The copyright © of this thesis belongs to its rightful author and/or other copyright owner. Copies can be accessed and downloaded for non-commercial or learning purposes without any charge and permission. The thesis cannot be reproduced or quoted as a whole without the permission from its rightful owner. No alteration or changes in format is allowed without permission from its rightful owner.

AN ENHANCED TERM WEIGHTING SCHEME METHOD OF IDENTIFYING AND EXTRACTING TERMS FOR ONTOLOGY LEARNING AND DEVELOPMENT

MUHAMMAD BIN MAHMUD

Universiti Utara Malaysia

DOCTOR OF PHILOSOPHY
UNIVERSITI UTARA MALAYSIA
2023

Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

MUHAMMAD MAHMUD

calon untuk Ijazah
(candidate for the degree of) **PhD**

telah mengemukakan tesis / disertasi yang bertajuk:
(has presented his/her thesis / dissertation of the following title):

"AN ENHANCED TERM WEIGHTING SCHEME METHOD OF IDENTIFYING AND EXTRACTING TERMS FOR ONTOLOGY LEARNING AND DEVELOPMENT"

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : ***20 Februari 2022.***
That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on:
20 February 2022.

Pengerusi Viva: *(Chairman for VIVA)*	**Prof. Dr. Wan Rozaini Sheik Osman**	Tandatangan *(Signature)*
Pemeriksa Luar: *(External Examiner)*	**Prof. Dr. Shahrul Azman Mohd Noah**	Tandatangan *(Signature)*
Pemeriksa Dalam: *(Internal Examiner)*	**Assoc. Prof. Dr. Siti Sakira Kamaruddin**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Assoc. Prof. Dr. Azman Yasin**	Tandatangan *(Signature)*
Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)*	**Assoc. Prof. Dr. Mazni Omar**	Tandatangan *(Signature)*

Tarikh:
(Date) ***20 February 2022***

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Universiti Utara Malaysia

Dean of Awang Had Salleh Graduate School of Arts and Sciences

UUM College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

Abstrak

Media sosial adalah penting dalam menambahbaik komunikasi proses Pengurusan Bencana (DM). Walau bagaimanapun, perwakilan maklumat dalam DM Media Sosial (DMSM) masih tidak mencukupi terutamanya dalam perwakilan ontologi. Disebabkan wujudnya jumlah data teks tidak berstruktur yang besar berkenaan DMSM, pengekstrakan maklumat untuk pembangunan ontologi boleh dicapai melalui perlombongan teks. Walau bagaimanapun, kajian sedia ada berkenaan pembangunan ontologi berasaskan perlombongan teks menggunakan skim tanpa seliaan yang dikenali, TF-IDF mengabaikan pengagihan dokumen dan membawa kepada ciri berdimensi tinggi. Objektif utama kajian ini adalah menambah baik pembangunan ontologi melalui penambahbaikan skim pemberat istilah (TWS) yang terselia dan membangunkan ontologi DMSM. Penambahbaikan ini direalisasikan dengan mengenal pasti TWS diselia sedia ada dan memberikan pemberat yang lebih tinggi kepada kategori positif dan bukannya kategori negatif, yang mengakibatkan penyingkiran istilah yang tidak berkaitan. Kajian ini dijalankan dengan mengumpul penerbitan saintifik DMSM, melaksanakan pra-pemprosesan, dan mengira lapan TWS yang terselia yang terpilih. Semua skim memperoleh pemberat yang tinggi pada kategori negatif, dan bukannya kategori positif. Penambahbaikan dilakukan dengan memperkenalkan nisbah kekerapan jangkaan positif dan nisbah kategori positif, di mana skim ditambah baik dengan mengekstrak terma yang relevan kepada kategori positif. Ontologi DMSM dibangunkan dan dinilai melalui kaedah penilaian ontologi berasaskan kaedah piawaian unggul bagi perbandingan sintaksis, mereka bentuk ontologi, dan menilai ontologi yang dipelajari. Berdasarkan keputusan, keputusan yang baik telah dicapai untuk TF.IDFEC-based.Enhanced dan TF.RF.Enhanced dengan 93.33% dan 91.03% untuk kejituan, 80.8% dan 78.02% untuk dapatan semula, dan 0.87 dan 0.84 untuk F-skor. Secara teorinya, kajian ini menyumbang kepada penambahbaikan TWS yang diselia dengan menekankan maklumat pengelasan korpus, justeru ciri dimensi boleh dikurangkan dan meningkatkan kepentingan perkataan yang diseimbangkan antara kelas positif dan negatif. Secara praktikal skim yang ditambah baik menyediakan teknik yang lebih baik untuk pembangunan ontologi bagi mengekstrak terma yang relevan daripada teks penerbitan saintifik yang tidak berstruktur terutama untuk domain DMSM.

Kata kunci: Pengurusan bencana, Media sosial, Skim pemberat istilah terselia, Perlombongan teks, Pembangunan ontologi.

Abstract

Social media is crucial in facilitating the Disaster Management (DM) communication process. However, the knowledge representation of DM Social Media (DMSM) is inadequate especially in ontology representation. Given to huge volume of DMSM unstructured text, information extraction for ontology development is achieved through text mining. However, existing works on text mining-based ontology development utilizes a well-known unsupervised scheme, TF-IDF that ignore document distribution and leads to high dimensionality of features. The main objectives of the study are to improve ontology development by enhancing supervised term weighting scheme (TWS) and developing DMSM ontology. The enhancement is achieved by identifying the existing supervised TWS and giving higher weightage to the positive category instead of the negative one, which results in the removal of irrelevant terms. The study is conducted by gathering DMSM scientific publications, performing pre-processing, and calculating the eight selected supervised TWS. All the schemes obtained high weightage on the negative category, instead of the positive category. An enhancement is performed by introducing a positive term frequency ratio and positive category ratio, whereby the enhanced schemes extract relevant terms to the positive category. The DMSM ontology is generated and evaluated using a gold-standard-based evaluation method for syntactic comparison, designing the ontology, and evaluating the learned ontology. From the results, it is found that good score is achieved for TF.IDFEC-based.Enhanced and TF.RF.Enhanced with 93.33% and 91.03% for precision, 80.8% and 78.02% for recall, and 0.87 and 0.84 for F-measure, respectively. Theoretically, this study contributes an enhanced supervised TWS by emphasizing the classification information of a corpus, hence features dimensionality can be reduced and boosts the importance of words that are distributed between the positive and the negative class. Practically the enhanced scheme provides an improved technique for ontology developers to extract relevant terms from unstructured scientific publication text especially for DMSM domain.

Keywords: Disaster management, Social media, Supervised term weighting scheme, text mining, Ontology development.

Acknowledgement

It is my pleasure to acknowledge several individuals who were instrumental in completing my PhD research. This thesis is my gift for all of them. First of all, I would like to express my highest gratitude to my supervisors Professor Dr. Azman Yasin and Associate Professor Dr. Mazni Omar for playing pivotal roles in my PhD study. Without them I might not be where I am today. They have given me thorough guidelines and constant supports since day one. I am grateful that even in the worst circumstances, both of them still believes in me. It is true that great teacher inspires. These acknowledgements would not be complete without mentioning all members of the Awang Had Salleh Graduate School administration for providing a pleasant and lively environment in nurturing me to learn and grow. Special thanks go to my PhD-mates as it was a pleasure working with all of them.

My deepest appreciation belongs to my family for their patience and understanding. For my mother (Che Amah Binti Mat), late father (Mahmud Bin Zaharuddin), beloved wife (Nurizzatul Atikha Binti Rahmat), late grandmother (Pok Binti Man), daughter (Sina Zehra Binti Muhammad), parent-in-laws, and family, their prayers and undying supports help pushing me towards the end of this journey even though they do not really apprehend the reality of the pursuit. Thank you so much, for cheering me all the way to the finishing line. I do not think I can ever repay the deeds I owe. Last but not least, this work would not materialize without the financial support from the Ministry of Higher Education (KPT) Malaysia. I would like to recognize the important role of the organizations in completing my academic work.

I believe this is not the end. Please support and wish the best of luck for my future academic endeavours. My passion for research will go on as I believe somewhere, something incredible is waiting to be discovered. Thank you.

20 February 2023

Muhammad Bin Mahmud

Universiti Utara Malaysia

Table of Contents

Permission to Use ..... **ii**

Abstrak ..... **iii**

Abstract ..... **iv**

Acknowledgement ..... **v**

Table of Contents ..... **vi**

List of Tables ..... **xv**

List of Figures ..... **xviii**

List of Appendices ..... **xx**

CHAPTER 1 INTRODUCTION **1**

1.1 Introduction 1

1.2 Background of Study 1

1.3 Problem Statement 7

1.3.1 Insufficient Disaster Management Social Media Knowledge Representation 7

1.3.2 Limitation on Unsupervised Term Weighting Scheme 8

1.4 Research Questions 11

1.5 Research Objectives 12

1.6 Research Scope 12

1.6.1 Text Mining to Ontology Term Weighting Scheme Scope 12

1.6.2 Ontology as the Knowledge Representation Scope 13

1.7 Research and Theoretical Framework 13

Universiti Utara Malaysia

1.8 Significance of Study 16

1.8.1 Body of Knowledge 16

1.8.2 Knowledge Representation for Ontology Engineer 16

1.9 Thesis Organization 17

CHAPTER 2 LITERATURE REVIEW 19

2.1 Introduction 19

2.2 Disaster Management Social Media 19

2.2.1 Disaster 20

2.2.2 Disaster Management 21

2.2.3 Social Media Role in Disaster Management 25

2.2.4 Disaster Management Knowledge Representation 28

2.3 Text Mining 32

2.3.1 Text Mining Techniques 34

2.3.2 N-gram 35

2.3.3 Term Weighting Schemes 36

2.3.4 Existing Text Mining Based Ontology Development Works 38

2.4 Classifications 45

2.4.1 Classification Techniques 48

2.4.2 Naïve Bayes (NB) 49

2.4.3 Support Vector Machine (SVM) 50

2.4.4 *k*-Nearest Neighbour (*k*-NN) 51

2.4.5 Comparison of Classification Techniques 52

Universiti Utara Malaysia

2.5 Supervised Term Weighting Schemes 53

2.6 Ontology 62

2.6.1 Structure of Ontology 63

2.6.2 Ontology Classifications 64

2.6.3 Importance of Ontology 66

2.6.4 Ontology Learning from Text 67

2.6.5 Ontology Evaluation Methods 69

2.7 Summary 73

CHAPTER 3 RESEARCH METHODOLOGY 75

3.1 Introduction 75

3.2 Design Science Research 75

3.3 Research Framework 76

3.4 Phase 1 - Problem Identification 78

3.5 Phase 2 - Identification of Existing Supervised Term Weighting Schemes 79

3.5.1 Step 1: Corpus Collection and Preliminary Domain Study 79

3.5.2 Step 2: Identifying Sets of Document to Suit Research Tasks 80

3.5.3 Step 3: Classify Documents to Binary Classification 82

3.5.4 Step 4: Extraction of General Concepts by using the Text Mining Technique 86

3.5.5 Step 5: Build Corpus Confusion Matrix 89

3.5.6 Step 6: Perform Unsupervised Term Weighting Schemes Computation 90

Universiti Utara Malaysia

3.5.7 Step 7: Perform Existing Supervised Term Weighting Schemes Computation Based on the Candidate Terms 92

3.6 Phase 3 - Construction of an Enhanced Supervised Term Weighting Scheme for Disaster Management Social Media Ontology Development Based On Binary Classification 95

3.6.1 Step 1: Analyse Unsupervised Term Weighting Schemes Result 96

3.6.2 Step 2: Analyse Existing Supervised Term Weighting Schemes Result 96

3.6.3 Step 3: Perform Enhanced Supervised Term Weighting Schemes Computation Based on the Candidate Terms 96

3.6.4 Step 4: Analyse Enhanced Supervised Term Weighting Schemes Results 102

3.6.5 Step 5: Construct Structured Database with all Extracted Terms and Weightage 103

3.7 Phase 4 - Evaluation of the Enhanced Supervised Term Weighting Schemes through the Developed Ontology 104

3.7.1 Step 1: Perform Term Filtering to the Extracted Terms 104

3.7.2 Step 2: Calculate Distance for Syntactic Comparison 105

3.7.3 Step 3: Design the Learned Ontology 108

3.7.4 Step 4: Evaluate the Learned Ontology through Gold-standard Based Evaluation Method 108

3.8 Summary 111

CHAPTER 4 EXPERIMENTS 112

Universiti Utara Malaysia

4.1 Introduction 112

4.2 Identification of Existing Supervised Term Weighting Schemes 112

4.2.1 Step 1: Corpus Collection and Preliminary Domain Study 112

4.2.2 Step 2: Identifying Sets of Documents to Suit Research Tasks 113

4.2.3 Step 3: Classify Documents to Binary Classification 114

4.2.4 Based on the algorithm ran, in total there are 117 documents classified as positive category with 112 included in the True Positive while 5 documents included in the False Positive category. The result continues with 22 publications categorised as False ositive category and 253 categorised as False Negative because of irrelevant content to DMSM. The next step of this experiment, with the purpose of extracting the related general concepts, pre-process the identified positive and negative publications as shown in Step 3: Classify Documents to Binary Classification 116

4.2.5 Step 4: Extraction of General Concepts by using the Text Mining Technique 118

4.2.6 Step 5: Build Corpus Confusion Matrix 120

4.2.7 Step 6: Perform Unsupervised Term Weighting Scheme Computation 121

4.2.8 Step 7: Perform Existing Supervised Term Weighting Schemes Computation Based on the Candidate Terms 121

4.3 Construction of an Enhanced Supervised Term Weighting Scheme for Disaster Management Social Media Ontology Development Based on Confusion Matrix 122

Universiti Utara Malaysia

4.3.1 Step 1: Analyse Unsupervised Term Weighting Schemes Result 122

4.3.2 Step 2: Analyse Existing Supervised Term Weighting Schemes Results 122

4.3.3 Step 3: Perform Enhanced Supervised Term Weighting Schemes Computation Based on the Candidate Terms 123

4.3.4 Step 4: Analyse Enhanced Supervised Term Weighting Schemes Results 124

4.3.5 Step 5: Construct Structured Database with all Extracted Terms and Weightage 124

4.4 Summary 124

CHAPTER 5 RESULTS AND DISCUSSIONS 126

5.1 Introduction 126

5.2 Identification of Existing Supervised Term Weighting Schemes 126

5.2.1 Chi-square (x^2) 126

5.2.2 Information Gain (IG) 128

5.2.3 Odds Ratio (OR) 130

5.2.4 Gain Ratio (GR) 132

5.2.5 Relevance Frequency (RF) 134

5.2.6 Mutual Information (MI) 135

5.2.7 Inverse Document Frequency Excluding Category-Based (IDFEC-based) 136

5.2.8 Short-text Term Weighting (SW) 138

Universiti Utara Malaysia

5.2.9 Term Frequency-Inverse Document Frequency (TF-IDF) 139

5.2.10 Results Comparisons between Selected Existing Supervised Term Weighting Schemes..... 141

5.3 Construction of an Enhanced Supervised Term Weighting Scheme for Disaster Management Social Media Ontology Development based on Binary Classification..... 143

5.3.1 TF.x^2.Enhanced..... 144

5.3.2 TF.IG.Enhanced..... 145

5.3.3 TF.OR.Enhanced..... 147

5.3.4 TF.GR.Enhanced..... 148

5.3.5 TF.RF.Enhanced..... 150

5.3.6 TF.MI.Enhanced..... 152

5.3.7 TF.IDFEC-based.Enhanced..... 153

5.3.8 TF.SW.Enhanced..... 155

5.3.9 Results Comparisons between Constructed Enhanced Supervised Term Weighting Schemes..... 157

5.4 Summary..... 159

CHAPTER 6 EVALUATION OF THE ENHANCED SUPERVISED TERM WEIGHTING SCHEMES THROUGH THE DEVELOPED ONTOLOGY.. 160

6.1 Introduction..... 160

6.2 Evaluation of the Enhanced Supervised Term Weighting Schemes through the Developed Ontology..... 160

6.2.1 Step 1: Perform Term Filtering to the Extracted Terms..... 160

Universiti Utara Malaysia

6.2.2 Step 2: Calculate Distance for Syntactic Comparison 164

6.2.3 Step 3: Design the Learned Ontology 165

6.2.4 Step 4: Evaluate the Learned Ontology through Gold-standard Based Evaluation Method 166

6.3 Results and Discussion 168

6.3.1 DMSM Ontology Development based on TF.RF.Enhanced 169

6.3.2 DMSM Ontology Development based on TF.IDFEC-based.Enhanced 176

6.3.3 DMSM Ontology Development based on TF.RF 182

6.3.3 DMSM Ontology Development based on TF.IDFEC-based 186

6.3.4 DMSM Ontology Development based on TF-IDF 189

6.3.5 Results Comparison between All Evaluations 192

6.4 Summary 195

CHAPTER 7 CONCLUSION AND FUTURE WORKS 196

7.1 Introduction 196

7.2 Study Recapitulation 196

7.2.1 Objective 1: To Identify Existing Supervised Term Weighting Schemes for Ontology Development 196

7.2.2 Objective 2: To Construct an Enhanced Supervised Term Weighting Scheme for Disaster Management Social Media Ontology Development based on Binary Classification 197

7.2.3 Objective 3: To Evaluate the Enhanced Supervised Term Weighting Schemes through the Developed Ontology 198

Universiti Utara Malaysia

7.3 Contributions 199

7.3.1 Text Mining to Ontology Development for Binary Classification 199

7.3.2 DMSM Ontology Development 200

7.4 Limitations and Future Works 201

7.5 Conclusions 202

REFERENCES 203

APPENDIX A EXAMPLE LIST OF EXTRACTED TERMS 227

TF.IDFEC-based.Enhanced 227

TF.RF.Enhanced 231

Universiti Utara Malaysia

List of Tables

Table 2.1 *Disaster categorised by natural and human influence (Slamet et al., 2018; Olteanu, Vieweg, & Castillo, 2015; Othman et al., 2014)* 21
Table 2.2 *Summary of DM Phases, the definitions, and examples of activities* 25
Table 2.3 *Summary of Social Media Role in DM* 28
Table 2.4 *Examples of the existing conceptual modelling.* 31
Table 2.5 *Comparison between existing text mining-based ontology development works.* 41
Table 2.6 *Classifications of the confusion matrix* 47
Table 2.7 *Comparison of supervised TWS* 61
Table 2.8 *Differences between lightweight and heavyweight ontology (Biagetti, 2020).* 65
Table 2.9 *Comparison between the four ontology evaluation methods* 73
Table 3.1 *Development Tools Required for Text Mining and Concepts Extraction Process* 82
Table 3.2 85
Table 3.3 *Development Tools Required for Text Mining and Concepts Extraction Process* 87
Table 3.4 *The algorithm to perform unsupervised TWS by using TF-IDF* 91
Table 3.5 *The algorithm to perform supervised TWS by using the existing supervised term frequency schemes* 92
Table 3.6 *The algorithm to perform by using enhanced supervised TWS* 98
Table 3.7 103
Table 3.8 *Example of structured database for the enhanced TF.RF* 103
Table 3.9 *List of extracted terms from the gold-standard taxonomy* 109
Table 3.10 *The algorithm to perform mapping between learned and referenced ontology concepts* 110
Table 4.1 *Data from all databases for retrieved, positive, and negative categories for DMSM.* 114
Table 4.2 *Binary classification constructed for DMSM publications.* 116
Table 4.3 *Examples of the extracted terms and the calculated term frequency.* 119
Table 4.4 *Examples of the extracted terms and the TP, TN, FP, and FN values.* 120
Table 5.1 The ten highest-ranked terms and the calculated TF.x^2 values 127

Table 5.2 The ten highest-ranked terms and the calculated TF.IG values 129
Table 5.3 *The ten highest-ranked terms and the calculated TF.OR values* 131
Table 5.4 *The ten highest-ranked terms and the calculated TF.GR values* 133
Table 5.5 *The ten highest-ranked terms and the calculated TF.RF values* 134
Table 5.6 *The ten highest-ranked terms and the calculated TF.MI values* 136
Table 5.7 *The ten highest-ranked terms and the calculated TF.IDFEC-based values* 137
Table 5.8 *The ten highest-ranked terms and the calculated TF.SW values* 139
Table 5.9 *The ten highest-ranked terms and the calculated TF-IDF values* 140
Table 5.10 *Definition of category information based on document classifications.* 141
Table 5.11 *Comparison between the findings for identification of existing supervised TWS.* 142
Table 5.12 *The ten highest-ranked terms and the calculated TF.x2.Enhanced values* 144
Table 5.13 *Comparison between TF.x2 and TF.x2.Enhanced values* 145
Table 5.14 *The ten highest-ranked terms and the calculated TF.IG.Enhanced values* 146
Table 5.15 *Comparison between TF.IG and TF.IG.Enhanced values* 146
Table 5.16 The ten highest-ranked terms and the calculated TF.OR.Enhanced values 148
Table 5.17 Comparison between TF.OR and TF.OR.Enhanced values 148
Table 5.18 The ten highest-ranked terms and the calculated TF.GR.Enhanced values 149
Table 5.19 Comparison between TF.GR and TF.GR.Enhanced values 149
Table 5.20 The ten highest-ranked terms and the calculated TF.RF.Enhanced values 151
Table 5.21 Comparison between TF.RF and TF.RF.Enhanced values 151
Table 5.22 *The ten highest-ranked terms and the calculated TF.MI.Enhanced values* 152
Table 5.23 *Comparison between TF.MI and TF.MI.Enhanced value* 153
Table 5.24 *The ten highest-ranked terms and the calculated TF.IDFEC-based.Enhanced values* 154

Universiti Utara Malaysia

Table 5.25 *Comparison between TF.IDFEC-based and TF.IDFEC-based.Enhanced values*..... 154
Table 5.26 *The ten highest-ranked terms and the calculated TF.SW.Enhanced values* 156
Table 5.27 *Comparison between TF.SW and TF.SW.Enhanced values*..... 156
Table 5.28..... 157
Table 6.1 *Summary of term filtering performed for TF.RF.Enhanced term weighting* 161
Table 6.2 *Summary of term filtering performed for TF.IDFEC-based.Enhanced term weighting*..... 161
Table 6.3 *Summary of term filtering performed for TF.RF term weighting* 162
Table 6.4 *Summary of term filtering performed for TF.IDFEC-based term weighting* 162
Table 6.5 *Summary of term filtering performed for TF-IDF term weighting*..... 163
Table 6.6 *List of extracted terms from the gold-standard taxonomy* 167
Table 6.7 *Precision and recall for mapped learned and gold-standard ontologies for TF.RF.Enhanced supervised TWS.*..... 173
Table 6.8 *Precision and recall for mapped learned and gold-standard ontologies for TF.IDFEC-based.Enhanced supervised TWS.* 178
Table 6.9 *Precision and recall for mapped learned and gold-standard ontologies for TF.RF supervised TWS.* 184
Table 6.10 *Precision and recall for mapped learned and gold-standard ontologies for TF.IDFEC-based supervised TWS.* 187
Table 6.11 *Precision and recall for mapped learned and gold-standard ontologies for TF.IDF supervised TWS.*..... 190
Table 6.12 *Summary of the evaluation results*..... 192

Universiti Utara Malaysia

List of Figures

Figure 1.1. Present study research and theoretical framework 15
Figure 2.1. The four main phases of DM (Mikulsen & Diduck, 2016) 22
Figure 2.2. Text Mining with other interrelated fields (Miner et al., 2012) 33
Figure 2.3. Existing text mining-based ontology development architecture (Liu & Zhang, 2020; Wang et al., 2020; Salatino et al., 2019). 44
Figure 2.4. Examples of document distribution of classes TN, TP, FP, and FN 47
Figure 2.5. Naïve Bayes classification 49
Figure 2.6. Support Vector Machine classification 50
Figure 2.7. k-NN classification 51
Figure 2.8 Example of lightweight ontology structure 65
Figure 2.9 Example of heavyweight ontology structure 65
Figure 2.10 Ontology learning layer cake (Asim et al., 2018a). 68
Figure 2.11. Summary of ontology evaluation methods 72
Figure 3.1. Research framework for present study 77
Figure 3.2. Classification process developed in Rapidminer. 85
Figure 3.3. Enhanced Supervised TWS 97
Figure 3.4. System architecture for the enhanced supervised TWS 100
Figure 3.5. Hierarchy of terms by using hierarchical clustering method performed in Orange Data Mining. 107
Figure 5.1. TF.x^2 ten highest ranked TWS result 127
Figure 5.2. TF.IG ten highest-ranked TWS result 129
Figure 5.3. TF.OR ten highest-ranked TWS result 131
Figure 5.4. TF.GR ten highest-ranked TWS result 132
Figure 5.5. TF.RF ten highest-ranked TWS result 134
Figure 5.6. TF.MI ten highest ranked TWS result 135
Figure 5.7. TF.IDFEC-based ten highest-ranked TWS result 137
Figure 5.8. TF.SW ten highest-ranked TWS result 138
Figure 5.9. TF-IDF ten highest-ranked TWS result 140
Figure 5.10. TF.x^2.Enhanced ten highest-ranked TWS result 144
Figure 5.11. TF.IG.Enhanced ten highest-ranked TWS result 146
Figure 5.12. TF.OR.Enhanced ten highest-ranked TWS result 147
Figure 5.13. TF.GR.Enhanced ten highest-ranked TWS result 149

Figure 5.14. TF.RF.Enhanced ten highest-ranked TWS result..... 150
Figure 5.15. TF.MI.Enhanced ten highest-ranked TWS result..... 152
Figure 5.16. TF.IDFEC-based.Enhanced ten highest-ranked TWS result 154
Figure 5.17. TF.SW.Enhanced ten highest-ranked TWS result 156
Figure 6.1. Hierarchy of terms by using hierarchical clustering method for TF.RF.Enhanced result..... 164
Figure 6.2. The designed learned ontology based on TF.RF.Enhanced result. 165
Figure 6.3. The designed learned ontology based on TF.IDFEC-based.Enhanced result. 166
Figure 6.4. Ontology developed from TF.RF.Enhanced results. 172
Figure 6.5. Ontology developed from TF.IDFEC-based.Enhanced results..... 177
Figure 6.6. Summary of the evaluation precision & recall results 193
Figure 6.7. Summary of the evaluation F-measure results 193

Universiti Utara Malaysia

List of Appendices

Appendix A Example List of Extracted Terms……227

Universiti Utara Malaysia

CHAPTER 1

INTRODUCTION

1.1 Introduction

The reason of performing this study is elaborated in this chapter, to provide an understanding of its importance. This chapter is arranged in an organized manner with the background of study, problem statement, research questions, research objectives, research scope, and significance of study, described to support the justification.

1.2 Background of Study

Xu et al. (2018) define disaster is an unforeseen man-made or natural hazards with high complexity to control and predict, ultimately impacting massive environmental, economic, human, material, or deprivation. Katutu et al. (2019) and Nishida, Yokoyama, Wagstaff, & Callomon (2017) explain that several causes have been identified related with this natural phenomenon for instance increased sea water levels, major storms, irresponsible human activities, and other unanticipated climate shift, harming the environment and community. The number of recorded global natural disasters has increased from 1970 to 2018 with the highest was recorded in 2005 (Ritchie, Hannah & Roser, 2019).

In order to minimized the impact caused by disasters, a high-quality Disaster Management (DM) process is proposed by Horney et al. (2016) that lessen the aftermath of the disaster. Throughout the DM operation, Abdalla & Esmail (2018) stated that casualties and property damage is reduced or eliminated, essential needs is supplied, and business continuation is secured.

Universiti Utara Malaysia

During the DM process, the making of informed and fast actions and decisions are essential (Yin et al., 2015). This capability can be reach with a continuous communication channel. DHS First Responder Group and Virtual Social Media Working Group (2014) listed the traditional communication methods including mobile phones, television, land mobile radio, satellite phones, radio, email, paper maps, landlines, mobile data, and traffic cameras. These methods however, experience some limitations particularly in facilitating real-time information sharing, enabling rich content, in addition to allow photo imagery, video, and profile updates along additional channels. These issues is expected to be solved by the emergence technology, social media by the virtue of a rapid and reliable channel of communication, a channel for information gathering and sharing for people to search and exchange information updates, and unlimited access for users to engage and interact (Simon et al., 2015; Huang & Xiao, 2015; Allaire, 2016).

Due to these circumstances, it is vital to have a structured knowledge representation for Disaster Management Social Media (DMSM) domain so that actions and decisions can be carried out seamlessly and efficiently, as it is observed that this domain knowledge representation is limited to conceptual modelling. Conceptual modelling is a representation in supporting a complex human design operation to strengthen the understanding and communication of a specific domain among users (Verdonck et al., 2019). For this study domain, the conceptual modelling exists but is limited to metamodel (Othman & Beydoun, 2016), User Requirement Notation (Khzam & Mussbacher, 2018), Business Process Model & Notation (Horita et al., 2017), and Unified Modelling Language (Chaawa et al., 2017). Even though there are several conceptual modelling types for this domain, as stated by Verdonck et al. (2019), conceptual modelling lacks sufficient semantic specifications of the terminology of the

Universiti Utara Malaysia

primary models, affecting the interpretation and uses of knowledge to be inconsistent. To address this drawback, ontology has become increasingly widespread to improve conceptual modelling since the creation is acknowledged theoretically or through a system of thought. Moreover, ontology is helpful to provide reasoning on the conceptual model contents (Karagiannis & Buchmann, 2018). By structuring DMSM knowledge representation through domain ontology, various related DMSM issues can be addressed. This can be the solution for many DMSM users.

With that being said, the knowledge repository for the DMSM domain contains significant information from structured (i.e. numerical), as well as unstructured data as observed in digital libraries which returns a huge number of results. Constructing the knowledge representation due to the exponential growth of digital data, can be conducted through data mining which is a procedure that engage exploring significant, novel, and potential valuable findings from huge collection of data; and executing algorithms in identifying of unrevealed pattern (Chen et al., 2015). Data mining is extensively used for identifying beneficial finding from the massive volume of data (Vijayarani et al., 2015).

Universiti Utara Malaysia

As categorized by Vijayarani, Ilamathi, & Nithya (2015) in data mining several types of interconnected areas are available, which are text, image, web, spatial, multimedia, graph, structure, sequential pattern, and medical. Although Data Mining focuses on structured information, Text Mining, on the other hand, is a technique to process unstructured data, particularly textual data. This technique is non-trivial since statistically, larger than 80 percent data available contains of semi-structured or unstructured data as an information source (Talib et al., 2016). Unstructured data is easier to understand and processed to humans, but not interpretable for machine. For

this reason, Text Mining operation is essential to obtain substantial pattern from unstructured text.

Depending on distinct characteristics in each practice area, text mining comprises of seven areas. All of these areas are interrelated, even though each is unique. In general, text mining projects necessitate multiple techniques from the seven areas. The seven areas are Information Retrieval (IR) (Baeza-Yates & Ribeiro-Neto, 1999), Document Clustering (Steinbach et al., 2000), Document Classification (Seo et al., 2016), Web Mining (Kosala & Blockeel, 2000), Information Extraction (IE) (Cowie & Lehnert, 1996), Concept Extraction (Jonnalagadda et al., 2012), and Natural Language Processing (NLP) (Chowdhury, 2003), described as follows.

Information Retrieval is a technique to acquire significant information resources related to information need to be extracted from a huge volume database (Guo et al., 2020). Secondly, Document Clustering is to segregate text documents into different clusters based on thorough selective information features by using suitable evaluation method (Abualigah et al., 2017), while Document Classification systematize documents for retrieval, analysis, and maintenance, commonly performed by using classifiers as an examples, Support Vector Machine and Decision Trees. Web Mining emphasises on gathering data from the web and mobile-based platform by implementing the latest computational techniques (Correa et al., 2019). The fifth technique, Information extraction (IE) is the task of structuring information from unstructured or semi-structured text automatically (Allahyari et al., 2017). Concept Extraction is a technique in which concepts for the named entities are derived from free text extraction (Tulkens et al., 2019). Finally compared to other areas, in this study, NLP is discovered the most appropriate due to the data from unstructured text can be structurally extracted. NLP process text for analysis in acquiring information useful in

Universiti Utara Malaysia

representing equivalent various information (Conneau et al., 2017). In conducting NLP to a specific document, an important transformation method must be carried out in reconstructing the unstructured or semi-structured text into a structured form, which is important in our motivation in which to construct the knowledge representation in the DMSM domain through ontology from the unstructured data.

Vector Space Model (VSM) is the most prominent structured representation used for textual data (Slamet, Atmadja, et al., 2018; Weggenmann & Kerschbaum, 2018). In said method, each document is represented using bag of words or a space of highly dimensional vector with every dimension of the space indicates a term or word related. Concerning this matter, it is necessary to perform the text pre-processing step in extracting the terms out of certain text data and assigning the term weightage in regard to the values and significance in the corpus. These weights are assigned depending on the preferred term weighting scheme (TWS).

Universiti Utara Malaysia

Two kinds of approaches are available to categorize TWS, namely unsupervised and supervised term weighting approaches. An unsupervised TWS omits the information of training documents in the measurements (Alsmadi & Hoon, 2019). In this way, the algorithm analyses and classifies the unlabelled data sets to extract the underlying pattern without the need of human intervention. One of the classic and extensively used unsupervised TWS (Carvalho & Guedes, 2020; Tang et al., 2020), named Term Frequency-Inverse Document Frequency (TF-IDF) is on the basis of the notion that the significance of a term is determined by occurrences along with the level of infrequency across the particular data. On the other hand, the supervised TWS focuses on labelled data, with the label classification utilizes preliminary knowledge regarding statistical information and document categories of the textual documents corresponding to these groups (Alsaeedi, 2020). The examples of supervised TWS

constructed according to the results of the extraction are Chi-square (x^2) (Debole & Sebastiani, 2003), Mutual Information (Yiming & Jan, 1997), Information Gain (Domeniconi, Moro, Pasolini, 2016), and Odds Ratio (Mladenic & Grobelnik, 1998). However, most studies that focuses on text mining to ontology development applies the unsupervised TWS as part of the process, Term Frequency-Inverse Document Frequency (TF-IDF) (Qiu, Qi, Wang, & Zhang, 2018; Ertek, Chi, Zhang, & Asian, 2017; Kuntarto, Moechtar, Gunawan, Santoso, & Ahmadin, 2017) which experience the main shortcoming, namely, ignoring the classification information (Jalilifard, Caridá, Mansano, Cristo, & da Fonseca, 2021; Wang, Xu, Yan, & Li, 2019; Wu & Yuan, 2018; Fauzi, Arifin, & Yuniarti, 2017).

These limitations can be addressed by employing a supervised TWS to label the data set in representing the document distribution, which is not applicable to be used in TF-IDF since it is an unsupervised TWS. Several research works (Alsmadi & Hoon, 2019; Noviantho et al., 2017; Sabbah et al., 2017) have employed this approach in conducting text mining process, however the goal for these previous studies is to perform text classification operation, not the intention of constructing knowledge representation through ontology as discussed before.

Therefore, to address the above issues, i.e. limited DMSM knowledge representation utilizing unstructured data, present work utilizes the latest and reliable scientific publications as the knowledge source to develop DMSM domain ontology. Text mining technique is applied due to its capability to process unstructured information especially textual data and to facilitate this technique, the present work aims to identify existing supervised TWS to develop DMSM knowledge representation through ontology. Based on the schemes identified, binary classification is employed to

Universiti Utara Malaysia

construct an enhanced term weighting for DMSM ontology development, followed by an evaluation process.

1.3 Problem Statement

Text mining-based ontology development is seen facing a limitation in the aspect of term weighting aspect, particularly utilising unsupervised TWS, TF-IDF. TF-IDF did not consider the classification information, observing that all documents' distribution is from the same class. Wu & Yuan (2018) stated that this is a concern since ignoring the dissimilarity between the classes of the data collection and the information distributed of the text categories, the classification effect on the information disparity is unable to obtain the ideal outcome.

Universiti Utara Malaysia

1.3.1 Insufficient Disaster Management Social Media Knowledge Representation

Knowledge representation for DM is limited to conceptual modelling such as metamodel (Othman & Beydoun, 2016), User Requirement Notation (Khzam & Mussbacher, 2018), Business Process Model & Notation (Horita et al., 2017), and Unified Modelling Language (Chaawa et al., 2017). Even though conceptual modelling is a method or instrument necessary to capture knowledge of interdisciplinary field used in various discipline (Fettke, 2020), this representation lacks sufficient semantic specifications Verdonck et al. (2019). The terms used for the modelling is unable to be expressed the meaning for the domain concepts semantically. Therefore, the comprehension of the knowledge conveyed is difficult to be grasped effectively and tends to lose the appropriate information that needs to be delivered. The application of the knowledge also may lead to inconsistencies and causes the

action carried out to be ineffective and did not achieve the expected results. Furthermore, conceptual modelling experiences time, cost, and effort consuming since manual reasoning need to be conducted, due to lack of capability to be processed by software application. It is necessary to address these issues with a better representation, ontology, to cater beyond what graphical representation could affords (Karagiannis & Buchmann, 2018).

1.3.2 Limitation on Unsupervised Term Weighting Scheme

Most of the ontology development process is utilising the TF-IDF technique as a component of the text mining process, which is well known as a classic TWS and considered as a baseline in most literature (Tang et al., 2020). Various previous works have measured the unsupervised TWS, TF-IDF to generate ontologies, as regards of its precision, recall, and F-measure; examples of the previous works include Liu & Zhang (2020), Wang et al. (2020), and Salatino et al. (2019).

Universiti Utara Malaysia

Liu & Zhang (2020) built ontology to address issues in understanding knowledge in the social security domain, brought from massive policy documents and complicated management. By using TF-IDF to perform unsupervised term weightings, an average accuracy of 53.16% and 63.73% recall are found. Wang et al. (2020) developed an e-Government domain ontology for knowledge sharing, with the adoption of TF-IDF to measure the importance of the concept properties. The term weightings are evaluated with government domain experts for correctness review and achieved more than 70% of the concept accuracy. Another study done by Salatino et al. (2019) achieved a precision of 40.4%, recall of 24.1%, and 30.1% F-measure for using TF-IDF as the unsupervised TWS. In this study of Salatino et al. (2019), they investigated the classification of research publications in accordance with the Computer Science

Ontology (CSO), which is an extensive ontology for area of research in the field of Computer Science. These results indicated that the text mining process can be further improved to achieve better accuracy, precision, recall, and F-measure rate.

As described above and mentioned by Guo & Yang (2016), even though the most frequently used TWS is TF-IDF, this measurement experiences some drawbacks mainly regarding document distribution, specifically, ignoring the distribution information between the feature words, ignoring the case of the word incomplete classification, and ignoring the distribution of information within the word class. Haddoud, Mokhtari, Lecroq, & Abdeddaïm (2016) also stated that TF-IDF ignores category information in measuring the importance of terms in the documents. At the same time, Mazyad, Teytaud, & Fonlupt (2018) also elaborated that the TF-IDF scheme can be improved by introducing a supervised TWS i.e. existing term frequency multiplied by a factor, which applies preliminary data on the number of documents associated to each category.

