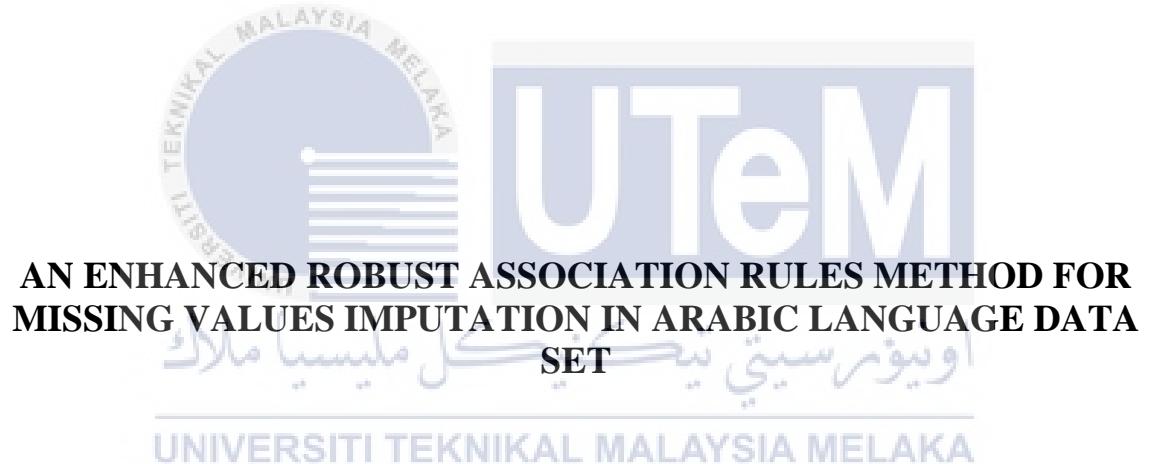




Faculty of Information and Communication Technology



Awsan Thabet Salem

Doctor of Philosophy

2023

**AN ENHANCED ROBUST ASSOCIATION RULES METHOD FOR MISSING
VALUES IMPUTATION IN ARABIC LANGUAGE DATA SET**

AWSAN THABET SALEM



UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2023

DECLARATION

I declare that this thesis entitled “An Enhanced Robust Association Rules Method for Missing Values Imputation in Arabic Language Data set” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.



Signature :

Name : Awsan Thabet Salem

Date : 02/06/2023

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of Doctor of Philosophy.


Signature :
Supervisor Name : Associate Professor Dr. Nurul Akmar Emran
Date : 02/06/2023


اونيورسيتي تيكنيكل مليسيا ملاك
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

DEDICATION

To my beloved mother, father, and my family



ABSTRACT

In data quality, missing values is one form of data completeness problem faced by people who deal with data. The failure to handle missing values usually causes unwanted consequences such as misleading analysis and decision-making. Thus, to deal with missing values, data imputation methods were proposed with the aim of improving the completeness of the data sets of concern. Data imputation's accuracy is a common indicator of a data imputation method's efficiency. However, the efficiency of data imputation in nominal data sets can be affected by the nature of the language in which the data set is written. Thus, there is a pressing need to deal with the problem, especially in non-Latin languages such as the Arabic language. In this thesis, the Enhanced Robust Association Rules (ERAR) method for missing values imputation is proposed. ERAR will improve the way to handle the Arabic language's complexity in terms of morphology and misspellings by adding an Arabic preparation step. The preparation step consists of Normalization, Error Detection, and Error Correction processes. ERAR is an extension of the Iterative method that adds filtering of frequent items. This method deals with high missing value rates by adjusting the support threshold in every iteration of the algorithm. This research aims to test the hypothesis that Arabic preparation and the filtering steps will improve the imputation processes in terms of accuracy, speed, and memory used. The findings discovered that with different missing value rates, ERAR was able to offer the highest accuracy percentage value reached 99% in the Arabic poetry data set, and speed as compared to the Iterative method in English and Arabic data sets at most MV rates, unfortunately not against the DT method. Nevertheless, the ERAR consumed the highest memory usage as compared to other methods during the imputation processes. In threshold values, the ERAR, Iterative methods are affected by different threshold values, where the accuracy decreases by reducing the support values, the same goes for elapsed time. In terms of memory usage, there is no clear effect. In the future, the research can be extended by covering the numerical data and other Arabic language issues. There is also room to improve ERAR in terms of memory use and speed.

