International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-024-00582-1

REGULAR PAPER

Identifying missing data handling methods with text mining

Krisztián Boros[1,2] · **Zoltán Kmetty**[2,3]

Received: 4 June 2023 / Accepted: 30 May 2024

Abstract
Missing data is an inevitable aspect of every empirical research. Researchers developed several techniques to handle missing data to avoid information loss and biases. Over the past 50 years, these methods have become more and more efficient and also more complex. Building on previous review studies, this paper aims to analyze what kind of missing data handling methods are used among various scientific disciplines. For the analysis, we used nearly 50.000 scientific articles published between 1999 and 2016. JSTOR provided the data in text format. We utilized a text-mining approach to extract the necessary information from our corpus. Our results show that the usage of advanced missing data handling methods, such as Multiple Imputation or Full Information Maximum Likelihood estimation, is steadily growing in the examination period. Additionally, simpler methods, like listwise and pairwise deletion, are still in widespread use.

Keywords Missing data · Imputation · FastText · Text mining

1 Introduction

Missing data is an immanent part of every empirical research. Every time a patient drops out of a clinical study or a respondent does not answer a question in a survey; we encounter missing data. Researchers have developed various techniques to account for these scenarios to reduce information loss. Such techniques spread from deletion of missing cases to complex algorithms that replace missing information with a predicted value [1–6]. During the past five decades, these techniques continuously evolved. Despite the improvements, many research uses more conservative methods such as listwise deletion (complete case analysis) or mean imputation. Numerous types of research suggest that the more advanced missing data handling techniques, such as multiple imputation, are more flexible and reliable than the older and simpler ones [6, 7]. Of course, there are scenarios where a simple deletion method could perform nearly as well as a more modern approach [7], but in general, it is recommended to use an advanced technique [2, p. 39]. Our goal in this study is to identify missing data handling methods in scientific papers during the period 1999–2016 in order to examine the usage of advanced missing data handling methods. Throughout our research, we utilize a text-mining approach to extract the necessary information from the articles. The start of our examination period coincides with the publication of "Statistical Methods in Psychology Journals: guidelines and explanations" by [8]. This paper discusses the recommended methodologies for missing data in Psychology journals and highlights the importance of proper documentation of data analysis. As [2, p. 39] mentions, several studies support the insights of [8] on missing data handling practices and warn about the disadvantages of deletion methods. Besides Psychology, researchers in other fields preferred simpler techniques to account for missing data, especially listwise and pairwise deletion [9–11]. Even though a comprehensive survey of missing data handling methods is yet to be made, there were a considerable amount of reviews about techniques for handling missing data. Most of the studies were conducted in the educational, psychological, and medical research areas, probably because of the discipline-specific origins of the missing data paradigm and its applications in survey-type designs [12]. Roth [13] examined randomly

✉ Krisztián Boros
soirkorob@gmail.com

Zoltán Kmetty
kmetty.zoltan@tk.hu

[1] Graduate School of Economics, Waseda University, Totsukamachi 1-104, Shinjuku 169-8050, Tokyo, Japan

[2] Centre for Social Sciences, HUN-REN, Tóth Kálmán str. 4, Budapest 1097, Budapest, Hungary

[3] Faculty of Social Sciences, Eötvös Loránd University, Pázmány Péter sétány 1/A, Budapest 1117, Budapest, Hungary

Published online: 17 June 2024

selected articles from two Psychology journals between 1989 and 1991 and concluded that the usage of deletion methods is pervasive in these journals. Later studies on various fields supported the results of [13] [14–17]. [16] reviewed Political Science papers in a five-year period between 1993 and 1997 and found similar trends as [13]. Peugh and Enders [11] compared articles from 1999 and 2003 in Educational Research and, on the one hand, concluded that the popularity of deletion methods did not change between 1999 and 2003. On the other hand, however, they noted that the reporting of missing data increased significantly: *"In 1999, 33.75% of the studies that we identified as having missing data explicitly reported the problem, whereas this number more than doubled, to 74.24%, in 2003."* [11, p. 30]. Additionally, [18] also reviewed the practices of handling missing data in this field between 1998 and 2004. Their findings are consistent with the conclusions of [11]. In Psychology, [19] found similar trends concerning missing data methods between 2000 and 2006. In the Medical Research field between 2001 and 2002, [20] with the analysis of seven cancer journals, and [21] with the examination of four medical journals found similar results concerning listwise and pairwise deletion. [9] reviewed articles from four medical journals in 2013 and provided a comparison of previous studies that examined missing data methods. They compared their results with the findings of [22–24] and [21]. The main conclusion was that the usage of deletion methods remained unchanged during the period 1997–2013, but there was a slight increase in the usage of advanced imputation methods. By and large, studies from Psychology, Educational Research, Medical Research show that the usage of deletion methods remained unchanged during 1989–2013, but the usage of advanced missing data handling methods increased slightly. This means that more and more research use an advanced technique when missing data occurs. To the best of our knowledge, there are no recent studies discussing missing data handling trends after 2016 [25–27]. Instead of analyzing trends in handling missing data, most recent literature examines and compares more advanced imputation methods and machine learning techniques for handling missing data [28, 29]. The cited studies examined randomly selected papers by manually reviewing and evaluating them. Most studies analyzed a small set of documents from a specific scientific field. Studies focus on many scientific fields, and more extended periods are missing. Our study aims to fill this research gap by examining changes in data imputation techniques over an extensive database, over many years and across several disciplines. In this study, we train classification models to classify the articles on a large scale regarding how they handle missing data. With this machine-learning approach, researchers could rapidly classify a vast number of papers with still high reliability. Although this method has a slightly lower accuracy than manual coding, it is perfectly suitable for detecting trends.

Our main contributions are:

- We identify missing data handling methods in scientific papers during the period 1999–2016 to examine the usage of advanced missing data handling methods. We employ a novel text mining approach for our analysis.
- Our approach is time-efficient and scalable, allowing us to classify thousands of papers without seeing them, which saves time and permits us to apply this approach to an arbitrarily large sample. This is in contrast with previous research done with random sampling.
- Our analysis provides an interdisciplinary overview of trends in the usage of missing data handling methods and may facilitate the application of text mining in future research in this area.

2 Data collection

Our data were provided by JSTOR's Data for Research service (DfR) [30, 31]. Since we received the data from JSTOR on 2020.07.20, the service (and platform) went through a few changes. This means that the way of our data request is no longer available, but this does not affect our data in any way. Currently, JSTOR offers a sophisticated text-mining platform named "Constellate" in order to help researchers create databases and perform basic text-mining tasks. This platform was not available at the time of our research; therefore, a simple keyword search might give different results now, mostly because of the extended data sources. The original data request and collection procedure was the following. At first, a search query had to be made with the required parameters to access the list of articles. This resulted in a search URL and a search syntax, which can be accessed in our repository. Our search parameters included the time interval of publications (1999.01.01.-2016.12.31.)[1], the keywords/expressions ("missing data," "missing observations," "incomplete data," "imputation"), and language of the articles (English). In the case of keywords/expressions we had to specify which keywords/expressions should the corpus and title of the articles contain. Our request was that the corpus

[1] The availability of data largely determined the period selected for our research. When we started this study in 2020, JSTOR's Data for Research service was still under development. We decided that our cut-off year would be 2016 based on the available articles in the database at that time. The years from 2016 to 2020 had significantly fewer articles in JSTOR's database, and including these years with their limited number of papers could have biased our final results. Since then, JSTOR has significantly improved its service, including developing the Constellate platform for text mining. Although we could have requested an additional batch of data from 2017, it would not have been compatible with our previous data from 1999 to 2016.

of articles must contain at least one of the following keywords/expressions: "missing data," "missing observations," "incomplete data," "imputation." Our research focused on studies that analyzed empirical data and reflected at least in some way that there was some missingness in the data. However, after the keyword-based selection of studies, there were still some methodological papers in the database that focused on how missing data should be addressed. We tried to exclude these papers from the analysis. This filtering was partly done by deleting papers with any of the following keywords/expressions mentioned in the study title: "missing data," "incomplete data," and "imputation.") Later, during preprocessing, we make an extra step to ensure that only those papers stay in our corpus that could have used some kind of missing data technique and not about missing data handling. As a result, we have collected 49.603 articles with metadata, uni-, bi-, tri-grams, and article content. The database contained nearly 2000 different journals. These journals were assigned the highest Q-rank of the journal in the 2016 SCIMAGO database[2] and the journal's SJR (Scientific Journal Rank). We linked the JSTOR and SCIMAGO databases through the journal names. We were able to assign Q and SJR values to 67 percent of the articles in the database.