Universiti Utara Malaysia

The supervised TWS employed a binary classification in labelling the data set to reduce the high dimension, with the chosen category classified as positive and the others as a negative category. Each of the categories is further categorised as the predicted and observed classes, with given the term t_j within class C_i, documents, the number of documents from category c_i (positive) which include the term t_j is referred to as True Positive (TP). True Negative is the number of documents under the category c_i (positive), which do not contain the term t_j**.** For the negative category, False Positive is the number of documents that do not belong to class c_i but comprise the term t_j, while False Negative is the number of documents that are not belong to class c_i and do not include the term t_j. Based on this definition, the TWS is computed and analysed for the outcome.

Binary classification is a two-class classification that set apart the input set into two opposite classes, to provide a correct distinction (Canbek at al., 2017). Essentially, it can be implied that the DMSM domain knowledge representation experiences binary classification problem, whereas the corpus collection consists of Disaster Management and Social Media documents, respectively. Since Social Media is the subclass of Disaster Management, it is perceivable the mismatch between expected results, Social Media and the results, Disaster Management which is exhibited in a learning event deliver irrelevant documents that may belong to the class.

In this way, the measurement on every gathered document is categorised into two distinct, binary classification problems. This implies, in every experiment, the predicted class, Social Media is classified as the positive category and the other categories in the similar corpus are classified as the negative category, Disaster Management (Alsaeedi, 2020).

The existing text mining-based ontology development implements unsupervised term weighting technique, accordingly, did not address the binary classification problem, consequently the discriminating power of a term is not being able to be specified either positive or negative class. The term weightages for this scheme only considers the occurrence of term in all the document positive as well as negative category, therefore does not accurately represent the discriminating power of the term. In reviewing other works, it is found that a supervised TWS is able to outperform unsupervised TF-IDF performance, as discussed before (Tang et al., 2020; Alsaeedi, 2020; Alsmadi & Hoon, 2019).

As an example in a study conducted by Alsmadi & Hoon (2019) shows that unsupervised TF-IDF obtained only in average 62% of accuracy if compared with other supervised TWS which are Relevance Frequency (RF), and Information Gain

(IG) with an average accuracy of 77.75% and 73.25%, respectively. Following the results obtained by the abovementioned studies, supervised TWS outperform the existing unsupervised TF-IDF performance.

Based on these benchmarks, this study must identify the supervised TWS available instead of unsupervised TF-IDF to ensure the effectiveness of the text mining process to construct ontology in the DMSM domain. The unsupervised TWS that did not consider documents classification categories as the factor, becomes a drawback of unsupervised TWS for ontology development. Nonetheless, the binary classification problem in representing DMSM domain classification also becomes the motive for this study to be conducted. The research questions and objectives for this study are defined in the following section to address all of the issues described.

Universiti Utara Malaysia

1.4 Research Questions

In accordance with the elaboration of the problem in the earlier section, the issue that needs to be addressed in this study is how to develop a text mining-based ontology by using an enhanced supervised TWS. Below are the following questions that should be answered:

1. What are the existing supervised term weighting schemes for ontology development and what are their performances?
2. How to construct an enhanced supervised term weighting scheme for Disaster Management Social Media ontology development?
3. How to evaluate the enhanced supervised term weighting scheme?

1.5 Research Objectives

As regards to the research questions defined, to solve the explained problems, this research is conducted. Mainly, a text mining-based ontology development by using an enhanced supervised TWS is established. Following objectives should be accomplished to address the goal:

1. To identify and evaluate existing supervised term weighting schemes for ontology development.
2. To construct an enhanced supervised term weighting scheme for Disaster Management Social Media ontology development by using binary classification.
3. To evaluate the enhanced supervised term weighting schemes through the developed ontology.

Universiti Utara Malaysia

1.6 Research Scope

The research scope of this work consists of the enhanced TWS and DMSM ontology development. The following sections are detail descriptions of the text mining and the ontology development process.

1.6.1 Text Mining to Ontology Term Weighting Scheme Scope

The text mining operation was conducted in this study to achieve the significant concepts to be identified as terminologies for the DMSM ontology development. The process is restricted to the concept extraction, whereas other component for instance relationship is identified by means of manual involvement. Through deep modelling expertise, the selection is decided to make sure that the element quality is established. In addition, the supervised TWS selected for this study include Chi-square (x^2),

Information Gain (IG), Odd Ratio (OR), Gain Ratio (GR), Relevance frequency (RF), Mutual Information (MI), TF.IDFEC-based, and Short-text Term Weighting (SW), with the justification provided in Chapter Two.

1.6.2 Ontology as the Knowledge Representation Scope

The introduced framework concerns on knowledge representation that applies ontology as the data modelling language. While the ontology development can be categorised as fully and semi-automatic, this study focuses on a semi-automatic approach in which requires certain human intervention to complete the whole process. Other than that, lightweight ontology is selected to be developed in this study considering the term weighting scheme process extracted the terminologies only to represent the domain knowledge without the axiom and constraint components.

Universiti Utara Malaysia

1.7 Research and Theoretical Framework

Figure 1.1 illustrate the study's research and theoretical framework. Four main phases are depicted in the framework consisting of Problem Identification (Phase 1), Identification of Existing Supervised Term Weighting Schemes for Ontology Development (Phase 2), Construction of an Enhanced Supervised Term Weighting Schemes for Disaster Management Social Media Ontology Development (Phase 3), and Evaluation of the Enhanced Supervised Term Weighting Schemes through the Developed Ontology (Phase 4). As explained in Chapter Three, this research is performed based on the Design Science Research methodology.

Phase 1 comprises the preliminary steps through related literature reviews in examining the gap or problem of the research domain mainly on DMSM knowledge

representation, supervised TWS, and ontology evaluation methods. The outcome for this process is the existing knowledge representation for DMSM, eight existing supervised TWS, and four different types of ontology evaluation methods. Phase 2 conducts the required data gathering as the knowledge base and set up the identification of existing supervised TWS, involving text mining activities, therefore achieving the first objective defined for this research.

Phase 3 performs the construction of enhanced supervised TWS, related to this research second objective. In this phase, the supervised TWS are enhanced based on the binary classification. Finally, Phase 4 carry out the evaluation of the ontology developed through the gold-standard based method the final objectives of this research. In the end, the results are presented and analysed based on the precision, recall, and F-measure obtained.

Universiti Utara Malaysia

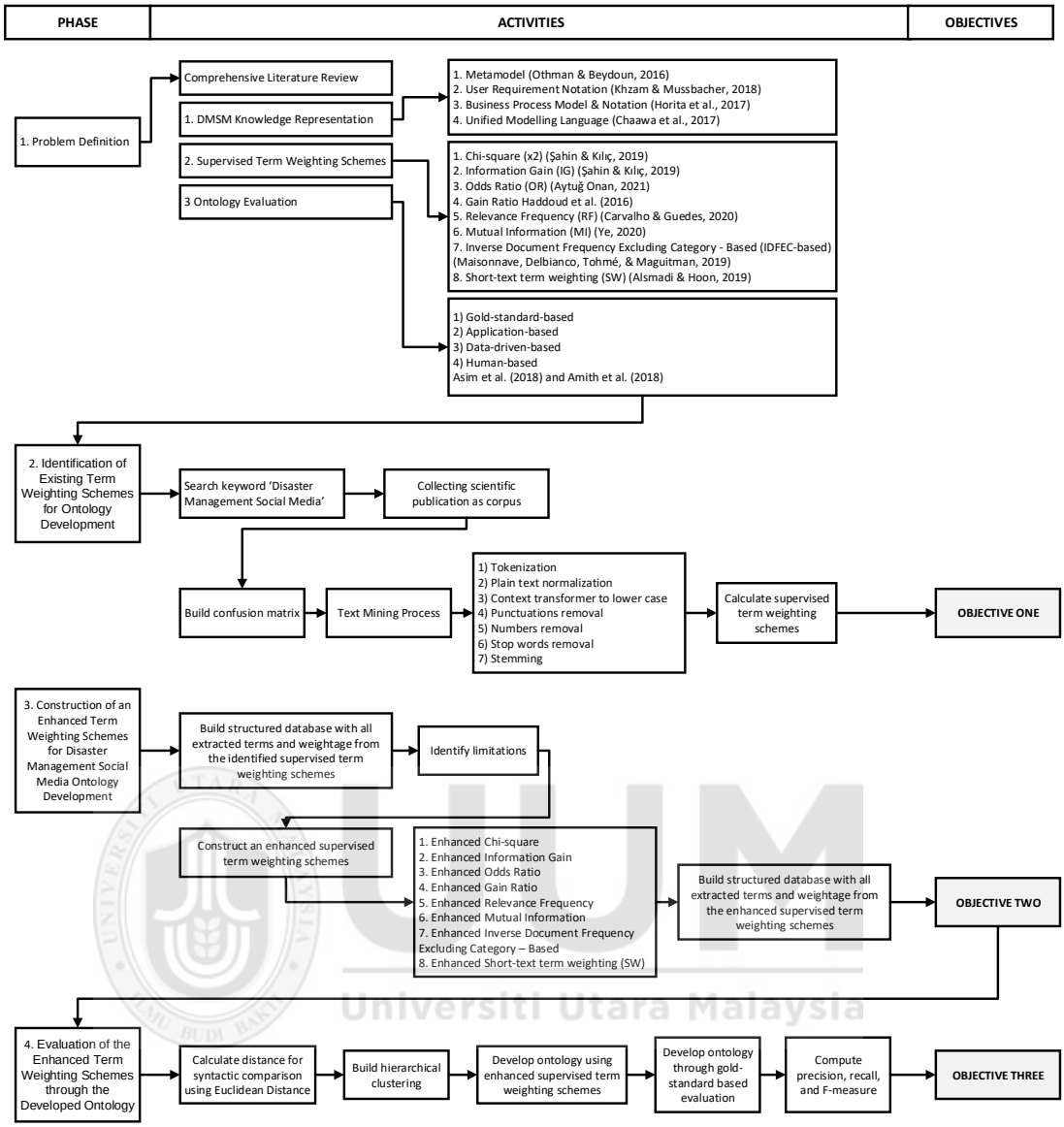

Figure 1.1. Present study research and theoretical framework

Universiti Utara Malaysia

1.8 Significance of Study

By addressing the objectives of this study, a technique to facilitate the construction of the enhanced TWS for DMSM ontology development is offered to the ontology engineers. Furthermore, this research contributes to the knowledge structure of specific domain as previously explained.

1.8.1 Body of Knowledge

The primary objective of this study is to deliver the construction of the enhanced supervised TWS for DMSM ontology development, which is having binary classification problem. To achieve this, the existing supervised TWS have been identified and enhanced based on the issue identified from the existing schemes. At the end of this study, a framework for text mining to ontology by using a supervised TWS is presented to address unsupervised TWS limitation on ignoring document distribution. Accordingly, by addressing this objective, this research supports the field of text mining area for ontology development with binary classification problem, which is a latest & enhanced approach to be applied by other researchers.

Universiti Utara Malaysia

1.8.2 Knowledge Representation for Ontology Engineer

Practically, the development of Disaster Management Social Media is facilitated in this research to present a beneficial method in delivering the domain concepts to new learners. Furthermore, this approach is helpful in comprehending the issue, enable active communication among domain experts, increase the adaptability of representation through considerate modelling tools, and assisting the ontology evaluation process. Nonetheless, the ontology developed is perceived as knowledge

representation to essentially enable reusability, better conceptualization, and information sharing within the selected domain.

1.9 Thesis Organization

There are seven chapters including this chapter organized in this thesis. The subsequent six chapters of thesis are outline are as below:

- **Chapter Two: Literature Review**

This chapter provides a review of the related analysis in various areas such as ontology and ontology learning with the focus being more on the supervised TWS. In addition, the evaluation method to evaluate the enhanced supervised TWS is also emphasized by using ontology.

- **Chapter Three: Research Methodology**

Chapter Three elaborates the research methodology carried out in performing this research. The descriptions related to the four phases of the methodology that were performed to enhance the supervised TWS.

- **Chapter Four: Experiments**

Chapter Four provides an in-depth discussion on the experiment performed to fulfil the first objective of this research to identify existing supervised TWS for ontology development and the second experiment involving the enhancement of supervised TWS for DMSM based on confusion matrix in achieving the second objective is presented.

Universiti Utara Malaysia

- **Chapter Five: Results and Discussions**

Based on Chapter Four results, the outcome of the experiment is discussed. The findings are presented and analysed in this chapter. The next chapter elaborates the third experiment which was conducted to evaluate the enhanced supervised TWS.

- **Chapter Six: Evaluation of the enhanced term extraction technique through the developed ontology**

This chapter reposts the evaluation of the enhanced supervised TWS. The evaluation is performed by developing the ontology and calculating the precision, recall, and F-measure.

- **Chapter Seven: Conclusion and Future Works**

Chapter Seven outlines all finding, and contributions are highlighted. In addition, the research work limitations are explained including several future works to be conducted in related field.

Universiti Utara Malaysia

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The research fundamental knowledge is elaborated here, namely Disaster Management Social Media, Disaster Management Social Media Domain knowledge representation, Text Mining, and Ontology Evaluation. The review initiated with the concept of disaster in Section 2.2, which the understanding of disaster, Disaster Management, Social Media Role in Disaster Management, and Disaster Management Knowledge Representation are elaborated. Afterwards, a detailed review of the text mining technique is followed to process unstructured information from the knowledge repository of the DMSM domain, in Section 2.3. In Section 2.4, ontology as the knowledge representation for Disaster Management Social Media is described, where its structure, types of languages, importance are explained. In addition, the ontology evaluation method is also reviewed in this subsection. Section 2.5 ends this chapter by providing the summary.

Universiti Utara Malaysia

2.2 Disaster Management Social Media

This section will further elaborate the focus of the domain in this study Disaster Management Social Media. The discussion is categorised based on disaster, Disaster Management (DM), and Social Media Role in Disaster Management.

2.2.1 Disaster

The disaster domain is a large and expanding field, investigated to further comprehend the knowledge base. The disaster field is investigated in this section, according to its structure and significance. Afterwards, the review on DM is examined to perceive a good insight of the whole concept. As specified according to The Malaysian National Security Council (MNSC) Directive 20, disaster is a crisis with high degree of complication leading to human casualties, environmental, and property damage, as well as obstructing economic and social processes (Shaluf & Ahmadun, 2006). As stated by Ritchie, Hannah & Roser (2019), the number of recorded global natural has increased throughout the years, as observed from 1970 to 2018 with the year 2005 recorded as the highest, and this trend is expected to continue. Disasters is classified into two categories depending on their reasons (Slamet, Rahman, et al., 2018).

The first category is categorised by natural disaster cause. This disaster category relates to catastrophic consequences happened impacted from natural hazard such as flood, landslide, and mudslide that happened without human being control. This type of events are usually believed with the act of Mother Nature. Shaluf & Ahmadun (2006) categorised the second category disaster caused by humans, also referred to man-made disaster, as well as technological disasters in addition to public location, transport, and production breakdown.

Table 2.1 itemize the disaster occurrences categorized with disaster caused by natural influence such as geological, meteorological, hydrological, climatological, and biological, while the disasters caused by human happened accidentally or intentionally. The next section discussed Disaster Management to facilitate the process of reducing the loss and damage of disaster events.

Universiti Utara Malaysia

Table 2.1

Disaster categorised by natural and human influence (Slamet et al., 2018; Olteanu, Vieweg, & Castillo, 2015; Othman et al., 2014)

Natural					Human Influence	
Meteoro-logical	**Hydro-logical**	**Geophysical**	**Climato-logical**	**Bio-logical**	**Intentional**	**Accidental**
Storm, hurricane, tornado, and winter storm	Water-based movements (mud/landslides, avalanches, flood, subsidence of hydrological origin, and rock falls)	Tsunami, volcano, earthquake, shifting land, erosion, and wet mass movement	Wildfire, droughts, heat, and cold waves	Insect infestation, animal violence, and epidemic	Shootout and bombing	Building sinking, dangerous chemicals. bridge collapsed, tunnel collapsed, derailment, radiology leak, dam incidents, nuclear explosion, and train hit

2.2.2 Disaster Management

Disasters affect many individuals and pose an increasing danger all over the world. Based on a report, from the year of 1995 until 2015, the top ten countries with the highest affected population are China, India, Philippines, Thailand, Pakistan, Brazil, Kenya, Vietnam and Ethiopia, with the reported values in million, 2274m, 805m, 131m, 130m, 76m, 55m, 51m, 47m, 44m, and 41m, respectively (Wallemacq et al., 2015).

Due to severe weather since 1995, eighty-seven million homes were damaged, as well as with more than one hundred thousand educational institution and medical facilities affected as reported by Centre for Research of the Epidemiology of Disasters (CRED), a global organization for investigating population health throughout national emergencies. In addition, the UNISDR's 2015 Global Assessment Report on Disaster Risk Reduction estimated overall financial loss is increasing to an average of two hundred fifty to three hundred billion dollar per year. For this reason, it is suggested

Universiti Utara Malaysia

that a projected of more than three hundred billion dollars to be prepared in preserving the environment and to cover future losses, must be reserved as annual average losses.

High level importance must be given on global initiatives in conserving these infrastructures from disaster harm, with the emphasis on health and education for sustainability in the future. The early warning system will be beneficial once devised to guarantee reliable communication to the communities as an effort of saving and reducing more casualties. Furthermore, CRED and UNISDR advise that Disaster Management (DM) should be handled as high priority for volunteering and humanitarian aid. On account of major issues particularly health effects and socio-economy, along with increasingly disaster events and recent global warming, having better mitigation and risk management will be an advantage to lessen he number of fatalities and other severe facilities damage.

In managing environmental hazards, DM is an initiative to reduce or avert any possible harms, making sure fast and appropriate support to affected communities of disaster, and expediting recovery (Seneviratne et al., 2010). As shown in *Figure 2.1*, Mikulsen & Diduck (2016) explains the operation involves four main phases, namely Mitigation, Preparedness, Response, and Recovery.

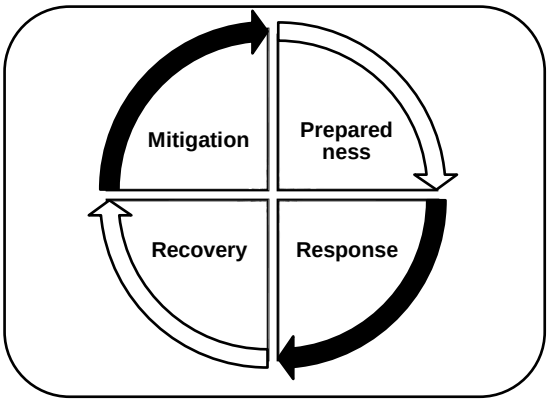

Figure 2.1. The four main phases of DM (Mikulsen & Diduck, 2016)

Firstly, Mitigation phase activate the action to lower the impact of the risk of a disaster happening whether pre, during, or post disaster (Kastolani & Mainaki, 2018). Kastolani & Mainaki (2018) indicated that involved disaster population faced various threats and hazards that require specific awareness and supports of DM. The affected victim needs to be instructed to be prepared of the disaster risks that happens in their living community. It is helpful to educate the people to be more ready and attentive in the events of disasters. Mitigation phase is observed to be effective in reducing the calamity particularly if the disaster cannot be avoided, therefore the affected areas to be prepared with preventive measures and required to be alert to the warning signals (Gupta, Starr, Farahani, & Matinrad, 2016). This action is necessary to prevent accidents in the industry and terrorist threats. Alternatively, other examples of the Mitigation efforts comprised of instituting buildings code to reinforce constructions and building levees in flood-prone areas, and bridge constructions.

Secondly, all operation to reduce its possible hazardous impact that can be carried out by the communities, relief organizations, and government earlier than the event of a disaster is known as Preparedness phase (Kunz et al., 2014). Appropriate Preparedness activities can extraordinarily boost Disaster Management operation. As an instance, as reported by Oscar, Albores, & Brewster (2018) in India, by preparing emergency plan, installing early warning system, and meteorological radars, disaster involving cyclones have causing least number of death compared to the previous events high number of death tolls of around twenty thousand victims affected by violent cyclones.

Kunz, Reiner, & Gold (2014) classified the Preparedness operation into two categories, namely physical and intangible. Physical operations include efforts using various resources that are tangible around the impacted population which embrace infrastructure establishment and inventory. For instance, the authorities are advised to

Universiti Utara Malaysia

relocate emergency aids within the disaster related locations. Other physical activities include tsunami-resistance residence in protected site development, earthquake-resistance infrastructure development, the infrastructure operation utilising communication tools and information technology installation that are required for DM, and the development of before disaster infrastructure, namely, hospitals, road systems, and backup electrical stations. For the intangible activities, the operation incorporate the systems for early warning instalment, awareness education on specific pre-disaster events, and financial resources preparation for rapid disaster response.

The third phase, as defined by Steigenberger (2016) Response phase allow emergency help and support, reducing the factor of additional damage and minimize problems for the next phase, Recovery. This phase is crucial for the authority in enforcing and regulating productive response management for disaster. Deen (2015) given example of operations of response, involving organizing public health facilities, coordinating rescue operations, establish relief camps, catering rapid relief efforts, and executing evacuation.

Universiti Utara Malaysia

Finally for the fourth phase, Recovery conducts short and long duration recovery operations such as receiving and providing updates on relief activities, recovery and reconstruction, emotional or mental support, and communicating with family and friends (Alexander, 2016; Mikulsen & Diduck, 2016; Houston, Hawthorne, Perreault, & Halliwell, 2014). Other alternatives than the recovery operation, Muskat & Nakanishi (2014) added rebuilding of financial and infrastructure as well as placement of victims from the disaster affected population.

The understanding of the four main phases in DM supports the comprehension of DM knowledge in this study. Understanding DM unified view is crucial in designing and developing ontology throughout this study.

Table 2.2

Summary of DM Phases, the definitions, and examples of activities

DM Phases	Definition	Example of Activities
Mitigation	Consist of steps to lessen the impact of the possibility of a disaster taken whether pre, current, and post disaster (Kastolani & Mainaki, 2018).	Public education, build embankment in flood-occurrence locations and regulating infrastructure development policy (Gupta, Starr, Farahani, & Matinrad, 2016).
Preparedness	Includes all actions that can be conducted by the government, communities, and relief organizations before the disaster event, reducing its predicted harmful impact (Kunz et al., 2014).	The activities are categorised as physical and intangible activities. The physical activities include pre-position relief supplies, communication equipment, and information technology installation, tsunami-resistant residence in protected location development, earthquake-resistant infrastructure development, and pre-disaster infrastructure development. In terms of the intangible preparedness activities, the examples include the installation of early warning systems, financial aid preparation, and educating awareness on certain pre-disaster events (Kunz, Reiner, & Gold, 2014).
Response	Comprises of steps to provide emergency assistance and aid, reducing the probability of additional damage and minimizing problems for recovery operations (Steigenberger, 2016).	The activities involve rapid relief efforts, building public health facilities, establishing relief camps, executing evacuation, and perform rescue operations, (Deen, 2015).
Recovery	Comprise of short and long term rehabilitation activities such as mental or behavioural health support, providing and receiving information on disaster response, reconnection with community members, and recovery and reconstruction (Alexander, 2016; Cozzolino, 2012; Cutter, 2003; Houston et al., 2014; Mikulsen & Diduck, 2016; Mileti & Noji, 1999).	Emphasis on preparation of financial and infrastructure, rehabilitation activities, including placement of victims from the impacted population (Muskat & Nakanishi, 2014).

Universiti Utara Malaysia

Table 2.2 summarises the four main phases of DM, Mitigation, Preparedness, Response, and Recovery with the definitions and examples of activities. The next section discusses the main aspect of this study domain which is social media role and usage in DM.

2.2.3 Social Media Role in Disaster Management

Social media usage increased exponentially throughout the last decade and shift the coordination of different processes regarding communication, information sharing, politics, community development, and news updates (Perrin, 2015). Zhang et al. (2015) describes social media as exchanging information emotionally between people concerning building and sharing information and ideas in online communities.

Furthermore, the medium is primarily utilised for information dissemination and gathering, as well as collaborating to solve problem and coping (Jurgens & Helsloot, 2018).

As stated by Procopio and Procopio (2007), public convey high concern and demand for a reliable and quick information source, which is flood maps, aerial footage or images, and event data during a disaster hazard. In reality, unfortunately, the real information source is usually limited during and post the disaster events. Accordingly, the high social media utilisation is important in getting the increased demand for information. As example, in the course of Katrina hurricane, the disaster victims reach out to social media to acquire the current news based on their geographic proximity. Procopio & Procopio (2007) explains the most valuable information found during the tragedy is updates regarding family and friend's safety and house damage.

A channel for citizens to collect situational information of disaster hazard is through information dissemination. Prior the rising trends of social media and internet, people relies on traditional media platforms for instance television, radio, and telephone to disseminate data on current situation and exchanging updates to family and friends. Centers for Disease Control and Prevention (2012) stated that communication is enabled through social media among people and allowing them for information exchange in media such as tweets, videos, images, and application postings seamlessly by using mobile devices.

The usage of social media allows collaboration on problem solving during and post disasters widely. Working together among users are facilitated with various communication lines for support systems and various skill sets exchange to obtain desired objectives. Apart from that, social media foster the knowledge of collaborating

Universiti Utara Malaysia

together among volunteers and provide assistance to affected victims in dire needs (Scanlon et al., 2014).

Besides information dissemination and gathering, including working together for problem solving, coping is another vital self-resilience. Technically, coping is defined as individual physical and mental action in dealing with distress, which in this case disaster event (Jurgens & Helsloot, 2018). It depends on the individual condition to be either bad or good in reacting to the coping behaviour. The reaction in the act of coping can be through emotional venting, emotional support, positive thinking, and instrumental support (Jin, 2010). An interesting finding, people refers to social media nowadays to provide and received comfort to cope with the adversities through collective coping (Edson & Takahashi, 2016). Social media provides a medium for three strategies of coping, namely facing and embracing certainty, sharing indication of life, and remembering and accepting losses. Through this phase, the proper management and expression of emotion and can be delivered with the wide platform of social media. As an example, an affected victim can convey their anger, fear, and emotion, while others can instantly express their support emotionally through social media platform.

Universiti Utara Malaysia

Based on the discussion above, the Disaster Management Social Media knowledge representation is known to be limited. Therefore, it is vital to have the knowledge representation for the DMSM domain, looking at its importance and how social media can influence and give impacts to disaster during, as well as after occurrence.

Table 2.3

Summary of Social Media Role in DM

Role of Social Media in DM	Purpose	Description
Information dissemination and gathering	Public to collect information regarding the disaster hazard (Jurgens & Helsloot, 2018).	Communication is shared among citizens through social media and provide them a platform to deliver information such as tweet, videos, images, and blog postings immediately by using mobile phones ("Centers for Disease Control and Prevention", 2012).
Collaborative problem solving	Users of social media work together in providing support systems and sharing diverse skill sets through various networks to achieve valuable objective (Jurgens & Helsloot, 2018).	Social media facilitate volunteers to work together to provide help to those requires help to overcome this hazardous condition together (Scanlon et al., 2014).
Coping	Physical and mental action of an individual in dealing distress (Jurgens & Helsloot, 2018).	To provide a platform for people to express the stress endured during disaster through collective coping (Edson & Takahashi, 2016).

Table 2.3 presents the summary of social media role in DM, mainly on information dissemination and gathering, collaborative problem solving, and coping, including the purpose and description. The next section discusses the knowledge representation for the DM domain.

Universiti Utara Malaysia

2.2.4 Disaster Management Knowledge Representation

Knowledge Representation (KR) is a field in Artificial Intelligence, to structure entities into classes with the dedicated aim of exchanging information, regarding the property and relationship components that related to the selected domain and specific objectives (Borgo, Ferrario, Masolo, & Vieu, 2019); Gayathri & Uma, 2018); and Shoeleh, Majd, Hamzeh, & Hashemi, 2015). KR is the primary action in recognising the environmental condition and supporting the communication among the related system and environment.

Knowledge representation can be applied through conceptual modelling. Various conceptual modelling types for DM are available based on several studies as an

example metamodel (Nasir et al., 2018; Othman & Beydoun, 2016; Benaben et al., 2016), User Requirement Notation (Khzam & Mussbacher, 2018), Business Process Model & Notation (Horita et al., 2017), Unified Modelling Language (Chaawa et al., 2017), and Domain Model (Ha, 2017). Othman & Beydoun (2016) in their work presented a metamodel for DM activities referring to the four main phases, Mitigation, Preparedness, Response, and Recovery. Nasir et al. (2018) also presented a metamodel focusing on mitigation phase in disaster management with geospatial aspects application. Another metamodel is presented by Benaben et al. (2016) which includes multiple point of views of disaster management with collaboration issue and data management issue as the key point. A metamodel is a concrete representation of a specific area or domain to assist in elucidating the elements related through constructs or concepts (Llesol & Intal, 2019). In their study, metamodeling is performed in synthesizing and validating models involving DM.

In another work by Khzam & Mussbacher (2018), User Requirement Notation (URN) is used to model DM concepts in terms of emergency processes. URN is a type of modelling that facilitates details of the requirements for workflow notations managing them (Khzam & Mussbacher, 2018). In their study, the model captures the domain concepts of the DM system, in which the instances in describing the expected real-life entities of a specific emergency. Business Process Model and Notation (BPMN) modelling in Horita et al. (2017) study, is carried out in defining the process concepts for an organization to assist decision making in DM. BPMN is an established model that specifies a collection of conceptual elements for representing these activities in a standard manner. Chaawa et al. (2017) in their study, introduced a Unified Modelling Language (UML) for DM to assess the strength as well as the weakness. UML is applied to conceptualise visually, comprehend, and exchange the structure and

Universiti Utara Malaysia

behavior of a system (Ahmad, Iqbal, Ashraf, Truscan, & Porres, 2019). A domain model concerning disaster management is introduced by Ha (2017) to provide a comprehensive framework consists of three domain model, disaster, human behavior, and disaster environment. This model simplifies the complexity of the disaster management framework by having a single representation to present the interrelationships.

Even though there are various conceptual modelling types for the DM domain, conceptual modelling experiences a drawback which did not comprise sufficient semantic specifications of the terminology of the primary models, affecting the interpretation and uses of knowledge to be inconsistent (Verdonck et al., 2019).

Ontology has become an increasingly widespread solution to address the aforementioned issue and improve conceptual modelling since the creation is acknowledged theoretically or through a system of thought. Compared to conceptual modelling, ontology provides a vocabulary of terms interpretable to humans and machines. Ontology ensures semantic consistency on data for multiple systems since these entities will use the same ontology as knowledge reference. In addition, through ontology, all of the experts' knowledge is encapsulated in formal or computational logics to enable reasoning, in which conceptual modelling only offers a visual and plain representation of data (Karagiannis & Buchmann, 2018). From these advantages of ontology and limitations of conceptual modelling, the knowledge representation must be further extended into ontology design.

Based on the reviews above, despite its urgency and importance, it is clear that knowledge representation for the DM domain only exists for conceptual modelling, without ontology although ontology provides various advantages namely, human and machine interpretable, ensuring data semantic consistency, and enable knowledge

Universiti Utara Malaysia

reasoning. In addition, based on the discussion on social media role in DM, social media gives a huge impact on the DM domain, however, the knowledge representation, in particular ontology for the DMSM domain, unfortunately is not yet available.

Table 2.4

Examples of the existing conceptual modelling.

Author(s)	Year	Modelling Types	Description	Modelling Purpose
Muzamil et al.	2022	Framework	This study proposes a flood management framework in Malaysia to improve the disaster management system and practices in the case of floods.	To provide a representation in assisting stakeholders to make informed decisions to save human lives and substantial financial loses (Muzamil et al., 2022).
Nasir et al.	2018	Metamodel	A study by Nasir et al. (2018) developed a metamodel for mitigation phase in disaster management by applying geospatial aspects.	To provide a concrete representation of a specific area or domain to assist in elucidating the elements related through constructs or concepts (Llesol & Intal, 2019).
Khzam & Mussbacher	2018	User Requirement Notation (URN)	The model captures the domain concepts of the DM system, in which the instances in describing the expected real life entities of a specific emergency Khzam & Mussbacher (2018).	To facilitate detail of the requirements for workflow notations managing them (Khzam & Mussbacher, 2018).
Seeger et al.	2018	Conceptual Model	The model is developed for public health in collaboration with the Centers for Disease Control and Prevention (CDC).	To inform practice and to guide evaluations of emergency risk communication (Seeger et al., 2018).
Horita et al.	2017	Business Process Model and Notation (BPMN)	In Horita et al. (2017) study, this model is carried out in defining the process concepts for an organization to assist decision making in DM.	To specify a collection of conceptual elements for representing these processes in a standard manner Horita et al. (2017).
Chaawa, Thabet, Hanachi, & Ben Said	2017	Unified Modelling Language (UML)	Chaawa et al. (2017) in their study, introduced a Unified Modelling Language (UML) for DM to assess the strength as well as the weakness.	To conceptually visually, comprehend, and exchange the structure and behavior of a system (Ahmad, Iqbal, Ashraf, Truscan, & Porres, 2019).
Ha	2017	Domain Model	This study by Ha (2017) provides a comprehensive framework, which is a three-domain model, for the goal of mitigating various disasters. This model take into account the three domains which are disaster, human behavior, and disaster environment.	To present the interrelationships among the three domains and factors directly or indirectly interact with each components (Ha, 2017).
Othman & Beydoun	2016	Metamodel	Othman & Beydoun (2016) in their work presented a metamodel for DM activities based on the four main phases, Mitigation,	To provide a concrete representation of a specific area or domain to assist in elucidating the elements

Universiti Utara Malaysia

			Preparedness, Response, and Recovery.	related through constructs or concepts (Llesol & Intal, 2019).
Bénaben et al.	2016	Metamodel	In this study by Bénaben et al. (2016), the research works focus on a crisis management metamodel, focusing on to provide a way to classify incoming data by connecting it to concepts of the metamodel.	To provide a concrete representation of a specific area or domain to assist in elucidating the elements related through constructs or concepts (Llesol & Intal, 2019).

The examples of existing conceptual modelling within the DM domain as the knowledge representation are described and summarised in Table 2.4. The descriptions are categorised based on the authors, year, modelling types, modelling purpose, and description. The next section focuses on the major aspect of this study which is text mining with the purpose to facilitate the steps for ontology development.

2.3 Text Mining

In order to process unstructured text to identify meaningful information, a mining process is necessary to be conducted. The technique that encompasses exploring unique, substantial, and potential valuable information from huge sets of data; and executing algorithms to the identification of unrevealed patterns is known as data mining (Chen et al., 2015). This technique is extensively applied for extracting significant findings from the huge volume of data (Vijayarani et al., 2015). In data mining, there are several categories of intersected fields, which are text, image, spatial, multimedia, web, structure, pattern, medical, and graph (Vijayarani, Ilamathi, & Nithya, 2015). The database in this work is gathered as text-based, therefore the text mining process is chosen. Text mining consists of activities for identifying and mining important patterns from unstructured text data.

Through the advancement of digital information, extraction of knowledge and data mining attract considerable needs to transform this type of information into

informative data that supported various systems, namely, criminal justice, commercial extraction systems, information retrieval, business management, and market analysis (Weiss, Indurkhya, & Zhang, 2015; Vijay Gaikwad, Chaugule, & Patil, 2014). At present, it is found that unstructured or semi-structured data is comprised of larger than 80 percent of data (Talib et al., 2016).

Although Data Mining concerns on numerical or structured information, text mining is a technique to process unstructured data, particularly text-based data. Processing unstructured data is easier and human interpretable, however not interpretable for machine. Accordingly, the text mining operation is essential, to identify valuable information based on the textual data, more importantly, in this study, scientific publications.

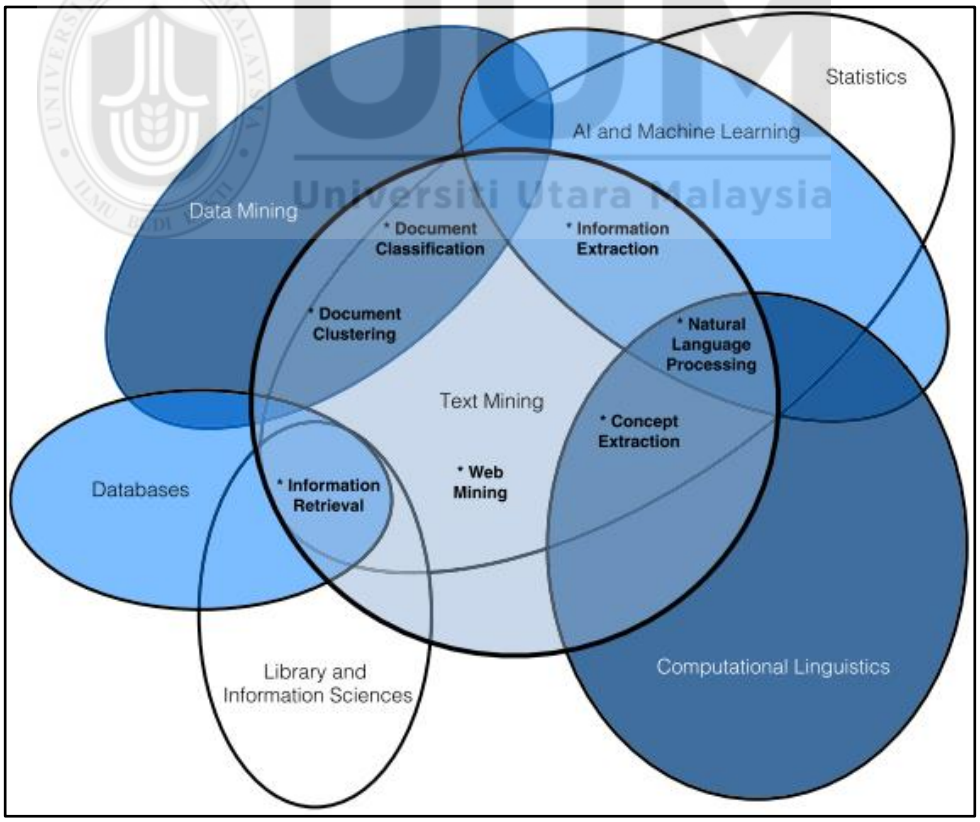

Figure 2.2. Text Mining with other interrelated fields (Miner et al., 2012)

In Figure 2.2 the Venn diagram depicts the intersection of Text Mining with other six fields known as Databases, Library and Information Sciences, Computational

Linguistics, Machine Learning, Artificial Intelligence, Statistics, and Data Mining. This indicated that in general, text mining fields involved all these fields in performing the process to transform unstructured data to meaningful results. For the next section, the techniques for performing the text mining process are presented and discussed.

2.3.1 Text Mining Techniques

Text mining can be classified into seven distinct but interrelated areas. Usually, text mining projects require multiple techniques from these fields. As shown in Figure 2.2, the seven fields are known as Concept Extraction, Natural Language Processing (NLP), Information Extraction (IE), Web Mining, Document Classification, Document Clustering, and Search and Information Retrieval (IR). Taking into account the requirements of this study is to specify the concepts that are associated to social media, also in accordance to the guideline by Miner et al. (2012), NLP is selected as the most suitable since the desired information from the text data can be extracted structurally. NLP is a technique for text mining with the intention of analysis in acquiring useful data to depict corresponding information differently (Conneau et al., 2017).