KAEDAH PETUA SEKUTUAN TEGAP YANG DIPERTINGKATKAN UNTUK IMPUTASI NILAI HILANG DALAM DATA SET BAHASA ARAB

ABSTRAK

Dalam kualiti data, nilai yang hilang adalah salah satu bentuk masalah kesempurnaan data yang dihadapi oleh orang yang berurusan dengan data. Kegagalan untuk mengendalikan nilai yang hilang biasanya menyebabkan akibat yang tidak diinginkan seperti analisis yang mengelirukan dan membuat keputusan. Oleh itu, untuk menangani nilai yang hilang, kaedah imputasi data telah dicadangkan dengan tujuan untuk meningkatkan kesempurnaan set data yang menjadi perhatian. Ketepatan imputasi data ialah penunjuk biasa bagi kecekapan kaedah imputasi data. Walau bagaimanapun, kecekapan imputasi data dalam set data nominal boleh dipengaruhi oleh sifat bahasa di mana set data ditulis. Oleh itu, terdapat keperluan mendesak untuk menangani masalah tersebut, terutamanya dalam bahasa bukan Latin seperti bahasa Arab. Dalam tesis ini, kaedah Enhanced Robust Association Rules (ERAR) untuk imputasi nilai hilang dicadangkan. ERAR akan menambah baik cara mengendalikan kerumitan bahasa Arab dari segi morfologi dan salah ejaan dengan menambah langkah penyediaan bahasa Arab. Langkah penyediaan terdiri daripada proses Normalisasi, Pengesanan Ralat dan Pembetulan Ralat. ERAR ialah lanjutan daripada kaedah Iteratif yang menambah penapisan item yang kerap. Kaedah ini menangani kadar nilai hilang yang tinggi dengan melaraskan ambang sokongan dalam setiap lelaran algoritma. Penyelidikan ini bertujuan untuk menguji hipotesis bahawa penyediaan bahasa Arab dan langkah-langkah penapisan akan meningkatkan proses imputasi dari segi ketepatan, kelajuan, dan ingatan yang digunakan. Penemuan mendapati bahawa dengan kadar nilai hilang yang berbeza, ERAR mampu menawarkan nilai peratusan ketepatan tertinggi mencapai 99% dalam set data puisi Arab, dan kelajuan berbanding kaedah Iteratif dalam set data Inggeris dan Arab pada kebanyakan kadar MV, malangnya, tidak bertentangan dengan kaedah DT. Namun begitu, ERAR menggunakan penggunaan memori yang paling tinggi berbanding kaedah lain semasa proses imputasi. Dalam nilai ambang, kaedah ERAR, Iteratif dipengaruhi oleh nilai ambang yang berbeza, di mana ketepatan berkurangan dengan mengurangkan nilai sokongan, perkara yang sama berlaku untuk masa berlalu. dari segi penggunaan memori, tiada kesan yang jelas. Pada masa hadapan, penyelidikan boleh diperluaskan dengan meliputi data berangka dan isu bahasa Arab yang lain. Terdapat juga ruang untuk menambah baik ERAR dari segi penggunaan memori dan kelajuan.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the almighty ALLAH swt, for giving me the strength, courage, health, creativity and ability to complete this thesis successfully, I am would like to extremely grateful to my supervisors, Associate Professor Dr. Nurul Akmar Binti Emran and Professor Dr. Azah Kamilah Draman for their invaluable advice, continuous support, and patience during my Ph.D. study. Their immense knowledge and experience have encouraged me throughout my academic research and daily life.

Finally, I would like to express my gratitude to my parents, wife, children, brothers and sisters. Without their tremendous understanding and encouragement over the past few years, it would be impossible for me to complete my study.