3 Methodology and data preparation

Our research aimed to categorize academic articles based on their application of missing data handling methods using complex text mining techniques. To the best of our knowledge, no previous studies have attempted this approach. Consequently, we had to develop a practical methodological framework to carry out our analysis. In the following section we give an overview about this process and highlight the most important aspects. In order to identify the usage of missing data handling methods we had to make a proper corpus to work with. Although the data we have gathered from JSTOR is filtered by the predefined keywords, it does not guarantee that only those articles get into our analysis that are eligible for our research. We had to make sure that only those papers are present in our corpus which use missing data handling methods, and are not about missing data handling methods. After defining our corpus, we performed several classifications to distinguish between various missing data handling techniques. Overall, our approach consisted of three main levels regarding classification (see Fig. 1). On the first level, we separated the papers according to their relation to missing data handling methods: If a given article was about missing data handling methods, then it was classified as "1" and was removed from the analysis. Otherwise, we kept the article in our corpus. The second level was destined to separate the usage of imputation methods from other techniques such as deletion, and from those cases where no technique was used. Accordingly, if an article used any type of imputation technique (multiple imputation, regression imputation, etc.), then it was classified as "1," otherwise "0." The third and the last level was divided into two parts. On the one hand, we checked whether a paper—that was classified into the "imputation" category on the previous level—used an advanced imputation method or not. On the other hand, if a paper did not use any imputation technique—according to the second level—then we checked if it has used any deletion method or not. Before heading toward the description of the preprocessing and classification, we must clarify what we mean by "imputation," "advanced imputation," and "deletion." The importance of this clarification lies in the fact that we had to have a solid definition of each method to be able to label our training set correctly. We heavily relied on the taxonomy of [5] since it gives a comprehensive overview of the techniques. Based on this paper, we have treated any form of substitution, replacement, and imputation as "imputation." For example, "mean substitution," "hot/cold-deck imputation," and "regression imputation" were treated as "imputation." The "advanced imputation" methods were the "full information maximum likelihood estimation" (FIML) and all variants of "multiple imputation" (MI). The definition of "deletion" is quite straightforward: every technique that includes the deletion of cases/observations, such as listwise/pairwise deletion.

4 Preprocess

To be able to extract the necessary information from the articles, we had to clean the text from meaningless symbols and noises, since the body of each article contained, for example, LaTeX markups, numbers, and Optical Character Recognition (OCR) errors. To remove LaTeX markups and to collect additional keyword lists and functions, we have created a small auxiliary package called *jprep*. All preprocessing and cleaning script can be accessed on GitHub. Every part of the preprocessing was conducted in R, and we used Python for the classification models. We would like to emphasize some of the preprocessing steps because they have significant impact on the results. As a general preprocessing step, we have removed stopwords from our corpus. The standard English stopword list provided by the *tm* package includes words such as "were," "not," and "at" [32]. However, these words were crucial for our analysis, as they appear in important expressions related to missing data, such as "data were missing," "missing at random," and "observations were discarded." In order to prevent any loss of information and ensure accurate analysis of our corpus, it was necessary to retain these words and exclude them from our stopword

[2] https://www.scimagojr.com/journalrank.php?openaccess=true&year=2016.

list. The next crucial preprocessing step was to remove long texts from the corpus. One would assume that since we are querying only articles, there are no outliers in terms of article length. Unfortunately, there were several documents that got into our query which were not articles. After examining the distribution of the number of tokens in the corpus, we have chosen a reasonable threshold (20000 tokens, which is equivalent to about 30–33 pages) for cutting the "tail" of our distribution to remove outliers (407 articles). Another step concerning the length of the documents was the removal of references and bibliography. These sections were unnecessary for our analysis and most likely would have biased the classification results. If an article used some kind of missing data technique, then probably referenced it afterward. This means that we would have missing data-related keywords outside the main contents. As we will discuss shortly, our models used a small piece of information from the article bodies, therefore an additional noise—such as the detailed references of other papers—could have shifted the focus of the classifier. As an example, let us suppose that an article mentions only once the EM-Algorithm in the body and cites one of the works of Little and Rubin. In the body, the classifier identifies the context in which the "EM-Algorithm" is mentioned, but since the respective paper is referenced at the end of the article, the classifier gets a further, unnecessary context. Instead of one meaningful context, we end up with two, from which one is absolutely useless. Lastly, we have applied another technique to boost our classification accuracy by further trimming the content of the documents. We "snipped out" the context of some predefined keywords[3] from each document in order to focus only the key parts of texts. During qualitative examination of some randomly selected papers we have seen that only a small fraction of the body of a paper deals with or even mentions the missing data handling method. So trimming down the papers only discards unnecessary noise from the texts—what we do not need for our classification task. For the sake of example, let us assume that there is an article about a clinical experiment where the researchers have decided to remove some observations due to their ineligibility and documented their decision with the statement "[. . .] 12 *cases were deleted from the analysis due to missingness."*. This is the only part of the article that would contain information about the missing data handling method, but this sentence is only a small piece of text compared to the whole paper. In order to identify the missing data handling method, our model should be able to correctly classify this article based on one sentence. To bypass this difficulty, we snip out the context around "missingness" and discard the remaining part of the article. As our results showed, using a small but meaningful fraction of each paper not only produced better classification performance, but it decreased the time required to train our models.

5 Classification

Because of the nature of our analyzed corpus, it was quite hard to find a classification model that can handle our special setup. The aim of the classification was to separate papers based on minimal information which—among other things—consisted of rare words. To train a supervised model, we had to make labeled training sets for each level of classification (see Table 1 and Fig. 1). Therefore, we hand-coded 200 papers in each level according to the respective goal and then trained a model with these training sets. The papers for annotation were randomly selected from the corpus. At the end of the annotation, we had 4 x 200 = 800 labeled papers for each level of classification. Our very first attempt was to use popular supervised models such as Support Vector Machine (SVM), Kernel Logistic Regression (KLR), or Naïve Bayes (NB). All of these models have failed, most likely because of the small and imbalanced (10–90 ratio) training samples and unusual classification task. After these models, we tried several semi-supervised models [33–36] in order to utilize the unlabeled cases. It was a small step forward in terms of classification performance, but far from ideal. Not only the training times were extremely long, but the accuracy of the semi-supervised models was not that much improvement that we anticipated. Furthermore, the implementation of the models made it tedious to use them effectively. All the aforementioned supervised and semi-supervised models used GloVe embeddings [37] for classification, so we assumed that it may have some effect on the performance of the models. Based on this assumption, we changed to an all-in-one FastText model [38]. FastText is not only extremely efficient and fast, but it has a huge advantage over traditional word-embedding models, since it handles out-of-vocabulary and rare words better [39]. Like GloVe, most embedding techniques create a word vector for each word in the training corpus, hence ignoring the morphological details. FastText, on the contrary, uses character-level vectorization, i.e., creates character n-grams. These character n-grams are then added together to represent the respective word. For example, GloVe gives a 5-dimensional vector representation[4] for the word "imputation" such as [0.34, 0.8, −0.12, −0.45, 0.77]. FastText, on the other hand, creates the word vector as the sum of the following character n-grams[5]: < im, imp, mpu, put, uta, tat, ati, tio, ion, on >. This way, even if "imputation" is not in the model's training corpus, the character n-grams are. FastText's main advantage in our research is its character n-gram

[3] The keywords were: "miss," "missing," "imput," "impute," "imputation," "imputed," "imputing"

[4] The numbers are arbitrary.