Universiti Utara Malaysia

When NLP is performed to specific text data, essential actions need to be carried out in transforming the semi-structured and unstructured text into a structured form, before performing further process of text mining. As introduced by Miner et al. (2012) the steps are, 1) the text data scope is set to be processed, 2) performing tokenization, 3) removing stop words, 4) performing stemming, 5) normalizing spelling, 6) identifying sentence boundaries, and 7) normalizing term case.

The integral part of NLP is text pre-processing due the characters, words, and sentences that are identified at this stage describe the core units prepared for the next processing phase ranging from analysis and tagging components, such as

morphological analysers and Part-of-Speech (POS) taggers, via applications known as Information Retrieval (IR) and machine translation systems (Gurusamy & Kannan, 2015). Main challenges of the pre-processing text is within the tokenization step. As defined by Gurusamy & Kannan (2015), categorising a sequence of text into words, phrases, symbols, or other valuable elements known as tokens is called tokenization technique. This collection of tokens will then be the input in text segmentation format processed in the step of this study (Uysal & Gunal, 2014).

The primary issue in this research concerns on the words segmentation that needs additional lexical and morphological intervention while conducting the tokenization (Gurusamy & Kannan, 2015; Verma et al., 2014). As an example, for the term 'social media' is tokenized into 'social' and 'media', the results are least meaningful to be analysed because the terms when compared to 'social media' will be interpreted differently. To remedy this problem, applied in this study is N-grams technique. The elaboration and description on this process is discussed in the next section.

Universiti Utara Malaysia

2.3.2 N-gram

Schonlau & Guenther (2016) stated that N-grams is a prominent method in the text mining technique and is commonly referred to as bag-of-words model considering it enumerate the sum of word occurrences regardless the word order and grammar. As defined by Jurafsky & Martin (2019), the N-gram method is given weightage by a sequence of N words such as bigram or a 2-gram, which indicates a two-word sequence of words, while trigram or 3-gram is a three-word sequence of words. Capturing the term 'social media' example with the N-gram model, the segments of 'social', 'media', and 'social media' are appropriately created.

Furthermore, Bamman, Eisenstein, & Schnoebelen (2014) stated that N-gram has proved to be applicable in various supervised learning tasks. The bigram or 2-gram is used in this experiment because of the processing consistency and effective weightage if compared to the 1-grams. Other than that, the bigram contains a larger number of tokens which is up to double. For this study, it is excessive to apply the 3-grams weightage because it is unnecessary and may cause bottleneck, especially when the performance concern and objective are already accomplished by using the 2-grams weightage.

The next section explains the TWS applied in the text mining process. Mainly there are two kinds of schemes available which are unsupervised and supervised TWS. The limitation exists that contributes as the motivation for this study is analysed and described in detail.

Universiti Utara Malaysia

2.3.3 Term Weighting Schemes

Vector Space Model (VSM) is the most well-known method applied for structured representation of textual data (Domeniconi et al., 2015). Every document is formatted using a high dimensional space vector or bag-of-words with each dimension of the space denoting a word or term respectively. For this reason, the text pre-processing step is crucial to extract the words or terms from a specific document and assigning them with weights related to their significance and value in each document. These weights are assigned based on the selected TWS.

Various TWS have been constructed based on the preference of an appropriate scheme that influence the results of the classification. Two types of TWS depending on class information are known as unsupervised and supervised TWS. Unsupervised TWS is a type of strategy to concerns with machine learning without human intervention and

performing classification by automatic means (Krallinger et al., 2017). One of the prominent and established TWS is known as the Term Frequency and Inverse Document Frequency (TF-IDF) and have been developed as many variants (Chen, Zhang, Long, & Zhang, 2016; Domeniconi et al., 2015; Singh & Singh, 2015; Albitar, Fournier, & Espinasse, 2014; Aphinyanaphongs et al., 2014).

TF-IDF is constructed attributed to the notion the term weightage is determined by the number of occurrences and the rate of infrequency through the text document. As defined by Wu & Yuan (2018), the Term Frequency (TF) component is the calculation of frequency of a particular term in a document *n*, against the total number of terms in the document as denoted in equation (2.1). Inverse Document Frequency (IDF) which is the second component, stress on unique words is more meaningful than higher frequency terms. In this way, IDF obtained higher weightage for terms that are rare in the document and assign lower weightage for regular terms. As denoted in equation (2.2), the IDF is divided by the total number of documents, *N*, with the number of documents containing the specific term, n, and then the logarithm of its quotient is acquired. In avoiding the denominator from being zero, in which the results will obtain an error, the denominator is usually added to 1. From the defined rules, TF-IDF multiplies each term within the document according to both factors as shown in equation (2.3). In the final result, the specific term's frequency and the rareness throughout the whole corpus achieved a balance weightage (Wu & Yuan, 2018).

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2.1}$$

$$IDF_i = \log \frac{N}{n} \tag{2.2}$$

$$TF_{i,j} * IDF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} * \log \frac{N}{n+1} \tag{2.3}$$

Universiti Utara Malaysia

To illustrate the TF-IDF expression, consider this example. Assume 5000 terms exists in a document, *j*, and the terms 'communication', *i*, occurs in *j* 600 times. In this case, the frequency of the term can be calculated by the ratio of total number of terms within the document, or according to the example, for *i* in *j*, 600 divided by 5000, obtained 0.1200 for term frequency. It is observed from this computation that even for the term 'could' probably achieved high term frequency rate even though it is less significant.

On account of the limitation, the IDF formula is applied by setting lower weightage to high occurrences term and vice versa for low occurrences term. For instance, consider a 50-document corpus, *N*, whereas the term *i* occurs 10 times throughout the corpus, df_i. Accordingly, the IDF is calculated with, $log\ (N/\ df_i)$ or $log\ (50\ /\ 10)$ = 0.6990, as the IDF value for the term. From the explanation before for TF formula, by having higher frequency of term in a document, higher term frequency is achieved, in contrast with IDF, whereas the lower frequency of the term, express higher rate of importance. To normalize the calculation, previous example is multiplied, with TF (0.12) * IDF (0.6990), resulting 0.084 TF-IDF weightage, indicating higher relevancy value compared to other terms with lesser TF-IDF weightage.

Universiti Utara Malaysia

2.3.4 Existing Text Mining Based Ontology Development Works

The analysis on the text mining-based ontology development for this study focuses on designing and building ontology based on concepts extracted automatically from unstructured text data. The reason for this procedure is because the traditional method of constructing ontology is done manually, therefore involvement of domain experts is necessary and expected (Wang et al., 2020). The main drawback of building ontology manually is it is time and resource exhaustive, due to the complexity of ontology in identifying the components as the knowledge on the domain might be

limited. By having an automated way of developing ontology, the needs of having domain experts or human intervention can be reduced, whereas the role can be shifted to validation phase. Taking advantage of the massive unstructured text data on DMSM domain, an approach by employing text mining is introduced in this study.

In extracting the concepts for the ontology development, TWS as described in previous section is utilized. Several text mining-based ontology developments from existing works have been identified in this section. The first related works for this study is conducted by Liu & Zhang (2020), focusing on enormous policy documents in the social security domain.

In China, social security policy comprises of information on pensions, health insurance, and other social schemes (Huang & Cai, 2021). The information for this type of data is huge, considering the database can store the data for many individuals. For this reason, Liu & Zhang (2020) conducted this study in order to develop an ontology as a way of storing knowledge and representation in this domain. In order to perform the keyword extraction process, TF-IDF unsupervised TWS is applied by calculating the term's weight and the terms that obtained high weightage is selected as the concept candidates. This study also compares the scheme with other schemes to assess the accuracy and recall rate. It is found that the results for TF-IDF is lower than the other scheme in term of accuracy and recall, with 59.26% and 67.81% for 720 texts and 47.05% and 59.64% for 1210 texts. The observed finding is that the higher the number of texts is increased, the lower the accuracy and recall are acquired for TF-IDF.

Second related works concerning the study conducted by Wang et al. (2020), builds the large scale e-Government ontology based on encyclopaedia and thesaurus. Basically, the domain ontology is built first from Chinese general e-Government thesaurus as rough ontology. Afterwards, the rough ontology is merged with the

Universiti Utara Malaysia

encyclopaedia knowledge, Baidubaike knowledge base, for the ontology improvement. Similar with Liu & Zhang (2020)'s work, TF-IDF is applied to calculate the term weightage of the concept properties. The importance is measured with the assumption that the higher the TF-IDF value, the larger the semantic feature of the concept is expressed. However, this assumption is not entirely correct since co-occurrences in different documents must be considered, instead of considering all documents are having similar distribution. Based on the ontology generated, the domain expert analyses the concepts extracted by using the TF-IDF scheme for correctness review. The evaluation shows that more than 70% of the properties are correctly mapped to its semantic features and relationships.

Salatino et al. (2019) in their work investigate the classification of research papers based on Computer Science ontology. Unsupervised TWS, TF-IDF helps ranking the concepts according to the calculated weightage scores. Gold-standard based evaluation method is employed in order to perform the evaluation. Among other classifiers, TF-IDF scheme shows an average performance in terms of precision, recall, and F-measure, with 40.40%, 24.10%, and 0.301 score, respectively.

The comparison of the discussed related works for text mining-based ontology development is shown in Table 2.5. The comparison is categorised by the types of domain ontology, authors, types of unstructured text as the data set, TWS applied, results, and limitations. Based on the review conducted, it is necessary for the TWS applied to be improved by using supervised TWS instead of unsupervised. In this way, enhancement is carried out by considering the document distribution aspect of the corpus.

Universiti Utara Malaysia

Table 2.5

Comparison between existing text mining-based ontology development works.

Domain Ontology	Author(s)	Types of Unstructured Texts	Term Weighting Scheme	Results	Limitations
Social security policy	Liu & Zhang (2020)	Corpus from social security policies of Fujian, Guangdong, Jilin, Liaoning, Shandong, Tianjin, and Zhejiang	TF-IDF	59.26% accuracy and 67.81% recall for 720 texts and 47.05% accuracy and 59.64% recall for 1210 texts.	Obtained low accuracy and recall rate which is less than 70% score.
e-Government	Wang et al. (2020)	Chinese general e-Government thesaurus and Baidubaike knowledge base encyclopaedia	TF-IDF	70% of the properties are correctly mapped.	The TWS did not considers document distribution as the basis of term weightage for concept extraction.
European space systems	Berquand et al. (2020)	56 feasibility reports provided by the European Space Agency, 40 books on space mission design publicly available and a collection of 273 Wikipedia pages.	TF-IDF	The commonality peak, 23% is reached for the combined corpora lexicon.	Obtained lower accuracy of 23% compared to other methods.
Biomedical	Shanavas et al. (2020)	Medical documents from Ohsumed and Medical Notes dataset.	TF-IDF	80.09% for average precision, 79.08% for average recall, and 78.9% for average F1 scores.	The terms extracted can be seen less relevance to the domain intended compared to other methods. The obtained average score of precision, recall, and F1 score is lower than 85% indicating improvement can be made to achieve higher results.
Indonesian Quran translation	Utomo et al. (2020)	Indonesian Quran translation verses	TF-IDF	70.75% for average accuracy and 71.55% precision.	The obtained average score of accuracy and precision is lower than 80% indicating improvement can be made to achieve higher results.
Computer Science	Salatino et al. (2019)	Research papers	TF-IDF	40.40% precision, 24.10% recall, and 0.301 F-measure.	Obtained an average score of precision, recall, and F-measure, with 40.40%, 24.10%, and 0.301 rate, respectively, compared to other classifiers.
Computer Science	Chughtai et al. (2019)	Web of Science articles published from 2014 to 2016 and the cited articles for Computer science subject.	TF-IDF	No quantitative results are discussed especially on the term weighting scheme, however the qualitative results indicated that the method is able to accomplish the desired results.	This study focuses on category classification of an article in which different objective for this study,

Universiti Utara Malaysia

Arabic text	Hawalah (2019)	Arabic textual data from various domains.	TF-IDF	81.20% for average precision, 81.92% for average recall, and 0.8135 for average F1 scores.	The obtained average score of accuracy and precision is lower than 85% indicating improvement can be made to achieve higher results.
Generic domains	(C et al., 2018)	Various generic documents	TF-IDF	No quantitative results are discussed especially on the term weighting scheme, however the qualitative results indicated that the method could accomplish the desired results.	This study focuses on clustering of an article in which different objective for this study,
Financial fraud	Attigeri et al. (2018)	Fraud case documentations	TF-IDF	No accuracy measure evaluated in this study such as precision, recall, and F-measure.	Based on reviews, other techniques than TF-IDF can be applied to improve the result.
Crime Scene Investigation	Liu et al. (2018)	Over 1000 Crime Scene Investigation images from different types of crimes provided by the Criminal Investigation Bureau in Shaanxi province, China.	TF-IDF	No quantitative results are discussed especially on the term weighting scheme, however the qualitative results indicated that the method could accomplish the desired results.	This study focuses on different scope of data particularly images, but TF-IDF TWS is utilised.
Bali tourism	Kuntarto et al. (2017)	Multimedia corpus from websites	TF-IDF	0.189 for precision.	Obtained precision lower than 0.2 and recall compared to other method.
Arabic technical texts	Benabdallah et al. (2017)	Arabic text documents containing definitions of technical terms in electronics and computer sciences in Arabic.	TF-IDF	90% for average precision, 71% for average recall, and 0.79 for average F1 scores.	The obtained results are good but from the extracted terms shown, improvement can be made to extract more relevant terms to the domain.

Figure 2.3 shows the general unsupervised TWS framework for text mining-based ontology development in sequence. Initially, the process begins with Step 1 to consolidate corpus collection of documents for the intended domain. Afterwards, in Step 2, the text mining pre-processing operation is carried out such as tokenization, plain text normalization, context transformer to lower case, punctuations removal, numbers removal, stop words removal, and stemming. The definition and the detail process of each step are discussed in Chapter Three.

The pre-processing operation is performed in order to transform noisy texts into categorised form of input for the algorithm to solve, in this case computing the term frequency and term weightage in the corpus (Sinyak et al., 2021). The next step, Step 3 and Step 4, computes the term frequency and term weightages for each token by using TF-IDF in which an unsupervised TWS, whereby the computation results facilitate the ontology development activities in Step 5. The generated ontology can be assessed in Step 6 by using various methods but one of the evaluation is by using gold-standard based evaluation method as performed by (Salatino et al., 2019; Kuntarto et al., 2017).

Universiti Utara Malaysia

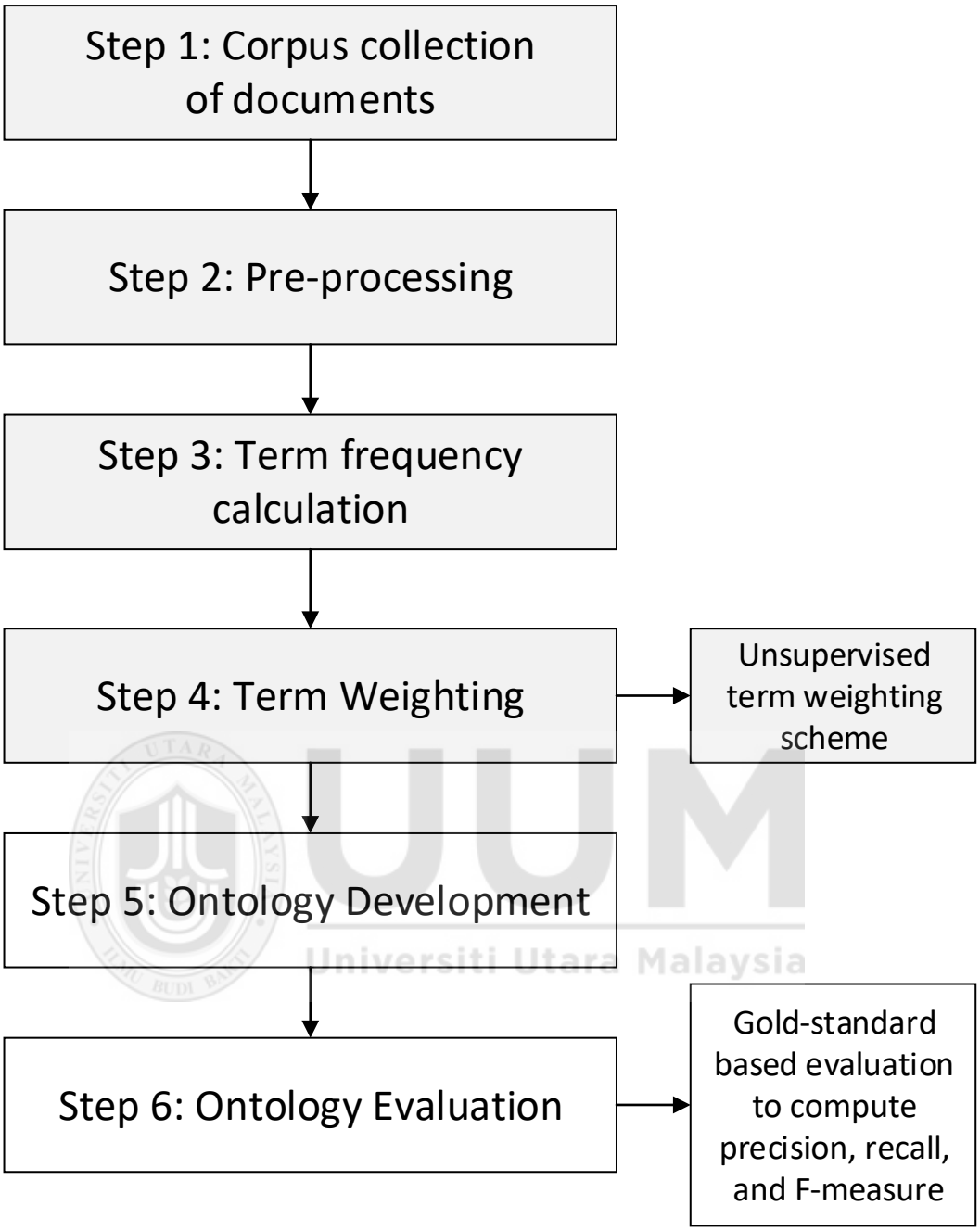

Figure 2.3. Existing text mining-based ontology development architecture (Liu & Zhang, 2020; Wang et al., 2020; Salatino et al., 2019).

Different from an unsupervised TWS, a supervised TWS requires entity-labelled text or class information in the weighting process. Alsaeedi (2020) consolidates several supervised TWS over the years consisting of Chi-square (x^2), Correlation Coefficient (CC), Information Gain (IG), Odds Ratio (OR), Gain Ratio (GR), Bio-normal separation, Relevance frequency, Mutual Information (MI), Mutual Information using

Sample Variance (MIUSV), Delta TF.IDF, Inverse Gravity Moment and TF (TF.IGM), Inverse Category Frequency and TF (TF.ICF), Inverse Term Entropy and TF (TF.ITE), Inverse document frequency excluding category (IDFEC), TF.IDFEC-based, Short-text Term Weighting (SW), Term Relevance Ratio (TRR), Prob-based, and $logtf \cdot rf_{max}$.

Among these supervised TWS, x^2, IG, OR, GR, RF, MI, IDFEC-based, and SW are evaluated in this study. Based on reviews by Maisonnave, Delbianco, Tohmé, & Maguitman (2019) and Mazyad, Teytaud, & Fonlupt (2018), x^2, IG, OR, GR, RF, and MI is the well-established supervised TWS. Most experimental works utilized these schemes for evaluation purposes. In addition, IDFEC-based, and SW are included in this study as they are the latest supervised TWS developed, but has been well received and included for evaluation for comparison among supervised TWS (Maisonnave, Delbianco, Tohmé, & Maguitman, 2019; Mazyad, Teytaud, & Fonlupt, 2018).

To understand supervised TWS better, the concept of classification must be further comprehended. In the next section, classifications, binary classification, multi-class classification, and multi-label classification are analysed and discussed.

Universiti Utara Malaysia

2.4 Classifications

The process of segregating multiple entities into a number of classes, defined by business rules, class boundaries, or some mathematical function, is known as classification (Nisbet et al., 2018). The classification can be performed based on the relationship between a recognized class assignment and the characteristics of the entity to be classified. If the known examples of a class is available, this type of classification is called supervised, while without the example of classes, the classification is performed as unsupervised. Three types of supervised classification are 1) binary

classification, 2) multi-class classification, and 3) multi-label classification (Nisbet et al., 2018; Mazyad et al., 2018).

Binary classification is the most studied classification type, with two classes classification that divides a provided input into two different classes, or two possible outcomes, such as ‘exist’ with ‘non-existent’ of a condition, ‘spam’ with ‘non-spam’ for email, and ‘malign’ for software data (Canbek et al., 2017). Furthermore, this type of classification is normally visualised by using a 2 x 2 confusion matrix or contingency table.

Multi-class classification, on the other hand, deals with classification with more than two classes. With multi-class, there are various possibilities for which class has to be predicted (Grandini et al., 2020). Each sample is mapped to one and one target label, as an example, a disaster can be flood or earthquake but not both at that same time.

To have a sample being assigned to a set of target labels or more than one class, multi-label classification can be used. Multi-label classification is a kind of classification where an instance belongs to several labels at the same time (Hadj-Kacem & Bouassida, 2021). To achieve this, the strategy is to simplify the multi-labelled data set into more than one distinct single-label binary data set (Mazyad et al., 2018). An example of a sample that can be multi-labelled is a news article, whereby the article can relate to sports, person, or location simultaneously.

In this study, the supervised binary classification is applied since the data set is categorised into two classes, documents relevant to Disaster Management Social Media and irrelevant to the domain selected. Most research works that apply supervised TWS use a confusion matrix in classifying the document distribution (Alsmadi & Hoon, 2019; Mazyad, Teytaud, & Fonlupt, 2018; Noviantho, Isa, & Ashianti, 2017). In a usual classification problem, the outcome is either true or false,

but with the confusion matrix, the probability of occurrences and the non-occurrences in the data set can be calculated. Through this method, the term weightage can be easily identified in the text mining process.

Confusion matrix classifies information on the occurrences of a specific behaviour is recognized correctly as the intended class and what is the occurrences identified as different behaviour for the other classes (Ruuska et al., 2018). This classification is applied in the supervised TWS. In the matrix, the columns indicate the instances in an actual class, while the rows indicate the instances in a predicted class. The confusion matrix consists of a binary problem, one class is labelled as positive while the other is negative and further categorised with actual and predicted classes (Alsmadi & Hoon, 2019).

Universiti Utara Malaysia

Table 2.6

Classifications of the confusion matrix

N = Number of documents in the corpus, sums of TP, TN, FP, and FN, with FN >> TP, TN, FP		**ACTUAL**	
		POSITIVE	**NEGATIVE**
PREDICTED	**POSITIVE**	TP	FP
	NEGATIVE	FN	TN

TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative

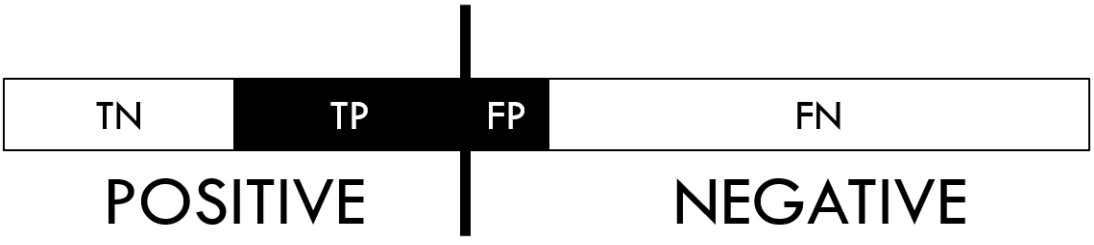

Figure 2.4. Examples of document distribution of classes TN, TP, FP, and FN

As depicted in Table 2.6, altogether, the classification consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), with TP as the number of documents from the positive class which observed relevant to the domain, TP is the number of documents belonging to the positive class, which is observed not relevant to the domain. FP is the number of documents that do not belong to the negative class but observed relevant to the domain, while FN is the number of documents that are not from the negative class and also observed does not relevant to the domain. In Figure 2.4, the document distribution of class TP, TN, FP, and FN are visually represented, whereby class TP and FP are closely related but distinguished whether the documents that relevant to the domain belong to positive or negative classes. Class TN and FN are independent since both classes indicate documents that does not relevant to the domain but belong to positive or negative classes.

To build the binary classification with an automated approach, the next section discusses the classification approach in detail.

Universiti Utara Malaysia

2.4.1 Classification Techniques

Classification is the task of categorizing objects into one or more predefined classes which often referred to as target, label, or categories, with the specific purpose (Kadhim, 2019). These classes establish a relationship between the terms and terms within the knowledge domain including the terms and the classes.

Classification is a type of supervised machine learning technique, to perform predictions for future results by taking account previous dataset as the reference (Al-janabi et al., 2018). Based on review conducted, there are several most widely used classification techniques known as Naïve Bayes (NB), Support Vector Machine

(SVM), and *k*-Nearest Neighbour (*k*-NN), and all these techniques are further elaborated in the next subsections.

2.4.2 Naïve Bayes (NB)

Naïve Bayes classifier is built from the family of probabilistic classifiers based on Naïve Bayes theorem (Aldrraji, 2022). This model performs the sentiments classification by learning from the labels of the training part. It assumes the existence of specific feature within a class is independent of the existence of other features in similar class (Al-janabi et al., 2018). In this theory, NB determines the probability of a particular event to have occurred based on the probabilistic related distributions of other specific events.

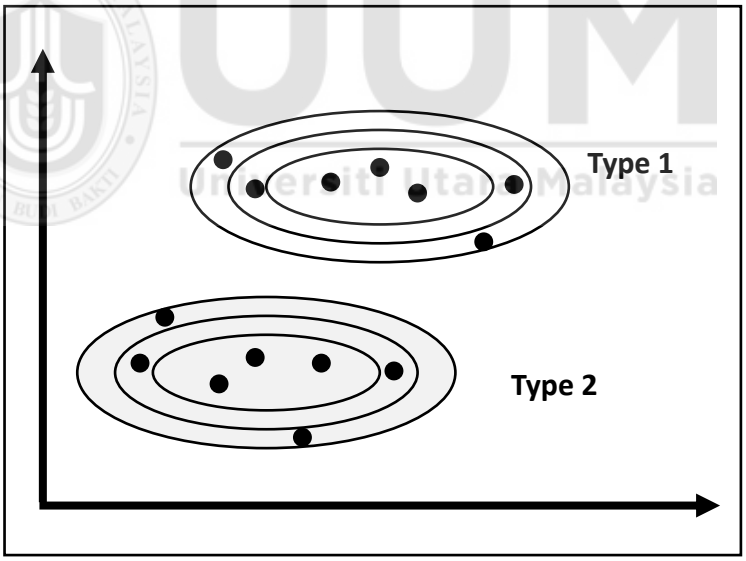

Figure 2.5. Naïve Bayes classification

As shown in Figure 2.5, the objects can be classified as in Type 1 or Type 2, with the task is to classify new objects to which class objects are belonged to, based on the currently existing objects. In the Bayesian analysis, the belief of the previous

experience is known as prior probability, and normally used to predict outcomes before the event happens.

2.4.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is an algorithm that allows to determine the best decision boundary between vectors which belong to a particular group and those that are not belong (Hasan et al., 2022). This theory consists of the concept of decision planes that set the decision boundaries, whereby the decision planes refer to the one that distinguish between a set of objects having separate class memberships.

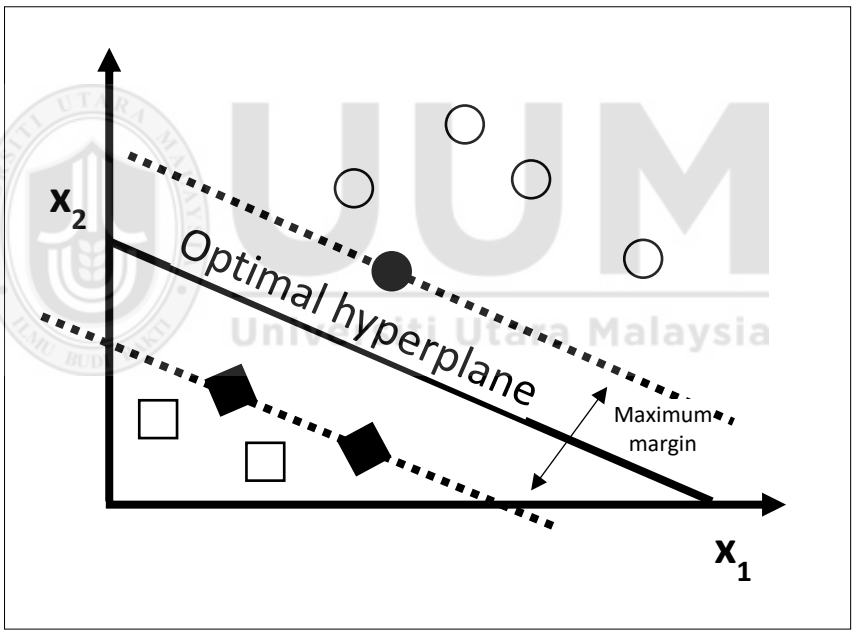

Figure 2.6. Support Vector Machine classification

Figure 2.6 shows the schematic example of the SVM theory, with the separating line defines a boundary or known as the hyperplane on the right side of which all objects are Type 1 and to the left of which all objects are Type 2. In this case, for any new object falling to the right is labelled as Type 1 or classified as Type 2 should it fall to the left of the separating optimal hyperplane.

2.4.4 *k*-Nearest Neighbour (*k*-NN)

k-NN is a simple and robust algorithm for classification and regression and applied in various problems such as text categorization, object recognition and pattern recognition ranking of models. This includes application in many different areas, such as bioinformatics and medicine (Ehsani & Drabløs, 2020). *k*-NN is a non-parametric and considered as lazy learning classifier. Lazy learning indicated that it delays the processing of training datasets until it receives the test instance for the class prediction. The *k*-NN algorithm relies on neighbourhood of close patterns relative to a conditional pattern.

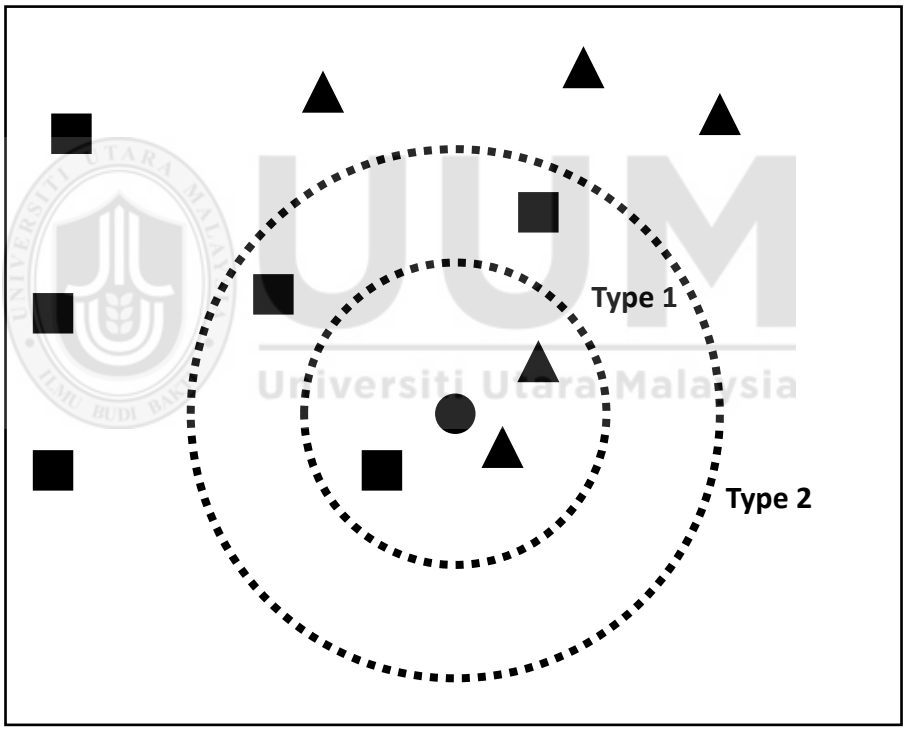

Figure 2.7. k-NN classification

Universiti Utara Malaysia

Consider an example in Figure 2.7 the outcome of KNN based on 1-nearest neighbour, whereby it is clear that in this case *k*-NN predicted the outcome of the query point with Type 1. After increasing the number of nearest neighbours to 2 or 5, *k*-NN will assign a Type 2 sign to the outcome of the query point.

2.4.5 Comparison of Classification Techniques

The review conducted shows that between these three algorithms, Naïve Bayes is most suited to be utilised in classifying and building the scientific journal publications confusion matrix during the experimental work in this study. SVM and *k*-NN is observed to be more suitable in solving multiclass classification problem, compared to NB whereby the algorithm is usually used for binary classification in which applied in this study as elaborated in section 2.4 (Artur, 2021; Blanco et al., 2020).

Binary classification in this study's context refers to predicting whether a document grouped into either two class labels, namely Disaster Management or Social Media. Therefore, it is unnecessary to use multiclass classification algorithm in this study, as the labels consists only two classes and not having multiple classes to be assigned to the documents. This conveys that NB algorithm is more feasible and suitable to be implemented during the experimental work. By having a compatible algorithm to perform the classification process, a faster result can be achieved to make a real-time prediction.

Other than that, based on study conducted by Zhang et al. (2019) NB is proven to be able to compute the posterior probability and reduce the dimension of feature words effectively with less computation complexity compared to the other algorithms. In a study conducted by Zaw et al. (2019) regarding brain tumour detection based on Naïve Bayes classification, the data requires binary classification to be constructed.

Segmented tumour images are trained and tested with the class labelled as it is a tumour or not. From the Naïve Bayes classification applied, the experimental results yield a high accuracy of 81.25% detection rate on tumour images and 100% detection rate on non-tumour images with the overall accuracy of 94%. In another study performed by Le et al. (2019) for linguistic classification based on the analysis of sentiments expressed within text written in Vietnamese and English language, NB classifier is utilised to perform the binary classification. Through the experimental work, binary classification is built to label the lexicons containing words and phrases that have been identified into positive or negative sentiments. At the end of the experiment, the results indicates that the NB classifier achieved the highest score, signifying that this algorithm is well suited for text or natural language classification systems because it performed better than other methods.

Results from these findings demonstrate a strong and consistent evidence that NB classifier is ideal to be implemented in this study to construct the binary classification for the gathered documents, to be labelled as either Disaster Management or Social Media. Based on the classification technique, the binary classification utilised for the supervised TWS can be performed. The next section discusses the eight supervised TWS selected for this study.

Universiti Utara Malaysia

2.5 Supervised Term Weighting Schemes

In this section, eight supervised TWS selected for this study are analysed and discussed. The selected schemes are Chi-square (x^2), Information Gain (IG), Odds Ratio (OR), Gain Ratio (GR), Relevance frequency (RF), Mutual Information (MI), TF.IDFEC-based, and Short-text Term Weighting (SW). In this section each of the schemes are explained with Table 2.7 summarises the comparison of the supervised TWS, based

on the authors, definition, and limitations as explained before. From the outcome of the term weightages, ontology for the DMSM domain is developed. The next section describes the structure of ontology, the importance of ontology, and ontology evaluation methods.

2.5.1.1 Chi-square (x^2)

Chi-square is a metric to compare the test of independence between two random variables, whereby a value is generated based on the association between term and the category (Şahin & Kılıç, 2019). For a value of zero, it indicates that the term and category are dissimilar in terms of relationship, meanwhile a higher computed CHI score expresses a closer relationship between the term and category exists. The mathematical equation for x^2 is shown in equation (2.4).

$$TF \cdot \chi^2 = TF(t_k, d_j)\ x\ N\ x\ \frac{(TP \cdot TN - FP \cdot FN)^2}{(TP+FN) \cdot (FP+TN) \cdot (TP+FP) \cdot (FN+TN)} \quad (2.4)$$

One of the applications of the Chi-square supervised TWS is seen in Alshaer, Otair, Abualigah, Alshinwan, & Khasawneh (2021) work. In this study, the Chi-square is utilized as the base for the supervised TWS. The purpose of this work is to study the effect of the scheme in processing the huge number of documents and pages in the Arabic language. The experimental results show that the average precision obtained is 80.40%, average recall is 75.93%, and average F-measure is 78.23%.

Another work by Shalaby & Zadrozny (2018) presented utilisations of supervised TWS, Chi-square for patent retrieval taking advantage of distributed representations of concepts and entities extracted from the patent's text. The Chi-square is utilized where weights are given based on how well each term is at discriminating the relevant

Universiti Utara Malaysia

and irrelevant candidates. The result shows that significant improvement is achieved with 4.6% in recall over the keyword search baseline.

Chi-square formulation is seen leveraging both document distribution for positive and negative categories. Based on (1), the numerator computes the differences of both classes, while in the denominator, *TP + FP* and *TN + FN* is to compute the distribution for positive and negative classes, with *TP + FN* and *TN + FP* enumerate the distribution of both classes. Chi-square emphasizes document distribution as a whole, which is seen as problematic when the expectation is to accentuate the positive class more than the negative classification. By using the formulation, it is expected that the result will indicate term extractions in a broader sense, namely Disaster Management instead of Disaster Management Social Media.

Universiti Utara Malaysia

2.5.1.2 Information Gain (IG)

Information Gain (IG) computes the term weightage by looking at the presence or absence of a term in categories by using entropy and information theory (Şahin & Kılıç, 2019). Entropy is a measurement of the uncertainty included in a probability distribution for random variables (H. Liu et al., 2018) as expressed in equation (2.5).

$$\omega_i = 1 + \Sigma_j \frac{(f_{ij}/g_i)*\log_2(f_{ij}/g_i)}{\log_2(n)} \qquad (2.5)$$

In equation (2), f_{ij}/g_i represents the probability of an event. In terms of TWS, f_{ij}/g_i is the term frequency divided by g_i, the number of times that term occurs in the corpus, and the number of documents in the corpus is denoted with *n*. By such a method, higher weight is given to the terms that occur infrequently in the corpus. From this expression, IG of a term is acquired by deducting the term entropy over the whole data set from

the sum amount of entropy values that exists to each category. The general equation of IG is annotated as in equation (2.6).

$$TF \cdot IG = TF(t_k, d_j)\, x - \frac{TP+FP}{N} \cdot log\left(\frac{TP+FP}{N} + \frac{TP}{N}\right) \cdot log\left(\frac{TP}{TP+FN} + \frac{FP}{N}\right) \cdot log\left(\frac{FP}{FP+TN}\right) \quad (2.6)$$

IG is used in measuring multi-label classification tasks done by Mazyad, Teytaud, & Fonlupt (2018). Through the experimental work, data set Reuters-21578 and Oshumed were used and micro-averaged and macro-averaged F_1 measures were selected as the evaluation metrics. The experimental results presented that the IG obtained an average F_1 measure of 79.28% micro-averaged and 44.0% macro-averaged rate.

Based on literature, even though IG is able to identify the features with the most information, IG experiences one issue mainly on unable to eliminate redundant features (Win et al., 2019). Therefore, the weighting has become biased in favour of features with more values even when they are not more meaningful (Zar Phyu & Oo, 2016). In text mining, it is necessary to exclude redundant features, since having irrelevant features is possible to increase computational time, hence impacting the system accuracy (Alhaj et al., 2016).