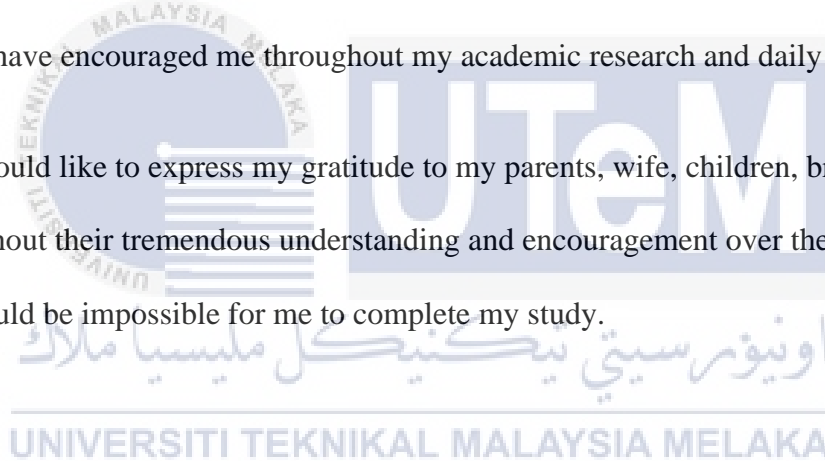


TABLE OF CONTENTS

	PAGE
DECLARATION	
APPROVAL	
DEDICATION	
ABSTRACT	i
ABSTRAK	ii
ACKNOWLEDGEMENTS	iii
TABLE OF CONTENTS	iv
LIST OF TABLES	vii
LIST OF FIGURES	ix
LIST OF APPENDICES	xii
LIST OF ABBREVIATIONS	xiii
LIST OF PUBLICATIONS	xiv
 CHAPTER	
1. INTRODUCTION	1
1.0 Chapter Overview	1
1.1 Background	1
1.1.1 Arabic Language	3
1.1.2 Ways to Deal with Missing Values	5
1.1.3 The Limitations of AR-based Missing Values Imputation Algorithms	8
1.2 Research Problems and Research Questions	10
1.3 Research Objectives	11
1.4 Research Motivation	12
1.5 Research Contributions	12
1.6 Scope of research	13
1.7 Thesis Outline	13
 2. LITERATURE REVIEW	17
2.0 Chapter Overview	17
2.1 Background of Arabic language	17
2.1.1 Orthography	18
2.1.2 Morphology in Arabic	22
2.1.3 Correction in Arabic language	25
2.1.4 Arabic Text Error Detection and Correction Methods	26
2.2 Background on Missing values	30
2.2.1 Causes of Missing Values	31
2.2.2 Missing Values Mechanisms	32
2.2.3 Missing Values Pattern	35
2.3 Methods to Handle Missing Values	39
2.3.1 Deletion	39
2.3.2 Imputation	42
2.3.2.1 Single imputation (SI)	44
2.3.2.2 Multiple Imputation (MI)	51
2.3.3 Association Rules (AR)	52
2.3.3.1 Interestingness Measures of Association Rules	55
2.3.3.2 Association Rule Mining Problems	56