[5] If we take n=3.

Table 1 Performance of FastText classification models by classification levels

Level	F1-score	Recall	Specificity	MCC
1	0.98	0.98	0.75	0.73
2	0.94	0.93	0.67	0.57
3.1	0.88	0.88	0.81	0.69
3.2	0.95	0.91	0.42	0.61

Levels: 1—About missing data or not, 2—Imputation or not, 3.1—Advanced Imputation or not, 3.2—Deletion or not

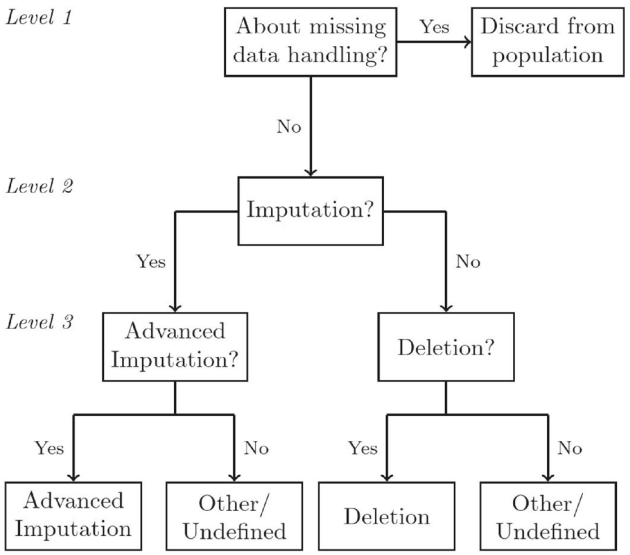

Fig. 1 Levels and stages of classification. The classification level corresponding to Table 1 is indicated on the left side

approach. We found many OCR errors in the article bodies. The FastText model handled these errors better than the GloVe embeddings.

Before discussing the results of our analysis, we briefly present the performance of the classification models at each level. As we have mentioned, we used FastText models in all levels. To measure performance, we used the F1-score, Recall, Specificity, and MCC (Matthews Correlation Coefficient) metrics. It may seem unnecessary to present all of these metrics in order to assess the performance, but our imbalanced training set requires us to consider a more in-depth evaluation.

As we can see from Table 1, the "easiest" task was to decide whether an article was about missing data handling methods or not (Level 1). This coincides with our intuition: If an article discusses missing data handling methods, then it includes a lot of sentences which contain keywords like "imputation" or "missing data." It gives the model more information to identify and distinguish these articles from others. On Level 2, however, we can see that despite the high F1-score and Recall values, the Specificity dropped to 0.67. It means that it was more difficult for the model in this level to find the articles that used imputation. If we recall our preprocessing steps again, we can conclude that the several meanings of the word "imputation" might affect the performance. Level 3.1 was quite consistent: Our model could safely identify if an article used advanced imputation or not. On the contrary, on Level 3.2 the model had more trouble finding the papers that used some kind of deletion technique. This result is consistent with the observation of [11], namely that researchers tend to omit the reporting of deletion in their papers. Peugh and Enders [11] says, moreover, that sometimes only the tables or degrees of freedoms imply that some cases were deleted from the database. Of course, our models are not able to identify such subtle details. To further verify the performance of our models, we conducted a small-scale qualitative evaluation. For this evaluation, we selected 50 papers for each of the three missing data handling methods: imputation, advanced imputation, and deletion. Within each group of 50 papers, we divided them evenly into 25 positive samples (where the respective missing data handling method was used) and 25 negative samples. We then manually reviewed each paper and assigned a class label accordingly. For identifying imputation, the qualitative assessment showed slightly weaker performance (recall: 0.72, Specificity: 0.63). The biggest challenge for our models was distinguishing between the various meanings of "imputation." This issue was particularly prevalent in Economic papers. As a result, the model designed to detect imputation methods made mistakes due to this ambiguity. For identifying advanced imputation, the qualitative assessment score was close to the model result, and it showed good performance (recall: 0.81, Specificity 0.72), and for deletion, the qualitative assessment result was even higher than what we got as the output of machine learning classification (recall: 0.92, Specificity 0.76) Overall, the qualitative evaluation showed that the automatic classification was highly accurate.

6 Corpus description

As a result of the first classification level (about missing data or not), we discarded 1243 papers from our initial corpus (after removing outliers, the corpus consisted of 49.196 papers). The interpretation of imputation is problematic sometimes because of the polysemantic nature of the words "imputation" and "impute." Besides the statistical meaning of "imputation," as per the Cambridge dictionary [40],[6] it has the following meaning: "a suggestion that someone is guilty of something or has a particular bad quality." This definition implies that in certain disciplines a bias could occur. Indeed, the discipline categories "Criminology & Law" and

[6] https://dictionary.cambridge.org/us/dictionary/english/imputation (last accessed 2023.05.28.)