Universiti Utara Malaysia

2.5.1.3 Odds Ratio (OR)

Odds ratio (OR) calculates the ratio of the probability of occurrence of a term in one category to the probability of occurrence in another category (Onan, 2021). A positive score is given to terms that exist more in one category than others, and a negative score if it exists in the other (Hamdan et al., 2016). Whenever there is a score of zero, the term that exists in one category is absolutely similar as the odds for it to exist in the other. The equation of OR is shown in equation (2.7).

$$TF \cdot OR = TF(t_k, d_j)\ x\ log\left(\frac{TP \cdot TN}{FP \cdot FN}\right) \quad (2.7)$$

OR is used in a study by Onan & Tocoglu (2021) to investigate the predictive performance of term weighted schemes on sarcasm identification. OR in the experimental work achieved 71.50% of mean accuracy rate which was higher than TF-IDF TWS, with 71.30% mean accuracy rate.

This TWS like others also experience limitations, whereby OR main concern is data exaggeration. George, Stead, & Ganti (2020) stated that OR value is applicable to everywhere but still not always a useful measurement, particularly data exaggeration, in which OR is a relative measure, therefore there are times, a huge OR value can be falsely assumed as a little difference between odds.

Universiti Utara Malaysia

2.5.1.4 Gain Ratio (GR)

Gain Ratio is a type of decision tree method, whereby the choosing of attributes is based on GR. This scheme is an extension of Information Gain (IG) supervised TWS, in which the improved scheme addresses the limitation of IG through the reduction of its bias. The enhancement is accomplished by taking the intrinsic information from each attribute. The GR formula notation is defined in (5).

$$TF \cdot GR = TF(t_k, d_j)\ x\ \frac{IG}{-\frac{(TP+FP)}{N} \cdot log\frac{(TP+FP)}{N} - \frac{(FN+TN)}{N} \cdot log\frac{(FN+TN)}{N}} \quad (2.8)$$

GR is seen used in a study conducted by Haddoud et al. (2016) to develop text classification to help human indexers by suggesting thematic categories for biomedical articles. In this experimental work, IG is applied among other schemes, to find identify term representation. Through this work, the GR computation obtained a micro-averaged and macro-averaged of 0.518 and 0.50, respectively.

Based on Pintas, Fernandes, & Garcia (2021) GR takes account stressing on rare terms, however, rare terms are irrelevant to classification tasks. The reason is these terms have a very low likelihood of to exist in new documents. To address this issue, eliminating rare terms before initiating the term weighting process and adapting the relevance metric formula are suggested.

2.5.1.5 Relevance Frequency (RF)

Relevance Frequency considers term distribution on positive and negative classes (Carvalho & Guedes, 2020). Specifically, only the documents including the term are considered. As displayed in equation (6), the minimal denominator is set 1 to prevent division by zero.

$$TF \cdot RF = TF(t_k, d_j)\ x\ log\left(2 + \frac{TP}{\max(1, FN)}\right) \quad (2.9)$$

Universiti Utara Malaysia

A study by Carvalho & Guedes (2020), aims to present a comparative study on different TWS within Sentiment Analysis which is a branch of Affective Computing that usually considers a binary classification task. The experimental results by Carvalho & Guedes (2020) reveal that RF is almost showing unsatisfactory results on all four datasets.

The drawback of this RF is the scheme is discriminative between the terms. RF did not consider intra-class and inter-class distribution since the two of them are even taken as equally important in a supervised TWS.

2.5.1.6 Mutual Information (MI)

The objective of Mutual Information is based on representation learning, to discover useful representations. A single idea being explored is to train a representation-

learning function, for example an encoder, to maximize the Mutual Information (MI) between its inputs and outputs (Ye, 2020). In this case, it can be seen that the more correlation the two representation has, the higher mutual information is gained. The formula notation is shown in equation (2.9).

$$TF \cdot MI = TF(t_k, d_j)\ x\ log\left(\frac{TP \cdot N}{(TP+FP)\cdot(TP+FN)}\right) \quad (2.10)$$

Gao et al. (2018) evaluated the MI performance from a series of experiments. The results indicated that MI is underperformed with other schemes. In observing MI, the main issue it is very difficult to calculate MI when the data to be processed is high-dimensional as in the case of images (Ye, 2020). In addition, (Gao et al., 2018) elaborate that, MI suffers from an overestimation issue, in which unable to accurately quantify direct association among variables.

Universiti Utara Malaysia

2.5.1.7 Inverse Document Frequency Excluding Category - Based (IDFEC-based)

The basic idea of Inverse document frequency excluding category - Based (IDFEC-based) terms that happened in documents that reside to a similar class should be assigned with higher weights, which means, these terms is strong, with the equation shown in (2.11).

$$TF \cdot IDFEC - based(t_k, d_j) = TF(t_k, d_j)\ x\ log\left(2 + \frac{TP+FN+TN}{\max(1,FN)}\right) \quad (2.11)$$

A study related to IDFEC-based is presented by Maisonnave, Delbianco, Tohmé, & Maguitman (2019). In this study, the set of documents labelled as relevant or irrelevant to the domain is analysed. From the labelled document collection, a supervised technique that weights terms based on their discriminating and descriptive power is introduced. It is found that the schemes achieve the highest possible precision value for certain data.

IDFEC takes only negative examples, therefore, to remedy this, term that exists in the document, with the document associated to category c_k are taken into consideration. Hence, a new term-weighting scheme that they presented as IDFEC-based includes both negative and positive documents.

2.5.1.8 Short-text Term Weighting (SW)

Short-text term weighting (SW) introduced by Alsmadi & Hoon (2019) identifies and addresses the weighting issue of words in short documents and assess the result using the proposed scheme against existing methods, and the weightage of each word is compared in its document. The scheme is evaluated against Twitter tweet posts, with the equation shown in (2.12).

$$TF \cdot SW(t_k, d_j) = TF(t_k, d_j)\ x\ log\left(1 + \frac{TP}{FP+FN+1}\right) \quad (2.12)$$

Universiti Utara Malaysia

SW is seen that it can address the weaknesses of short texts and remediate the limitations of conventional approaches in the literature, thus enhancing performance to 80.83 and 90.64 (F-measure) on the Sanders dataset. This term weightage is newly introduced; therefore, the performance is yet to be measured. The formula indicated that ratio of the positive distribution, TP calculated against the other classification, without considering negative class documents that did not contains the term, TN. It is suggestive that term weightage to consider all document distribution class in determining the weightage.

2.5.1.9 Comparison of Supervised Term Weighting Schemes

Table 2.7 summarises the comparison of the supervised TWS, based on the authors, definition, and limitations as explained before. From the outcome of the term weightages, ontology for the DMSM domain is developed. The next section describes the structure of ontology, the importance of ontology, and ontology evaluation methods.

Table 2.7

Comparison of supervised TWS

Supervised Term Weighting Schemes	Author(s)	Definition	Limitation
Chi-square (x^2)	Debole and Sebastiani (2004)	Used to assess how independent the term and class are. This approach is quite heavy to execute and is better for classifiers to select the terms with highest (x^2) scores.	x^2 scheme did not concentrate either positive or negative classes. Therefore, the distribution is ambiguous.
Information Gain (IG)	Domeniconi et al. (2016), Deng et al. (2004), and Debole and Sebastiani (2004)	IG is utilised as if the provided term is present or absent for the predicted category in a document.	Unable to eliminate redundant features.
Odds Ratio (OR)	Mladenic and Grobelnik (1998)	The OR assigns higher scores to terms which frequently occur in one class, but rarely exists in the other class.	Data exaggeration.
Gain Ratio (GR)	Lan et al. (2009)	An improvement to IG scheme, to reduce of its bias issue. This is done by taking the intrinsic information from each attribute.	Stressing on rare terms, but, rare terms are irrelevant to classification tasks.
Relevance frequency (RF)	Lan et al., (2009, 2006)	Term with the highest frequency and distributed more in a positive class, the more possibility positive samples from negative ones is identified and vice versa. This factor is given to increase the positive and negative documents discrimination.	Discriminative between the terms since RF did not consider intra-class and inter-class distribution.
Mutual Information (MI)	Yang and Pedersen (1997)	Mutual information (MI) assesses the relationship concerning terms and classes.	MI suffers from an overestimation issue, in which unable to accurately quantify direct association among variables.
Inverse Document Frequency Excluding Category (IDFEC-based)	Domeniconi et al. (2016)	TF·IDFEC takes only negative samples, hence, the term-weighting scheme is improved known as TF·IDFEC-based to include both negative and positive documents.	IDFEC takes only negative examples, therefore, the documents appeared to a category c_k and consists of the term is considered. A new term-weighting scheme is introduced IDFEC-based to have negative and positive documents.

Universiti Utara Malaysia

Short-text Term Weighting (SW)	Alsmadi & Hoon (2019)	Term weighting method to perform classification on short text.	SW ignores the documents in negative distribution that did not contains the particular term, It is suggestive that all class information must be considered for the term weighting.

2.6 Ontology

Ontology is selected in this study as the knowledge representation, as described in the previous section, for the DMSM domain, on account of its benefit compared to normal conceptual modelling. The benefits of ontology are enabled knowledge reasoning, ensuring data semantic consistency, and human and machine interpretable. The description, structure of ontology, importance of ontology, and ontology evaluation methods are explained in this section.

As defined by Genesereth & Nilsson (1987), ontology is described as an explicit specification of a set of objects, concepts, and other entities that are predicted to occur in a several area of interest as well as their relationships. The function of ontology consists of specifying the domain concepts and primarily, it support the machine to examine the definition of concepts for the domain. This includes establishing the relationships between the concepts in ontology.

To illustrate the definition, taking an example of staff in a company with their relationships. For the corporation domain, given the concept *People*, *Supervisor*, and *Staff* will be relevant in which *People* is the super concept. A relevant relationship for these corresponding concepts is *Works-with* and a concrete person working in a company can be considered as an instance. The subsequent sections discuss the structure of ontology, the importance of ontology, and the ontology evaluation methods.

2.6.1 Structure of Ontology

Ontology is an approach of representing the relationships between entities within a knowledge domain. Iroju et al. (2012) stated that ontology comprises of several elements known as concepts, instances, relations, and axioms. The main part of ontology is called Concept, also can refer with class, and is usually associated by several attributes. For this context, attributes give concrete data fields which is data type properties. Taking into example in the disaster ontology, the concept Victim can be identified by *name* attribute. From the concept, an instance shows the element in the domain that assigned to particular concept. The concrete value is expressed in this way. Taking another example, *Ali* belongs to class *Victim* as an instance. In ontology, relation is assigned to show an interaction among the domain concepts and attributes.

The term *purchase* is used to convey a relationship for someone who bought an item from a seller. Axiom models sentences that are normally true. The possible meaning of its symbols is restricted through axiom by setting the constraint in which necessary to follow. In addition, axiom connects a symbol with other symbols of the ontology and restrict the potential interpretations of that symbol. Other than that, axiom is beneficial for other purposes whereas, information restriction, correctness validation, and new information introduction. Another container, a tuple maps a number of variables to some RDF terms (Cyganiak, 2005). A tuple comprises of three components which are subject, predicate, and object. A formal notation of ontology is given as follows based on the definition provided above.

An ontology is a tuple when,

Notation: O = <C; A; I;P>

This ontology is a tuple such that:

Universiti Utara Malaysia

One set of concepts C = {c_1, c_2……c_k} and

A = {A (c_1), A (c_2) ……A (c_k)} with A (c_1) = {a_{11}……a_{1n}} is a set of attributes assigned to concept c_1.

A set of relations R = {r_1, r_2……r_n} and

A set of instances I = {i_1, i_2……i_n} assigned to a concept C.

2.6.2 Ontology Classifications

Ontology is generally distinguished as lightweight and heavyweight ontologies in term of expressiveness (Ajami & Mcheick, 2018). As described by Eiden et al. (2021) lightweight ontology is an informal or semi-formal descriptions of ontology structures, focusing on representing domain knowledge rather than the creation of an all-encompassing and well-structured model of the world. Software engineering methods such as Entity Relationship Diagrams (ERD) and Unified Modelling Language (UML) can be used in developing lightweight ontologies.

In comparison with lightweight ontologies, even heavyweight ontologies contain similar structures, axioms and constraints are added to give the representation a deeper and meaningful information as a formal description as shown in Table 2.8. With these components, semantic interpretation of the presented entities, rules, and class constraints, and present multiple relations between concepts can be defined (Biagetti, 2020). The resulting structures are usually expressed as a separate complex model as an example by using semantic web language known as Web Ontology Language (OWL).

Universiti Utara Malaysia

Table 2.8

Differences between lightweight and heavyweight ontology (Biagetti, 2020).

	Lightweight Ontology	Heavyweight Ontology
Terminology Labels	Single or simple noun phrases expressing atomic concepts as the most typical case	Often, compound noun phrases denoting complex concepts
Relationship	Define "is-a" relation	Define "is-a", "part-of", or, more generally, "intersection" relation
Axioms/Constraints	Not defined	Defined to provide deeper meaning

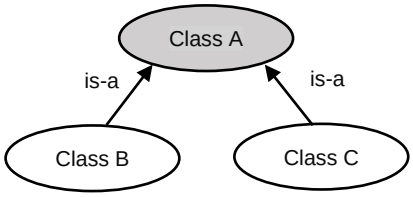

Figure 2.8 Example of lightweight ontology structure

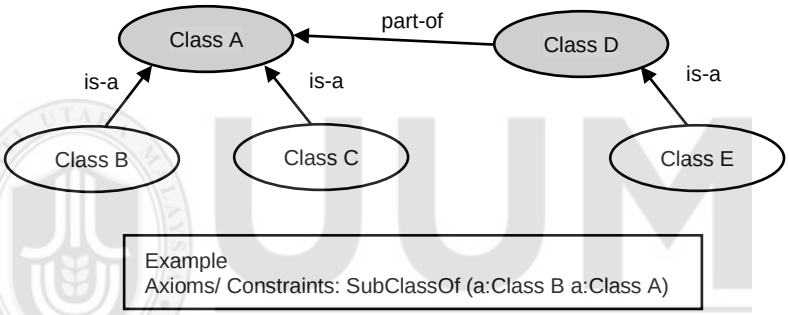

Figure 2.9 Example of heavyweight ontology structure

As shown in *Figure 2.8* and *Figure 2.9*, the difference of lightweight and heavyweight ontologies is shown structurally. For lightweight ontology comprising a set of concepts written in a natural language and organised in a hierarchy. Based on the definition in Table 2.8 lightweight ontology relationship can be defined with 'is-a' association and axioms and constraints are not defined, indicating the distinct difference with heavyweight ontology. Compared to heavyweight ontology, the structure is more complex with the relationships expressed with "is-a", "part-of", or, more generally, "intersection" relation. In addition, axioms or constraints is utilised to express semantic interpretation of the knowledge domain.

In this study, lightweight ontology is selected to be developed since the term weighting scheme performed extracted the terms to represent the domain knowledge without the axiom and constraint components to construct a heavyweight ontology. Based on the extracted terms, lightweight ontology is constructed, with the term expresses the nature and structured of the domain and "is-a" relation defined as the edge relationship (Guizzardi et al., 2021).

2.6.3 Importance of Ontology

Sugumaran & Storey (2006) and Noy & McGuinness (2001) explains that for Semantic Web developed ontology due to certain advantages. First of all, sharing of a unified view of the information structure between people or software applications is possible through ontology (Chandrasekaran et al., 1999). Once design of the concepts and relationship of the domain knowledge has completed, the ontology is reusable for others. This is helpful for other systems to reduce time and resources in building the ontologies.

Universiti Utara Malaysia

Ontology caters flexibility in altering explicit domain assumptions underlying an operation. Information describing a domain is expected to evolve depending on the situation, therefore this flexibility is very beneficial (Shadbolt et al., 2006). Another major point is that ontology presents domain different compared to operational knowledge. This separation involves an action to setup a product from its components based on a specified requirement. The program execution that implements this configuration did not depends on the products and components (Tankelevičienė, 2008). In addition, ontology also provides the analysis of the domain knowledge based on the terms declared in a document. All users have the flexibility to define their own attributes and relationships with other users.

Ontology can be considered as the basis of Semantic Web. From the concept of ontology, the Semantic Web is able to translate the accepted data into machine-interpretable language (Rubin et al., 2007). Ontology facilitates user involvement during the development which helps to acknowledge various users depending to their own perspectives related to the particular domain. Numerous ontology language editors known as Altova SemanticWorks, Protégé, Onto Lingua, and DOME and are able to aid engineers in constructing the Semantic Web.

To summarise, ontology ease information sharing, enable reusability, as well as domain knowledge separation and analysis. It also caters flexibility in domain changes and defining self attributes and relationships, as well as supports the Semantic Web technology and enables user cooperation in a domain. Additionally, there are numerous tools that can be used to build ontologies.

Universiti Utara Malaysia

2.6.4 Ontology Learning from Text

Ontology learning is a set of methods and techniques used for developing an ontology from beginning, making improvement, and adapting existing ontology through semi-automatic method by utilising several resources (Konys, 2019). The motivation behind of ontology learning is due to the huge cost of ontology building manually, constantly change in knowledge source generally, vast amount of existing text with numbers exponentially increased and the comprehensive need for various ontology type resources (Somodevilla García et al., 2018).

The process of ontology development can be understood from the ontology learning layer cake (Asim et al., 2018a). The process known as ontology learning layer cake involves term extraction and the synonyms from underlying text. Based on the identified terms and synonyms, both are combined to build concepts. Afterwards,

taxonomic, and non-taxonomic relations between these concepts are identified. At the end, from the unstructured text, axiom schemata are instantiated, and general axioms are extracted. This whole process of ontology learning layer cake is illustrated in *Figure 2.10*.

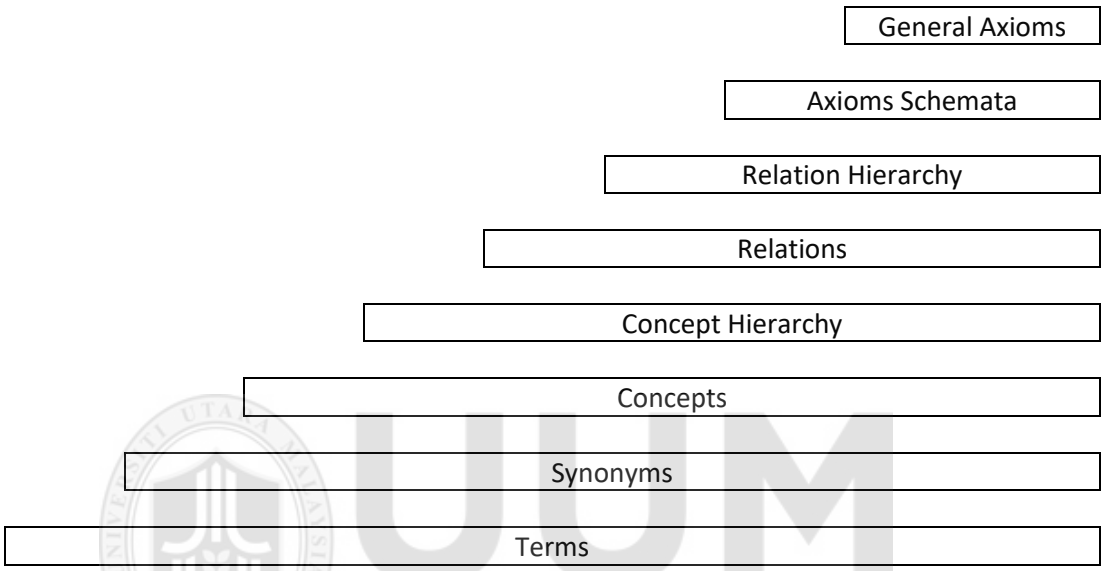

Figure 2.10 Ontology learning layer cake (Asim et al., 2018a).

Universiti Utara Malaysia

Somodevilla García et al. (2018 and Asim et al. (2018a) summarises that ontology learning is achieved with four fundamental approaches classified as text, linked data mining, concept learning, and crowdsourcing. Ontology learning from text utilizes automatic and semi-automatic construction of lightweight ontologies with the purpose of text mining and information extraction. Linked data mining is a type of resource that detect meaningful patterns in RDF graphs through statistical schema induction or statistical learning. Meanwhile, conceptual learning consists of learning schema axioms from the existing ontologies and instance data logically. Finally, ontology learning from crowdsourcing combines the speed of computers with the accuracy of humans.

This study focuses on the development of ontology learning from text with scientific publication as the main source, elaborated in section 3.5.1. Based on **Error! Reference source not found.**, the experimental work concentrates on the Terms. Synonyms, and Concepts layer in developing the lightweight ontology for DMSM domain, with the relationship expresses with 'is-a' as defined by Guizzardi et al. (2021) for lightweight ontology.

2.6.5 Ontology Evaluation Methods

Ontology evaluation is a method that improves the quality and validity of an ontology according to a set of evaluation standard, with respect to what aim of the ontologies are being evaluated and the objective (Amith et al., 2018). According to Asim et al., (2018) and Amith et al., (2018) review, there are four types of ontology evaluation methods that can be performed, namely, 1) gold-standard-based, 2) application-based, 3) data-driven-based, and 4) human-based evaluation. Gold-standard-based evaluation is an evaluation method of the learning ontology with an ideal ontology selected as the predefined benchmark (Amith et al., 2018). However, the main challenges with this method are to identify the suitable ontology, since the referenced ontology is expected to follow an aligned motive. This is to ensure that this evaluation fulfils all of the ontology evaluation criteria, which are completeness, conciseness, and accuracy.

Secondly, application-based method is to perform the evaluation by assessing the ontology in a particular application in performing desired tasks. The competencies of the learning ontology are assessed based on the outcome of the task performed. Nevertheless, the application-based evaluation has several drawbacks, one of which is the outcome of the evaluation performed is difficult to be generalized since the learning ontology is only evaluated after the task ended. Additionally, it is possible

Universiti Utara Malaysia

that the learning ontology is a small part of an application, therefore the effect over the results is perhaps indirect and minor.

Thirdly, ontology can also be evaluated through the existing domain-specific knowledge repository, by examining the knowledge extent in a specific domain. This evaluation enables the process of assessing of one or more learning ontologies with a certain corpus. Similar with the gold-standard-based evaluation, the main challenge is to gather for domain-specific corpus but a bit easier compared to seeking for domain-specific benchmark ontology.

Finally, for the human-based evaluation, the ontology is basically based on defining and executing numerous decision criteria for the selection of the best concept candidate. This method however since requires manual intervention by humans, is time and cost consuming.

Based on this review, gold-standard-based evaluation is selected as the evaluation method. This is because based on all four types of evaluations, the gold-standard-based is the best suited for this study considering an idealized outcome for the DMSM domain exists with a taxonomy introduced by Phengsuwan et al. (2021). Figure 2.11 and Table 2.9 show the summary and comparison between the ontology evaluation methods as discussed before.

Gold-standard-based evaluation has been applied in evaluating ontology in several works. Mohamed, Ellatif, & Farhan (2018) addresses requirement engineering issues through constructions of ontology based on requirements that have been gathered in a natural language form. The evaluation of the learned ontology is performed by assessing it with the gold standard ontology, built manually with the assistance of the domain experts in the research field. These differences between the learned and gold

standard ontology are considered errors, with precision and recall differences can be calculated, whereby the final results are 90% precision and 94% recall rate.

Qiu, Qi, Wang, & Zhang (2018) introduced a framework to build simple ontology from Chinese domain Web documents corpus. According to the framework, an evaluation is undertaken by using gold-standard-based evaluation build manually, with the news of football games as the target domain. Finally, by taking into account the differences between the learned and gold ontology, precision, recall, and F-measure are calculated with 77.26%, 76.04%, and 76.65% are achieved respectively.

A framework has been developed by Alobaidi, Malik, & Sabra (2018), to correctly identify concepts based on biomedical literatures. An Alzheimer's disease ontology developed by domain experts is selected as the gold ontology to compare with the generated ontology. The results show a successful evaluation with 72.48% F-measure rate for concepts identification.

Studies conducted by Alobaidi et al. (2018), Mohamed et al. (2018) and Qiu et al. (2018) indicated that the gold-standard-based evaluation is feasible and effective in evaluating the generated ontology based on the domain selected. The extraction process outcome can be evaluated against a gold standard to compare the capability of the introduced process and remove irrelevant candidate terms while preserving the relevant terms and expressions.

Precision, recall, and F-measure can be computed to express a formal measurement in assessing the supervised term weighting process effectiveness quantitatively. From the results, either improvement or continuation can be decided to obtain the best performance for the framework developed.

Universiti Utara Malaysia

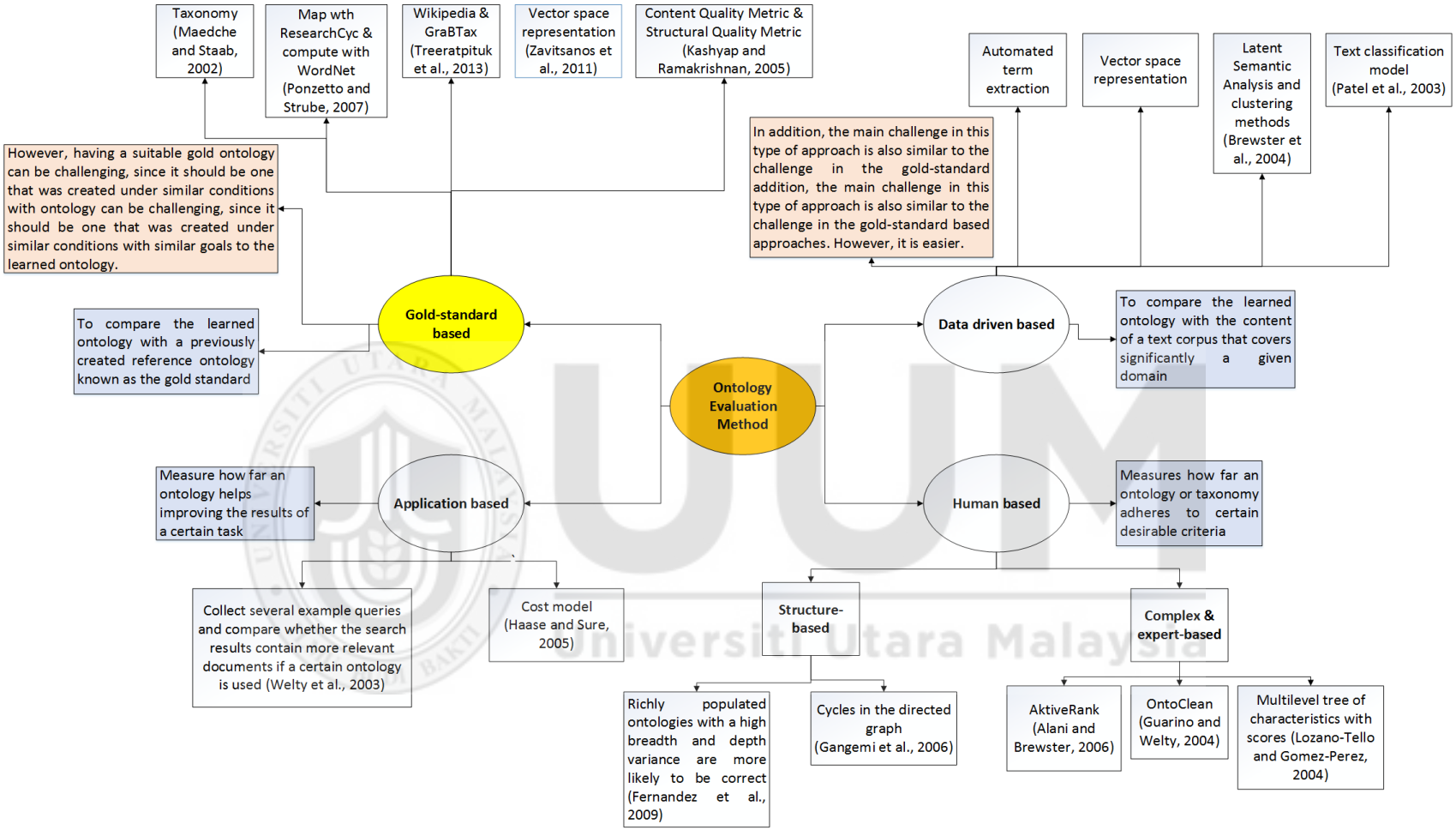

Figure 2.11. Summary of ontology evaluation methods

Universiti Utara Malaysia

Table 2.9

Comparison between the four ontology evaluation methods

Evaluation Method	Related Works	Definition	Limitations
Gold-standard-based	Alobaidi, M., Malik, K. M., & Sabra, S. (2018). Qiu, J., Qi, L., Wang, J., & Zhang, G. (2018). Mohamed, K. A., Abd Ellatif, M., & Farhan, M. S. (2017, December).	To evaluate the learning ontology with an ideal ontology selected as the predefined benchmark.	It is challenging to identify the suitable ontology, since the referenced ontology is expected to follow an aligned motive.
Application-based	Guo, B. H., & Goh, Y. M. (2017).	To perform the evaluation by assessing the ontology in a particular application in performing desired tasks.	The outcome of the evaluation performed is difficult to be generalized since the learning ontology is only evaluated after the task ended. Additionally, it is possible that the learning ontology is a small part of an application, therefore the effect over the results is perhaps indirect and minor.
Data-driven-based	Knoell, D., Atzmueller, M., Rieder, C., & Scherer, K. P. (2017). Araújo, W., Lima, G., & Pierozzi Jr, I.	To evaluate ontology through the existing domain-specific knowledge repository, by examining the knowledge extent in a specific domain.	Difficult to gather for domain-specific corpus.
Human-based	Jarosław, W. (2018).	To define and formulate numerous decision criteria for the selection of the best concept candidate.	Requires manual intervention by human which is time and cost consuming.

Universiti Utara Malaysia

2.7 Summary

This chapter has discussed the existing work found in the literature concerning the basic concepts of disaster, Disaster Management, Disaster Management Knowledge Representation, text mining, and ontology. The discussion begins with the overview of disaster and disaster management, social media role in DM, and DM Knowledge Repository. In this section, it is found that the Disaster Management Social Media domain still lacks ontology as the knowledge representation. Therefore, this study intends to develop a DMSM ontology as the knowledge representation, with the concepts for the ontology are identified through the text mining process.

Next, the discussion is continued with the concept of text mining and defining the gaps identified in the existing work which are answered by this study. The TWS used in the

text mining process consists of two types, unsupervised and supervised. Based on the review conducted, most text mining to ontology development utilised an unsupervised TWS, TF-IDF, which experiences the main limitation, namely, ignoring the classification information. To address this gap, this study applies supervised TWS, in order to enhance the text mining process. An analysis based on eight existing supervised TWS has been performed.

In addition, ontology and the formal evaluation of the completed ontology are discussed. Mainly, the ontology evaluation methods are analysed and discussed, from the four methods, 1) gold-standard-based, 2) application-based, 3) data-driven-based, and 4) human-based evaluation, gold-standard-based evaluation has been selected in this study for evaluating the ontology developed from the term weighting results. The following chapter describes how the proposed framework was developed.

Universiti Utara Malaysia

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In this chapter, the research methodology approaching the construction of enhanced supervised TWS. Section 3.2 provides a brief explanation to Design Science Research (DSR) presented as the fundamental in performing this research. In Section 3.3, the research framework is depicted and described. Phase 1 of this research is explained in Section 3.4 related to the problem identification, while Section 3.5 elaborates the identification of existing supervised TWS.

The construction of the enhanced supervised TWS for Disaster Management Social Media ontology development based on binary classification or Phase 3 is described in Section 3.6 and the evaluation of the enhanced supervised TWS through the developed ontology of Phase 4 is explained in Section 3.7. At the end, in Section 3.8, this chapter is ended with a summary.

3.2 Design Science Research

Design is a method with a sequence of ideas and operations by which an artefact is construct and realized (Andreasen et al., 2002). This concept supports an actualized research paradigm that grasps for the development of innovative artefacts to answer real-life problems (Cross, 2007; Simon, 1996), that include the construction of new design techniques, methods, and procedures (Cross, 2007).

From the definition, Design Science Research (DSR) is a research methodology that involves the development of an innovative and goal-based artefact for a distinct

domain issue (Hevner et al., 2004). Based on Hevner, March, Park, & Ram (2004), DSR involves several phases as guidance which are Problem Identification, Data Preparation, Development, and Evaluation. The purpose of this study was to identify existing supervised TWS, in which planning the design was necessary. The process of the design led to the construction of the enhanced TWS, for DMSM ontology development. The design research prepares the necessary understanding for performing the experimental process.

3.3 Research Framework

As briefly described in Chapter One, in this research framework, there are four phases specified as Problem Identification (Phase 1), Identification of Existing Supervised Term Weighting Schemes for Ontology Development (Phase 2), Construction of an Enhanced Supervised Term Weighting Scheme for Disaster Management Social Media Ontology Development (Phase 3), and Evaluation of the Enhanced Supervised Term Weighting Schemes through the Developed Ontology (Phase 4), as shown in Figure 3.1. To specify the problem for this research, in Phase 1 several reviews to investigate the appropriate theory, method, approaches, and practice have been conducted. With these steps, the research problem, objective, scope, and design can be derived as described in Chapters One and Two.

Universiti Utara Malaysia

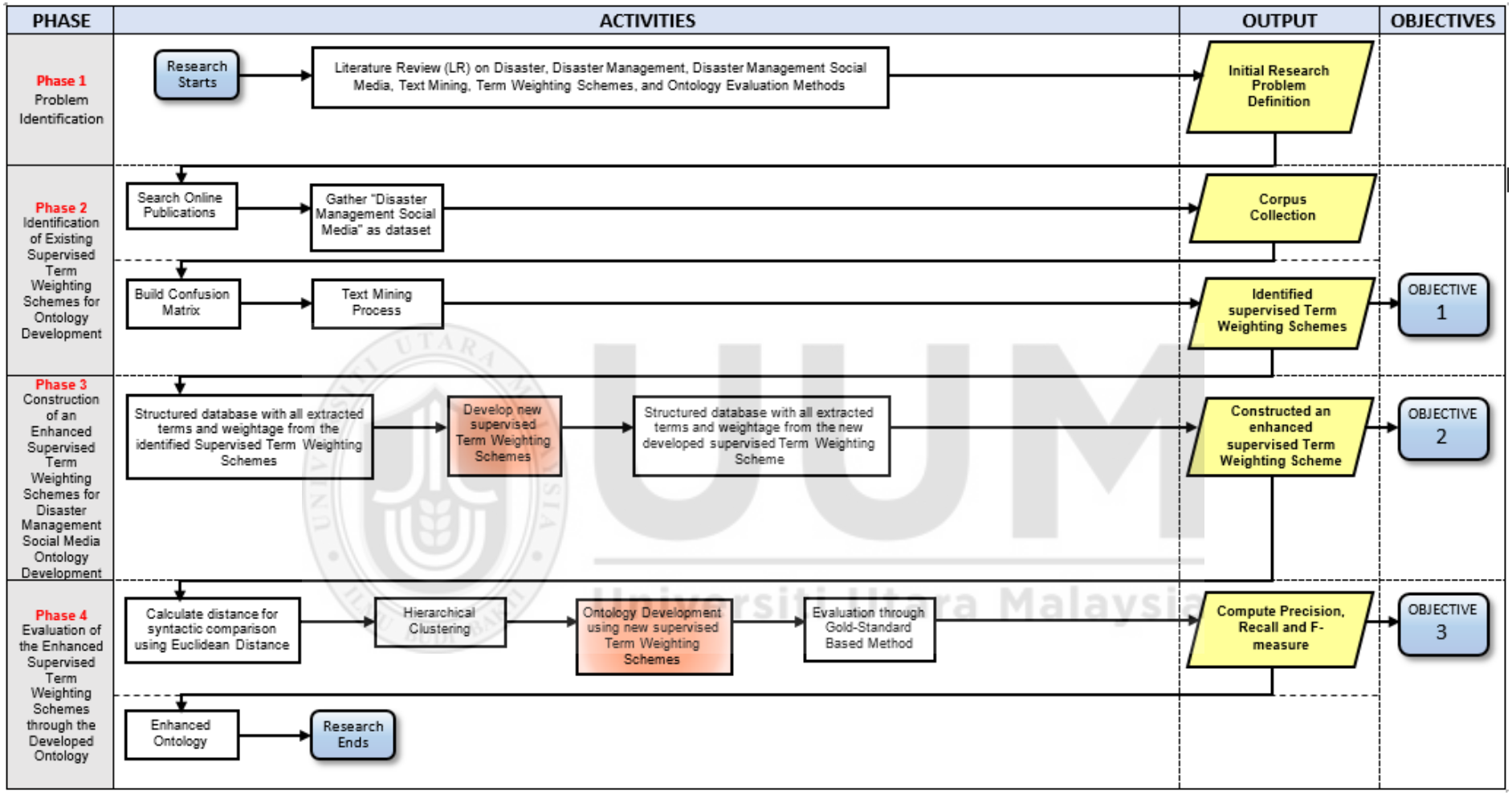

Figure 3.1. Research framework for present study

Phase 2 involves the identifying of existing supervised TWS for ontology development. In this phase based on the review conducted, the identified supervised term weighing schemes are assessed to observe the outcome from the experiment. This step answers the first objective of this study and detail explanation of this phase is elaborated in the next section.

Based on the results of Phase 2, Phase 3 involves the gathering of corpus collection by conducting publications search and downloading from several digital databases such as ACM DL, Google Scholar, ScienceDirect, Scopus, and Wiley Online Library. The construction of the enhanced supervised TWS was conducted by using the text mining approach to identify the concept candidate of the DMSM ontology development. This phase helps to address the second objective of this study.

The enhanced supervised TWS was evaluated through ontology development in Phase 4. From the generated ontology, the precision, recall, and F-measure is calculated through gold-standard-based ontology evaluation to address the third objective.

Universiti Utara Malaysia

3.4 Phase 1 - Problem Identification

To achieve the aim of the study, Chapters One and Two described Phase 1 in detail. All the key research concepts and those in the domain of interest; Disaster Management Social Media and text mining have been reviews on both chapters. The review concerns on the examination of the existing supervised TWS and in specifying the relevant components that can be further constructed for enhancement.

Previously, rigorous reviews were performed on the existing literature by focusing on the supervised TWS to identify the current techniques to extract relevant concepts.

Furthermore, the ontology evaluation methods were also reviewed to study the various methods in performing evaluation to ontology learning process.

3.5 Phase 2 - Identification of Existing Supervised Term Weighting Schemes

In this phase, the identification of existing supervised TWS for ontology development was undertaken. To identify the schemes, several works have been reviewed and analysed to be selected as the candidate for research work as described in Chapter Two. From the selected TWS, the below steps are coordinated to observe the performance of each of the schemes.

3.5.1 Step 1: Corpus Collection and Preliminary Domain Study

A number of 392 publications focusing on Disaster Management Social Media domain are gathered based on five online libraries which are ACM DL, IEEE, Google Scholar, ScienceDirect, Scopus, and Wiley Online Library that go through expert-reviewed and high-standard journals in the subject area. In this section, by using the knowledge source and scientific journal publications from the public domain the data collection process was carried out. Several criteria were defined for the knowledge source selection, which is based on reliability, prominent, and well-known among researchers. Furthermore, Ehnis (2018), Salih, Omar, & Yasin (2018), Stephenson et al. (2018), and Zaveri et al. (2016) stated that the databases contains a massive and extensive archive of scientific publications for corpus collection.

