Evolutionary Data measures: Understanding the Difficulty of Text Classification Tasks. arXiv preprint arXiv:1811.01910.

Damerau, F. J. 1964. "A Technique for Computer Detection and Correction of Spelling Errors." *Communications of the ACM* 7, no. 3: 171–176.

Davis, S. 1987. "Future perfect." In *Reading, MA: Addison*.

Day, A. 2022. "Exploratory Analysis of Text Duplication In Peer-Review Reveals Peer-Review Fraud and Paper Mills." *Scientometrics* 127, no. 10: 5965–5987.

Dumont, B., S. Maggio, G. S. Said, and Q. T. Au. 2019. Who Wrote This Book? A Challenge for E-Commerce. arXiv preprint arXiv:1905.01973.

Dwork, C., and A. Roth. 2014. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends in Theoretical Computer Science* 9, no. 3–4: 211–407.

European Commission. 2012. Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation). https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX:52012PC0011.

Evans, J. A., and P. Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42, no. 1: 21–50.

Fellegi, I. P., and A. B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64, no. 328: 1183–1210.

García, S., S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera. 2016. "Big Data Preprocessing: Methods and Prospects." *Big Data Analytics* 1: 9.

Ghauri, P. 2004. "Designing and Conducting Case Studies in International Business Research." In *Handbook of Qualitative Research Methods for International Business*, Edited by R. Piekkari and C. Welch, 109–124. Elgar online.

Hudák, M., E. Kianičková, and R. Madleňák. 2017. "The Importance of E-Mail Marketing In E-Commerce." *Procedia Engineering* 192: 342–347.

Jain, A., H. Patel, and L. Nagalapatti, et al. 2020. "Overview and Importance of Data Quality for Machine Learning Tasks." In *In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3561–3562.

Jani, D. 2024. US Consumers are Fed Up with Excessive Texts and Emails, But 4 Marketing Tactics can Keep Them Engaged. GetApp. https://www.getapp.com/resources/digital-content-consumers-unsubscribe-from-marketing/.

Jaro, M. A. 1989. "Advances In Record-Linkage Methodology As Applied to Matching the 1985 Census of Tampa, Florida." *Journal of the American Statistical Association* 84, no. 406: 414–420.

Jürgensmeier, L., and B. Skiera. 2024. "Generative AI for Scalable Feedback to Multimodal Exercises." *International Journal of Research in Marketing* 41, no. 3: 468–488.

Karami, A., M. Shemshaki, and M. Ghazanfar (2024). Exploring the Ethical Implications of AI-Powered Personalization in Digital Marketing. Data Intelligence, In-Press. https://doi.org/10.3724/2096-7004.di.2024.0055.

Kiawkaew, T. A., N. Kaothanthong, and T. Theeramunkong. 2023. "A Practical Technique for Thai-English Word Mapping Using Phonetic Rules: Person Name Matching Case Study." In *2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 1–6. IEEE.

Kumar, V., and R. S. Rathore. 2016. "A Review on Natural Language Processing." *International Journal of Engineering Development and Research* 4, no. 3: 380–381.

Kumar, V., X. Zhang, and A. Luo. 2014. "Modeling Customer Opt-In and Opt-Out In a Permission-Based Marketing Context." *Journal of Marketing Research* 51, no. 4: 403–419.

Levenshtein, V. I. 1966. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." In *Proceedings of the Soviet Physics Doklady*.

Lim, W. M., T. Rasul, S. Kumar, and M. Ala. 2022. "Past, Present, and Future of Customer Engagement." *Journal of Business Research* 140: 439–458.

Mariani, M. M., R. Perez-Vega, and J. Wirtz. 2022. "Ai In Marketing, Consumer Research and Psychology: A Systematic Literature Review and Research Agenda." *Psychology & Marketing* 39, no. 4: 755–776.

Martin, K. D., A. Borah, and R. W. Palmatier. 2017. "Data Privacy: Effects on Customer and Firm Performance." *Journal of Marketing* 81, no. 1: 36–58.

Matzner, T. 2019. "The Human Is Dead–Long Live the Algorithm! Human-Algorithmic Ensembles and Liberal Subjectivity." *Theory, Culture & Society* 36, no. 2: 123–144.

Medhat, D., A. Hassan, and C. Salama. 2015. "A Hybrid Cross-Language Name Matching Technique Using Novel Modified Levenshtein Distance." In *In 2015 Tenth International Conference on Computer Engineering & Systems (ICCES)*, 204–209. IEEE.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. 2021. "A Survey on Bias and Fairness In Machine Learning." *ACM Computing Surveys* 54, no. 6: 1–35.