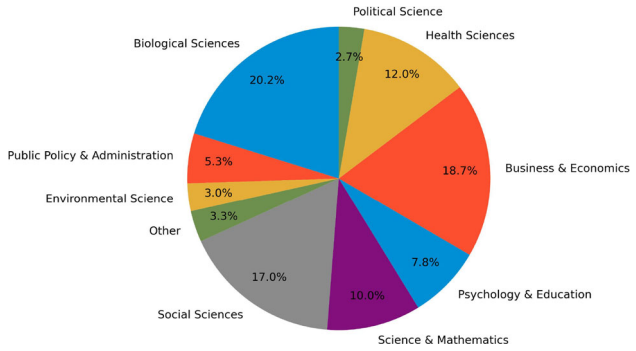

Fig. 2 The distribution of scientific disciplines in our corpus

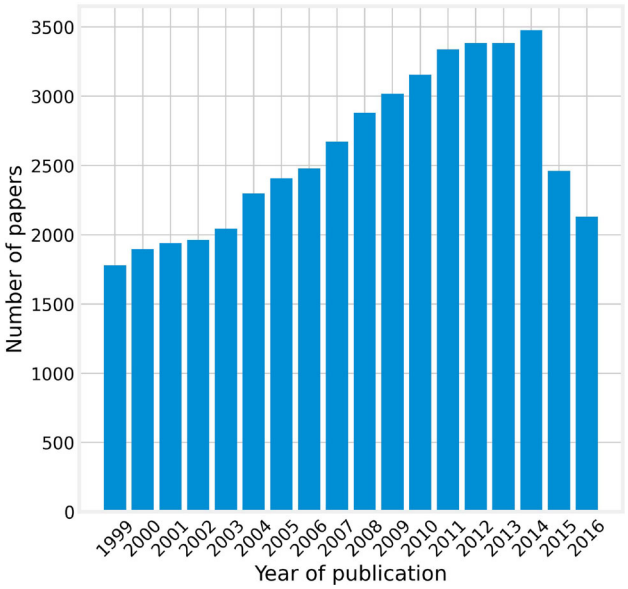

Fig. 3 Number of papers in our corpus by year

"Humanities & Arts" presented high relative frequency of imputation, but low frequency of other missing data handling method. For this reason, we decided to exclude all papers from these two discipline categories from our analysis (2464 papers). Therefore, our actual working corpus consisted of 45.489 articles. We have divided the articles into 12 major discipline categories based on the journal and scientific discipline information from their metadata. The majority of the papers are from Biological Sciences (20.2%), Business & Economics (18.7%), Social Sciences (17.0%), and Health Sciences (12.0%). As we have mentioned before, previous research on missing data handling methods focused mainly on Social- and Educational Sciences; therefore, our research may provide a more widespread perception of the applied missing data techniques. There is a caveat, however, since we do not know exactly which paper used empirical data during their respective research. This is a major difference between ours and the previous studies' approach. Accordingly, we need to assume that there are studies in our corpus that used some kind of empirical data and that our model can identify them. Of course, it is not a plausible assumption in the case of for example Humanities & Arts. One must keep in mind that the distribution of disciplines we show in Fig. 2 is only a description of scientific disciplines in our initial corpus. It does not carry any information about the distribution of missing data handling methods among these academic fields in general. Our results about missing data handling methods cannot be generalized to the full set of articles since we do not know the exact distribution of them.

The time interval of our study was from 1999.01.01. to 2016.12.31. This means that only those articles could get into our corpus that were published in this period. Figure 3 shows the number of papers by publication years in our corpus. There were 22 articles where no publication year was documented.

7 Results

In the first sub-section of the results, we present the temporal trends in how missing data was handled in the selected papers. In the following sub-section, we analyze the variation in missing data handling techniques by disciplines. The last subsection presents the independent and joint effect of publication year and discipline with binomial logistic regression models. We also analyze how the journal's quality correlates with the ratio of the imputation method used.

7.1 Missing data handling methods by year

Overall, the three main categories of our classification were imputation, advanced imputation, and deletion. We focus primarily on advanced imputation, but we also highlight some critical trends from the other two categories. Figure 4 displays the change in the usage of missing data handling methods over years. Since the amount of articles differs year by year, we did not use the raw frequencies. Instead, we made relative frequencies for missing data handling methods in each year (and later, in each discipline). There is a significant increasing trend in the case of imputation and advanced imputation. The usage[7] of advanced imputation methods grew from 2.4% to almost 10% over the years. It even surpassed the relative frequency of deletion methods. The turning point between these two techniques was the period 2009–2011. The change in the usage of imputation methods is similar to the trend of advanced imputation. From 10.3%, it almost reaches 19% at

[7] We do not know whether a missing data handling method was *actually* used—one of the limitations of text mining.

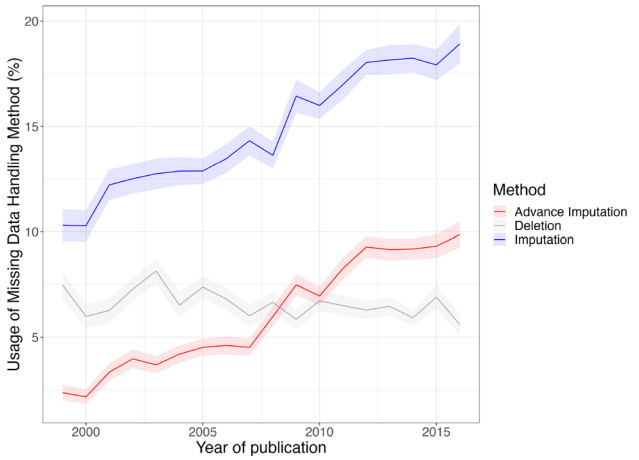

Fig. 4 Usage of missing data handling methods by year (1999–2016) with bootstrap error. The vertical axis shows the percentage of papers that used the respective missing data handling methods in a given year

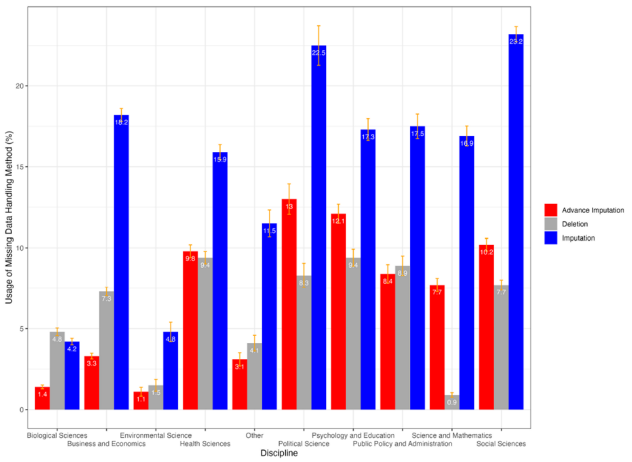

Fig. 5 Usage of missing data handling methods per scientific disciplines with bootstrap error bar

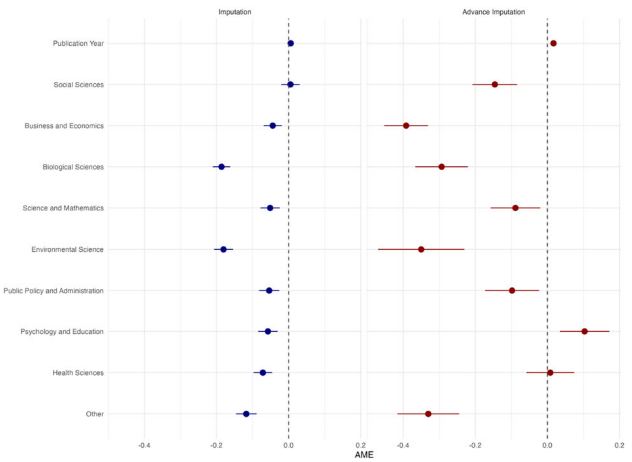

Fig. 6 Predicted average marginal effect (AME) of year and disciplines based on binomial logistic regression model

the end of the interval. The usage of deletion methods is stagnating with a little fluctuation over the period. It constantly stays between 5.8% and 8.1%.

There are several factors that could have influenced the usage of advanced missing data handling techniques over the years. Maybe the most straightforward to assume would be the spread of modern statistical software, packages, and other analytic tools. As the software used in data analysis became more advanced, more missing data handling options were implemented. For example, in the case of the R programming language, the packages MICE and Amelia offer sophisticated and easily applicable methods to deal with missing data [41–43]. This claim is further supported by the observation of [43, p. 83]: *"Both reviews* [referring to the articles of [20] and [44]—note by K.B.] *indicate that there is a considerable gap between statistical methodologies and methods that are commonly used in practice. Flexible comprehensive implementations of these methods may spur their use."* Our findings imply that advancements in implementations of missing data handling techniques may have increased their usage.

7.2 Missing data handling methods by disciplines

Significant differences existed between the disciplines in the missing data management techniques and whether an imputation solution was even mentioned (see Fig. 5). In both Environmental Science and Biological Sciences, there was notably low mention of data imputation techniques. In most fields, simple imputation was the most common method over the whole period, with deletion not being the leading solution in any field. There were also significant differences between the use of simple and advanced imputation. In Business and Economics, the proportion of simple imputation was relatively high, but there were hardly any studies using advanced imputation. In contrast, in Political Science and Health Sciences, the rate of advanced imputation exceeded that of simple imputation over the whole period, and in Psychology and Education, the difference in favor of advanced imputation was markedly significant. This is perhaps not a surprising result because, as mentioned in the introductory chapter, these three fields are the ones from which most methodological studies on data imputation have been carried out.