Scopus store publications from over five thousand publishers that are mandatory to be investigated and chosen by an independent board before being indexed on Scopus. Based on the 2017 content coverage guide, the Scopus database comprises of over

twenty thousand peer-reviewed journals and includes over three thousand six hundred full open access journals (Scopus, 2017). A reliable search strategy is provided in extracting a meaningful proportion of the published material concerning to the Disaster Management Social Media domain.

In making sure that the documents represented by the latest research works, the selection was restricted to publications within the last five years. Most notably, in the last five years, social media application has reached its peak of two billion users (Singh, 2019). This signifies that the knowledge base is quantitatively established from large data sets, thus ensuring well-grounded and validated information. All of the gathered publications afterwards are referred as corpus.

Step 1.1: Formulate Research Questions: The Research Questions in this experiment have been defined with the aims of establishing a clear boundary and limit for the corpus collection process. The Research Questions are formulated for this step and are expected to be addressed in this experimental works.

Universiti Utara Malaysia

Step 1.2: Define Search Keywords: The search strategy was executed based on the search keywords associated with DMSM. The keyword which is derived from the research questions is entered to gain the desired results.

3.5.2 Step 2: Identifying Sets of Document to Suit Research Tasks

Research questions are articulated to help analysing the quality scientific studies. As mentioned by Kitchenham (2004), the research questions need to be comprised of, 1) significance and importance to researchers, 2) impact of changes or boosting up the confidence in the value of the current software engineering process, and 3) distinct

dissimilarities between reality and normal beliefs. These criteria were adhered in preparing the research questions to collect documents for this experimental work.

Step 2.1: Perform Studies Selection Procedures: A search strategy was defined formally from the formulated research questions, and examination of all available materials relevant to the objective was conducted. In this step, the search terms were defined, literature resources were gathered, studies selection procedures were consolidated, and quality assessment was conducted. To set the limitation for the conducted systematic review, studies selection procedures were listed to differentiate the documents to positive and negative categories. In this experiment, all the documents gathered from the search process is consolidated. As elaborated in section 2.4, based on the search results, the documents that predicted relevant to social media in disaster management is predicted as positive class, while documents that predicted to be irrelevant to social media in disaster management is predicted as negative class.

The next step is to carefully inspect the documents in both categories, in which the content of the publications is examined thoroughly. The documents are categorised mainly based on predicted and observed categories, whereby predicted is the first assumption made to categorise the document either relevant to DMSM domain or not by going through the title and abstract section of the documents manually. After performing a thorough inspection, then the documents can be truly categorised as relevant or not, hence the observed category. Documents in positive category identified correctly relevant to social media in disaster management is categorised in the True Positive (TP) while remaining documents incorrectly predicted categorised in the False Positive (FP) category. Similar process is performed for the Negative category in which the correctly identified documents relevant to social media in disaster management is categorised in the True Negative (TN) category, while the

Universiti Utara Malaysia

other documents categorised as the False Negative (FN) category. Based on the confusion matrix in Table 3.2, the binary classification is categorised into four categories, TP, FP, TN, and FN, with N value indicating the total number of documents gathered.

3.5.3 Step 3: Classify Documents to Binary Classification

To perform the classification process, the RapidMiner application is used, due to its reputation as a prominent open source framework for information mining, (Gupta & Malhotra, 2015) with Mendeley Desktop produced by Elsevier to manage the research publications as summarized in Table 3.1.

Table 3.1

Development Tools Required for Text Mining and Concepts Extraction Process

Tool Name	Description	Version
RapidMiner Studio	Visual workflow designer for data science software platform.	Version 10.0.000
Mendeley Desktop	Reference management software for managing and sharing research publications.	Version 1.19.8
Microsoft Windows	Operating system for development machine.	Windows 11 Enterprise 64 bit

RapidMiner is capable as a standalone application for information investigation and as a data mining engine for the integration into self-products. It is built using Java programming language that provides an interactive Graphical User Interface (GUI) in designing and executing analytical workflows. The GUI helps in lessening the learning gap to operate this application. The workflows are called *Process* and consist of multiple *Operators*. Every operator performs a particular task within the process, whilst the output of each operator constructs the input of the next one. These

workflows and operators were used in this experiment to perform the text mining activities.

In addition, the RapidMiner application provides several features that suit this experiment. The first refers to the open source and operating system independence. Hence, this application is cost effective without demanding high software licensing fee.

Steps 3.1 until 3.9 list all the steps taken in performing the steps to each of the documents gathered. This process mainly utilising Naïve Bayes algorithm as discussed in section 2.4.5. Figure 3.2 depicts the entire steps developed by using Rapidminer application.

Step 3.1: Read Excel: This operator reads a training set from the specified Excel file. In this Excel, the document abstract is specified with the training label of the documents achieved in Step 2 before.

Step 3.2: Nominal to Text: This operator changes the type of selected nominal attributes to text. It also maps all values of these attributes to corresponding string values. The Nominal to Text operator converts all nominal attributes to string attributes. Each nominal value is simply used as a string value of the new attribute. If the value is missing in the nominal attribute, the new value will also be missing.

Step 3.3: Process Documents from Data: Generates word vectors from string attributes.

Step 3.4: Select Attributes: This Operator selects a subset of Attributes of the data set and removes the other Attributes. The Operator provides different filter types to make Attribute selection easy.

Step 3.5: **Set Role:** This Operator is used to change the role of one or more Attributes. The role of an Attribute describes how other Operators handle this Attribute. The default role is regular, while other roles are classified as special.

Step 3.6: **Cross Validation:** This Operator performs a cross validation to estimate the statistical performance of a learning model. It is mainly used to estimate how accurately a model will perform in practice. The Cross Validation Operator is a nested Operator which have two subprocesses, namely a Training subprocess and a Testing subprocess. The Training subprocess is used for training a model while the trained model is then applied in the Testing subprocess. The performance of the model is measured during the Testing phase. The evaluation of the performance of a model on independent test sets yields a good estimation of the performance on unseen data sets.

Step 3.7: **Naïve Bayes:** Naive Bayes is a high-bias, low-variance classifier, and it can build a good model even with a small data set. It is simple to use and computationally inexpensive. Typical use cases involve text categorization, including spam detection, sentiment analysis, and recommender systems.

Universiti Utara Malaysia

Step 3.8: **Apply Model:** A model is first trained on the data set by another Operator, which is often a learning algorithm. Afterwards, this model can be applied on another data set. Usually, the goal is to get a prediction on unseen data or to transform data by applying a pre-processing model. The data set upon which the model is applied, must be compatible with the Attributes of the model. This means, that the data set has the same number, order, type and role of Attributes as the data set used to generate the model.

Step 3.9: **Performance:** This operator can be used for all types of learning tasks. It automatically determines the learning task type and calculates the most common criteria for that type.

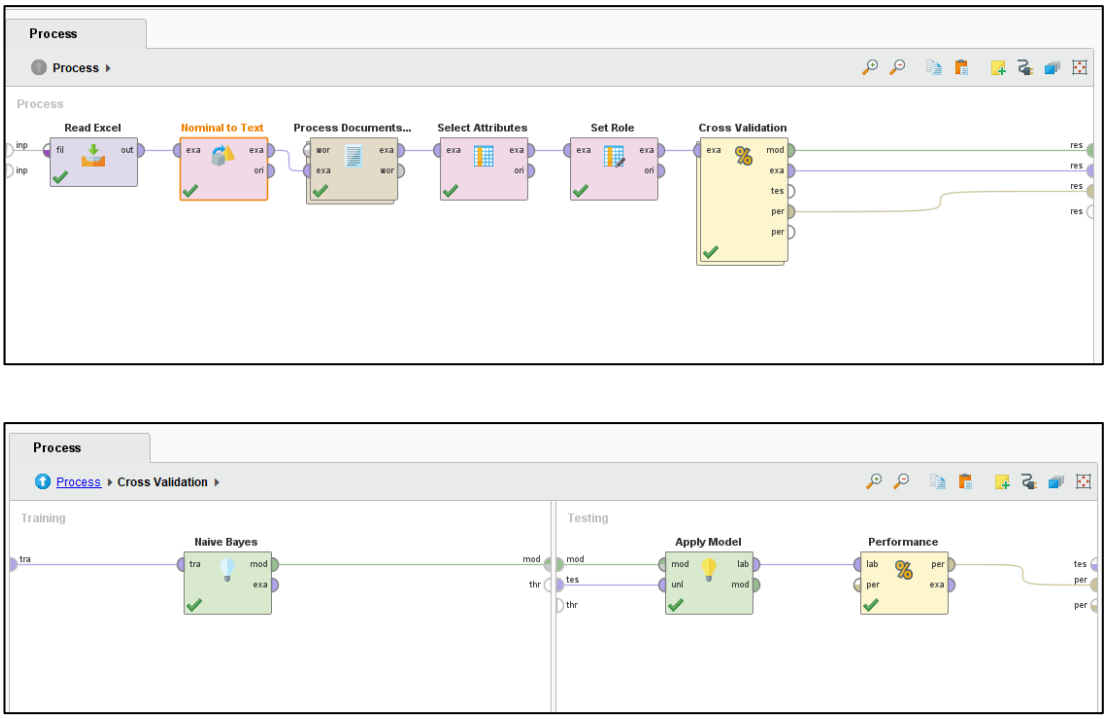

Figure 3.2. Classification process developed in Rapidminer.

Table 3.2

Confusion matrix constructed for DMSM publications.

N = 200		PREDICTED			
		POSITIVE		NEGATIVE	
OBSERVED	POSITIVE	10	TP	25	TN
	NEGATIVE	15	FP	150	FN

TP: True Positive, **TN:** True Negative,
FP: False Positive, **FN:** False Negative

The documents are assigned based on the automated categorisation by using the classification techniques Naïve Bayes as discussed in Chapter Two. As an example, the total number of documents gathered is 200, with 25 documents predicted as relevant to DMSM domain by classification techniques using abstract section of the publications and categorised in positive category. Remaining 175 documents are predicted to be irrelevant to the domain categorised in the negative category. From the 25 documents predicted in the positive category, after analysing the content, only 10

documents are correctly relevant to DMSM therefore labelled with TP category, while the remaining documents incorrectly identified are labelled as FP category. Similar in the negative category, after inspecting the content, observed documents correctly identified relevant to DMSM is labelled with TN, while remaining documents labelled with FN category.

3.5.4 Step 4: Extraction of General Concepts by using the Text Mining Technique

In order to accurately record the information acquired, this step is necessary to plan the data extraction process. Once the study protocol has been planned, the data extraction procedures must be performed to avoid the chances of being biased. The text mining process was performed to extract relevant terms to be included in the DMSM ontology developed to maintain the data integrity. As defined by Vijayarani & Janani (2016), text mining is an operation to identify meaningful information, knowledge, or pattern from unstructured document with different contents of the sources.

Structured information is constructed through text mining operation by transforming the words and phrases in unstructured information into quantitative values in the database and examined with traditional data mining techniques (Vijayarani et al., 2015). Verma et al. (2014) list several techniques from different areas undertaken by text mining which are information retrieval, data mining, information extraction, and natural language processing. With combination of these various stages of text mining operation, a workflow can be customised.

The information retrieval phase was focused on in this stage, in identifying and extracting terms, since the other stages such as data mining, information extraction,

and natural processing as mentioned previously was more adaptable for massive text mining process. Information retrieval at this stage refers to the matching of a user's query request for documents from a database repository. The text mining process is the initial step to obtain the document contents that are relevant to the research questions.

R Studio is useful as an open-source application, using R programming language, for statistical analysis and graphic production stage (Kaya et al., 2019). Two main packages have been imported to carry this research work which are 'pdftools' for text extraction, rendering, and converting of PDF documents and 'tm' for the purpose of performing text mining processes.

Table 3.3

Development Tools Required for Text Mining and Concepts Extraction Process

Tool Name	Description	Version
R Studio	Integrated Development Environment (IDE) for R, a programming language for statistical computing and graphics.	Version 1.4.1103
Mendeley Desktop	Research publications reference management software for managing and sharing.	Version 1.19.8
Microsoft Windows	Development machine operating system.	Windows 11 Enterprise 64 bit

As an open-source application, R Studio by using R programming language is capable for statistical analysis and graphic production stage (Kaya et al., 2019). 'pdftools' and 'tm' are the two main packages imported to carry this research work. 'pdftools' package is for text extraction, rendering, and converting of PDF documents, while 'tm' package is necessary for the purpose of performing text mining process. All the steps taken in performing the text mining process to each of the documents gathered is listed in Steps 3.1 until 3.4.

Step 3.1: Process PDF documents for conversion to text streams: The initial step in the text mining operation is to prepare the corpus, a collection of text documents, from unstructured into structured text. Entire texts in this corpus are transformed in a list and fully kept in memory, or known as volatile corpus, with all changes reflected to the corresponding object.

Step 3.2: Pre-processing for text mining: Once the corpus is transformed to structured text, pre-processing steps for text mining is carried out programmatically. As mentioned in the review, the general method of pre-processing is necessary consists of 1) tokenization, 2) plain text normalization, 3) context transformer to lower case, 4) punctuations removal, 5) numbers removal, 6) stop words removal, and 7) stemming.

Tokenization splits the text into streams of words or tokens in a document, whereby term weightage can be obtained by calculating the frequency of occurrences. Vijayarani & Janani (2016) defines tokenization as an operation in constructing tokens by categorising a stream of texts into phrases, words, symbols, and other meaningful elements. In Linguistics and Computer Science field, this technique is helpful as it constructs the part of the lexical analysis (Ramasubramanian & Ramya, 2013).

By default, data in the text format is considered as a set of characters generally. Only words that are available in the data set is required by the text analysis process. The requirement for a parser is for the document to be tokenized based on the reason. The reliability of the documents is also provisioned by tokenization technique.

Plain text normalization displays a compact representation instead of the content. In avoiding confusion between similar words in the lower- or upper-case form, context transformer to lower case is added to transform all characters in a document to lower or upper case. The word ‘Social’ and ‘social’ as an example, was transformed to lower case form to be normalized. As frequently punctuation symbols and numbers might

Universiti Utara Malaysia

exist in the text streams, punctuation and number removal is performed to remove all punctuation marks from the corpus. The Filter Stop words (English) operator was used to eliminate the most common words or stop words that are not helpful to text mining such as prepositions, articles, and pronouns. Since these types of words were not beneficial for the text mining process, the stop words were removed, as an example 'a', 'an', 'the', 'is', and 'are'. Finally, stemming maps different patterns of similar word to a common stem form. As an example, the English stemmer maps 'direction', 'directions', 'directive', 'directed', and 'directing' to connect. So, searching for 'directed' would also find documents that only have the other forms.

Step 3.3: Build term document matrix: A term document matrix is a way to represent the words in the text in a matrix form of numbers. The rows of the matrix indicate the text responses to be analysed, and the columns of the matrix represent the words from the text that will be used in the analysis.

Step 3.4: Compute term frequency by documents: The next step is to calculate the term frequency, whereby the term frequency is a dictionary, and each token is assigned to a weight. The weightage is indicated by the number of existences for a term from one document.

$$TF(t_i, d_j) = \frac{f_{ij}}{\sum_{t \in T_j} f_{tj}} \quad (3.1)$$

3.5.5 Step 5: Build Corpus Confusion Matrix

Based on the corpus categories, the documents are further classified into positive and negative. As described in the literature, documents that are related to the topic of DMSM will be classified as positive, while the rest are categorised as negative. In the positive category, further screening is done with documents that contain the term is

categorised as True Positive, with others as True Negative. Similar in the negative category, documents that contain the term is categorised as False Positive while others as False Negative. From the outcome of Step 3 by using Excel Worksheet, the list of terms and frequencies are consolidated into one workbook with two different worksheets, Positive and Negative.

The next step is to merge both lists into one master list to identify the TP, TN, FP, and FN values for each term, anyhow the process is time and effort consuming to be conducted manually since the existence of the massive data. Therefore, an algorithm has been developed to perform the merging with the use of Excel Visual Basic for Applications (VBA). In this algorithm, basically each term in the Positive category is mapped with the terms in the Negative category, on row by row between both worksheets, once identified will set the FP and FN value in the Positive category respectively.

Universiti Utara Malaysia

3.5.6 Step 6: Perform Unsupervised Term Weighting Schemes Computation

Previous works to construct ontology by using text mining process normally is using TF-IDF unsupervised TWS as explained in Chapter Two, section 2.3.3. Therefore, as an evaluation, the present study performed a text mining process by using TF-IDF in order to create a benchmark as a comparison between unsupervised and supervised TWS in constructing ontology. The algorithm to perform unsupervised TWS is shown in Table 3.4.

Table 3.4

The algorithm to perform unsupervised TWS by using TF-IDF

Algorithm TF-IDF: An unsupervised TWS by using TF-IDF

Input: A corpus of scientific publication documents $C_w = \{d_1, d_2, \ldots d_n\}$ for domain DMSM.

For every documents d_i in C_w **do**
    Add the d_i to VectorCorpus $VC_w = \{d_1, d_2, \ldots d_k\}$
End for

For every documents d_i in VC_w **do**
    Parse and pre-process the documents for term weighting process
        Parse d_i as plain text document
        Transform d_i to lower case
        Remove any punctuations symbol in d_i
        Remove number in d_i
        Remove all English stop words in d_i
        Perform stemming for all words in d_i to common stem forms
    Add to term document matrix $TDM_w = \{d_1, d_2, \ldots d_k\}$
End for

Parse the D_w to collect a set of terms $T_w = \{t_1, t_2, \ldots t_k\}$

For each term t_i in T_w **do**
    Parse d_i with bigram sequence of words
    Calculate the TF-IDF w_i of term t_i as,
    $w_t = [\sum (\text{Count of } t \text{ in } d_i) / \sum (\text{Count of terms in } d_i)] * [\log (n / d_f)]$
    where n indicates the count of documents in D_w
    d_f indicates the count of documents with term t_i
End for

Output: A set of terms with the calculated unsupervised TWS, TF-IDF values, $S_w = \{(t_1, w_1), (t_2, w_2), \ldots (t_k, w_k)\}$

Universiti Utara Malaysia

Generally, the algorithm begins with a corpus of scientific publications for the DMSM domain as the input. As regards the input, each document in the corpus is parsed and stored into the VectorCorpus which is the volatile memory storage. For every document stored in the memory, it will be pre-process based on the standard text mining process for the term weighting operation, which consists of parsing document as plain text, transforming document to lower case characters, removing any punctuation symbols, removing all numbers and all English stop words, and performing stemming for all words.

Once the pre-processing has been completed, all terms are stored as a term-document matrix, which is the document vectors in matrix form. The rows indicate the terms in

the document while the columns correspond to the documents in the corpus, with the cells indicates the weightage of the terms. Afterwards, as described in Chapter Two, all terms are parsed with bigram sequence of words, and the term weightage is calculated by using the TF-IDF formula. The outcome of this algorithm produced the extracted terms and the calculated unsupervised TWS values.

3.5.7 Step 7: Perform Existing Supervised Term Weighting Schemes Computation Based on the Candidate Terms

Existing supervised term weightage schemes were implemented in this research, as explained in Chapter Two, to calculate the term frequency weightage for each of the terms selected as the candidate to represent the DMSM domain. In Chapter Two also the explanations and descriptions regarding this scheme's formula have been explained. From the explanation provided in Chapter Two, the selected supervised TWS for computation are Chi-square (x^2), Information Gains (IG), Odds Ratio (OR), Gain Ratio (GR), Relevance Frequency (RF), Mutual Information (MI), Inverse Document Frequency Excluding Category (IDFEC-based), and Short-text Term Weighting (SW). All of the results from these computations are presented in Chapter Four. Table 3.5 show the algorithm developed in this study to compute all of the supervised TWS.

Table 3.5

The algorithm to perform supervised TWS by using the existing supervised term frequency schemes

Algorithm: Supervised TWS by using TF.x^2, TF.IG, TF.OR, TF.GR, TF.RF, TF.MI, TF.IDFEC-based, and TF.SW

STEP 1: Input: A corpus of scientific publication documents $C_w = \{d_1, d_2, \ldots. d_n\}$ for positive category.
For every documents d_i in C_w **do**
Add the d_i to VectorCorpus $VC_w = \{d_1, d_2, \ldots.d_k\}$
End for

For every documents d_i in VC_w **do**
Parse and pre-process the documents for term weighting process
Parse d_i as plain text document
Transform d_i to lower case
Remove any punctuations symbol in d_i
Remove number in d_i
Remove all English stop words in d_i
Perform stemming for all words in d_i to common stem forms
Add to term document matrix $TDM_w = \{d_1, d_2, \ldots d_k\}$
End for

Parse the D_w to collect a set of terms $T_w = \{t_1, t_2, \ldots t_k\}$

For each term t_i in T_w **do**
Parse d_i with bigram sequence of words
Calculate the term frequency for positive category tfp_i of term t_i as,
$w_t = [\sum$ (Count of t *in* d_i) / $\sum$ (Count of terms in d_i)]
End for

Output: A collection of terms with the calculated term frequency for positive category, $SP_w = \{(t_1, tfp_1), (t_2, tfp_2), \ldots (t_k, tfp_k)\}$.

STEP 2: **Input**: A corpus of scientific publication documents $C_w = \{d_1, d_2, \ldots d_n\}$ for negative category.

For every documents d_i in C_w **do**
Add the d_i to VectorCorpus $VC_w = \{d_1, d_2, \ldots d_k\}$
End for

For every documents d_i in VC_w **do**
Parse and pre-process the documents for term weighting process
Parse d_i as plain text document
Transform d_i to lower case
Remove any punctuations symbol in d_i
Remove number in d_i
Remove all English stop words in d_i
Perform stemming for all words in d_i to common stem forms
Add to term document matrix $TDM_w = \{d_1, d_2, \ldots d_k\}$
End for

Universiti Utara Malaysia

Parse the D_w to collect a collection of terms $T_w = \{t_1, t_2, \ldots t_k\}$

For each term t_i in T_w **do**
Parse d_i with bigram sequence of words
Calculate the term frequency for negative category tfn_i of term t_i as,
$w_t = [\sum$ (Count of t *in* d_i) / $\sum$ (Count of terms in d_i)]
End for

Output: A collection of terms with the calculated term frequency for negative category, $SN_w = \{(t_1, tfn_1), (t_2, tfn_2), \ldots (t_k, tfn_k)\}$.

STEP 3: **Input**: A set of calculated term frequency for positive $SP_w = \{(t_1, tf_1), (t_2, tf_2), \ldots (t_k, tf_k)\}$ *and negative* $SN_w = \{(t_1, tf_1), (t_2, tf_2), \ldots (t_k, tf_k)\}$ category.

For each terms t_i in SP_w **do**
Count d_i which contains t_i for TP value, tp_i in SP_w
Count d_i which did not contains t_i for TN value, tn_i in SP_w
Count d_i which contains t_i for FP value, fp_i in SN_w
Count d_i which did not contains t_i for FN value, fn_i in SN_w
Add tp_i, tn_i, fp_i, fn_i, tpf_i, and tfn_i to the confusion matrix CM_w
End for

Output: A set of confusion matrix, $CM_w = \{(t_1, tp_1, tn_1, fp_1, fn_1, tpf_1, tfn_1), (t_2, tp_2, tn_2, fp_2, fn_2, tpf_2, tfn_2), \ldots (t_k, tp_k, tn_k, fp_k, fn_k, tpf_k, tfn_k)\}$.

STEP 4: **Input**: A set of confusion matrix $CM_w = \{(t_1, tp_1, tn_1, fp_1, fn_1, tpf_1, tfn_1), (t_2, tp_2, tn_2, fp_2, fn_2, tpf_2, tfn_2), \ldots (t_k, tp_k, tn_k, fp_k, fn_k, tpf_k, tfn_k)\}$.

For every term t_i in CM_w **do**

Calculate the term weightage TF.x^2 of term t_i as,
$w_t = [(tpf_i + tfn_i) * \sum (\text{Count of } d_i \text{ in } C_w) * ((tp_i * tn_i - fp_i * fn_i)^2 / ((tp_i + fn_i)*(fp_i + tn_i)*(tp_i + fp_i)*(fn_i + tn_i)))]$

End for

For every term t_i in CM_w **do**
Calculate the term weightage TF.IG of term t_i as,
$w_t = [(tpf_i + tfn_i) * -(tp_i + fp_i) / \sum (\text{Count of } d_i \text{ in } C_w) * \log((tp_i + fp_i) / \sum (\text{Count of } d_i \text{ in } C_w) +$
$(tpf_i / \sum (\text{Count of } d_i \text{ in } C_w)) * \log((tp_i / (tp_i + fn_i)) + fp_i / \sum (\text{Count of } d_i \text{ in } C_w)) * \log((fp_i / (fp_i + tn_i)))]$

End for

For every term t_i in CM_w **do**
Calculate the term weightage TF.OR of term t_i as,
$w_t = [(tpf_i + tfn_i) * ((tp_i * tn_i / fp_i * fn_i))]$

End for

For every term t_i in CM_w **do**
Calculate the term weightage TF.GR of term t_i as,
$w_t = [(tpf_i + tfn_i) * (-(tp_i + fp_i) / \sum (\text{Count of } d_i \text{ in } C_w) * \log((tp_i + fp_i) / \sum (\text{Count of } d_i \text{ in } C_w) +$
$(tpf_i / \sum (\text{Count of } d_i \text{ in } C_w)) * \log((tp_i / (tp_i + fn_i)) + fp_i / \sum (\text{Count of } d_i \text{ in } C_w)) * \log((fp_i / (fp_i + tn_i)))) /$
$-((tp_i + fn_i) / \sum (\text{Count of } d_i \text{ in } C_w)) * \log((tp_i + fn_i) / \sum (\text{Count of } d_i \text{ in } C_w)) *$
$-((fn_i + tn_i) / \sum (\text{Count of } d_i \text{ in } C_w)) * \log((fn_i + tn_i) / \sum (\text{Count of } d_i \text{ in } C_w))]$

End for

For every term t_i in CM_w **do**
Calculate the term weightage TF.RF of term t_i as,
$w_t = [(tpf_i + tfn_i) * \log(2 + (tp_i / (\max(1, fn_i))))]$

End for

For every term t_i in CM_w **do**
Calculate the term weightage TF.MI of term t_i as,
$w_t = [(tpf_i + tfn_i) * \log(tp_i * \sum (\text{Count of } d_i \text{ in } C_w) / ((tp_i + fp_i) * (tp_i + fn_i)))]$

End for

For every term t_i in CM_w **do**
Calculate the term weightage TF.IDFEC-based of term t_i as,
$w_t = [(tpf_i + tfn_i) * \log(2 + (tp_i + fn_i + tn_i) / (\max(1, fn_i)))]$

End for

For every term t_i in CM_w **do**
Calculate the term weightage TF.SW of term t_i as,
$w_t = [(tpf_i + tfn_i) * \log(1 + (tp_i / (fp_i + fn_i + 1)))]$

End for

Output: A set calculated term weightage based on TF.x^2, TF.IG, TF.OR, TF.GR, TF.RF, TF.MI, TF.IDFEC-based, and TF.SW supervised TWS TF.x^2_w | TF.IG_w | TF.OR_w | TF.GR_w | TF.RF_w | TF.MI_w | TF.IDFEC-based$_w$ | TF.SW_w = *{(t_1, tw_1), (t_2, tw_2), ….(t_k, tw)}*.

The algorithm for this step is exhibited in Table 3.5, with the steps consist of distinct difference from unsupervised TWS with the term frequencies are calculated for both positive and negative corpus. The algorithm consists of four main components, 1) text mining pre-processing for positive category, 2) text mining pre-processing for negative category, 3) building corpus confusion matrix, and 4) supervised term weighting

calculation. In Step 1, the input of the algorithm is the corpus for positive documents, $C_w = \{d_1, d_2, ... d_n\}$ performed with text mining pre-processing process such as tokenization, plain text normalization, context transformer to lower case, punctuations removal, numbers removal, stop words removal, and stemming. The term frequency for each term are calculated later in this step. Similar process is performed for negative category in Step 2.

With the term frequency calculated, Step 3 continues to construct the confusion matrix for the particular corpus. Based on the classification, the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values are calculated, in which all of these values are utilised for the supervised TWS computation later. In the final step, Step 4, all of the selected supervised TWS in this study, Chi-square (x^2), Information Gain (IG), Odds Ratio (OR), Gain Ratio (GR), Relevance frequency (RF), Mutual Information (MI), TF.IDFEC-based, and Short-text Term Weighting (SW), with all of the schemes' formula are applied respectively. From the results of this step, the outcome is analysed for the construction of an enhanced supervised TWS in the next section, Phase Three.

Universiti Utara Malaysia

3.6 Phase 3 - Construction of an Enhanced Supervised Term Weighting Scheme for Disaster Management Social Media Ontology Development Based On Binary Classification

In this phase, the construction of the enhanced supervised TWS of this research was undertaken. To enhance the scheme, five steps were taken to further enhance the existing supervised TWS.

3.6.1 Step 1: Analyse Unsupervised Term Weighting Schemes Result

The computation for the unsupervised TWS has been computed from the previous step. Based on the result, analysis on the scheme, in particular TF-IDF is analysed as a comparison between unsupervised and supervised schemes. This is to evaluate how well the term weighting outcome against the current technique as the benchmark.

3.6.2 Step 2: Analyse Existing Supervised Term Weighting Schemes Result

From the computation applying existing schemes in Phase 2, the results for each calculation are analysed. This is to verify if the results conform to the objective and to identify the merit and fault of the existing supervised term weighting components.

3.6.3 Step 3: Perform Enhanced Supervised Term Weighting Schemes Computation Based on the Candidate Terms

Based on the existing supervised term weightage schemes, the construction of the enhanced TWS was implemented in this research. This is to retrieve the best representation for the DMSM domain candidate terms. Candidate terms refers to terminologies that used to represent the knowledge on DMSM domain.

Universiti Utara Malaysia

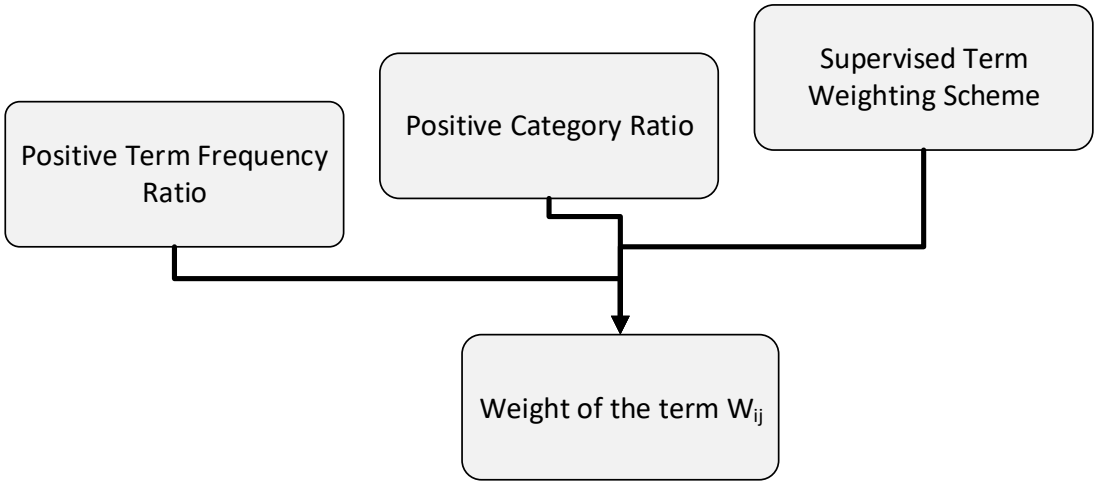

Figure 3.3. Enhanced Supervised TWS

From the previous result analysis, since most of the schemes are extracting concepts from the negative category, a positive category ratio is introduced. The positive category ratio takes the TP / TN value to compute the ratio of positive over negative observed class. The ratio is later multiplied with the expected positive class to obtain the overall positive values between the expected and observed class. The positive category ratio is multiplied with each of the existing TWS and new term frequency to discriminate the positive category of the classified documents as shown in Figure 3.3. The constructed enhanced supervised TWS is established formally in equation (3.2) and the algorithm in Table 3.6. The computation for this introduced enhanced TWS is presented in Chapter Four and the results are shown in Chapter Five.

$$TF.Enhanced = Positive\ Term\ Frequency\ Ratio * Positive\ Category\ Ratio * Supervised\ Term\ Weighting\ Schemes$$

$$= \frac{Term\ Frequency\ Positive}{Term\ Frequency\ Positive + Term\ Frequency\ Negative} * \left(\frac{TP}{TN} * FP\right) * Supervised\ Term\ Weighting\ Schemes \quad (3.2)$$

Universiti Utara Malaysia

Table 3.6

The algorithm to perform by using enhanced supervised TWS

Algorithm TF.Enhanced: A supervised TWS by using TF.Enhanced

STEP 1: Input: A corpus of scientific publication documents $C_w = \{d_1, d_2, \ldots. d_n\}$ for positive category.

For every documents d_i in C_w **do**
Add the d_i to VectorCorpus $VC_w = \{d_1, d_2, \ldots. d_k\}$
End for

For every documents d_i in VC_w **do**
Parse and pre-process the documents for term weighting process
Parse d_i as plain text document
Transform d_i to lower case
Remove any punctuations symbol in d_i
Remove number in d_i
Remove all English stop words in d_i
Perform stemming for all words in d_i to common stem forms
Add to term document matrix $TDM_w = \{d_1, d_2, \ldots. d_k\}$
End for

Parse the D_w to collect a set of terms $T_w = \{t_1, t_2, \ldots. t_k\}$

For each term t_i in T_w **do**
Parse d_i with bigram sequence of words
Calculate the term frequency for positive category tfp_i of term t_i as,
$w_t = [\sum$ (Count of t *in* d_i) / $\sum$ (Count of terms in d_i)]
End for

Output: A collection of terms with the calculated term frequency for positive category, $SP_w = \{(t_1, tfp_1), (t_2, tfp_2), \ldots. (t_k, tfp_k)\}$.

STEP 2: Input: A corpus of scientific publication documents $C_w = \{d_1, d_2, \ldots. d_n\}$ for negative category.

For every documents d_i in C_w **do**
Add the d_i to VectorCorpus $VC_w = \{d_1, d_2, \ldots. d_k\}$
End for

For every documents d_i in VC_w **do**
Parse and pre-process the documents for term weighting process
Parse d_i as plain text document
Transform d_i to lower case
Remove any punctuations symbol in d_i
Remove number in d_i
Remove all English stop words in d_i
Perform stemming for all words in d_i to common stem forms
Add to term document matrix $TDM_w = \{d_1, d_2, \ldots. d_k\}$
End for

Parse the D_w to obtain a collection of terms $T_w = \{t_1, t_2, \ldots. t_k\}$

For every term t_i in T_w **do**
Parse d_i with bigram sequence of words
Calculate the term frequency for negative category tfn_i of term t_i as,
$w_t = [\sum$ (Count of t *in* d_i) / $\sum$ (Count of terms in d_i)]
End for

Output: A set of terms with the calculated term frequency for negative category, $SN_w = \{(t_1, tfn_1), (t_2, tfn_2), \ldots. (t_k, tfn_k)\}$.

STEP 3: Input: A set of calculated term frequency for positive $SP_w = \{(t_1, tf_1), (t_2, tf_2), \ldots. (t_k, tf_k)\}$ *and negative* $SN_w = \{(t_1, tf_1), (t_2, tf_2), \ldots. (t_k, tf_k)\}$ category.

For every terms t_i in SP_w **do**
Count d_i which contains t_i for TP value, tp_i in SP_w
Count d_i which did not contains t_i for TN value, tn_i in SP_w

Universiti Utara Malaysia

Count d_i which contains t_i for FP value, fp_i in SN_w
Count d_i which did not contains t_i for FN value, fn_i in SN_w
Add tp_i, tn_i, fp_i, fn_i, tpf_i, and tfn_i to the confusion matrix CM_w

End for

Output: A set of confusion matrix, $CM_w = \{(t_1, tp_1, tn_1, fp_1, fn_1, tpf_1, tfn_1), (t_2, tp_2, tn_2, fp_2, fn_2, tpf_2, tfn_2), \ldots.(t_k, tp_k, tn_k, fp_k, fn_k, tpf_k, tfn_k)\}$.

STEP 4: Input: A set of confusion matrix $CM_w = \{(t_1, tp_1, tn_1, fp_1, fn_1, tpf_1, tfn_1), (t_2, tp_2, tn_2, fp_2, fn_2, tpf_2, tfn_2), \ldots.(t_k, tp_k, tn_k, fp_k, fn_k, tpf_k, tfn_k)\}$.

For every term t_i in CM_w **do**
Calculate the term weightage TF.Enhanced$_w$ of term t_i as,
$w_t = [(tpf_i / tpf_i * tfn_i) * ((tp_i / tn_i) * fp_i) *$ Supervised TWS]

End for

Output: A set calculated term weightage based on enhanced supervised TWS TF.Enhanced$_w$ = $\{(t_1, tw_1), (t_2, tw_2), \ldots.(t_k, tw)\}$.

In Table 3.6, the algorithm for the enhanced supervised TWS is presented. Basically, the pre-processing steps, Step 1, Step 2, and Step 3 are similar as the existing supervised TWS as discussed in section 3.5.7, however enhancement has been made for Step 4 by shifting the term weighting component with the enhanced scheme. Based on the introduced enhanced supervised TWS, the eight selected schemes are integrated into the formula in order to obtain the best result. These results are discussed and analysed in the next section to select the optimum result for the DMSM ontology development.