2.3.3.3	Solutions for the Association Rules problems	56
2.3.3.4	Association Rules in Missing Values Imputation	65
2.4	Discussion on Research Gaps	79
2.5	Conclusion	82
3.	RESEARCH METHODOLOGY	89
3.0	Chapter Overview	89
3.1	Research Design	89
3.1.1	Investigation Phase	91
3.1.2	Implementation Phase	92
3.2	Operational Procedure	97
3.2.1	Data Collection	98
3.2.2	Data Preparation	103
3.2.2.1	General Data set Preparation	103
3.2.2.2	Arabic Data set Preparation	104
3.2.3	Implementation of ERAR	110
3.2.3.1	Threshold Value	111
3.2.3.2	Filtering Frequent Items	112
3.2.3.3	Frequent Itemset Mining	113
3.2.3.4	Rules Generation	114
3.2.4	Missing Values Imputation	115
3.3	Evaluation of the method	116
3.3.1	Construct Hypothesis	117
3.3.2	Experiment Configuration Setup	117
3.3.3	Experiment Execution and Analysis	118
3.3.4	Experiment Result Analysis and Evaluation	118
3.3.5	Statistical Evaluation	120
3.4	Conclusion	124
4.	DEVELOPMENT OF THE ENHANCED ROBUST ASSOCIATION RULES (ERAR)	125
4.0	Chapter Overview	125
4.1	Experimental environment setup	125
4.2	Data Preparation	126
4.2.1	General data set preparation	126
4.2.2	Arabic data set preparation	127
4.2.2.1	Normalization	128
4.2.2.2	Error Detection	131
4.2.2.3	Error Correction	133
4.3	Implementation of ERAR method	138
4.3.1	Threshold value setup	138
4.3.2	Filtering and selecting the frequent itemsets	139
4.3.3	Rules Generation	140
4.3.4	Missing values imputation	141
4.4	Implementation Results	141
4.4.1	MV Imputation Accuracy	142
4.4.2	Imputation Performance: Elapsed Time Results	145
4.4.3	Imputation Performance: Memory Usage	148
4.5	Conclusion	152

5. RESULTS AND DISCUSSION	154
5.0 Chapter Overview	155
5.1 MV Imputation Accuracy	154
5.2 Statistical Evaluation	158
5.2.1 Correlation Coefficient Accuracy vs. Elapsed time	158
5.2.2 Correlation Coefficient Accuracy vs. Memory usage	161
5.3 Discussion	164
5.4 Conclusion	171
6. CONCLUSION AND FUTURE WORKS	173
6.0 Conclusions	174
6.1 Research Contributions	175
6.1.1 An enhanced missing values imputation algorithm that is not limited to the Arabic data set	175
6.1.1.1 Data preparation	175
6.1.1.2 Filtering frequent itemsets	176
6.1.2 Experimental evaluation of missing values imputation method	176
6.2 Limitations and future works	177
REFERENCES	179
APPENDICES	192



LIST OF TABLES

TABLE	TITLE	PAGE
1.1	Diacritics Arabic Letters (Saad, 2010)	4
1.2	Research Outline	15
2.1	Arabic Letters (Algarni, 2016)	21
2.2	Nominal Suffix (Algarni, 2016)	22
2.3	Current Arabic Stemmers Features and Limitations (Algarni, 2016)	24
2.4	Confusion Matrix (Hasan et al., 2021)	42
2.5	List Of The Commonly Applied Metrics (Hasan et al., 2021)	44
2.6	Example Of Basket Market Transaction (Tan, Kumar And Steinbach, 2018)	55
2.7	Comparison Of Ar Algorithms In Mv Imputation (Chavan And Verma, 2013)	79
2.8	Related Work	85
3.1	Summary Of The Subjects Under Study in The Investigation Phase	91
3.2	an Overall Research Strategy	93
3.3	Data Sets Size	99
3.4	Zomato Restaurant's Attributes	100
3.5	Yellow Pages's Attributes	101
3.6	Arabic Poetry's Data Set Attributes	102
3.7	List Of Prefixs	107
3.8	List Of Suffixs	107
3.9	Arabic Keyboard Letters	109
3.10	Similarly Sound Arabic Letters	110
4.1	Punctuation Marks	126
4.2	Sample of Zomato Restaurants Data Set	127
4.3	Sample of Poems Data Set	129
4.4	Tokenization One Word	129