7.3 Logistic models

To better understand the effect of publication year and disciplines on the use of imputation methods, we fit two binomial logistic regression models on our data: one where the outcome variable was whether there was an imputation or not; and one where the outcome variable was whether there was an

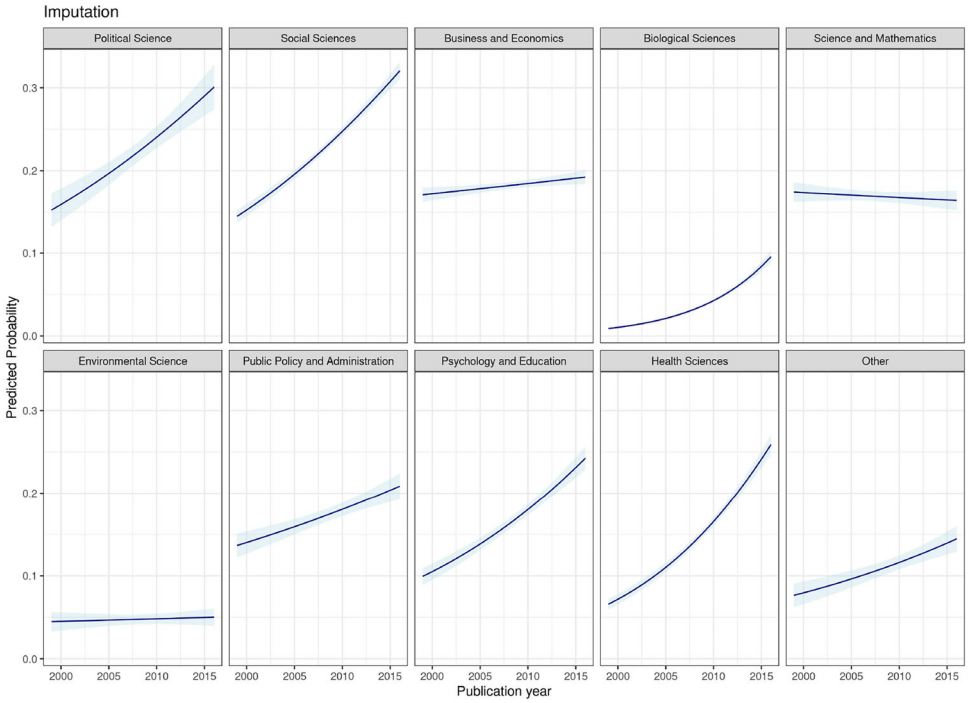

Fig. 7 Predicted marginal probabilities of imputation level per disciplines and year based on binomial logistic regression model

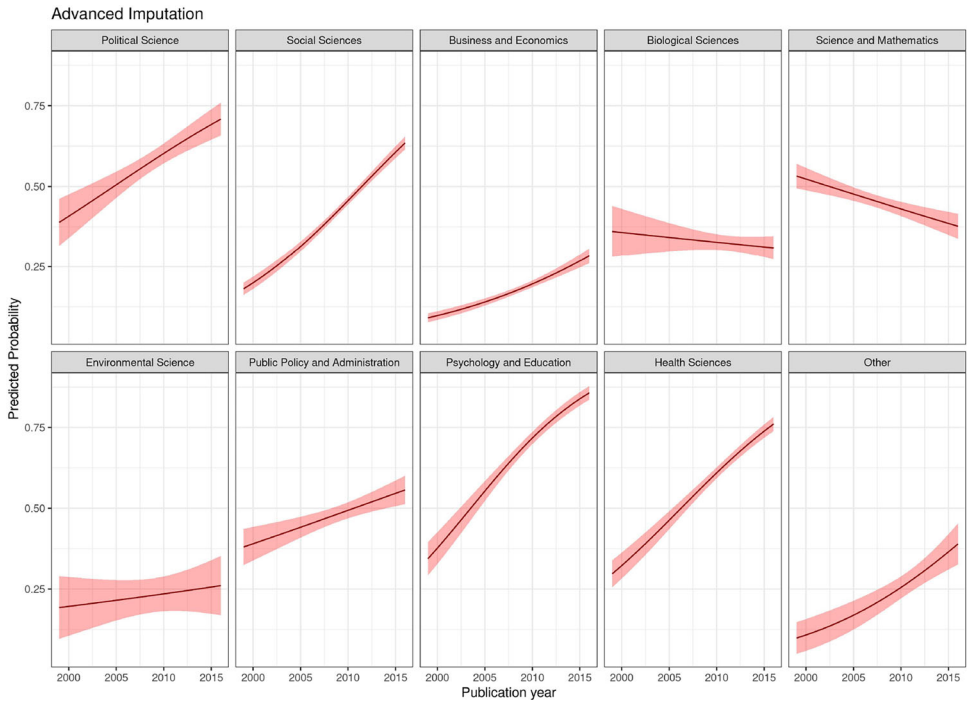

Fig. 8 Predicted marginal probabilities of advanced imputation level per disciplines and year based on binomial logistic regression model

advanced imputation or not[8]. For explanatory variables, we used publication year and discipline category. In the case of discipline category, we used political science as the reference category. We extended the base models with the interaction of time and discipline category to explore whether we find differences in the temporal variation of imputation in different disciplines. We plot the marginal effects of the base models and the predicted probability per year and discipline based on the interactions. Figure 6 shows the predicted average marginal effect (AME) of the year and disciplines based on the binomial logistic regression model. A positive AME value in the figures means that as the variable increases, the

[8] This paper mainly focuses on imputation and advanced imputation, so we did not include the regression models on deletion methods in this sub-section. Model results about deletion methods are available in the supplementary section of the paper.

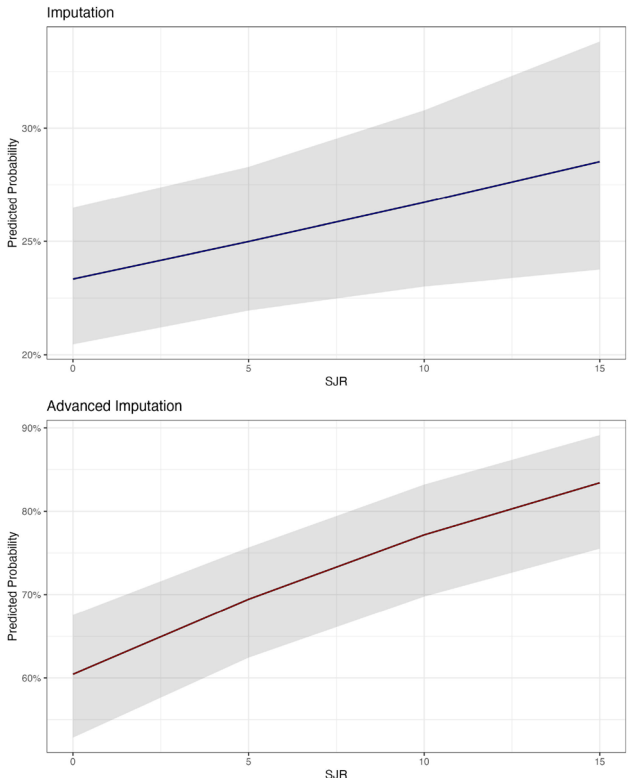

Fig. 9 Predicted marginal probabilities of imputation and advanced imputation level per SJR value based on binomial logistic regression model