Universiti Utara Malaysia

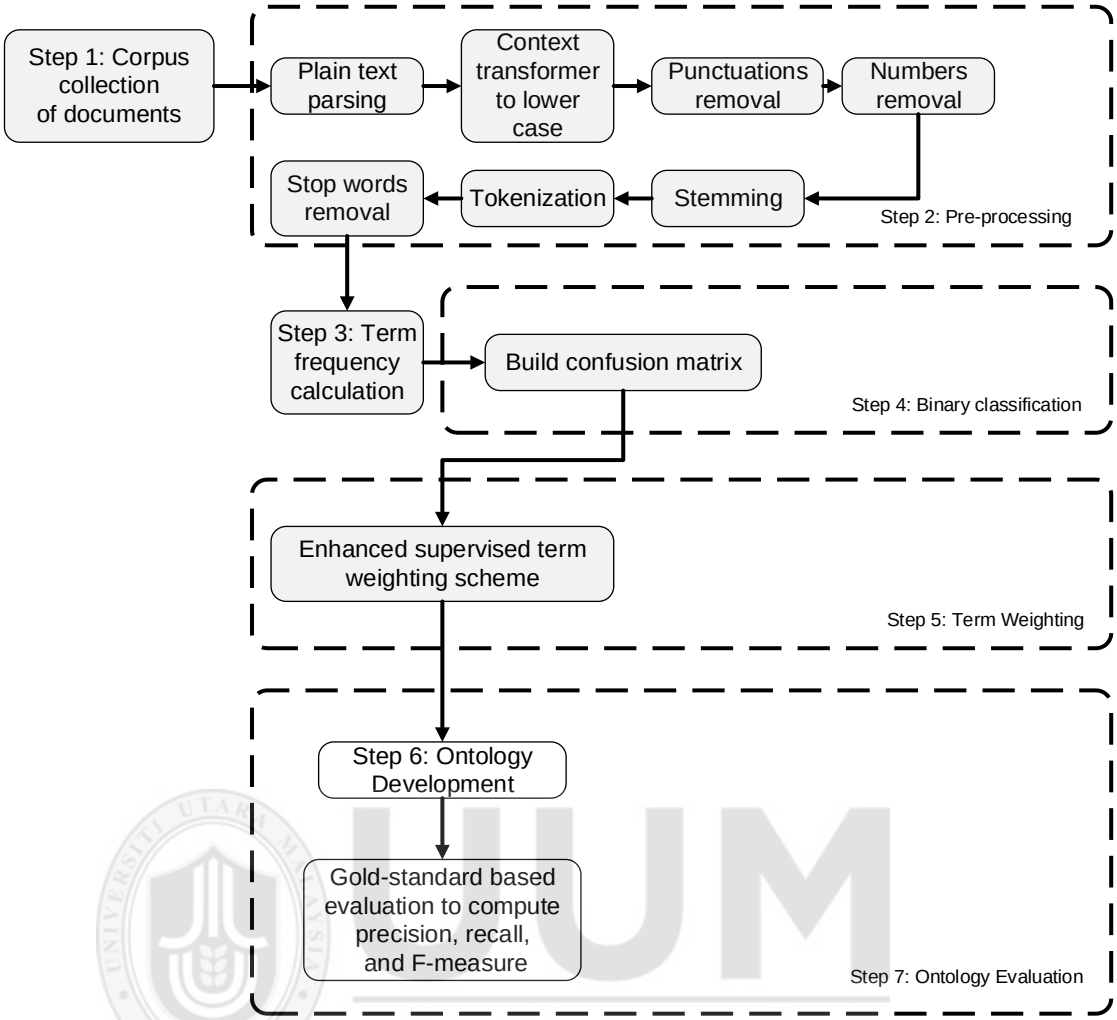

Figure 3.4. System architecture for the enhanced supervised TWS

Figure 3.4 depicts the system architecture as conducted in this study, in order to perform the enhanced supervised TWS. This architecture is an enhancement, whereby the existing architecture has been presented and discussed in Chapter Two, section 2.3.3. The operation begins with Step 1 by collecting a set of documents as a corpus and categorised by positive and negative categories. Positive category contains documents related to DMSM while negative category consists of document unrelated to DMSM. The next step, Step 2 is to perform pre-processing operation which are parsing document as plain text, transform document to lower case characters, remove any punctuation symbols, remove all numbers, remove all English stop words, and

performing stemming for all words. From the pre-processing step, the term frequency for each term is calculated in Step 3 so that the next step, Step 4, binary classification can be performed. Confusion matrix is built in the binary classification, so that True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values can be calculated. These values are crucial for the next step, Step 5, term weighting in order to compute the weightages by using the enhanced supervised TWS. Based on the results, the foremost scheme results are selected and used in to build ontology for the DMSM domain in Step 6. The generated ontology then is evaluated against the gold-standard ontology to compute the precision, recall, and F-measure in the final step, Step 7.

As compared to the existing architecture for text mining-based ontology development framework as shown in Chapter Two, section 2.3.3, enhancement have been made to the Step 1 of the process until Step 5, indicated with grey boxes, which is the term weighting calculation. Different to the existing architecture, instead of handling all document distribution as the same, the corpus is categorised by positive and negative class for the document's binary classification. In Step 2 and 3 also, the pre-processing and term frequency computation is performed based on the category classification, in which the positive and negative documents are processed separately. The confusion matrix is developed in Step 4, based on the computed term frequency, categorised by TP, TN, FP, and FN class in the binary classification.

Significant enhancement has been taken by using supervised TWS, as a substitute to existing works in which using unsupervised TWS, TF-IDF. Eight supervised TWS are selected for this study, which are Chi-square (x^2), Information Gain (IG), Odds Ratio (OR), Gain Ratio (GR), Relevance Frequency (RF), Mutual Information (MI), Inverse Document Frequency Excluding Category (IDFEC-based), and Short-text Term

Universiti Utara Malaysia

Weighting (SW), analysed and compared the results in order to be applied in the ontology development, Step 6. Finally, Step 7 assessed the generated ontology by using gold-standard based ontology evaluation method in order to examine the precision, recall, and F-measure rate of the mapping concepts.

3.6.4 Step 4: Analyse Enhanced Supervised Term Weighting Schemes Results

In conformity with the computation results, the outcome of the enhanced supervised TWS is analysed. This is to discover which factors for the enhanced formula are best suited for this research objective. The most suitable factors are determined on the basis that the TWS results must be towards the positive category and extracting terms related to DMSM domain. To interpret the result the definitions of category information are defined. The results of supervised TWS are categorised based on this information. As shown in Table 3.7, the category information is classified as their category, negative, neutral, and positive. Based on the document classifications, the negative category indicated that the term weighting results is distributed more on Disaster Management class, while the positive category is more on Disaster Management Social Media. The positive category is the aim of this study in order to obtain the most relevant extracted terms as representation for Disaster Management Social Media. There are extracted terms that are extracting ambiguous terms, which means neither belong to negative or positive categories, therefore assigned to a neutral class. The analysis for this step is presented in detail in Chapter Four and the results in Chapter Five.

Universiti Utara Malaysia

Table 3.7

Definition of category information based on document classifications

Category Information	Document Classification
Negative	Disaster Management
Neutral	-
Positive	Disaster Management Social Media

3.6.5 Step 5: Construct Structured Database with all Extracted Terms and Weightage

All of the extracted terms and weightages are consolidated in a form of a structured database. The database consists of rows and columns, with the rows consist of the data dimension or the extracted terms, and the columns indicate the assigned weightages based on the TWS calculated. This way, the relationships between the extracted terms and weightages constructed from the enhanced supervised TWS can be observed. The database is ranked based on the highest to the lowest term weightages.

Universiti Utara Malaysia

Table 3.8

Example of structured database for the enhanced TF.RF

No.	Extracted Terms	Weightages
1	social media	178.4146682
2	disaster	63.99642041
3	can use	62.00213299
4	case study	56.04065831
5	emergency management	53.12202623
6	use	50.2763298
...	...	...

In this study, Microsoft Excel is used as the platform to place the extracted values and weightages as a collection of connected items. The example of a structured database is shown in Table 3.8 with TF.RF extracted terms and weightages. The next section

presents the evaluation process in examining the term candidates for the DMSM ontology domain systematically.

3.7 Phase 4 - Evaluation of the Enhanced Supervised Term Weighting Schemes through the Developed Ontology

As compared to the other practices, one of the important processes of the DSR refers to the construction of an artefact that must be defined carefully, formally, logically, and consistently (Hevner et al., 2004). Therefore, the enhanced supervised TWS requires the evaluation stage to compute that the proposed scheme's precision, recall, and F-measure as discussed in Chapter Two. Four main sub-steps are performed in this phase in order to carry out the operation, which are, 1) perform term filtering to the extracted terms, 2) calculate the distance for syntactic comparison, 3) design the learned ontology, and 4) evaluate the learned ontology through gold-standard based evaluation method.

Universiti Utara Malaysia

3.7.1 Step 1: Perform Term Filtering to the Extracted Terms

Since the database might contain noise or irrelevant terms, this step is necessary to ensure the integrity of the data. Several steps are applied to filter the database, namely, steps 1.1 until 1.4 taken in performing the text mining process to each of the terms extracted.

Step 1.1: Illegal characters removal: Illegal characters that exist as the candidate term is expected to be removed since it did not provide useful information to the study. Examples of illegal characters are colons, brackets, and question marks.

Step 1.2: Verbs removal: Verbs refer to a word or phrase that describes an action, condition, or experience. The purpose of removing verbs from the candidate terms since ontology commonly utilizes nouns as the term selection. Examples of verbs that are identified to be removed are ‘use’, ‘need’, and ‘can’.

Step 1.3: Scientific publication terms removal: Since the corpus gathered is from collections of scientific publications, the scientific publication terms are expected to be removed to lessen the noise. Examples of the terms enlisted are ‘author’, ‘open access, and ‘journal.

Step 1.4: Irrelevant terms removal: In case there are terms that are considered irrelevant to the term extraction process, this step can be applied to omit the meaningless data. In this, more meaningful information can be perceived from the text mining process conducted.

Once all of the noise or insignificant terms have been eliminated, the extracted terms are equipped to be taken into account as candidates for ontology development. The next steps are carried out to determine the relationships among the concepts.

3.7.2 Step 2: Calculate Distance for Syntactic Comparison

Calculating distance is important to observe the association between one term to another. Two methods that can be applied to perform this step are by using Euclidean distance and hierarchy of terms.

Step 2.1: Compute Euclidean Distance: Euclidean distance is described as a direct straight-line interval between two entities in the Euclidean area. When performing the clustering, the Euclidean distance is used, in which when the interval among two points from the centroid is similar, but the points are either in the similar or opposite

directions, then also they may set into the similar cluster (Patel & Upadhyay, 2020). Based on this definition, the distance for each term is computed for one another by using this formula (3.3).

$$E = \sqrt{(x - y)^2} \tag{3.3}$$

As an example, the weightage for Relevance Frequency TWS, the term 'disaster recovery' and 'early warning' is 13.64 and 14.38, respectively, therefore the Euclidean distance will be 0.74.

Step 2.2: Perform hierarchy of terms using hierarchical clustering method: In order to identify the pattern for the extracted terms, two main techniques can be used based on two types of machine learning techniques, unsupervised and supervised. One of the commonly used methods for unsupervised learning is clustering (Allahyari et al., 2017), whereby clustering aims at grouping data objects (observations, examples) into groups or clusters (Ertek et al., 2017). For supervised learning, classification is used, in which there are many well-known techniques for classification, such as Naïve Bayes, Decision Tree, and Support Vector Machine.

However, for this study even though supervised TWS is applied, unsupervised machine learning technique or clustering is used because to develop the ontology, there are no predefined class for the terms to be assigned with, in case classification technique is used. Classification is more suitable to be performed in this step due the technique is able to identify similarities between the terms. One of the well-established clustering algorithms is hierarchical clustering, applied in this step to facilitate the clustering process.

Hierarchical clustering method is implemented in reference to the Euclidean distance computed. Clustering is an unsupervised machine learning technique, which targets at

categorising data objects into a grouping (Ertek et al., 2017). This step is performed such that assumptions within a cluster have high similarity, whereas they are dissimilar to the observations in other group. In this study context, the clustering is performed to define the 'is-a' relationship for the lightweight ontology concepts hierarchically. Orange Data Mining application which is an open-source data visual programming toolkit, is used in this step to perform the hierarchy of terms computation.

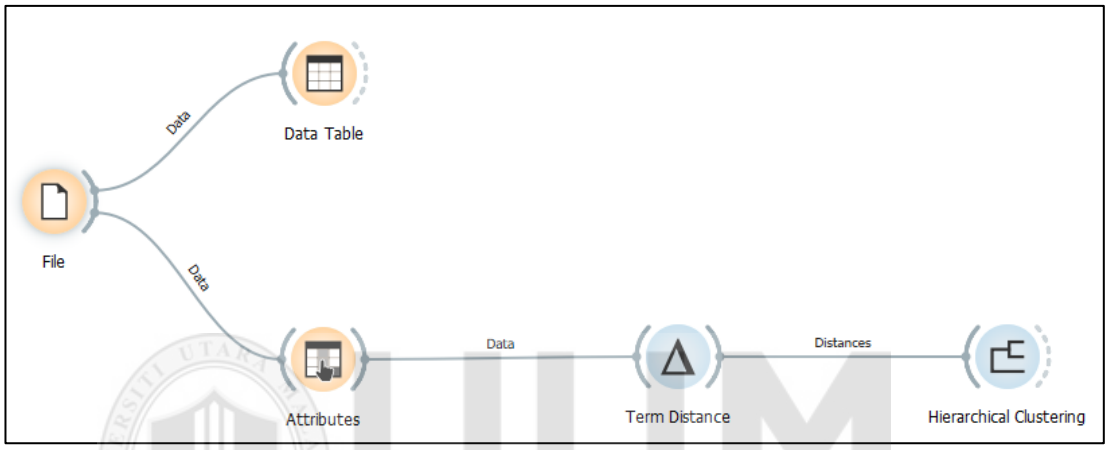

Figure 3.5. Hierarchy of terms by using hierarchical clustering method performed in Orange Data Mining.

Universiti Utara Malaysia

Figure 3.5 depicts the process programmed to carry out hierarchical clustering by using Orange Data Mining application. This application provides a selection of widgets for user to load and program desired tasks, in this case the widget 'File', 'Data Table', 'Attributes', 'Term Distance', and 'Hierarchical Clustering' are selected and executed. The widget 'File' enables the structured database to be uploaded, with 'Data Table' widget to display the information on the data input selected. In this way, user will be able to confirm and verify the data loaded is correct for the task. 'Attributes' widget allows user to select which column or attribute to be used from the structured database. The Euclidean Distance can be automatically computed by using the 'Term Distance' widget, and finally 'Hierarchical Clustering' widget provides the construction for

hierarchy of terms based on the structured database. The result for the step is shown in Chapter Four.

3.7.3 Step 3: Design the Learned Ontology

Based on the hierarchy of terms generated, the learning ontology can be designed. As discussed in section 2.6.2, the type of ontology developed in this study is stated to as lightweight ontology. However, the ontology design still requires human or manual intervention to ensure the concepts are properly mapped. Basically, based on the hierarchy of terms result, the terms are associated accordingly based on the clustering with 'is-a' definition for parent and child nodes. The developed ontology based on supervised TWS result is presented and explained in Chapter Six.

Universiti Utara Malaysia

3.7.4 Step 4: Evaluate the Learned Ontology through Gold-standard Based Evaluation Method

As discussed in Chapter Two, the selected method to conduct the evaluation for the learned ontology is through gold-standard evaluation method. Certain steps need to be taken in performing the evaluation as stated below.

Step 4.1: Extract gold-standard-based ontology concepts: Gold-standard reference is used in this evaluation as explained in Chapter Two by using the taxonomy introduced by Phengsuwan et al. (2021). As mentioned by Biagetti (2020), taxonomies is part of lightweight ontology, therefore this representation can be considered as comparison for evaluation. All the concepts identified from the referenced taxonomy are extracted to be utilised as the mapping with the learned ontology. The extracted

terms are presented in Table 3.9. Each of the referenced extracted terms is mapped with learned ontology concepts to compute the similarities.

Table 3.9

List of extracted terms from the gold-standard taxonomy

Extracted Terms from Phengsuwan et al. (2021) DMSM Taxonomy		
Data source	Social Media User	Government Authority
		Research/Academic Institution
		NGO
		Public
	Sensor	Human Sensor
		Physical Sensor
	Social Media Platform	
	Third Party	
Language	Mixed Script	
	Mixed Language	
	Local Language	
	Global Language	
Information Dimension	Temporal	Post-event
		Real-time
		Pre-event
	Spatial	Spatial Coverage
		User-defined
		Geo-tagging
Methodology	Data Management	
	Data Analysis	
Application	Disaster Management Type	
	Disaster Management Phase	

Universiti Utara Malaysia

<u>Step 4.2:</u> Perform mapping between learned and referenced ontology concepts: In this step, each concept from the learned ontology is mapped to the extracted concept of the referenced ontology as shown in Table 3.9. Any concept that exists in the referenced ontology will be counted as 1 and accumulated. The algorithm for this step is shown in Table 3.10 and the computed outcome of this step is presented in Chapter Four.

Table 3.10

The algorithm to perform mapping between learned and referenced ontology concepts

Algorithm Ontology Mapping: A mapping between learned and referenced ontology concepts

Input: Concepts for learned ontology $C_L = \{cl_1, cl_2, \ldots cl_i\}$ and referenced ontology $C_G = \{cg_1, cg_2, \ldots cg_j\}$.

For every concepts cl_i in C_L **do**
    If the concept cl_i exists in C_G **then**
        Add 1 to total mapped terms
End for

Output: A total mapped terms between learned and referenced ontology concepts.

Table 3.10 presents the algorithm in accumulating the total mapped terms between the learned and referenced ontology. The algorithm accepts the learned and referenced ontology concepts as input and iterated every concept to the gold-standard concepts. If the concept exists in the referenced terms, the total mapped terms are added with one. This goes until all of the terms in learned ontology have been iterated and checked. At the end of the operation, the total mapped terms between the learned and referenced are obtained, therefore the precision, recall, and F-measure can be computed and analysed.

Universiti Utara Malaysia

Step 4.3: Compute precision, recall, and F-measure: From the accumulated values of the mapped concepts, the precision, recall, and F-measure are computed referring to the formula as shown in equation (3.4), (3.5), and (3.6). The Semantic Cotopy of a concept is described to be the collection of all its super and sub-concepts (Zavitsanos et al., 2011). In formula (2), Probabilistic Cotopy Precision (PCP) is the total of learned ontology concepts in the cotopy set of C_L matched to concepts in the referenced ontology cotopy set of C_G, against by the number of concepts in the cotopy set of C_L. Probabilistic Cotopy Recall (PCR) IS the number of concepts in the learned ontology cotopy set of C_L matched to concepts in the referenced ontology cotopy set of C_G,

against by the number of concepts participating in the learned ontology cotopy set of C_G. Finally, the F-measure is commonly used as the combined measure of precision and recall (Sabbah et al., 2017).

$$PCP_i = \frac{||CS(C_L) \cap CS(C_G)||}{||CS(C_G)||} \quad (3.4)$$

$$PCR_i = \frac{||CS(C_L) \cap CS(C_G)||}{||CS(C_L)||} \quad (3.5)$$

$$F - measure = 2 * \frac{PCR_i * PCP_i}{PCR_i + PCP_i} \quad (3.6)$$

3.8 Summary

To summarize, this chapter describes the methodology for conducting this research. The Design Science has been selected as the research framework that comprised of five phases: Problem Identification (Phase 1), Identification of Existing Supervised Term Weighting Schemes for Ontology Development (Phase 2), Construction of an Enhanced Supervised Term Weighting Scheme for Disaster Management Social Media Ontology Development (Phase 3), and Evaluation of the Enhanced Supervised Term Weighting Schemes through the Developed Ontology (Phase 4). All of these phases were applied in the experiment conducted in Chapter Four to perform the identification of existing supervised TWS and the construction of an enhanced supervised TWS, Chapter Five presents the results and discussions, and Chapter Six describes the evaluation of the enhanced supervised TWS through the developed ontology.

Universiti Utara Malaysia

CHAPTER 4

EXPERIMENTS

4.1 Introduction

This chapter further discusses the first and second conducted experiments based on the methodology discussed before. In determining the identification of existing supervised TWS for ontology development, the first experiment is conducted to accomplish the first objective of this research. To address the second objective of this research, in particular the second experiment was performed, which is to construct an enhanced supervised TWS for Disaster Management Social Media ontology development based on a confusion matrix.

Universiti Utara Malaysia

4.2 Identification of Existing Supervised Term Weighting Schemes

The identification of the existing supervised TWS have been conducted well. By following the procedures critically designed in Chapter Three, the steps conducted from the experimental work are detailed out in the next parts.

4.2.1 Step 1: Corpus Collection and Preliminary Domain Study

Collecting an applicable and relevant corpus for this study as the knowledge source is initiated in this step, as described in Chapter Three. DMSM related scientific publications were consolidated from five digital database, in particular ACM DL, Google Scholar, ScienceDirect, Scopus, and Wiley Online Library, with a total number of 392 publications. The search strategy in deriving the keywords based on the research

question is defined in Steps 1.1 and 1.2 for the purpose of ensuring the results acquired is accurate.

Step 1.1: Formulate Research Questions: A clear boundary and limit for the corpus collection process is necessary to be established. For this reason, the Research Questions are formulated as described in Chapter Three. Step 1, Step 2, and Step 3 addresses all of these research questions respectively.

i) RQ1: How many previous research publications are related to Disaster Management Social Media activities from the last five years?
ii) RQ2: What are the most used concepts related to Disaster Management Social Media?

Step 1.2: Define Search Keywords: The research questions defined provide the context so that a search keyword can be derived, as defined and listed in Chapter Three. The keyword is useful in helping to find the most suitable results for the scope of this research. The search term is defined as *("disaster management social media")*. In response to the search keywords, a high amount of search results was retrieved. Inclusion and exclusion criteria to refine the appropriate documents is performed as the next activity in Step 2.

4.2.2 Step 2: Identifying Sets of Documents to Suit Research Tasks

To identify the most acceptable documents to be analysed, the search results were further refined in this step. The steps applied in retrieving an ideal set of documents is described in the steps below.

Step 2.1: Perform Studies Selection Procedures: The search keyword retrieved 392 findings from the five digital databases consists of ACM DL, Google Scholar,

ScienceDirect, Scopus, and Wiley Online Library. As mentioned in section 3.5.2 by going through the title and abstract section of the documents manually.

Table 4.1

Data from all databases for retrieved, positive, and negative categories for DMSM.

Database	Retrieved	Positive	Negative
ACM DL	53	20	33
Google Scholar	55	29	26
ScienceDirect	124	25	99
Scopus	96	21	75
Wiley Online Library	64	22	42
Total	392	117	275

Initially there are 117 documents classified as positive category with 275 included as the Negative category. This is shown in Table 4.1 and used as the training label for the next step to perform the classification algorithm.

Universiti Utara Malaysia

4.2.3 Step 3: Classify Documents to Binary Classification

The automated classification is performed through the use of Naïve Bayes algorithm as elaborated in section 2.4 and the result obtained from this classification process is shown in Table 4.2. With the help of the RapidMiner Studio, the classification technique, as described in Chapter Two, was applied to classify the documents that are commonly associated with social media for disaster management activities as binary classification. Steps 3.1 until 3.9 listed all the steps taken in performing the text mining process on each of the gathered documents. Detail descriptions of these steps are included in Chapter Three.

Step 3.1: Read Excel: An Excel file consists of the Abstract extracted from each of the documents and predefined label attribute is imported by using this operator. This operator helps to delivers the Excel file in tabular form along with the meta data.

Step 3.2: Nominal to Text: All the nominal attributes are converted to string data type. This is to ensure the next step can be executed seamlessly without having runtime error.

Step 3.3: Process Documents from Data: Word vector are generated from the string attributes namely Abstract. The vector creation method selected for this operator is TF-IDF.

Step 3.4: Select Attributes: The Operator provides different filter types to make Attribute selection easy. No specific selection is set for this operator, therefore all of the attributes are selected for the next step.

Step 3.5: Set Role: Attribute name Category is selected for this operator. Attribute Category contains the predefined label assigned for each of the documents with value Disaster Management or Social Media. In this way the classification algorithm will be able to use the information for the data training process.

Step 3.6: Cross Validation: In this operator the actual classification process is performed, namely Step 3.7, Step 3.8, and Step 3.9 for training and testing subprocess.

Step 3.7: Naïve Bayes: In this subprocess, the training process is performed by using Naïve Bayes algorithm to generate predictions.

Step 3.8: Apply Model: This operator expects the defined model in which the Naïve Bayes as the input that has been trained. From here, it will help to provide the labelled data to the next operator.

Step 3.9: Performance: This operator helps to provide information based on the training model applied namely, Naïve Bayes. The binary classification is generated based on the documents classified by the algorithm.

Table 4.2

Binary classification constructed for DMSM publications.

N = 392		PREDICTED			
		POSITIVE		NEGATIVE	
OBSERVED	POSITIVE	112	TP	5	TN
	NEGATIVE	22	FP	253	FN

TP: True Positive, **TN:** True Negative,
FP: False Positive, **FN:** False Negative

4.2.4 Based on the algorithm ran, in total there are 117 documents classified as positive category with 112 included in the True Positive while 5 documents included in the False Positive category. The result continues with 22 publications categorised as False Positive category and 253 categorised as False Negative because of irrelevant content to DMSM. The next step of this experiment, with the purpose of extracting the related general concepts, pre-process the identified positive and negative publications as shown in Step 3: Classify Documents to Binary Classification

The automated classification is performed through the use of Naïve Bayes algorithm as elaborated in section 2.4 and the result obtained from this classification process is shown in Table 4.2. With the help of the RapidMiner Studio, the classification technique, as described in Chapter Two, was applied to classify the documents that are commonly associated with social media for disaster management activities as binary classification. Steps 3.1 until 3.9 listed all the steps taken in performing the text mining

Universiti Utara Malaysia

process on each of the gathered documents. Detail descriptions of these steps are included in Chapter Three.

Step 3.1: Read Excel: An Excel file consists of the Abstract extracted from each of the documents and predefined label attribute is imported by using this operator. This operator helps to delivers the Excel file in tabular form along with the meta data.

Step 3.2: Nominal to Text: All the nominal attributes are converted to string data type. This is to ensure the next step can be executed seamlessly without having runtime error.

Step 3.3: Process Documents from Data: Word vector are generated from the string attributes namely Abstract. The vector creation method selected for this operator is TF-IDF.

Step 3.4: Select Attributes: The Operator provides different filter types to make Attribute selection easy. No specific selection is set for this operator, therefore all of the attributes are selected for the next step.

Step 3.5: Set Role: Attribute name Category is selected for this operator. Attribute Category contains the predefined label assigned for each of the documents with value Disaster Management or Social Media. In this way the classification algorithm will be able to use the information for the data training process.

Step 3.6: Cross Validation: In this operator the actual classification process is performed, namely Step 3.7, Step 3.8, and Step 3.9 for training and testing subprocess.

Step 3.7: Naïve Bayes: In this subprocess, the training process is performed by using Naïve Bayes algorithm to generate predictions.

Universiti Utara Malaysia

Step 3.8: Apply Model: This operator expects the defined model in which the Naïve Bayes as the input that has been trained. From here, it will help to provide the labelled data to the next operator.

Step 3.9: Performance: This operator helps to provide information based on the training model applied namely, Naïve Bayes. The binary classification is generated based on the documents classified by the algorithm.

Table 4.2.

4.2.5 Step 4: Extraction of General Concepts by using the Text Mining Technique

The text mining technique, is employed with the aim to extract all terminologies that are relevant with the DMSM domain, as described in Chapter Two, with the assistance of the R Studio and R programming language. As indicated in Table 4.1, the included 392 publications were processed, and the results were further refined using the existing supervised TWS. The steps carried out in performing the text mining operation on each of the collected documents is listed in steps 3.1 until 3.4. Chapter Three includes the detail descriptions of these steps.

Step 3.1: Process PDF documents for conversion to text streams: All 392 documents have been successfully parsed to text streams. The text streams are stored into an in-memory storage, or in this case VectorCorpus, so that transformation on the data can be performed.

Step 3.2: Pre-processing for text mining: Document texts is split into a stream of tokens at first to create tokenization. In total there are 426, 380 tokens were eventually

generated from the tokenization for the positive category and 750, 255 tokens for the negative category, respectively. The next operation of the subsequent step is to normalize the tokens into plain text after the document texts were tokenized. All tokens were normalized into lower case, to eliminate ambiguity between exact terms in the form of lower or upper case. One of the tokens, for instance, 'Tweet' was transformed to lower case, 'tweet', to avoid uncertainty of two similar terms. Punctuations and numbers are also removed to ensure integrity of the data. Afterwards, all stop words were eliminated by using the English stop words dictionary like 'a', 'an', 'the', 'is', and 'are'. The extraction results were more expressive by doing this and useful for the later selection process. Stemming is performed to make certain the terms are all in common stem forms as described in Chapter Three.

Step 3.3: Build term document matrix: In accordance with the pre-processing process, the term document matrix is constructed in the memory based on the tokens generated, transforming the unstructured to structured data for the next process. The term document matrix consists of the terms in row sections, the documents in the column sections, and term frequency values in the cell field. In this way, the relationships between the term and document can be expressed through the term frequency values.

Universiti Utara Malaysia

Step 3.4: Compute term frequency by documents: From each of the tokens in positive and negative categories, the term frequency is calculated. The calculation is based on the frequency of the term for a document.

Table 4.3

Examples of the extracted terms and the calculated term frequency.

Term	Frequency	Total number of tokens	Term Frequency

data	7346	35260	0.20833806
network	3384	35260	0.095972774
relief	964	35260	0.027339762
warning	495	35260	0.014038571

As shown in Table 4.3, the examples of the extracted terms and the calculated term frequency is displayed. From the frequency of the term, it will be divided by the total number of tokens to obtain the term frequency. As an example, for the term ‘data’, the frequency is 7346, therefore when dividing with the total number of tokens, 35260, the term frequency is computed as 0.2083. The calculated term frequencies are used for the step in the TWS computation.

4.2.6 Step 5: Build Corpus Confusion Matrix

By using the merging algorithm in Chapter Three, the 426, 380 tokens in positive are mapped against 750, 255 tokens in the negative category to build the confusion matrix, with TP, TN, FP, and FN values for each positive token. These values are crucial in computing the supervised TWS since this scheme considers the classification information.

Universiti Utara Malaysia

Table 4.4

Examples of the extracted terms and the TP, TN, FP, and FN values.

Term	TP	TN	FP	FN
data	113	4	1	274
network	103	14	1	274
relief	92	25	2	273
warning	83	34	1	274

The examples of the extracted terms and the TP, TN, FP, and FN values are listed in Table 4.4 for the term ‘data’, ‘network’, ‘relief’, and ‘warning’. From the corpus

confusion matrix, it can be seen that as an example for the term ‘data’, the TP, TN, FP, and FN values are 113, 4, 1, and 274, respectively. The values signify that in the positive category corpus, the term ‘data’ exists for 113 documents while 4 documents did not contain this term. Meanwhile, in the negative category, 1 document contain this term while in the remaining 274 documents, this term did not exist.

4.2.7 Step 6: Perform Unsupervised Term Weighting Scheme Computation

To provide a comparison between the existing text mining to ontology development method, the currently used unsupervised TWS is calculated by using TF-IDF. Compared to the supervised TWS, this scheme did not consider the classification information or the document distribution of the corpus. The result from this scheme is presented and analysed in Chapter Five in detail.

4.2.8 Step 7: Perform Existing Supervised Term Weighting Schemes Computation Based on the Candidate Terms

Selected existing supervised TWS, Chi-square (x^2), Information Gain (IG), Odds Ratio (OR), Gain Ratio (GR), Relevance Frequency (RF), Mutual Information (MI), Inverse Document Frequency Excluding Category (IDFEC-based), and Short-text Term Weighting (SW) are performed to compute the weightage for each token. The outcome for each scheme is discussed in the next Chapter Five.

4.3 Construction of an Enhanced Supervised Term Weighting Scheme for Disaster Management Social Media Ontology Development Based on Confusion Matrix

The construction of the enhanced supervised TWS of this research was undertaken in this phase and successfully performed. Five steps were taken to further enhance the existing supervised TWS as defined in Chapter Three.

4.3.1 Step 1: Analyse Unsupervised Term Weighting Schemes Result

The results of the unsupervised TWS that have been performed are analysed. This is to observe the quality of results from this scheme, whereby the result is compared with the supervised TWS. The analysis of this result is presented in the next Chapter Five.

4.3.2 Step 2: Analyse Existing Supervised Term Weighting Schemes Results

From the computation applying existing schemes in Phase 2, the results for each calculation are analysed and discussed in Chapter Five. This is to verify if the results conform to the objective and to identify the merit and fault of the existing scheme components. The experimental results showed that the existing supervised TWS obtained ambiguous performance in terms of various category distribution data sets and different learning algorithms. Most of the supervised term weighting methods which are based on Chi-square (x^2), Information Gain (IG), Odds Ratio (OR), Gain Ratio (GR), Relevance Frequency (RF), Mutual Information (MI), Inverse Document Frequency Excluding Category (IDFEC-based), and Short-text Term Weighting (SW) have not shown satisfactory performance since irrelevant terms not specifically related to Disaster Management Social Media are extracted from term weightages such as Chi-square and Gain Ratio produced top terms related to negative category information.

Therefore, this experiment aims to construct an enhanced supervised TWS for DMSM ontology development based on a confusion matrix.

4.3.3 Step 3: Perform Enhanced Supervised Term Weighting Schemes Computation Based on the Candidate Terms

The enhancement of the existing supervised TWS is based on the idea to discriminate the positive over the negative category so that the concepts extracted are reflected on the DMSM domain. There are two factors introduced to be enhanced in this research which are the term frequency and the term weightage.

Step 3.1: Perform enhancement on the term frequency: The term frequency is proposed to consider category information instead of ignoring it. The regular formula for term frequency is consider term t_i, the frequency that t_i appears in the provided j^{th} document is indicated by f_{ij}. By considering the document distribution, the term frequency is differentiated between both classes, positive and negative. In this case accordingly a given term t_i, the frequency that t_i happens in the given j^{th} positive document is indicated by f_{ij}, divided by the f_{ij} in the documents in the positive and negative categories. As an example, the term 'data' appears 5 times in a 100 words positive document, and 1 time in 50 words negative document, therefore the new term frequency for 'data' is calculated as 5/100 = 0.05 for positive class and 1/50 = for negative class. The total new term frequency will be 0.05 divided by the sum of 0.05 and 0.02, which is 0.71.

Step 3.2: Perform enhancement on the term weightage scheme: Based on the enhanced formula, each of the existing supervised TWS, is substituted as a factor to perform weighting for each term. The results from this enhancement are presented and analysed in Chapter Five, provided with a comparison with the existing TWS results.

4.3.4 Step 4: Analyse Enhanced Supervised Term Weighting Schemes Results

Each of the existing supervised TWS is factored in into the constructed enhanced TWS. From the computation, the generated results of the extracted terms are analysed, with the current study found that RF and IDFEC-based factors showed the best results. The discussion on each of the outcomes is elaborated in the Results and Discussion section.

4.3.5 Step 5: Construct Structured Database with all Extracted Terms and Weightage

A database with all extracted terms and weightages is constructed structurally. The structured database consists of a table format with association between the different rows and columns. Usual examples of structure databases are Microsoft Excel and Microsoft SQL Server. The database is built based on the terms extracted from the best constructed enhanced TWS, namely, RF and IDFEC-based. This database is organized with the purpose of preparing it for the next steps of this experimental process and contains the list of the extracted terms and the computed weightages. The steps are discussed and presented in Chapter Six of this thesis.

Universiti Utara Malaysia

4.4 Summary

The experiment conducted to identify the existing supervised TWS is discussed thoroughly in this chapter, thereby achieving this research first objective. To achieve this outcome, there are five steps required, which are corpus collection and preliminary domain study, identifying sets of documents to suit research tasks, extraction of general concepts by using text mining technique, building corpus confusion matrix, and finally, performing existing supervised TWS computation based on the candidate terms. Furthermore, the second objective is discussed focusing on the experimental

work conducted for this research. An enhanced supervised TWS is constructed and presented with a positive result.

Universiti Utara Malaysia

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Introduction

The identification of the existing supervised TWS and construction of an enhanced supervised TWS for Disaster Management Social Media ontology development based on confusion matrix has been successfully performed to address the first and second objectives of this research. The results of the experiments are elaborated and discussed and analysed in this chapter.

5.2 Identification of Existing Supervised Term Weighting Schemes

From the results that have been successfully acquired, the findings were presented and further analysed to obtain more meaningful results. All of the results based on the eight selected TWS are analysed and discussed in this section. The results are presented with the top ten highest term weightage from 426, 380 tokens as an indicator of what the results represent.

Universiti Utara Malaysia

5.2.1 Chi-square (x^2)

The results of TF.x^2 indicate that the term weightage is exhibiting higher concepts from the negative category. As seen in Figure 5.1, TF.x^2 provided the largest set of irrelevant concepts instead of the DMSM domain.

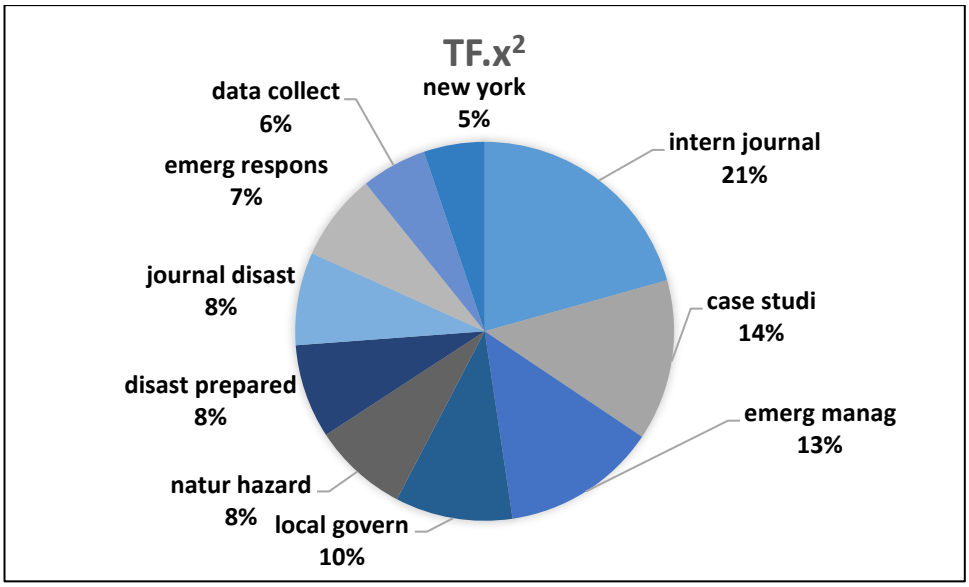

Figure 5.1. TF.x^2 ten highest ranked TWS result

Table 5.1

The ten highest-ranked terms and the calculated TF.x^2 values

No.	Terms	TP	TN	FP	FN	Term Frequency	x^2	TF.x^2
1	intern journal	64	53	152	123	0.035386586	3.82626E+15	1.35398E+14
2	case studi	78	39	161	114	0.02571519	3.50081E+15	9.00239E+13
3	emerg manag	80	37	139	136	0.028229925	3.07631E+15	8.68439E+13
4	local govern	35	82	108	167	0.021744696	3.01971E+15	6.56628E+13
5	natur hazard	59	58	121	154	0.018331472	2.90898E+15	5.33259E+13
6	disast prepared	63	54	120	155	0.019031022	2.76407E+15	5.26031E+13
7	journal disast	37	80	99	176	0.021650001	2.39947E+15	5.19485E+13
8	emerg respons	72	45	133	142	0.015996246	3.05981E+15	4.89455E+13
9	data collect	73	44	115	160	0.015476123	2.36655E+15	3.66251E+13
10	new york	56	61	128	147	0.010276568	3.31264E+15	3.40426E+13

As mentioned before, the values presented in Figure 5.1 are the ten highest-ranked terms and displayed in percentage for easier visual interpretation. The actual values are provided in Table 5.1 with the few highest terms are ‘intern journal’, ‘case studi’, and ‘emerg manag’ with TF.x^2 values 1.35398E+14, 9.00239e+13, and 8.68439E+13 respectively.