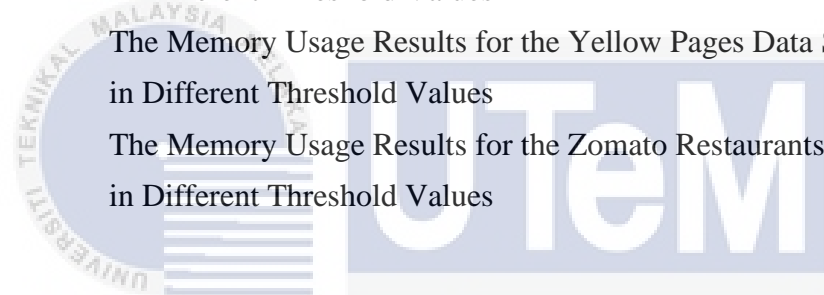
4.5	Tokenization More Than One Word	130
4.6	Normalization Steps	131
4.7	an Example The Results Of The Stemming Process by Affixes Removal	132
4.8	an Example Of Frequents Words In A Data Set	133
4.9	Example On Damerau–Levenshtein Result	134
4.10	Example On Distance Value Keyboard Related	135
4.11	an Example Of Sound Similarity Results	136
4.12	Selection Of The Correct Word	137
4.13	The Confusion Matrix Results for The Arabic Poetry Data Set	143
4.14	The Confusion Matrix Results for The Restaurants on Yellow Pages Data Set	144
4.15	The Confusion Matrix Results for The Zomato Restaurants Data Set	145
4.16	The Average PI Results for Elapsed Time Using The Arabic Poetry Data Set	146
4.17	The Average PI Results Using The Restaurants on Yellow Pages Data Set	147
4.18	The Average Pi Results Using The Zomato Restaurants Data Set	148
4.19	The Memory Used And Pi of The Arabic Poetry Data Set	149
4.20	The Memory Used And Pi of The Yellow Pages Data Set	150
4.21	The Average PI for Memory Usage Results Using The Zomato Restaurants Data Set	151
5.1	Results Of Correlation Coefficient Accuracy Vs. Elapsed Time in the Arabic Poetry Data Set	159
5.2	Correlation Coefficient Accuracy Vs. Elapsed Time in The Restaurants on Yellow Pages Data Set	160
5.3	Results of Correlation Coefficient Accuracy Vs. Memory Usage in the Arabic Poetry Data Set	162
5.4	Results Of Correlation Coefficient Accuracy Vs. Memory Usage in the Restaurants on Yellow Pages Data Set	163
5.5	Confusion Matrix Results for the Yellow Pages Data Set in Different Threshold Values	167
5.6	The Confusion Matrix Results for The Zomato Restaurants Data Set In Different Threshold Values	167

LIST OF FIGURES

FIGURE	TITLE	PAGE
1.1	History Line for AR Algorithm in MV Imputation	8
2.1	Distribution of Spelling Mistakes in Arabic (Said et al., 2013)	26
2.2	Univariate Pattern (Baraldi and Enders, 2010)	36
2.3	Unit Nonresponse Pattern (Baraldi and Enders, 2010)	36
2.4	Monotone Pattern (Baraldi and Enders, 2010)	37
2.5	Arbitrary Pattern (Baraldi and Enders, 2010)	37
2.6	Planned Mv Pattern (Baraldi and Enders, 2010)	38
2.7	Latent Variable Pattern (Baraldi and Enders, 2010)	39
2.8	Listwise Deletion	40
2.9	Pairwise Deletion	41
2.10	Multiple Imputation Method Steps (Chhabra, 2017)	52
2.11	a 2-Way Contingency Table for Variables X And Y (Abdur Rahman, 2012)	55
2.12	Counting the Support of Candidate Itemsets (Tan, Kumar and Steinbach, 2018)	57
2.13	Illustration of Frequent Itemset Generation Using The Apriori Algorithm (Tan, Kumar and Steinbach, 2018)	59
2.14	Construction of an FP-Tree (Tan, Kumar and Steinbach, 2018)	61
2.15	an FP Representation for The Data Set Shown in 2.14 With a Different Item Ordering Scheme (Tan, Kumar and Steinbach, 2018)	62
2.16	Example for Aprioritid (Agrawal and Srikant, 1994)	63
2.17	Elapsed Time for AR Algorithms (Nigam, Nigam and Dalal, 2017)	65
2.18	Memory Used for AR Algorithms (Nigam, Nigam and Dalal, 2017)	65