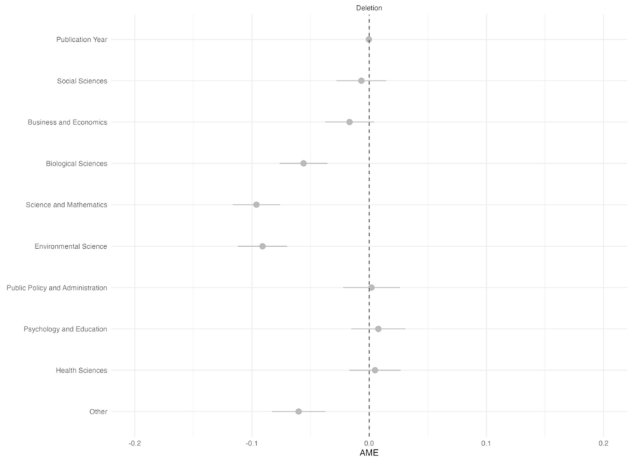

Fig. 10 Predicted average marginal effect (AME) of year and disciplines on deletion based on binomial logistic regression model

probability that imputation or advanced imputation was used increases. Since the disciplines are measured on a 0/1 scale, AME values can be interpreted as effect sizes in their case; the higher the absolute value, the greater the deviation from the reference variable, which in our case were the "political science" papers. If the confidence interval of an AME value "falls within" the dotted line (zero value), then the effect of that variable is not significantly different from zero. Figures 7 and 8 show the predicted marginal probabilities of the dependent variables per discipline and year.

We start the analysis with the imputation part of the regression (see left side of Fig. 6). The time variable was significant with a positive value, which confirms the univariate approach; there was a significant temporal increase in the use of imputation method. Political Science and Social Science papers used most frequently imputation methods. On the other side of the scale, we could find Biological and Environmental Sciences. The interaction presents the different temporal pattern of imputation trends across disciplines (Fig. 7). The figure shows how the imputation ratio has changed over time by discipline. Predicted probability indicates the percentage of papers that used data imputation techniques based on regression models. In political science, for example, the imputation rate has risen from around 40 percent to nearly 75 percent. It is worth comparing this field with other disciplines. The same increasing trend could be observed for Social Science and Psychology and Education. Compared to these fields, Biological and Health Science had a steeper increase in the use of imputation although Biological Sciences started from a very low value. Business and Economics and Science and Mathematics differed negatively from the increasing trend. For the latter based on the marginal predications we could observe a decrease in the use of imputation.

The second regression analyzes the factors behind the variance of advanced imputation level (right side of Fig. 6). Here, we can observe a positive trend value, as expected, so year by year, the advanced imputation was more and more popular. Our results also mean that advanced imputation usage was increased within imputation. However, we can observe high differences between the disciplines. We also used "political science" studies as a reference category in this model. In the Business and Economics, Biological Science, and Environmental Science domains, the use of advanced imputation methods was significantly lower than in the reference category, while in Psychology and Education, the model estimated a higher value than in the political science domain. The interaction terms reveal the temporal differences behind the spreading of advanced imputation (see Fig. 8). Psychology, Education, and Health Sciences had the most intense increase in advanced imputation compared to the other disciplines. Political and social science papers have also recently used advanced imputation techniques more frequently. On the other hand, we could observe a decline in these techniques in Science, Mathematics, and Biological Science.

In the final step of the analysis, we investigated whether higher academy-ranked papers are more likely to use imputation techniques. Two logistic regression models were fitted. In the first model, the dependent variable was whether imputation was used in the analysis, and advanced imputation

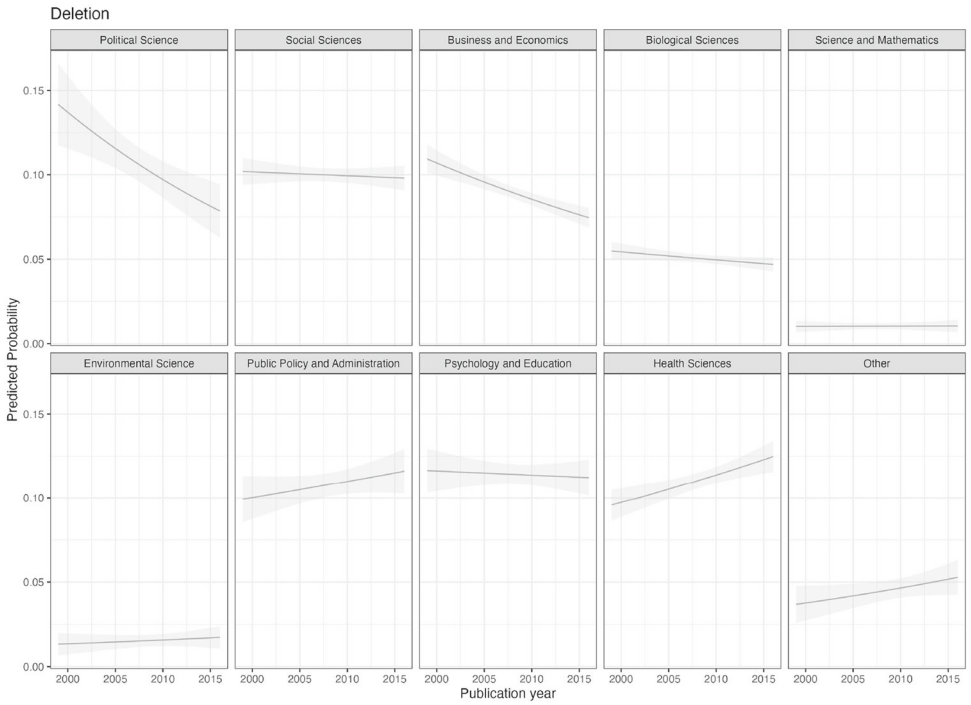

Fig. 11 Predicted marginal probabilities of deletion level per disciplines and year based on binomial logistic regression model

techniques were used in the second. As in the previous model, the independent variables were the year of publication and field of science, supplemented by the 2016 SJR of the papers. In Fig. 9, we have highlighted from the models how the SJR value is related to the probability of using imputation and advanced imputation solution ("predicted probability"). In the context of more general imputation, the SJR value had a weak ($p = 0.02$) positive effect (top part of the figure). Papers in journals with higher SJR had a slightly higher probability of using imputation. The difference between low- and high-ranked papers was much more pronounced for advanced imputation (lower part of the figure). Advanced imputation methods were more than 20 percent more likely to occur in high (SJR$\geq$15) papers than in low (SJR=0) papers. The analysis of the Q value of the journals also confirms this vast difference between low- and high-ranked journals. Advanced imputation was used in seven percent of the studies published in Q4 journals, 18 percent in Q3 papers, 33 percent in Q2, and 48 percent in Q1 journals.

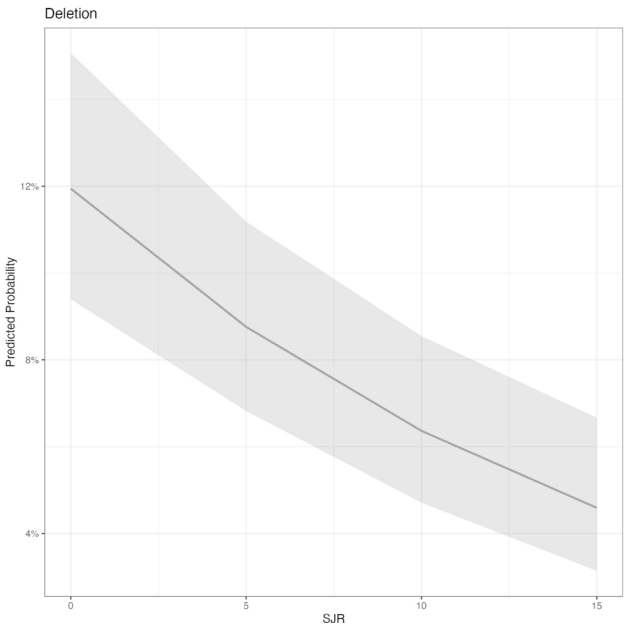

Fig. 12 Predicted marginal probabilities of deletion level per SJR value based on binomial logistic regression model