In analysing the equation (5.1), the numerator determines the effect on the result. The numerator will always have negative category as the main factor for TP * TN – FP * FN, taking into account FN value will always be the largest, FN >> TP, TN, FP as described in Chapter Two, section 2.4. The numerator is computed with the power of two in order to convert the value into positive state. For this reason, the highest ranked term extracted for x^2, mainly unrelated to DMSM corpus and consists of the negative category such as 'emerg manag', 'natur hazard', and 'emerg respons'.

This finding correlates with the intended purpose of this scheme which is to identify how independent the term and the class are. The intention correlates with the result of this scheme in which mostly the extracted terms are irrelevant to positive and negative class such as 'intern journal' and 'case studi' being the top two terms as shown in Table 5.1. It is perceived that this kind of terms have high occurrences and more related to the nature of the unstructured data set type for this study, in which scientific publications.

Universiti Utara Malaysia

5.2.2 Information Gain (IG)

The finding for TF.IG is contrary to previous TWS since the concepts extracted is neither belonging to positive or negative categories. In Figure 5.2, overall, it is deducted that the concepts discovered from the scheme such as 'abl', 'abraham', and 'abstract' did not provide meaningful information in representing the DMSM domain.

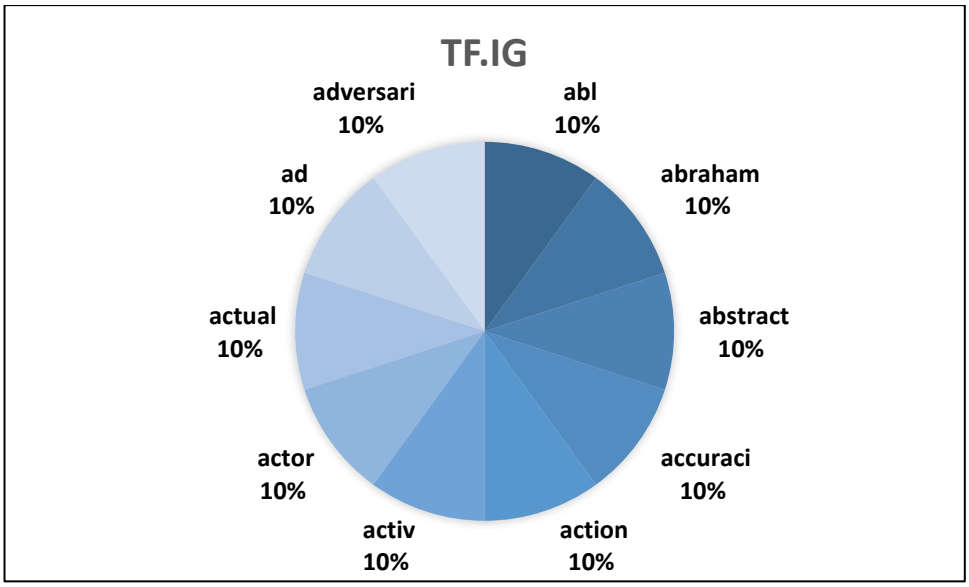

Figure 5.2. TF.IG ten highest-ranked TWS result

Table 5.2

The ten highest-ranked terms and the calculated TF.IG values

No.	Terms	TP	TN	FP	FN	Term Frequency	IG	TF.IG
1	abl	1	116	1	274	0.0000284	-0.0101208	-2.87431E-07
2	abraham	1	116	1	274	0.0000284	-0.0101208	-2.87431E-07
3	abstract	1	116	1	274	0.0000284	-0.0101208	-2.87431E-07
4	accuraci	1	116	1	274	0.0000284	-0.0101208	-2.87431E-07
5	action	1	116	1	274	0.0000284	-0.0101208	-2.87431E-07
423352	emerg plan	30	87	59	216	0.004953974	-9.91818158	-0.049134414
426380	data	113	4	1	274	0.299611578	-86.1319076	-25.80611675

An unanticipated finding is that the scheme gives higher weightage to the terms that do not exist on either distribution. As seen in *Figure 5.2*, the higher the TN and FN values, the higher the TF.IG values. This indicated that TF.IG concerns on irrelevant terms to the whole corpus. As an example, the term ‘abl’, obtains the highest TF.IG value with -2.87431E-07 since the TN and FN is high with value 116 and 274 respectively. The finding was unexpected to observe subsequent terms such as ‘emerg plan’ and ‘data’ with increased occurrences in the documents for TP and FP values resulting in lower TF.IG rate, -0.049134414 and -25.80611675, respectively.

The nature of equation (5.2) determine the outcome since the factors emphasized on variable N, in which giving more weight to the negative category, with N >> FP >> FN, TN, TP. This equation produced the TF.IG value in negative since the equation focuses on the irrelevancy of terms and also the equation by itself consider negative as a factor. The implication of the formula makes the equation unfit to the binary classification due to the inclination toward higher weightage on the irrelevant terms. Even the highest value is established from the sequence of lowest value, the terms are corresponding to the whole corpus which is the negative category. Thus, based on the results, the relevant terms with higher TF.IG values are more prone to negative category, which is unsuitable to binary classification in this study.

5.2.3 Odds Ratio (OR)

The outcome for TF.OR is similar with TF.IG as it is did not present relevant information for the DMSM domain even though the concepts revealed are slightly better than TF.IG. From this view, as shown in Figure 5.3, the concept ‘network’, ‘model’, ‘govern’, and ‘resili’ is distributed to the negative category, whereby semantically related to the DM domain.

Universiti Utara Malaysia

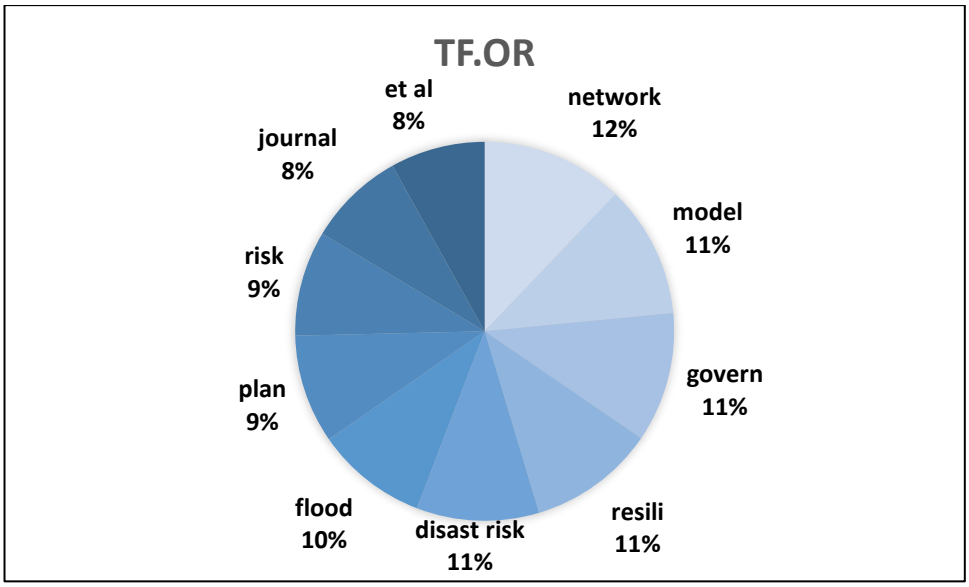

Figure 5.3. TF.OR ten highest-ranked TWS result

Table 5.3

The ten highest-ranked terms and the calculated TF.OR values

No.	Terms	TP	TN	FP	FN	Term Frequency	OR	TF.OR
1	network	103	14	1	274	0.160603577	0.721214698	0.11582966
2	model	102	15	1	274	0.144789178	0.746940868	0.108148954
3	govern	97	20	1	274	0.12439846	0.850051167	0.105745056
4	resili	83	34	1	274	0.102048967	1.012806447	0.103355852
5	disast risk	87	30	1	274	0.101982129	0.978889945	0.099829281
6	flood	96	21	1	274	0.10487337	0.866739965	0.090897941
7	plan	104	13	1	274	0.127868523	0.693226129	0.088641801
8	risk	105	12	2	273	0.236184562	0.363177902	0.085777014
9	journal	96	21	1	274	0.091411321	0.866739965	0.079229845
10	et al	105	12	2	273	0.212526592	0.363177902	0.077184962

Table 5.3 lists the extracted terms and the TF.OR values with the term ‘network’, ‘model’, and ‘govern’ are among the highest with the values i.e. 0.11582966, 0.108148954, and 0.105745056, respectively. From the equation (5.3), in general TF.OR computes the ratio between the positive and the negative outcome, which is incompatible for the binary classification, whereby each classifications of TP, TN, FP, and FN is stressed on. The nature of the formula also stated that the equation emphasized on the positive category only as the numerator, without considering the

terms that exists in the negative category, FP. In Table 5.3, from the TP, TN, FP, and FN values, the scheme is seen contemplating both the TP and FN distribution to obtain the highest value, however, the results are expressed as terms from the negative category as the candidate concepts for the DMSM ontology.

5.2.4 Gain Ratio (GR)

From the TF.GR TWS performed, the results suggest that the concepts identified are inclined towards the negative category. The terms extracted as seen in Figure 5.4, such as 'crossref', 'estat', and 'eoc' are likely not to be related to either positive or negative categories.

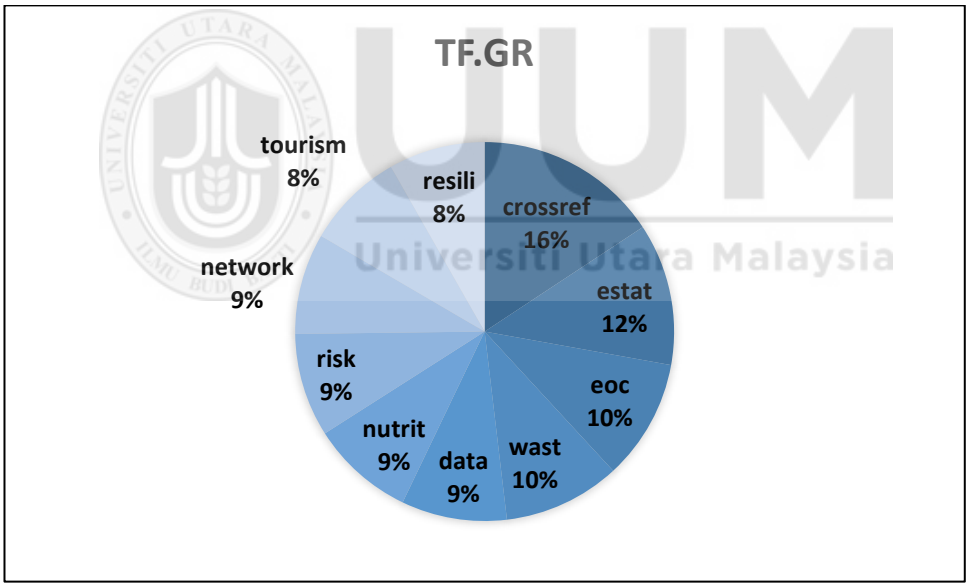

Figure 5.4. TF.GR ten highest-ranked TWS result

Universiti Utara Malaysia

Table 5.4

The ten highest-ranked terms and the calculated TF.GR values

No.	Terms	TP	TN	FP	FN	Term Frequency	GR	TF.GR
1	crossref	3	114	1	274	0.00765744	0.00765744	91.66360385
2	estat	3	114	1	274	0.005908557	0.005908557	70.72855004
3	eoc	3	114	1	274	0.005020314	0.005020314	60.09581188
4	wast	20	97	1	274	0.045910853	0.045910853	58.24423324
5	data	113	4	1	274	0.299611578	0.299611578	52.53344856
6	nutrit	3	114	1	274	0.004310877	0.004310877	51.60347604
7	risk	105	12	2	273	0.236184562	0.236184562	51.53283967
8	network	103	14	1	274	0.160603577	0.160603577	50.3189026
9	tourism	20	97	2	273	0.047987953	0.047987953	48.70729117
10	resili	83	34	1	274	0.102048967	0.102048967	48.01102778

Irrelevant terms are obtained from the TF.GR supervised TWS, with ‘crossref’, ‘estat’, and ‘eoc’ are examples of the highest terms as seen in Table 5.4, whereby the TF.GR values are 91.66360385, 70.72855004, and 60.09581188, respectively. Contradict with TF.OR, this scheme takes into account, terms that do not exist as the factor to obtain higher term weighting values. The term ‘crossref’, ‘estat’, and ‘eoc’ all obtained TN and FN values as high, with 114 and 274, respectively, together with low TF values, signifying the lower the term occurrences, the higher the term weighting will be. This indication is not aligned with the expected outcome as the terms that exist in the positive category are expected to express higher term weightages.

Even though the equation (5.4) utilizes TF.IG with negative value as a factor for the numerator, the TF.GR value produces positive value, since the denominator value is becoming larger when positive category is deducted with negative category, FN >> TP, TN, FP. So with the negative value for numerator divided by the negative denominator, positive value is produced. Generally, this equation seeks the ratio between the irrelevance terms with the ratio of irrelevancy in the negative category, therefore it does not contribute at all with the binary classification.

Universiti Utara Malaysia

5.2.5 Relevance Frequency (RF)

The TF.IG results demonstrated that the equation emphasizes weightage on the negative category. From the data in Figure 5.5, it is apparent that the terms extracted, 'disast', and 'manag' are related to the DM domain or negative category.

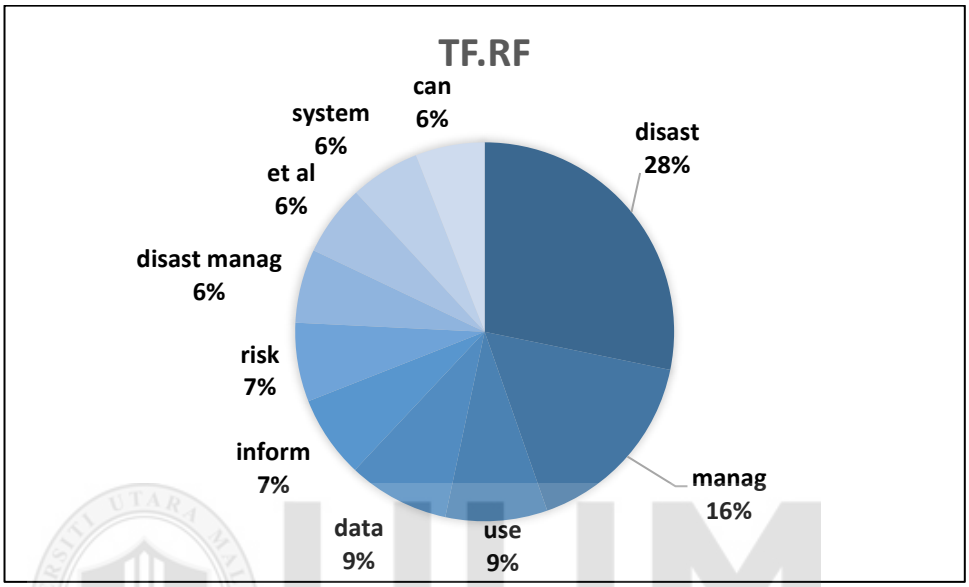

Figure 5.5. TF.RF ten highest-ranked TWS result

Universiti Utara Malaysia

Table 5.5

The ten highest-ranked terms and the calculated TF.RF values

No.	Terms	TP	TN	FP	FN	Term Frequency	RF	TF.RF
1	disast	115	2	7	268	0.968592271	0.968592271	0.373340205
2	manag	116	1	1	274	0.564911876	0.564911876	0.21716202
3	use	116	1	2	273	0.29900081	0.29900081	0.115024219
4	data	113	4	1	274	0.299611578	0.299611578	0.114586717
5	inform	116	1	1	274	0.242066567	0.242066567	0.093054629
6	risk	105	12	2	273	0.236184562	0.236184562	0.089140386
7	disast manag	114	3	1	274	0.217605043	0.217605043	0.083366108
8	et al	105	12	2	273	0.212526592	0.212526592	0.080211434
9	system	113	4	5	270	0.205587236	0.205587236	0.078852865
10	can	116	1	1	274	0.203185767	0.203185767	0.078108168

Table 5.5 presents the actual values of TF.RF supervised TWS with 'disast' and 'manag' is among the highest. Non-meaningful terms are also found to be ranked as high such as 'use' and 'can' for this scheme. From the trend, the higher the TP and FN

values, the higher the TF.RF rate is computed. This indicates that the scheme's equation (5.5), emphasizes term occurrences in the positive category and term absences in the negative category to determine the TF.RF rate. Results in Table 5.5 show a finding that the TF.RF values depends on the term frequency, in which the higher the frequency, the higher the TF.RF value. Based on the formula, this scheme emphasizes on the term that exists in positive category only, without considering the term that occurs in negative category. Although the result give more weight to the negative category, this scheme is relevant to be applied in the binary classification problem.

5.2.6 Mutual Information (MI)

On the TWS TF.MI, the results have noted the concepts found did not belong to either positive or negative category, being in a neutral state. In Figure 5.6, the concepts identified such as 'section', 'survey', 'wast', and 'tourism' generally did not be part of positive or negative categories.

Universiti Utara Malaysia

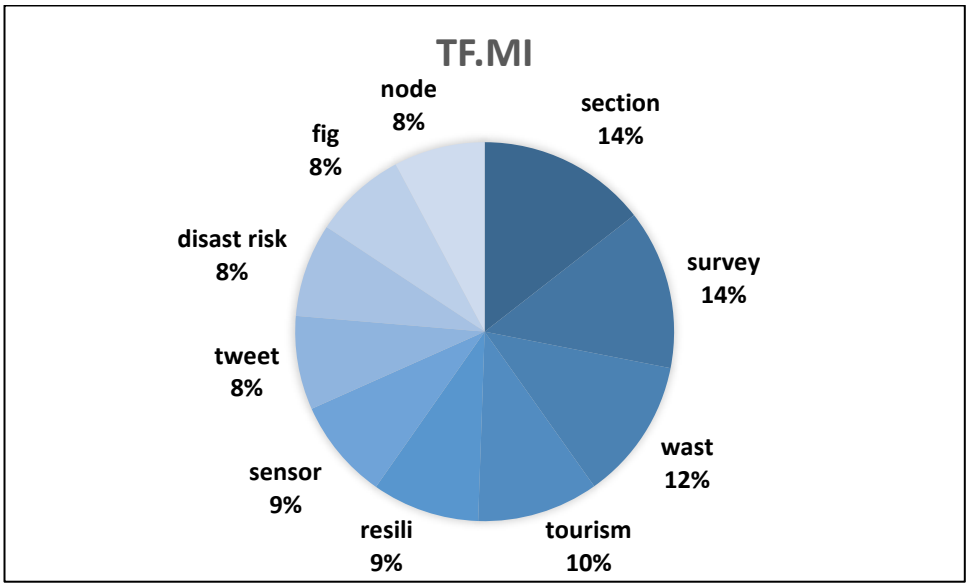

Figure 5.6. TF.MI ten highest ranked TWS result

Table 5.6

The ten highest-ranked terms and the calculated TF.MI values

No.	Terms	TP	TN	FP	FN	Term Frequency	MI	TF.MI
1	section	1	116	0	275	0.037362435	0.152376985	0.005693175
2	survey	1	116	0	275	0.035216169	0.152376985	0.005366134
3	wast	20	97	1	274	0.045910853	0.103749438	0.004763225
4	tourism	20	97	2	273	0.047987953	0.085025762	0.004080212
5	resili	83	34	1	274	0.102048967	0.035416657	0.003614233
6	sensor	52	65	1	274	0.047174954	0.071795941	0.00338697
7	tweet	38	79	1	274	0.035867002	0.087850463	0.003150933
8	disast risk	87	30	1	274	0.101982129	0.030815446	0.003142625
9	fig	63	54	1	274	0.052924401	0.058816742	0.003112841
10	node	35	82	3	272	0.043507796	0.070432139	0.003064347

Table 5.6 shows the list of the ten highest terms by using TF.MI scheme computation with ‘section’, ‘survey’, ’wast’, and ‘tourism’ among the highest weightage terms. Based on the trend, it is obvious that this scheme accentuates the non-occurrences rate in the document, whereby as seen in Table 5.6, the higher the FV value, the higher the TF.MI weightage. The results shown that this scheme is not suitable for binary classification since the numerator in equation (5.6) is multiplied with N which include FN category, or the negative class. The denominator also considers negative class, FP and FN. This is the possible explanation for the reason the high ranked term for this study is mostly irrelevant to the positive category. The finding correlates with TF.MI purpose, whereas it considers all corpus attributes to measure the relationship.

Universiti Utara Malaysia

5.2.7 Inverse Document Frequency Excluding Category-Based (IDFEC-based)

For TF.IDFEC-based, the result displays a strong relationship between the concepts extracted with the negative category. As shown in Figure 5.7, the terms ‘disast’, ‘manag’, and ‘data’ are pertaining to the DM domain.

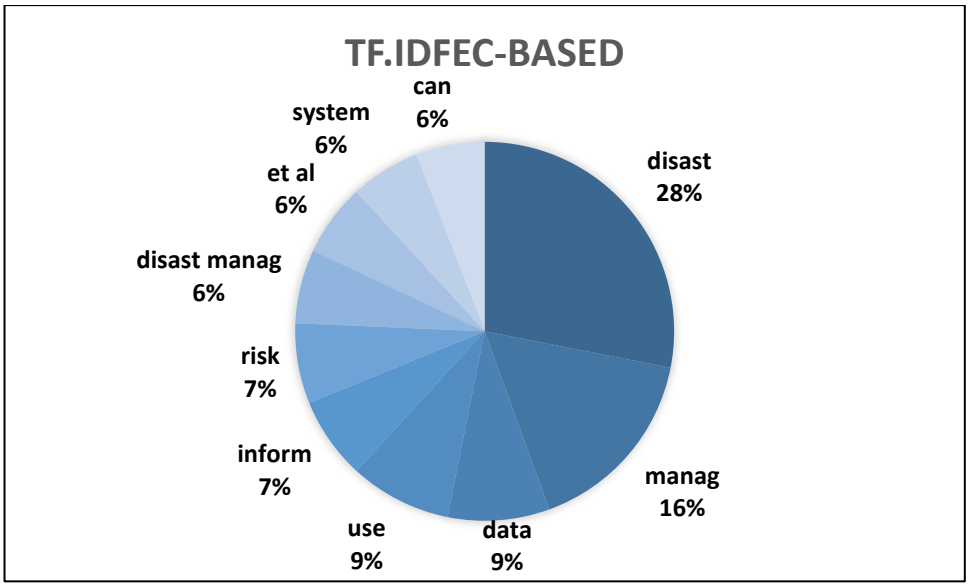

Figure 5.7. TF.IDFEC-based ten highest-ranked TWS result

Table 5.7

The ten highest-ranked terms and the calculated TF.IDFEC-based values

No.	Terms	TP	TN	FP	FN	Term Frequency	IDFEC-based	TF.IDFEC-based
1	disast	115	2	7	268	0.968592271	2.079181246	2.013878885
2	manag	116	1	1	274	0.564911876	2.079181246	1.174554178
3	data	113	4	1	274	0.299611578	2.079181246	0.622946774
4	use	116	1	2	273	0.29900081	2.079181246	0.621676877
5	inform	116	1	1	274	0.242066567	2.079181246	0.503300266
6	risk	105	12	2	273	0.236184562	2.079181246	0.491070512
7	disast manag	114	3	1	274	0.217605043	2.079181246	0.452440324
8	et al	105	12	2	273	0.212526592	2.079181246	0.441881304
9	system	113	4	5	270	0.205587236	2.079181246	0.427453126
10	can	116	1	1	274	0.203185767	2.079181246	0.422460036

Compared with other supervised TWS, this scheme expressed better term weightages, with ‘disast’. ‘manag’, and ‘data’ are among the highest as shown in Table 5.7. It is observed that this scheme obtains almost similar results as TF.RF, in which the formula is using similar factors except that TF.IDFEC-based considers FN and TN values. However, based on the trends, this scheme is also highly based on the term occurrences in the positive category and term absence in the negative category.

The results indicated that the outcome depends on the term frequency value, whereby higher RF value is achieved with higher term frequency. From the equation (5.7), the numerator consists of combination of TP, FN, and TN values with the denominator also considers FN values. This is the contribution on the reason TF.IDFEC-based extracts high term on negative category.

5.2.8 Short-text Term Weighting (SW)

The result of TF.SW also indicates that the concepts retrieved are inclined towards the negative category or DM domain. The concepts captured are ‘disast’, and ‘manag’ are generally related to the DM domain as seen in *Figure 5.8*.

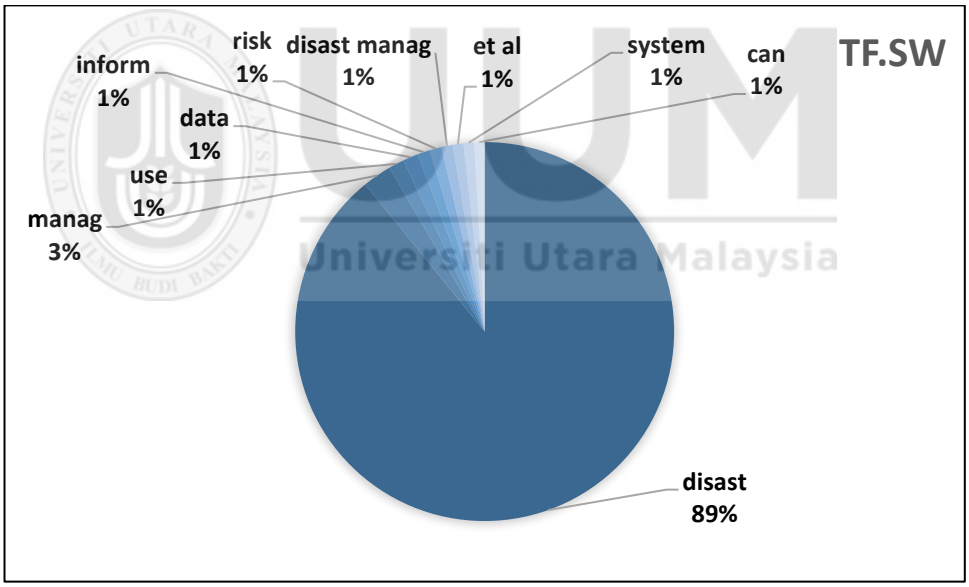

Figure 5.8. TF.SW ten highest-ranked TWS result

Universiti Utara Malaysia

Table 5.8

The ten highest-ranked terms and the calculated TF.SW values

No.	Terms	TP	TN	FP	FN	Term Frequency	SW	TF.SW
1	disast	115	2	7	268	0.968592271	8.049573561	7.796754736
2	manag	116	1	1	274	0.564911876	0.384417517	0.21716202
3	use	116	1	2	273	0.29900081	0.384695342	0.115024219
4	data	113	4	1	274	0.299611578	0.382450897	0.114586717
5	inform	116	1	1	274	0.242066567	0.384417517	0.093054629
6	risk	105	12	2	273	0.236184562	0.377418342	0.089140386
7	disast manag	114	3	1	274	0.217605043	0.383107427	0.083366108
8	et al	105	12	2	273	0.212526592	0.377418342	0.080211434
9	system	113	4	5	270	0.205587236	0.383549417	0.078852865
10	can	116	1	1	274	0.203185767	0.384417517	0.078108168

Table 5.8 presented the TF.SW rate with ‘disast’, ‘manag’, and ‘use’ are among the highest. This scheme obtains similar results as TF.RF and TF.IDFEC-based, whereby it computes the outcome heavily on TP and FN values. However, it is worth noting that the TF.SW rate is obtained a very high value on the term ‘disast’ and subsequent terms value below than 0.5. The main reason is because of the term ‘disast’ has a higher term frequency than others, in which it is expected that this term occurrence is the highest since the DMSM domain is discussing disaster as the most. The equation (5.8) applies denominator that use all negative classes, FP and FN which makes the denominator larger, therefore producing smaller ratio for TF.SW. With the smaller ratio to TP value, it means that terms are related to negative category, which is not suitable for binary classification.

5.2.9 Term Frequency-Inverse Document Frequency (TF-IDF)

The TF-IDF computation is performed as the comparison between the unsupervised and supervised TWS. *Figure 5.9* displays the result of the computation with the term ‘node’, ‘term’, and ‘wast’ are distributed neither in positive nor negative categories.

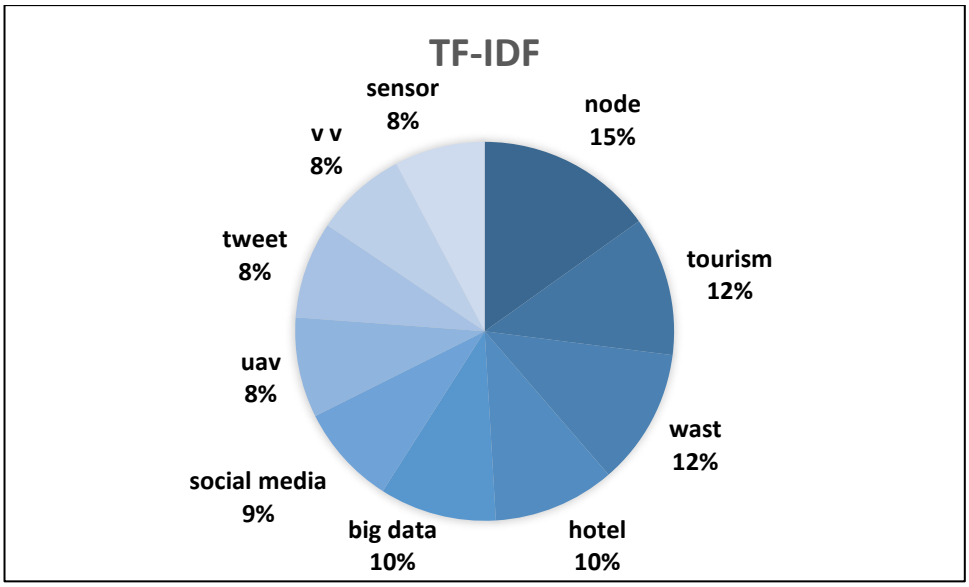

Figure 5.9. TF-IDF ten highest-ranked TWS result

Table 5.9

The ten highest-ranked terms and the calculated TF-IDF values

No.	Terms	Frequency	Total No. of Documents	Document Frequency	Term Frequency	IDF	TF-IDF
1	node	3662	392	92	9.341836735	4.260869565	5.88066978
2	tourism	1984	392	48	5.06122449	8.166666667	4.616063628
3	wast	2412	392	72	6.153061224	5.444444444	4.528367378
4	hotel	602	201	9	2.995024876	22.33333333	4.040149432
5	big data	972	191	33	5.089005236	5.787878788	3.880465358
6	social media	2804	392	134	7.153061224	2.925373134	3.334623156
7	uav	1137	392	28	2.900510204	14	3.324356063
8	tweet	1285	392	41	3.278061224	9.56097561	3.214146276
9	v v	346	191	4	1.811518325	47.75	3.041491037
10	sensor	1957	392	98	4.992346939	4	3.005692355

TF-IDF results show non-meaningful terms being ranked as high with the terms ‘node’, ‘tourism’, and ‘wast’ achieved 5.88066978, 4.616063628, and 4.528367378 TF-IDF rate, respectively. As expected, based on the formula definition in Chapter Two, this scheme relies heavily on the term frequency and the IDF value to compute the TF-IDF rate. As an example, in Table 5.9, the term ‘node’ obtained the highest weightage, 5.88066978 with a high term and IDF frequency of 9.341836735 and 4.260869565

values, respectively. Based on this result, it can be concluded that by using the unsupervised TWS, whereas ignoring the classification information of a corpus did not obtain the best result for the binary classification problem, or in this context DMSM domain.

5.2.10 Results Comparisons between Selected Existing Supervised Term Weighting Schemes

To interpret the result the definitions of category information are defined as explained in section 3.6.4 with the results of supervised TWS are categorised and analysed based on this information.

Table 5.10

Definition of category information based on document classifications.

Category Information	Document Classification
Negative	Disaster Management
Neutral	-
Positive	Disaster Management Social Media

A Chi-square scheme is applied to measure how much the term t_k and the category c_i is independent, whereby the factor explains why the result is emphasized on the neutral category. The usage of IG is to measure dependency between one random term and one random class. However, since the term t_k and the category c_i in this study are set to be specific instead of random, the discrepancy of the results are attributed to this case.

TF.OR is a scheme that measure the weightage of association between two random variables. Even though TF.OR displays slightly better results, this scheme can be further improved to obtain the desired outcome, particularly to obtain category

information as the positive category. TF.GR attempts to decrease the bias of Information Gain by introducing a normalizing term called Intrinsic Information, whereby the equation utilizes the IG to be divided with the Intrinsic Information. TF.GR is able to improve the TF.IG outcome from before but still the scheme extract terms that belong to the neutral category.

The main idea of TF.RF is to obtain terms with higher frequencies and concentration to the positive class to discriminate between the negative classes and vice versa. Based on the result, this scheme can be further improved to obtain more balanced results towards the positive category. MI derives the measure of information the one random term have about the category c_i, that is to say the amount of their mutual dependence. However, the scheme is unable to provide the distinct results of positive or negative classes.

The main idea behind TF.IDFEC-based is that weights set to term t_k occur in documents that related to the same category c_i must be higher; indicating these terms t_k are not low. Nevertheless, the equation is generating results based on the negative category. The scheme evaluates the importance of a term t_k in a document where it occurs, on top of the term t_k distribution within a category c_i and in the whole corpus. This scheme however obtained results that discriminated towards the negative category.

Table 5.11

Comparison between the findings for identification of existing supervised TWS.

Term Weightage	Category Information
TF.x^2	Neutral
TF.IG	Neutral
TF.OR	Negative
TF.GR	Neutral
TF.RF	Negative
TF.MI	Neutral

Universiti Utara Malaysia

TF.IDFEC-based	Negative
TF.SW	Negative

Table 5.11 presents the comparison between the identification of existing supervised TWS as discussed. The results of this analysis conclude that the existing supervised TWS did not answer the objective of this study, in order to retrieve the concepts that fully represent the DMSM domain. This is because all of the TWS present results that emphasized by giving higher weightage more on the negative category. Based on this observation, the existing TWS must be further improved to address the objective of this research.

5.3 Construction of an Enhanced Supervised Term Weighting Scheme for Disaster Management Social Media Ontology Development based on Binary Classification

With the constructed enhanced TWS applied to all existing supervised TWS, the findings were presented and further analysed to identify meaningful results. All of the results based on the eight selected TWS incorporated together with the constructed enhanced TWS are analysed and discussed in this section. For the subsequent sections, the integration is referred to as TF.x^2.Enhanced, TF.IG.Enhanced, TF.OR.Enhanced, TF.GR.Enhanced, TF.RF.Enhanced, TF.MI.Enhanced, TF.IDFEC-based.Enhanced, and TF.SW.Enhanced. To interpret these results, similar with the previous section, Table 5.10 is referred to classify the outcome with the category information.

5.3.1 TF.x^2.Enhanced

The results of TF.x^2.Enhanced indicate that the term weightage is exhibiting higher concepts from the negative category, slightly different but similar with the previous result. As seen in Figure 5.10, TF.x^2.Enhanced presents the highest set of significant concepts on Disaster Management instead of the Social Media domain. The concepts extracted such as 'emerg manag', 'emerg resposn', and 'climat chang' are semantically related to the DM domain.

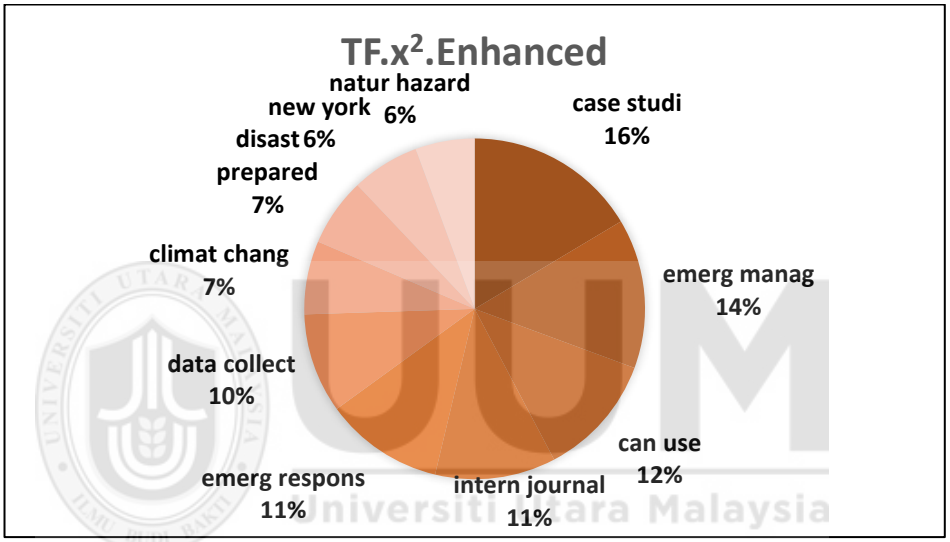

Figure 5.10. TF.x^2.Enhanced ten highest-ranked TWS result

Table 5.12

The ten highest-ranked terms and the calculated TF.x2.Enhanced values

No.	Terms	TP	TN	FP	FN	x2	Positive Term Frequency Ratio	Positive Category Ratio	TF.x2.Enhanced
1	case studi	78	39	161	114	3.50081E+15	322	0.40586	4.57509E+17
2	emerg manag	80	37	139	136	3.07631E+15	300.5405405	0.427974	3.95686E+17
3	can use	81	36	110	165	2.07282E+15	247.5	0.632039	3.2425E+17
4	intern journal	64	53	152	123	3.82626E+15	183.5471698	0.454425	3.19142E+17
5	emerg respons	72	45	133	142	3.05981E+15	212.8	0.482246	3.14004E+17
6	data collect	73	44	115	160	2.36655E+15	190.7954545	0.584583	2.63955E+17
7	climat chang	35	82	120	155	3.7881E+15	51.2195122	1	1.94025E+17
8	disast prepared	63	54	120	155	2.76407E+15	140	0.466444	1.805E+17

9	new york	56	61	128	147	3.31264E+15	117.5081967	0.460878	1.79403E+17
10	natur hazard	59	58	121	154	2.90898E+15	123.0862069	0.437831	1.56768E+17

Table 5.13

Comparison between TF.x2 and TF.x2.Enhanced values

No.	Terms	TF.x^2	Terms	TF.x^2.Enhanced
1	intern journal	1.35398E+14	case studi	4.57509E+17
2	case studi	9.00239E+13	emerg manag	3.95686E+17
3	emerg manag	8.68439E+13	can use	3.2425E+17
4	local govern	6.56628E+13	intern journal	3.19142E+17
5	natur hazard	5.33259E+13	emerg respons	3.14004E+17
6	disast prepared	5.26031E+13	data collect	2.63955E+17
7	journal disast	5.19485E+13	climat chang	1.94025E+17
8	emerg respons	4.89455E+13	disast prepared	1.805E+17
9	data collect	3.66251E+13	new york	1.79403E+17
10	new york	3.40426E+13	natur hazard	1.56768E+17

Table 5.12 presents the list of the extracted terms and the TF.x^2.Enhanced weighting rate, while Table 5.13 shows the comparison with TF.x^2 terms. From both tables, the ranked terms by the TF.x^2.Enhanced weightage, however, considers irrelevant terms such ‘case studi’, ‘emerg manag’, and ‘intern journal’, similar with TF.x^2 results. It is observed that the scheme did not improve after utilising the introduced scheme. This is determine by referring on the classification in Table 5.10, in which TF.x^2.Enhanced extracted terms are mostly contains high weightage from negative category. As explained in section 5.2.1, the main explanation for the result is due to the limitation of the formula in which the numerator considers negative category, FP and FN as the factor for the outcome.