2.19	Flow Chart to Show Filling of Missing Values (Chavan and Verma, 2013)	66
2.20	Pseudocode RAR Algorithm (Agrawal et al., 1996)	67
2.21	RAR Approach for Dealing With Mv (Ragel and Cremilleux, 1999)	69
2.22	Correction of Supports for Missing Values (Ragel and Cremilleux, 1999)	69
2.23	Robust Association Rules Performance (Ragel, 1998)	70
2.24	Completion of Prediction (Shen and Chen, 2003)	72
2.25	The Comparison of The Association Rules Number For MVC and FRCAR (Shen, Chang and Li, 2007)	73
2.26	The Comparison of Recovery Precision for MVC and FRCAR (Shen, Chang and Li, 2007)	74
2.27	The Comparison of Running Times for MVC and FRCAR (Shen, Chang and Li, 2007)	74
2.28	Comparison of Accuracy Between Iterative and MVC Algorithms	77
3.1	Research Design	90
3.2	The Proposed Operational Framework	97
3.3	Data Source Types	99
3.4	Arabic Data Set Preparation Modules	104
3.5	Similarities and Differences Between Iterative and ERAR Methods	111
3.6	Experimental Flow	116
4.1	Tokenization Steps	129
4.2	Selection of Frequent 1-Itemsets in Rows	140
5.1	F-Measure Scores for ERAR, Iterative and DT Methods for The Arabic Poetry Data Set	155
5.2	F-Measure Scores for ERAR, Iterative and DT Methods for The Restaurants on Yellow Pages Data Set	156
5.3	F-Measure Scores for ERAR, Iterative and Dt Methods for the Zomato Restaurants Data Set	157
5.4	Scatter Plot of Correlation Coefficient Accuracy Vs. Elapsed Time In the Arabic Poetry Data Set	159
5.5	Scatter Plot of Correlation Coefficient Accuracy Vs. Elapsed Time In the Yellow Pages Data Set	160

5.6	Scatter Plot of Correlation Coefficient Accuracy Vs. Memory Usage in the Arabic Poetry Data Set	162
5.7	Scatter Plot of Correlation Coefficient Accuracy Vs. Memory Usage in the Restaurants on Yellow Pages Data Set	163
5.8	Results of TNR And FPR on Arabic Poetry Data Set	165
5.9	Results of TNR And FPR Restaurants on Yellow Pages Data Set	166
5.10	F-Measure Scores for the Restaurants on Yellow Pages Data Set in Different Threshold Values	168
5.11	F-Measure Scores for the Zomato Restaurants Data Set in Different Threshold Values	168
5.12	The Elapsed Time Results for the Yellow Pages Data Set in Different Threshold Values	169
5.13	The Elapsed Time Results for the Zomato Restaurants Data Set in Different Threshold Values	170
5.14	The Memory Usage Results for the Yellow Pages Data Set in Different Threshold Values	170
5.15	The Memory Usage Results for the Zomato Restaurants Data Set in Different Threshold Values	171



اوتيمر ستي تيكنيكل مليسيا ملاك

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	Confusion matrix results for the Arabic poetry data set	192
B	Confusion matrix results for the Yellow pages data set	196
C	Confusion matrix results for the Zomato Restaurants data set	200



LIST OF ABBREVIATIONS

AR	-	Association Rules
DT	-	Decision Tree
ERAR	-	Enhanced Robust Association Rules
FN	-	False negatives
FP	-	False Positives
MV	-	Missing values
MVC	-	Missing Values Completion
NLP	-	Natural Language Processing
RAR	-	Robust Association Rules
TN	-	True negatives
TP	-	True Positives

LIST OF PUBLICATIONS

1. Awsan Thabet, Nurul A. Emran, Azah k. Muda, Zahriah Sahri, Abdulrazzak Ali., 2022. Missing Values Imputation in Arabic Data sets Using Enhanced Robust Association Rules (ERAR). *Indonesian Journal of Electrical Engineering and Computer Science*, 28(2), pp.1067-1075.
2. Awsan Thabet, Nurul A. Emran, Azah K. Muda., 2020. Enhanced robust association rules (ERAR) method for missing values imputation. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(4), pp.6036-6042.
3. Ali, Abdulrazzak, Nurul A. Emran, and Siti A. Asmai, and Thabet, Awsan., 2018. Duplicates Detection Within Incomplete Data Sets Using Blocking and Dynamic Sorting Key Methods. *International Journal of Advanced Computer Science and Applications*, 9(9), pp.629-636.