8 Discussion

The aim of our study was to identify the trends in the usage of missing data handling methods within various scientific disciplines from 1999 to 2016. Missing data is a pivotal element of many empirical research since gathering all the data we originally intended is practically impossible. This immanent information loss problem helped missing data handling methods emerge and evolve. During the past 50 years, more and more techniques have been developed for assessing missing data. Our results show that the usage of imputation and advanced imputation methods increased during the period 1999–2016. One plausible argument to explain this increase is that the documentation of missing data handling methods improved; therefore, it is much more easier to find them in the papers. We think this claim could be one of the

possible causes. The evolution and implementation of these techniques could have boosted their application. More and more statistical software implements complex missing data handling methods which makes these techniques more accessible for researchers. Furthermore, the growing tendency of item non-response in survey-type data collection [45] could also facilitate the usage of missing data handling methods. As mentioned above, it is also obvious that the type of data researchers analyze differs through disciplines. It is not evident that missingness appears at the same level. In Political Science and Social Science surveys are the main quantitative methods, and surveys always contain some level of missing data. From this point of view, it is not surprising that imputation is the most common in these fields. In disciplines where survey data is used, there will be more problems with missing data, so these disciplines have to use imputation more frequently. But we must highlight that we only include papers in this analysis where missing or incomplete data is mentioned. So, we could assume that this disciplinary difference is lower in our sample than in the scientific field. And when we find imputation, we could expect less disciplinary differences between advanced and not advanced techniques. Our result did not support this expectation. The disciplinary difference was huge, and we could also observe small differences between Political and Social Science, where the type of data sources is quite similar. But it is not just the discipline that matters; it is also the journal in which the paper is published. Journals with a higher academic output have a much higher proportion of advanced imputation solutions, which may indicate the higher demands that better journals place on authors. It is also important to evaluate our study from a methodological point of view. In contrast to previous research on this topic, we utilized a text-mining approach to extract the necessary information from the articles. This new methodology, however, comes with some advantages and disadvantages. On the one hand, it allows us to work with a much larger corpus than the previous studies. The actual analysis of the articles is less human resource intensive, and the whole research is very scalable: It does not matter whether we work with ten thousand papers or with one million—the increase in computational time will be negligible. Additionally, with a larger corpus, we are able to make comparisons of missing data handling methods among various disciplines and long time periods. On the other hand, text mining makes us researchers more distant from the papers. Since we have not seen each paper's contents, we can never be sure if a paper used a missing data handling method or only mentioned it—we need to trust the classification. Identifying the used methods was a problem even in previous studies: Often researchers neglected the appropriate documentation of methods they used to handle missing data. And if a human cannot decide whether there was any missing data handling, then how could a classification algorithm. For example, there were several instances, where the authors did not mention which kind of technique they used to handle missing data in their research; only the difference of the samples sizes in the models implied that a deletion technique was used [11]. Clearly, an algorithm is not able to notice such subtle detail, but a human can. All in all, this approach inevitably results in some kind of information loss and the underestimation of imputation usage. Overall, the qualitative evaluation of the machine learning models showed that the automatic classification was highly accurate. We used a limited number of keywords to detect those papers where missingness could be an issue. Our keyword choice might underrepresent some disciplines, where different phrases are also used to describe non-response (like item-non-response in social science). And there are also differences in data-generating processes between fields. Missingness is usually higher in surveys than in experimental designs. But our analysis could well present the temporal trends of applying imputation and advanced imputation within a field. For comprehensive studies of a similar nature, it may be worthwhile to start the evaluation one step earlier and first identify whether empirical (quantitative) data were used in the article and examine the proportion of these articles that used data. This distinction may point to further disciplinary differences. It may also be worthwhile to try models that do not estimate the type of study in separate steps but perform the entire categorization in one step. In our tests, this approach has been less effective. Still, with the development of prediction models and the advance of Large Language Models (LLM), conducting methodological tests along these lines may be worthwhile. However, these classifications are complicated because only the part dealing with imputation is usually very short compared to the entire length of the studies. This difficulty is why we used a data reduction approach in the papers to focus the attention of the models on the relevant texts. This approach may be helpful for other research on similar classification problems.

9 Conclusion

Based on our results, we can state that not only is the data type essential, but disciplines have their methodological canon, which strongly affects how scholars in different fields handle the missing data problem. It is hard to tell anything about general trends in how science deals with missing data. But we can see trends per discipline, most of which are linear. Papers in high-impact journals are the ones that apply the most advanced methodology in the field. And it seems top scholars in some fields have started to apply advanced multiple imputation techniques more and more frequently. Open science and strict pre-registration of studies might also boost this trend shortly. So, we can predict a further increase in using

(advanced) imputation techniques. Our paper also shows how Natural Language Processing (NLP) methods could be used to answer research questions on a large scale. The classification efficiency of NLP methods grows rapidly. This paper used a FastText approach, a state-of-the-art solution when we started this project. But transformer-based NLP models (like BERT or GPT) would offer even higher efficiency in these tasks. These models could solve many issues efficiently and be used for tasks we solved in this paper. These new tools could pave the road for future research in this field.

Appendix A: Supplementary

See Figs. 10, 11, and 12.

Funding Open access funding provided by HUN-REN Centre for Social Sciences. The work of Krisztián Boros was supported by the Japanese Government (Monbukagakusho: MEXT) Scholarship. The research was supported by the European Union within the framework of the RRF-2.3.1-21-2022-00004 Artificial Intelligence National Laboratory Program. The work of Zoltán Kmetty was supported by the Bólyai Scholarship, grant number: BO/834/22.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article