Moller, M. 2020. The ROI of Email Marketing. Litmus. https://www.litmus.com/resources/email-marketing-roi.

Montgomery, A. L., and M. D. Smith. 2009. "Prospects for Personalization on the Internet." *Journal of Interactive Marketing* 23, no. 2: 130–137.

Moon, S., M. Y. Kim, and D. Iacobucci. 2021. "Content Analysis of Fake Consumer Reviews by Survey-Based Text Categorization." *International Journal of Research in Marketing* 38, no. 2: 343–364.

Murthi, B. P. S., and S. Sarkar. 2003. "The Role of the Management Sciences in Research on Personalization." *Management Science* 49, no. 10: 1344–1362.

Navarro, G. 2001. "A Guided Tour to Approximate String Matching." *ACM Computing Surveys* 33, no. 1: 31–88.

Peppers, D., and M. Rogers. 1997. *The One-to-One Future*. Double Day Publications.

Perloff, R. M. 1993. *The Dynamics of Persuasion: Communication and Attitudes in the 21st Century*. Routledge.

Ponte, G. R., J. E. Wieringa, T. Boot, and P. C. Verhoef. 2024. "Where's Waldo? A Framework for Quantifying the Privacy-Utility Trade-Off In Marketing Applications." *International Journal of Research in Marketing* 41, no. 3: 529–546.

Rahm, E., and H. H. Do. 2000. "Data Cleaning: Problems and Current Approaches." *IEEE Technical Bulletin Data Engineering* 23, no. 4: 3–13.

Riemer, K., and C. Totz. 2003. "The many faces of personalization." In *The customer centric enterprise*, Edited by M. M. Tseng and F. T. Piller, 35–50. Springer.

Sáez-Ortuño, L., S. Forgas-Coll, R. Huertas-Garcia, and J. Sánchez-García. 2023. "Online Cheaters: Profiles and Motivations of Internet Users Who Falsify Their Data Online." *Journal of Innovation & Knowledge* 8, no. 2: 100349.

Sáez-Ortuño, L., J. Sanchez-Garcia, S. Forgas-Coll, R. Huertas-García, and E. Puertas-Prats. 2023. "Impact of Artificial Intelligence on Marketing Research: Challenges and Ethical Considerations." In *Philosophy of Artificial Intelligence and Its Place in Society*, Edited by L. Moutinho, L. Cavique, and L. Bigné, 18–42. IGI Global.

Sahni, N. S., D. Zou, and P. K. Chintagunta. 2017. "Do Targeted Discount Offers Serve As Advertising? Evidence From 70 Field Experiments." *Management Science* 63, no. 8: 2688–2705.

Shannon, C. E. 1951. "Prediction and Entropy of Printed English." *Bell System Technical Journal* 30, no. 1: 50–64.

Smith, W. R. 1956. "Product Differentiation and Market Segmentation As Alternative Marketing Strategies." *Journal of Marketing* 21, no. 1: 3–8.

Surprenant, C. F., and M. R. Solomon. 1987. "Predictability and Personalization In the Service Encounter." *Journal of Marketing* 51, no. 2: 86–96.

Syarafina, N. N., and J. F. Palandi. 2021. "Designing a Word Recommendation Application Using the Levenshtein Distance Algorithm." *Matrix: Jurnal Manajemen Teknologi dan Informatika* 11, no. 2: 63–70.

Tarbit, J., J. Wirtz, W. Kunz, and N. Hartley. 2023. "Interpretation of Corporate Digital Responsibility Risks and Concerns by Automated Service Technologies: An AI Co-Created Article." *ROBONOMICS: The Journal of the Automated Economy* 4: 52.

Winkler, W. E. 1990. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354–359.

Yeung, K. 2017. "'Hypernudge': Big Data as a Mode of Regulation by Design." *Information, Communication & Society* 20, no. 1: 118–136.

Yin, R. K., and R. K. Yin. 1994. "Discovering the Future of the Case Study. Method In Evaluation Research." *Evaluation practice* 15, no. 3: 283–290.

Yujian, L., and L. Bo. 2007. "A Normalized Levenshtein Distance Metric." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, no. 6: 1091–1095.

Zelenetska, K., N. Porplytsya, I. Stasiv, S. Stańczyk, A. Jankowiak, and L. Bilovus. 2023. "SEO-Optimization of Product Content on a Marketplace Platform." In *In 2023 13th International Conference on Advanced Computer Information Technologies (ACIT)*, 201–205. IEEE.

Zhang, X., V. Kumar, and K. Cosguner. 2017. "Dynamically Managing a Profitable Email Marketing Program." *Journal of Marketing Research* 54, no. 6: 851–866.