اوتنورسيتي تيكنيكل مليسيا ملاك

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

CHAPTER 1

INTRODUCTION

1.0 Chapter Overview

This chapter begins with Section 1.1 that consists of the background of missing values, the nature of Arabic data and its challenges, techniques used to deal with missing values, and limitations of Association Rules (AR) imputation techniques. Section 1.2 consists of the research problems and questions, whereas Section 1.3 consists of the motivation for this thesis. Section 1.4 presents the objectives of the research before research contributions in Section 1.5. The scope of this research can be found in Section 1.6, and finally, Section 1.7 has the description of the remaining chapters of this thesis.

1.1 Background

Usually, data is contaminated at the source. Data that has impurities like duplication, misspellings, and missing values are referred to as "dirty" in the context of data quality (MVs). The ratio of impurities in data sets varies and factors such as failure of monitoring, faults in data input process, equipment errors, interference with data collectors' ability to communicate with the central management system, and failure of the archiving system (hardware or software) or human errors contributing to the problem (Kaiser, 2014; Emmanuel et al., 2021). Dirty data must be cleaned before it can be useful in decision-making or analysis for an organization. In fact, the quality of data determines the analysis quality (Russom, 2011; Suthar et al., 2012; Zhang, and Thorburn, 2022).

MV is an example of a data completeness problem that causes dirty data. It has been a historical problem for decades of data analysis and a common problem in many fields such as Business, Industry, Healthcare, and E-Government (Zainuri, Jemain, and Muda, 2015). MV occurs when no values of data are stored for the variable in an observation (Suthar, Patel and Goswami, 2012). MV is a typical occurrence in many applications and can significantly affect the results that can be drawn from the data. MV is one of the data quality problems from the data completeness dimension.

MV usually occurs in many forms such as “NULL” values in the database or as empty cells in the spreadsheet. Some flat-file formats use various symbols for MV, such as Attribute-Relation File Format (arff) files uses “?” symbol for MV. No matter what the reason is, MV is a big problem faced by many statistical areas (Allison, 2001). A serious problem that occurs when handling the gap between the historical and the present data. For example, merging two or more data sets that have different structures (number of attributes) and records, causing the occurrence of MVs.

The effects of MV can lead to significant issues. First, cases with missing data are typically immediately excluded by statistical techniques. In the end, there might not be enough information to conduct the analysis. One could hardly, for instance, perform a factor analysis on a small sample of cases. Second, due to the limited amount of input data, the analysis might be performed but the outcomes might not be statistically significant. Third, if the examples being studied are not picked at random, the results might be deceptive. These consequences are not restricted to English data sets but also to Arabic data sets or other languages (Brown and Kros, 2003; Godfrey and Loots, 2014).

Addressing MV is essential to avoid unwanted consequences that affect the quality of the analysis. Data imputation is a term used to replace the MV with plausible values in the data set (Suthar, Patel and Goswami, 2012; Chhabra, 2017; Abidin, Ismail and Emran,

2018). Additionally, imputation is used to finish data sets in order to raise their quality. Due to Natural language processing, the imputation is a complex task, especially in the presence of noise (for example. misspellings and morphology complexity) (Alkhatib, Monem and Shaalan, 2020).

Most MV imputation approaches suffer from severe limitations. They are almost exclusively restricted to numerical data, and they either offer only simple imputation methods or are difficult to scale and maintain in production (Biessmann et al., 2018). Generally, most modern machine language (ML) applications that involve text data are based on rather sophisticated natural language models. Combinations of such models with tabular data are an important field of research, but they are beyond the scope of most imputation research so far (Yin et al., 2020; Jäger, Allhorn and Biebmann, 2021).