References

1. Dong, Y., Peng, C.-Y.J.: Principled missing data methods for researchers. Springerplus **2**(1), 222 (2013). https://doi.org/10.1186/2193-1801-2-222
2. Enders, C.K.: Applied Missing Data Analysis. Methodology in the social sciences. Guilford Press, New York (2010)
3. Graham, J.W., Cumsille, P.E., Shevock, A.E.: Methods for Handling Missing Data. In: Handbook of Psychology, 2nd edn., pp. 109–141. Wiley, Hoboken, NJ (2013). https://doi.org/10.1002/9781118133880.hop202004
4. Little, T.D., Jorgensen, T.D., Lang, K.M., Moore, E.W.G.: On the Joys of Missing Data. J. Pediatr. Psychol. **39**(2), 151–162 (2014). https://doi.org/10.1093/jpepsy/jst048
5. Little, T.D., Lang, K.M., Wu, W., Rhemtulla, M.: Statistical Issues: What Happens When Data Go Missing? In: Developmental Psychopathology, Third edition edn., p. 37. Wiley, Hoboken, NJ (2016). ISBN: 978-1-118-12179-5
6. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley series in probability and statistics, Third edition Wiley, Hoboken, NJ (2019)
7. Schafer, J.L., Graham, J.W.: Missing data: our view of the state of the art. Psychol. Methods **7**(2), 147–177 (2002). https://doi.org/10.1037/1082-989X.7.2.147
8. Wilkinson, L.: Task force on statistical inference: statistical methods in psychology journals: guidelines and explanations. Am. Psychol. **54**(8), 594–604 (1999)
9. Bell, M.L., Fiero, M., Horton, N.J., Hsu, C.-H.: Handling missing data in RCTs; a review of the top medical journals. BMC Med. Res. Methodol. **14**(1), 118 (2014). https://doi.org/10.1186/1471-2288-14-118
10. Cheema, J.R.: A review of missing data handling methods in education research. Rev. Educ. Res. **84**(4), 487–508 (2014). https://doi.org/10.3102/0034654314532697
11. Peugh, J.L., Enders, C.K.: Missing data in educational research: a review of reporting practices and suggestions for improvement. Rev. Educ. Res. **74**(4), 525–556 (2004). https://doi.org/10.3102/00346543074004525
12. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)
13. Roth, P.L.: Missing data: a conceptual review for applied psychologists. Pers. Psychol. **47**(3), 537–560 (1994). https://doi.org/10.1111/j.1744-6570.1994.tb01736.x
14. Bodner, T.E.: Missing data: prevalence and reporting practices. Psychol. Rep. **99**(3), 675–680 (2006). https://doi.org/10.2466/PR0.99.3.675-680
15. Fernandes-Taylor, S., Hyun, J.K., Reeder, R.N., Harris, A.H.: Common statistical and research design problems in manuscripts submitted to high-impact medical journals. BMC. Res. Notes **4**(1), 304 (2011). https://doi.org/10.1186/1756-0500-4-304
16. King, G., Honaker, J., Joseph, A., Scheve, K.: Analyzing incomplete political science data: an alternative algorithm for multiple imputation. Am. Political Sci. Rev. **95**(1), 49–69 (2001). https://doi.org/10.1017/S0003055401000235
17. Schlomer, G.L., Bauman, S., Card, N.A.: Best practices for missing data management in counseling psychology. J. Couns. Psychol. **57**(1), 1–10 (2010). https://doi.org/10.1037/a0018082
18. Peng, J., Harwell, M., Liou, S.-M., Ehman, L.H.: Advances in missing data methods and implications for educational research. In: Real Data Analysis. Quantitative Methods in Education and the Behavioral Sciences: Issues, Research, and Teaching, pp. 31–78. Information Age Publishing, Charlotte, NC (2006). https://www.researchgate.net/publication/292794490_Advances_in_missing_data_methods_and_implications_for_educational_research
19. Jeličić, H., Phelps, E., Lerner, R.M.: Use of missing data methods in longitudinal studies: the persistence of bad practices in developmental psychology. Dev. Psychol. **45**(4), 1195–1199 (2009). https://doi.org/10.1037/a0015665
20. Burton, A., Altman, D.G.: Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Br. J. Cancer **91**(1), 4–8 (2004). https://doi.org/10.1038/sj.bjc.6601907
21. Wood, A.M., White, I.R., Thompson, S.G.: Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clin. Trials: J. Soc. Clin. Trials **1**(4), 368–376 (2004). https://doi.org/10.1191/1740774504cn032oa
22. Fielding, S., Maclennan, G., Cook, J.A., Ramsay, C.R.: A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. Trials **9**(1), 51 (2008). https://doi.org/10.1186/1745-6215-9-51

23. Gravel, J., Opatrny, L., Shapiro, S.: The intention-to-treat approach in randomized controlled trials: are authors saying what they do and doing what they say? Clin. Trials **4**(4), 350–356 (2007). https://doi.org/10.1177/1740774507081223
24. Hollis, S., Campbell, F.: What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ (Clinical research ed.) **319**(7211), 670–674 (1999). https://doi.org/10.1136/bmj.319.7211.670
25. Khan, N.A., Torralba, K.D., Aslam, F.: Missing data in randomised controlled trials of rheumatoid arthritis drug therapy are substantial and handled inappropriately. RMD Open **7**(2), 001708 (2021). https://doi.org/10.1136/rmdopen-2021-001708
26. Ibrahim, F., Tom, B.D.M., Scott, D.L., Prevost, A.T.: A systematic review of randomised controlled trials in rheumatoid arthritis: the reporting and handling of missing data in composite outcomes. Trials **17**(1), 272 (2016). https://doi.org/10.1186/s13063-016-1402-5
27. Fielding, S., Ogbuagu, A., Sivasubramaniam, S., MacLennan, G., Ramsay, C.R.: Reporting and dealing with missing quality of life data in RCTs: has the picture changed in the last decade? Qual. Life Res. **25**(12), 2977–2983 (2016). https://doi.org/10.1007/s11136-016-1411-6
28. Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., Tabona, O.: A survey on missing data in machine learning. J. Big Data **8**(1), 140 (2021). https://doi.org/10.1186/s40537-021-00516-9
29. Duy Le, T., Beuran, R., Tan, Y.: Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare. In: 2018 10th International Conference on Knowledge and Systems Engineering (KSE), pp. 247–251. IEEE, Ho Chi Minh City (2018). https://doi.org/10.1109/KSE.2018.8573344 . https://ieeexplore.ieee.org/document/8573344/
30. Burns, J., Brenner, A., Kiser, K., Krot, M., Llewellyn, C., Snyder, R.: JSTOR - Data for Research. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science, vol. 5714, pp. 416–419. Springer, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04346-8_48
31. Boros, K., Kmetty, Z.: Identifying missing data handling methods with text mining. Ann Arbor, MI: Inter-Univ. Consort. Political Soc. Res. (2023). https://doi.org/10.3886/E185961V1
32. Feinerer, I., Hornik, K.: tm: Text Mining Package. R package version 0.7-8 (2020)
33. Abdel-Hady, M., Schwenker, F., Palm, G.: Semi-supervised learning for regression with co-training by committee. In: Artificial Neural Networks - ICANN 2009, 19th International Conference, Limassol, Cyprus, September 14–17, 2009, Proceedings, Part I, vol. 5768, pp. 121–130 (2009). https://doi.org/10.1007/978-3-642-04274-4_13
34. Bennett, K.P., Demiriz, A.: Semi-Supervised Support Vector Machines. In: Kearns, M.S., Solla, S.A., Cohn, D.A. (eds.) Advances in Neural Information Processing Systems, vol. 11, pp. 368–374. MIT Press, London (1999)
35. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-supervised Learning. Adaptive computation and machine learning. MIT Press, Cambridge, Mass (2006). ISBN: 978-0-262-03358-9
36. Cohen, I., Cozman, F.G., Sebe, N., Cirelo, M.C., Huang, T.S.: Semisupervised learning of classifiers: theory, algorithms, and their application to human-computer interaction. IEEE Trans. Pattern Anal. Mach. Intell. **26**(12), 1553–1566 (2004). https://doi.org/10.1109/TPAMI.2004.127
37. Pennington, J., Socher, R., Manning, C.: Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (2014). https://doi.org/10.3115/v1/D14-1162
38. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. (2016). arXiv: 1607.01759
39. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. (2017). arXiv: 1607.04606
40. Cambridge Dictionary. Cambridge: Cambridge University Press (2020). https://dictionary.cambridge.org/dictionary/english/imputation. Accessed 2023-05-28
41. Buuren, S.V., Groothuis-Oudshoorn, K.: mice: Multivariate Imputation by Chained Equations in R. J. Stat. Softw. **45**, 1–67 (2011). https://doi.org/10.18637/jss.v045.i03
42. Honaker, J., King, G., Blackwell, M.: Amelia II A Program for Missing Data. J. Stat. Softw. (2011). https://doi.org/10.18637/jss.v045.i07
43. Horton, N.J., Kleinman, K.P.: Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am. Stat. **61**(1), 79–90 (2007). https://doi.org/10.1198/000313007X172556
44. Horton, N.J., Switzer, S.S.: Statistical Methods in the Journal (research letter). N. Engl. J. Med. **353**, 1977–1979 (2005)
45. Luiten, A., Hox, J., Leeuw, E.: Survey nonresponse trends and fieldwork effort in the 21st century: results of an international study across countries and surveys. J. Off. Stat. **36**(3), 469–487 (2020). https://doi.org/10.2478/jos-2020-0025