From that perspective, this research investigates the problem of MV in nominal data sets, especially in the Arabic language data sets. This thesis attempts to develop a method by using the best features of the current approaches while at the same time minimizing their current drawbacks in handling the Arabic language issues. The next section provides an elaboration of the challenges in handling the Arabic language.

1.1.1 Arabic Language

Over 400 million people worldwide are native Arabic speakers (Kourdi, Bensaid and Rachidi, 2004; Wikipedia, 2018). One of the most commonly used mother tongues in the world, it is the primary language in several Arab states and is spoken as a second language in many other nations. In contrast to Latin-based alphabets, Arabic is written from right to left. Arabic words contain two genders, feminine and masculine, three grammatical cases nominative, accusative, and genitive as well as singular, dual, and plural forms. When a noun is the subject of a verb, it is in the accusative case; when it is the object of a preposition, it

is in the genitive case. There are three main categories of words: nouns (which include adjectives and adverbs), verbs, and particles (Al-Harbi et al., 2008).

Arabic is based on an alphabetical system that uses 28 basic letters. The following basic 28 letters make up Arabic alphabets (ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك أ) and Hamza (ء). Arabic letters differ from English ones where there is no upper or lower case. Arabic writing is distinctive where its orientation is from right to left (Al-Harbi et al., 2008; Duwairi, Al-refai and Khasawneh, 2009; Saad, 2010).

Arabic language has symbols called diacritical marks, or simply diacritics, which are also known as Harakat (Arabic name). The fundamental objective of Harakat is to offer a phonetic manual or help. There are several Arabic diacritics, including Fatha, Kasra, Damma, Sukn, Shadda, and Tanwin (Zitouni, Sorensen and Sarikaya, 2006). Table 1.1 shows the examples of the aforementioned diacritics pronunciation for the Arabic letter (ب).

Table 1.1: Diacritics Arabic Letters (Saad, 2010)

Double Constant	Not Vowel	Nunation			Vowel		
بْ /bb/	بُ /b/	بِ /bin/	بُنْ /bun/	بَنْ /ban/	بِي /bi/	بُو /bu/	بَا /ba/

Arabic is difficult to learn for a variety of reasons (Alansary, Nagi and Adly, 2007; Al-Harbi et al., 2008; Kwaik et al., 2018; Atwan et al., 2021). Which can be stated as follows:

- Orthography with diacritics is more phonetic and less confusing in Arabic; some character combinations can be expressed in various ways, such as [ة] , [هـ] and [ت] , [ة].
- Compared to the English language, Arabic morphology is extremely complicated such as prefixes, affix, and suffixes where one word comes in different forms.

- iii. Synonyms are common in the Arabic language, and it is considered as highly inflectional and derivational language.

The challenges in the Arabic language stated above pose unique challenges in the data imputation process in practical for nominal data sets. Precisely, the process that will be affected is the matching process during the computation of frequent itemsets, which leads to the production of inaccurate data. Therefore, there is a need to overcome data imputation issue caused by the complexity of the language.

1.1.2 Ways to Deal with Missing Values

Several techniques are proposed in handling MV based on the type of their MV mechanism. These techniques are classified into two approaches which are case deletion and imputation (Uenal, Mayer and du prel, 2014; Mirzaei et al., 2022).

Deletion methods, are the most common way to deal with MV. The most common techniques for handling MV are Listwise and Pairwise deletion (Baraldi and Enders, 2010; Emmanuel et al., 2021). With Listwise deletion, a record will be removed if one or more MV are found within the record. When data are Missing Completely at Random (MCAR), only unbiased parameter estimates will be produced by Listwise deletion, where neither observed data nor unobserved data is dependent on the MV of an attribute. Listwise deletion causes high data loss especially when the proportion of records with MV is high. To minimize the loss caused by listwise deletion, pairwise deletion is proposed. By eliminating instance variables (contains MV) instead of the entire record, this strategy is intended to decrease the number of cases that are eliminated in any given analysis. This will reduce the ratio of data lost with Listwise technique. Deletion methods will be elaborated further in Chapter 2.