Universiti Utara Malaysia

5.3.2 TF.IG.Enhanced

Consistent with the previous finding, for TF.IG.Enhanced the concepts extracted is neither belonging to positive nor negative categories. In *Figure 5.11*, generally, the result deducted that the concepts learned from the scheme such as ‘sediment’, ‘gic’, and ‘solomon’ did not give meaningful information in representing the DMSM domain.

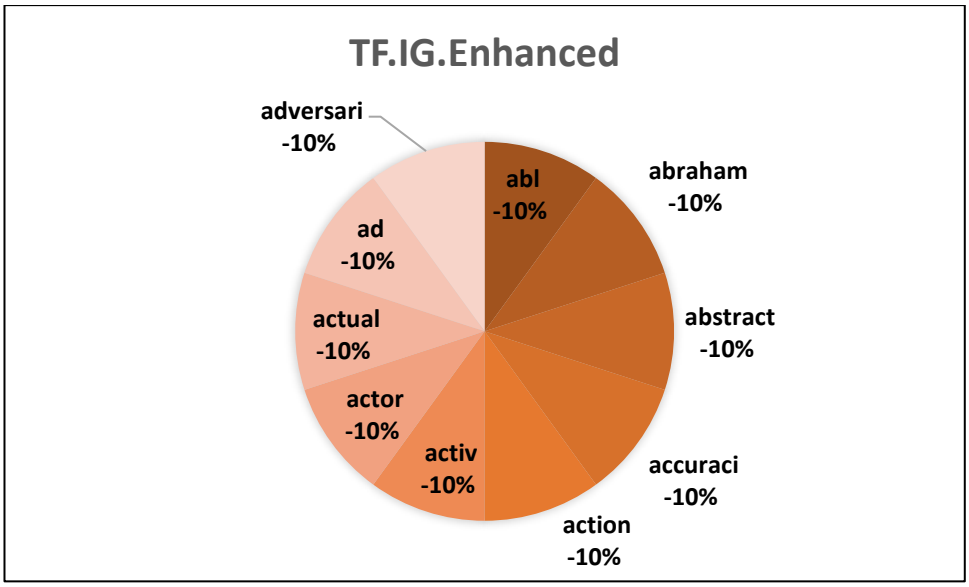

Figure 5.11. TF.IG.Enhanced ten highest-ranked TWS result

Table 5.14

The ten highest-ranked terms and the calculated TF.IG.Enhanced values

No.	Terms	TP	TN	FP	FN	IG	Positive Term Frequency Ratio	Positive Category Ratio	TF.IG. Enhanced
1	sediment	1	116	1	274	-0.0101208	0.008978	0.00862069	-7.83299E-07
2	gic	1	116	1	274	-0.0101208	0.0093	0.00862069	-8.1139E-07
3	solomon	1	116	1	274	-0.0101208	0.009721	0.00862069	-8.48152E-07
4	frm	1	116	1	274	-0.0101208	0.012064	0.00862069	-1.05252E-06
5	drm law	1	116	1	274	-0.0101208	0.01226	0.00862069	-1.06971E-06
6	sdn	1	116	1	274	-0.0101208	0.013117	0.00862069	-1.14444E-06
7	pharmaci	1	116	1	274	-0.0101208	0.013592	0.00862069	-1.18587E-06
8	commonwealth	1	116	1	274	-0.0101208	0.013971	0.00862069	-1.21897E-06
9	grenada	1	116	1	274	-0.0101208	0.015404	0.00862069	-1.34398E-06
10	app	1	116	1	274	-0.0101208	0.015893	0.00862069	-1.38665E-06

Table 5.15

Comparison between TF.IG and TF.IG.Enhanced values

No.	Terms	TF.IG	Terms	TF.IG.Enhanced
1	abl	-2.87431E-07	sediment	-7.83299E-07
2	abraham	-2.87431E-07	gic	-8.1139E-07
3	abstract	-2.87431E-07	solomon	-8.48152E-07
4	accuraci	-2.87431E-07	frm	-1.05252E-06
5	action	-2.87431E-07	drm law	-1.06971E-06
6	activ	-2.87431E-07	sdn	-1.14444E-06
7	actor	-2.87431E-07	pharmaci	-1.18587E-06
8	actual	-2.87431E-07	commonwealth	-1.21897E-06
9	ad	-2.87431E-07	grenada	-1.34398E-06
10	adversari	-2.87431E-07	app	-1.38665E-06

This finding is also demonstrated in Table 5.14 and Table 5.15. In the table the term ‘sediment’, ‘gic’, and ‘solomon’ obtained TF.IG.Enhanced rate of -7.83299E-07, -8.1139E-07, and -8.48152E-07, respectively, in which the term did not provide meaningful results to this study. In Table 5.15 also, comparing with TF.IG result, no improvement has been achieved by using the introduced positive category ratio and positive term frequency ratio. The factor contributing to this result is due to the equation focuses on determining the irrelevancy of the terms since negative variables are mainly the weightage for the formula, as discussed in section 5.2.2.

5.3.3 TF.OR.Enhanced

TF.OR.Enhanced presents almost similar results as before, which is more towards the neutral category. From this view, as shown in *Figure 5.12*, the concepts i.e., ‘intern’, ‘data’, ‘communic’, and ‘take’ are not related to the DMSM domain standpoint.

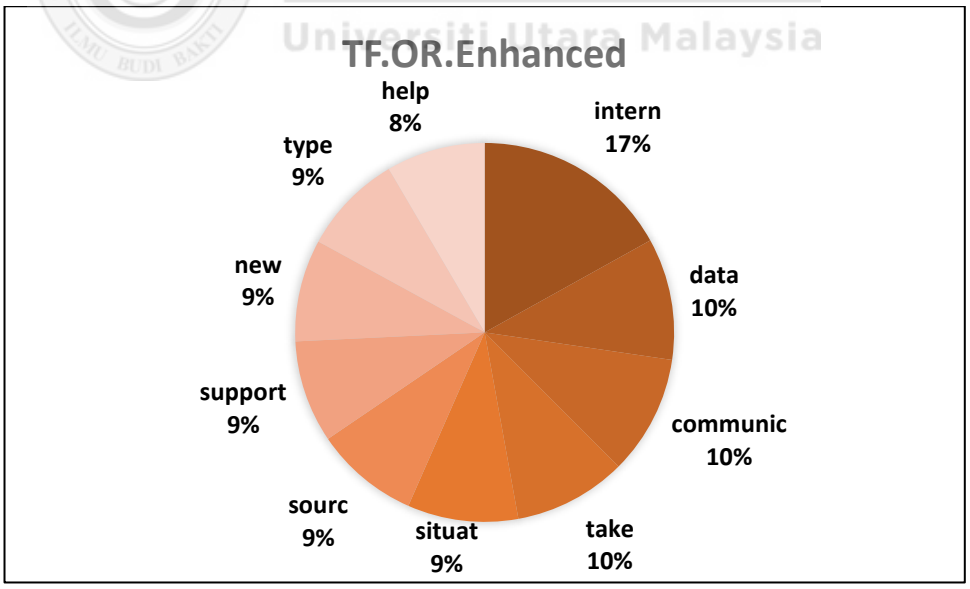

Figure 5.12. TF.OR.Enhanced ten highest-ranked TWS result

Universiti Utara Malaysia

Table 5.16

The ten highest-*ranked* terms and the calculated TF.OR.Enhanced values

No.	Terms	TP	TN	FP	FN	OR	Positive Term Frequency Ratio	Positive Category Ratio	TF.OR. Enhanced
1	intern	112	5	1	274	0.310437464	1	22.4	6.9537992
2	data	113	4	1	274	0.217387872	0.695361	28.25	4.2703531
3	communic	111	6	1	274	0.385723666	0.586108	18.5	4.1824041
4	take	96	21	1	274	0.866739965	1	4.571428571	3.9622399
5	situat	108	9	1	274	0.549915702	0.588494	12	3.8834641
6	sourc	105	12	1	274	0.662619982	0.630699	8.75	3.6567481
7	support	110	7	1	274	0.448740162	0.50925	15.71428571	3.5910416
8	new	111	6	1	274	0.385723666	0.501683	18.5	3.5799525
9	type	108	9	1	274	0.549915702	0.535449	12	3.5334197
10	help	113	4	1	274	0.217387872	0.564764	28.25	3.4683313

Table 5.17

Comparison between TF.OR and TF.OR.Enhanced values

No.	Terms	TF.OR	Terms	TF.OR.Enhanced
1	network	0.11582966	intern	6.9537992
2	model	0.108148954	data	4.2703531
3	govern	0.105745056	communic	4.1824041
4	resili	0.103355852	take	3.9622399
5	disast risk	0.099829281	situat	3.8834641
6	flood	0.090897941	sourc	3.6567481
7	plan	0.088641801	support	3.5910416
8	risk	0.085777014	new	3.5799525
9	journal	0.079229845	type	3.5334197
10	et al	0.077184962	help	3.4683313

Universiti Utara Malaysia

Table 5.16 shows the obtained result for using the enhanced scheme with TF.OR supervised TWS. In Table 5.17, similar with TF.x^2.Enhanced and TF.IG.Enhanced, the TF.OR.Enhanced result did not get better by using the enhanced scheme, since the formula takes the ratio between the positive and negative category. The scheme only considers the positive category ratio, TP and TN, even though FP also contains the terms as discussed in section 5.2.3.

5.3.4 TF.GR.Enhanced

With the TF.GR.Enhanced TWS performed, the results provide that the concepts specified are emphasized on neutral category. The terms extracted as seen in Figure

5.13, such as ‘sediment’, ‘gic’, and ‘solomn’ are considered as meaningless terms to the DM domain.

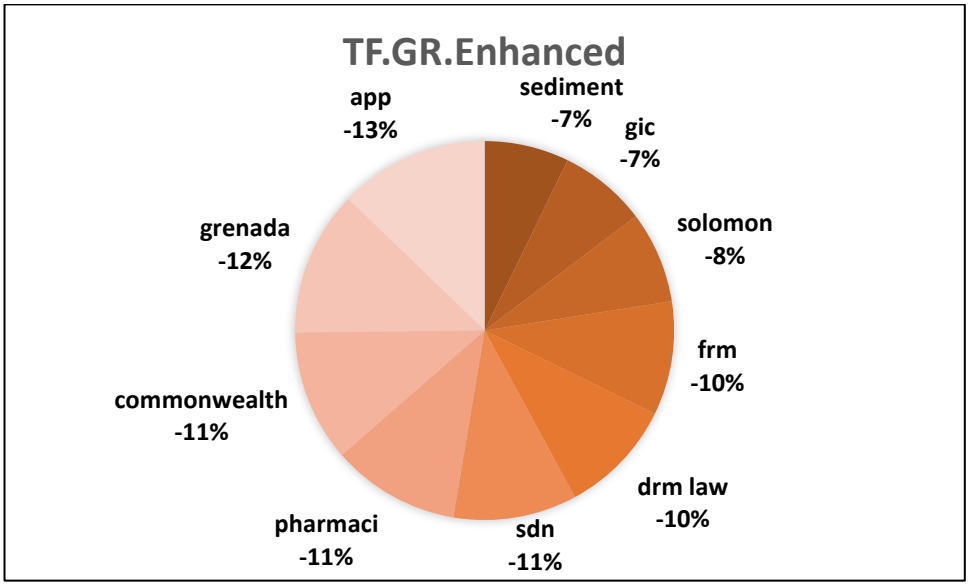

Figure 5.13. TF.GR.Enhanced ten highest-ranked TWS result

Table 5.18

The ten highest-ranked terms and the calculated TF.GR.Enhanced values

No.	Terms	TP	TN	FP	FN	GR	Positive Term Frequency Ratio	Positive Category Ratio	TF.GR. Enhanced
1	sediment	1	116	1	274	-4.59101229	0.008978	0.00862069	-7.83299E-07
2	gic	1	116	1	274	-4.59101229	0.0093	0.00862069	-8.1139E-07
3	solomon	1	116	1	274	-4.59101229	0.009721	0.00862069	-8.48152E-07
4	frm	1	116	1	274	-4.59101229	0.012064	0.00862069	-1.05252E-06
5	drm law	1	116	1	274	-4.59101229	0.01226	0.00862069	-1.06971E-06
6	sdn	1	116	1	274	-4.59101229	0.013117	0.00862069	-1.14444E-06
7	pharmaci	1	116	1	274	-4.59101229	0.013592	0.00862069	-1.18587E-06
8	commonwealth	1	116	1	274	-4.59101229	0.013971	0.00862069	-1.21897E-06
9	grenada	1	116	1	274	-4.59101229	0.015404	0.00862069	-1.34398E-06
10	app	1	116	1	274	-4.59101229	0.015893	0.00862069	-1.38665E-06

Table 5.19

Comparison between TF.GR and TF.GR.Enhanced values

No.	Terms	TF.GR	Terms	TF.GR.Enhanced
1	crossref	91.66360385	sediment	-7.83299E-07
2	estat	70.72855004	gic	-8.1139E-07
3	eoc	60.09581188	solomon	-8.48152E-07
4	wast	58.24423324	frm	-1.05252E-06
5	data	52.53344856	drm law	-1.06971E-06
6	nutrit	51.60347604	sdn	-1.14444E-06

7	risk	51.53283967	pharmaci	-1.18587E-06
8	network	50.3189026	commonwealth	-1.21897E-06
9	tourism	48.70729117	grenada	-1.34398E-06
10	resili	48.01102778	app	-1.38665E-06

As seen in Table *5.18* and Table *5.19*, the result indicates a similar outcome in which no improvement is achieved by using the enhanced scheme. Furthermore, the weighted term in the higher level is meaningless to be considered for the ontology development candidates. This is due to the nature of the equation, discussed in section 5.2.4, whereby the formula finds the ratio between irrelevancies with irrelevancy in negative classification. The finding become the explanation for the TF.GR.Enhanced outcome in which obtain irrelevant results for the enhanced scheme.

5.3.5 TF.RF.Enhanced

The most interesting finding is for the TF.IG.Enhanced. as the results show that the equation focuses on weightage on negative category. From the data in *Figure 5.14*, it is elucidated that the terms extracted, 'social media', 'disast', and 'emerg manag' are related to DMSM domain.

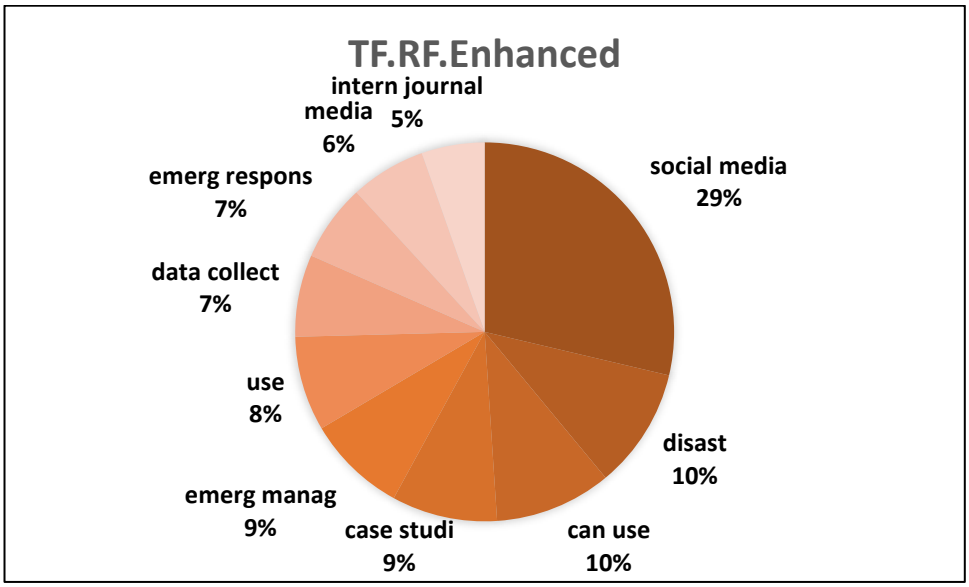

Figure 5.14. TF.RF.Enhanced ten highest-ranked TWS result

Table 5.20

The ten highest-ranked terms and the calculated TF.RF.Enhanced values

No.	Terms	TP	TN	FP	FN	RF	Positive Term Frequency Ratio	Positive Category Ratio	TF.RF. Enhanced
1	social media	112	5	22	253	0.387867954	0.933417542	492.8	178.414668
2	disast	115	2	7	268	0.385446195	0.412501978	402.5	63.9964204
3	can use	81	36	110	165	0.396357878	0.632039081	247.5	62.0021330
4	case studi	78	39	161	114	0.428816575	0.405859572	322	56.0406583
5	emerg manag	80	37	139	136	0.413003755	0.42797418	300.5405405	53.1220262
6	use	116	1	2	273	0.384695342	0.563324517	232	50.2763298
7	data collect	73	44	115	160	0.390272568	0.584583038	190.7954545	43.5293578
8	emerg respons	72	45	133	142	0.399161654	0.482245897	212.8	40.9627380
9	media	116	1	1	274	0.384417517	0.898216209	116	40.0536451
10	intern journal	64	53	152	123	0.401456582	0.454424877	183.5471698	33.4848513

Table 5.21

Comparison between TF.RF and TF.RF.Enhanced values

No.	Terms	TF.RF	Terms	TF.RF.Enhanced
1	disast	0.373340205	social media	178.414668
2	manag	0.21716202	disast	63.9964204
3	use	0.115024219	can use	62.0021330
4	data	0.114586717	case studi	56.0406583
5	inform	0.093054629	emerg manag	53.1220262
6	risk	0.089140386	use	50.2763298
7	disast manag	0.083366108	data collect	43.5293578
8	et al	0.080211434	emerg respons	40.9627380
9	system	0.078852865	media	40.0536451
10	can	0.078108168	intern journal	33.4848513

Contrary to expectations, referring to the supervised TWS results, TF.RF.Enhanced expresses an interesting improvement, in which the term ‘social media’ achieved the highest term 178.414668 as seen in Table 5.20. Compared to TF.RF results, the highest term ranked is ‘disast’ in which obtained 0.373340205 TF.RF rate shown in Table 5.21. The result of TF.RF.Enhanced indicated that the TWS emphasizes on the positive classification of the corpus. As discussed in section 5.2.5, the main reason this scheme able to obtain good result is due to the numerator considers TP value as the factor. With this principle in the equation, by improving the formula with positive term frequency ratio and positive category ratio, the formula is able to give higher

weightage to the positive category. In this way, high ranked terms is more related to the positive category.

5.3.6 TF.MI.Enhanced

For the TWS TF.MI.Enhanced, the results have remarked that the concepts found did not belong to either positive or negative category, being in the neutral state. In *Figure 5.15*, the concepts identified such as 'springer', and 'find', and 'relev', generally did not be part of positive or negative categories.

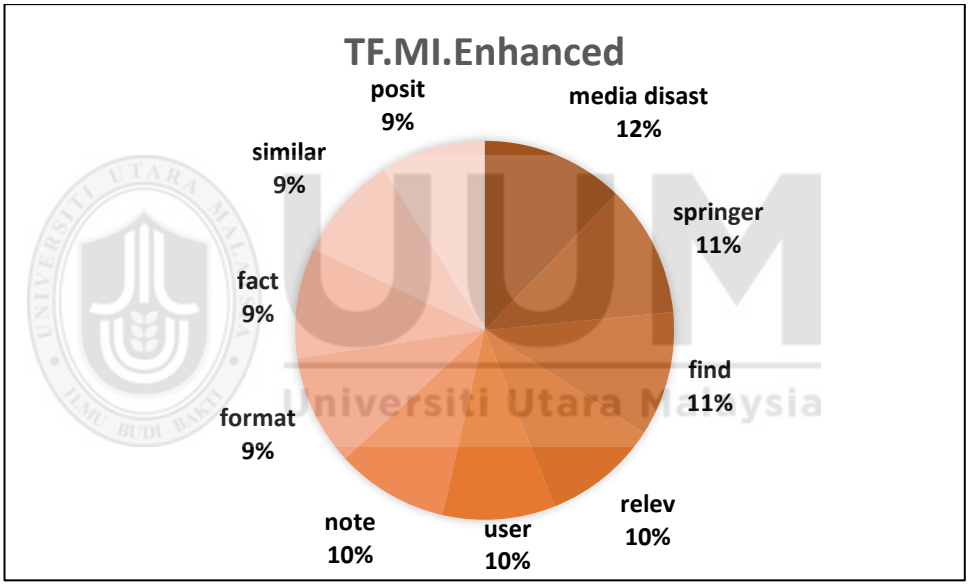

Figure 5.15. TF.MI.Enhanced ten highest-ranked TWS result

Table 5.22

The ten highest-ranked terms and the calculated TF.MI.Enhanced values

No.	Terms	TP	TN	FP	FN	MI	Positive Term Frequency Ratio	Positive Category Ratio	TF.MI. Enhanced
1	media disast	43	74	6	269	0.042403849	0.950959	3.486486486	0.14059018
2	springer	68	49	4	271	0.038262785	0.606249	5.551020408	0.12876583
3	find	88	29	5	270	0.015402764	0.52416	15.17241379	0.12249473
4	relev	91	26	3	272	0.019292981	0.558116	10.5	0.11306112
5	user	82	35	2	273	0.03259228	0.722791	4.685714286	0.11038335
6	note	73	44	5	270	0.029220204	0.454672	8.295454545	0.11021011 4

7	format	85	32	3	272	0.025554104	0.533143	7.96875	0.10856617
8	fact	71	46	5	270	0.030976444	0.453435	7.717391304	0.10839701
9	similar	92	25	4	271	0.014896036	0.470449	14.72	0.10315517
10	posit	85	32	5	270	0.01823413	0.425211	13.28125	0.10297411

Table 5.23

Comparison between TF.MI and TF.MI.Enhanced value

No.	Terms	TF.MI	Terms	TF.MI.Enhanced
1	section	0.005693175	media disast	0.14059018
2	survey	0.005366134	springer	0.12876583
3	wast	0.004763225	find	0.12249473
4	tourism	0.004080212	relev	0.11306112
5	resili	0.003614233	user	0.11038335
6	sensor	0.00338697	note	0.110210114
7	tweet	0.003150933	format	0.10856617
8	disast risk	0.003142625	fact	0.10839701
9	fig	0.003112841	similar	0.10315517
10	node	0.003064347	posit	0.10297411

Based on the extracted terms in Table 5.22 and the TF.MI values, the scheme is unable to deliver the expected outcome for the ontology development candidates since most of the extracted terms are irrelevant to the domain. Furthermore, compared to the current TF.MI scheme, the TF.MI.Enhanced result did not improve as many irrelevant terms are weighted as high candidates as shown in Table 5.23. As discussed in section 5.2.6, the equation takes into account negative classes, FP and FN as the main factor in the numerator and denominator, therefore the extracted terms mostly are irrelevant to the context.

5.3.7 TF.IDFEC-based.Enhanced

Another important finding is TF.IDFEC-based.Enhanced, with the result displays a strong relationship between the concepts extracted with the positive category. As shown in *Figure 5.16*, the terms ‘social media’, and ‘disast’ are clearly relevant to DMSM domain.

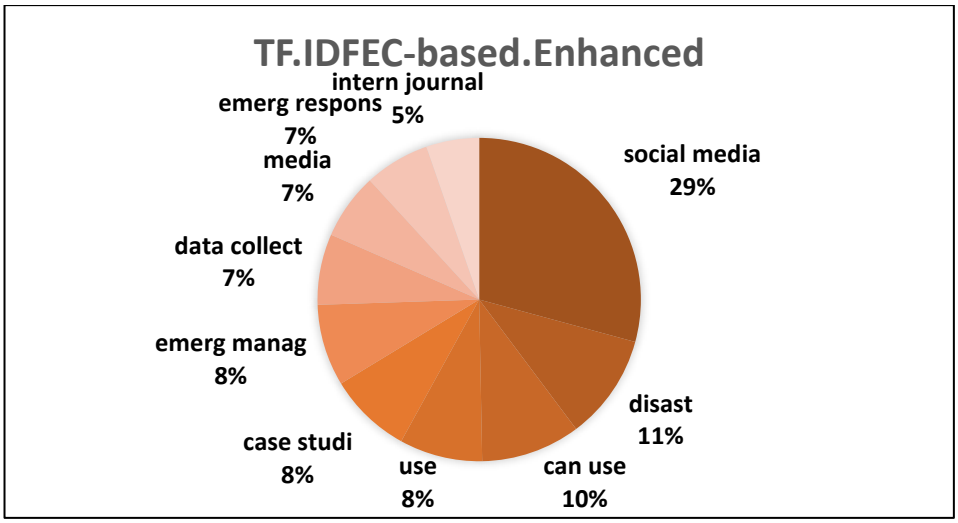

Figure 5.16. TF.IDFEC-based.Enhanced ten highest-ranked TWS result

Table 5.24

The ten highest-ranked terms and the calculated TF.IDFEC-based.Enhanced values

No.	Terms	TP	TN	FP	FN	IDFEC-based	Positive Term Frequency Ratio	Positive Category Ratio	TF.IDFEC-based. Enhanced
1	social media	112	5	22	253	2.079181246	0.933418	492.8	956.398765
2	disast	115	2	7	268	2.079181246	0.412502	402.5	345.210717
3	can use	81	36	110	165	2.079181246	0.632039	247.5	325.245641
4	use	116	1	2	273	2.079181246	0.563325	232	271.730875
5	case studi	78	39	161	114	2.079181246	0.40586	322	271.721506
6	emerg manag	80	37	139	136	2.079181246	0.427974	300.5405405	267.431759
7	data collect	73	44	115	160	2.079181246	0.584583	190.7954545	231.903116
8	media	116	1	1	274	2.079181246	0.898216	116	216.636298
9	emerg respons	72	45	133	142	2.079181246	0.482246	212.8	213.369586
10	intern journal	64	53	152	123	2.079181246	0.454425	183.5471698	173.421181

Table 5.25

Comparison between TF.IDFEC-based and TF.IDFEC-based.Enhanced values

No.	Terms	TF.IDFEC-based	Terms	TF.IDFEC-based.Enhanced
1	disast	2.013878885	social media	956.398765
2	manag	1.174554178	disast	345.210717
3	data	0.622946774	can use	325.245641
4	use	0.621676877	use	271.730875
5	inform	0.503300266	case studi	271.721506
6	risk	0.491070512	emerg manag	267.431759
7	disast manag	0.452440324	data collect	231.903116
8	et al	0.441881304	media	216.636298
9	system	0.427453126	emerg respons	213.369586
10	can	0.422460036	intern journal	173.421181

Universiti Utara Malaysia

As presented in Table 5.24, the term ‘social media’ has achieved high results with 956.398765 TF.IDFEC-based.Enhanced value, which is considered as a distribution to the positive category. This scheme is also seen as improved by using the positive term frequency ratio and positive category ratio if compared with the current TF.IDFEC-based result as shown in Table 5.25. It is observed that the TF.RF is also obtaining the similar result. Based on the explanation in section 5.2.7, the equation emphasizes TP value in the numerator as representation for positive category with the negative category. The nature of the equation is able to obtain terms related to positive category, whereby by adding the factor of positive term frequency ratio and positive category ratio introduced in this study, improve the extracted terms related to the positive category.

Universiti Utara Malaysia

5.3.8 TF.SW.Enhanced

With minimal changes, the result of TF.SW.Enhanced also indicates that the concepts retrieved are more towards the positive category or Disaster Management domain. The concepts captured are ‘disast’, ‘social media’, and ‘emerg manag’ are generally related to the DM domain as seen in *Figure 5.17*.

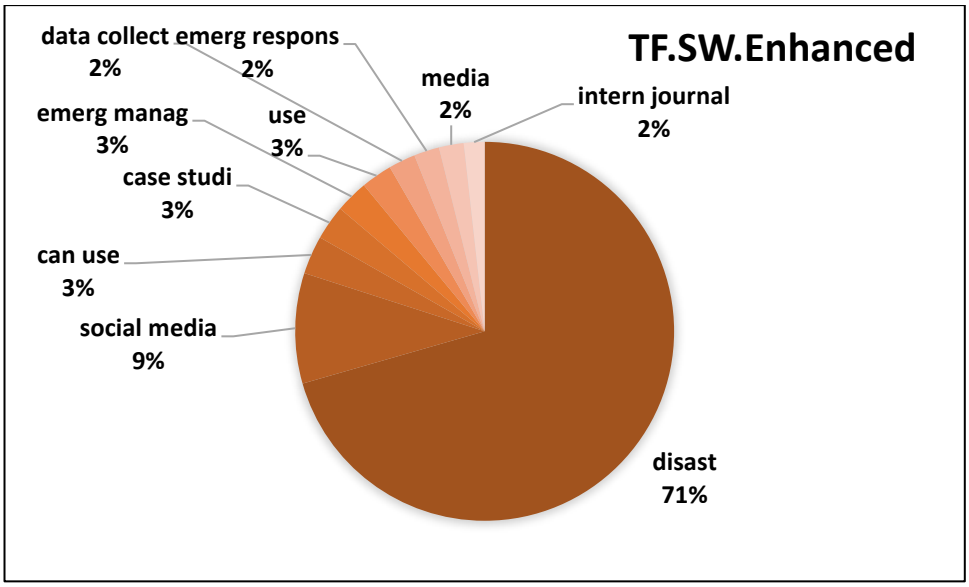

Figure 5.17. TF.SW.Enhanced ten highest-ranked TWS result

Table 5.26

The ten highest-ranked terms and the calculated TF.SW.Enhanced values

No.	Terms	TP	TN	FP	FN	SW	Positive Term Frequency Ratio	Positive Category Ratio	TF.SW. Enhanced
1	disast	115	2	7	268	8.049573561	0.412502	402.5	1336.487
2	social media	112	5	22	253	0.387867954	0.933418	492.8	178.4147
3	can use	81	36	110	165	0.396357878	0.632039	247.5	62.00213
4	case studi	78	39	161	114	0.428816575	0.40586	322	56.04066
5	emerg manag	80	37	139	136	0.413003755	0.427974	300.5405	53.12203
6	use	116	1	2	273	0.384695342	0.563325	232	50.27633
7	data collect	73	44	115	160	0.390272568	0.584583	190.7955	43.52936
8	emerg respons	72	45	133	142	0.399161654	0.482246	212.8	40.96274
9	media	116	1	1	274	0.384417517	0.898216	116	40.05365
10	intern journal	64	53	152	123	0.401456582	0.454425	183.5472	33.48485

Table 5.27

Comparison between TF.SW and TF.SW.Enhanced values

No.	Terms	TF.SW	Terms	TF.SW.Enhanced
1	disast	7.796754736	disast	1336.487
2	manag	0.21716202	social media	178.4147
3	use	0.115024219	can use	62.00213
4	data	0.114586717	case studi	56.04066
5	inform	0.093054629	emerg manag	53.12203
6	risk	0.089140386	use	50.27633
7	disast manag	0.083366108	data collect	43.52936

8	et al	0.080211434	emerg respons	40.96274
9	system	0.078852865	media	40.05365
10	can	0.078108168	intern journal	33.48485

TF.SW.Enhanced obtained an improvement by using the positive term frequency ratio and the positive category ratio. The highest term weightage achieved is for 'disast' with 1336.487 TF.SW.Enhanced rate listed in Table 5.26. However, this scheme did not obtain the highest term from positive document classification as shown in

Table 5.27. As discussed in section 5.2.8, the denominator utilises all negative category inclusive FP and FN which makes the denominator larger value, hence smaller ratio gained produced for the formula. With smaller ratio to the numerator, TP, convey the terms are related to negative category.

5.3.9 Results Comparisons between Constructed Enhanced Supervised Term Weighting Schemes

Table 5.28 presented the summary of comparison between the findings for constructed enhanced TWS that have been described in previous sections. Mostly, other supervised TWS obtained the term category information as neutral and negative except for TF.RF.Enhanced and TF.IDFEC-based.Enhanced.

Table 5.28

Comparison between the findings for constructed enhanced supervised TWS

Term Weightage	Category Information
TF.x^2.Enhanced	Neutral
TF.IG.Enhanced	Neutral
TF.OR.Enhanced	Neutral
TF.GR.Enhanced	Neutral
TF.RF.Enhanced	Positive
TF.MI.Enhanced	Neutral
TF.IDFEC-based.Enhanced	Positive
TF.SW.Enhanced	Negative

The TF.RF.Enhanced scheme is utilized to increase the discrimination of the positive and negative documents. By factoring with the constructed enhanced supervised TWS, the weightages are able to be shifted to the positive category. The principle behind TF.RF.Enhanced is to retrieve terms with higher frequencies and concentration to the positive class to discriminate between the negative classes and vice versa. With this basic idea, it is expected that TF.RF.Enhanced is able to obtain results based on the positive category concepts by factoring with the positive category ratio.

TF.IDFEC-based.Enhanced is also able to achieve the desired result since the weights assigned to term t_k occur in documents that related to the same category c_i is set to be higher. With the constructed enhanced supervised TWS factored in, the term categorised in the positive category will be to retrieve higher results.

Based on these findings, TF.RF.Enhanced and TF.IDFEC-based.Enhanced supervised TWS are selected for the next section. The next section discusses the evaluation performed in developing the DMSM domain ontology and assessed against the gold-standard-based evaluation method. Appendix A provide an example of list of the extracted terms and calculated weightages for TF.IDFEC-based.Enhanced and TF.RF.Enhanced as reference.

Universiti Utara Malaysia

5.4 Summary

Results from the TWS experimental work were achieved and analysed in this chapter. The first component of the results concerning on the identification of existing supervised TWS indicate that the schemes must be further improved since most of the extracted terms are given more weight on the negative category, in which must be enhance so that positive category will obtain higher weightage. In this way, more relevant and suitable terms can be extracted for the DMSM ontology development. The enhancement for the existing schemes results are presented, in which TF.IDFEC-based and TF.RF delivers the best results compared to others. The subsequent chapter discusses the experiment conducted to achieve the third objective of this research. A detailed discussion of the gold-standard-based evaluation method to evaluate the DMSM ontology developed from these experimental results is presented.

Universiti Utara Malaysia

CHAPTER 6

EVALUATION OF THE ENHANCED SUPERVISED TERM WEIGHTING SCHEMES THROUGH THE DEVELOPED ONTOLOGY

6.1 Introduction

The procedure of the construction of the enhanced supervised TWS evaluation was undertaken by performing the evaluation stage. In this process, the gold-standard-based evaluation method was selected to ensure a complete and comprehensive evaluation process. These stages are elaborated in detail throughout this chapter, hence addressing the final objective of this research.

Universiti Utara Malaysia

6.2 Evaluation of the Enhanced Supervised Term Weighting Schemes through the Developed Ontology

The evaluation was carried out to verify whether the constructed enhanced supervised TWS conforms to its purpose. In the present study, the evaluation stage was conducted to verify that, i) the text mining technique used to enhance the TWS, and ii) the concept included in the enhanced model properly accurately represents the social media aspect. To achieve this, the gold-standard-based evaluation method had been selected because it has been recognized as a reliable method to evaluate ontology.

6.2.1 Step 1: Perform Term Filtering to the Extracted Terms

In Chapter Five, as elaborated, all extracted terms from positive and negative categories are used to compute the term weightages by using the eight selected

supervised TWS. From all eight, it is found that the two TWS achieved the desired results which are TF.RF.Enhanced and TF.IDFEC-based.Enhanced. From these two sets of result consisting of ranked term based on the calculated term weightages, are further filtered in order to remove meaningless and irrelevant terms for ontology development candidate terms, in other words, to preserve the data integrity. The evaluation also includes existing supervised TWS, TF.RF and TF.IDFEC-based, and unsupervised TWS, TF-IDF, as benchmarks for comparison with the enhanced schemes in this study.

Table 6.1

Summary of term filtering performed for TF.RF.Enhanced term weighting

Steps	Example of Terms		Total Terms Removed	Total Remaining Terms
Step 1: Illegal characters removal	<U+FB01>rst © author © elsevi		3	497
Step 2: Verbs removal	also consid also help also includ		229	268
Step 3: Scientific publication terms removal	author author declar author email avail sciencedirect		99	169
Step 4: Irrelevant terms removal	fact survey limit resourc		6	163
Step 5: Perform merging	relief oper oper	relief oper	85	78
	emerg respons respons crisi respons	emerg respons		

Table 6.2

Summary of term filtering performed for TF.IDFEC-based.Enhanced term weighting

Steps	Example of Terms	Total Terms Removed	Total Remaining Terms
Step 1: Illegal characters removal	<U+FB01>eld <U+FB01>rst © author © elsevi	4	496
Step 2: Verbs removal	can use use	163	333

	need can can also			
Step 3: Scientific publication terms removal	author case studi isbn literatur review		110	223
Step 4: Irrelevant terms removal	fact base build		5	218
Step 5: Perform merging	Disast earthquak earthquak	earthquak	128	90
	natur natur hazard natur manmad hazard	natur hazard		

Table 6.3

Summary of term filtering performed for TF.RF term weighting

Steps	Example of Terms		Total Terms Removed	Total Remaining Terms
Step 1: Illegal characters removal	<U+FB01>rst © author © elsevi		5	495
Step 2: Verbs removal	also consid also help also includ		302	193
Step 3: Scientific publication terms removal	author journal framework		104	89
Step 4: Irrelevant terms removal	fact survey one		5	84
Step 5: Perform merging	system app	application	44	40
	manag disast manag	disast manag		

Table 6.4

Summary of term filtering performed for TF.IDFEC-based term weighting

Steps	Example of Terms		Total Terms Removed	Total Remaining Terms
Step 1: Illegal characters removal	<U+FB01>eld <U+FB01>rst © author © elsevi		4	496
Step 2: Verbs removal	also can follow		325	171
Step 3: Scientific publication terms removal	author et al journal		110	61
Step 4: Irrelevant terms removal	fact base build		5	56
Step 5: Perform merging	social media media	social media	17	39
	respond respons	respons		

Table 6.5

Summary of term filtering performed for TF-IDF term weighting

Steps	Example of Terms		Total Terms Removed	Total Remaining Terms
Step 1: Illegal characters removal	land<U+FB01>ll		1	499
Step 2: Verbs removal	can use use need		255	244
Step 3: Scientific publication terms removal	et al figure ieee		105	139
Step 4: Irrelevant terms removal	cbdrm drm drr		5	134
Step 5: Perform merging	risk manag risk	rusk	53	81
	social media media	social media		

As defined in Chapter Three, five sub-steps are applied to filter the database from trivial data, namely illegal characters removal, verbs removal, scientific publication terms removal, irrelevant terms removal, and perform merging. Table 6.1, Table 6.2, Table 6.3, Table 6.4, and Table 6.5 provided the summary of the five main steps in filtering the extracted term. This steps are applied for TF.RF.Enhanced, TF.IDFEC-based.Enhanced, TF.RF, TF.IDFEC-based, and TF-IDF TWS. The tables show the steps performed, example of terms, the total number of terms removed, and total number of remaining terms. The final number of terms is included in the next step of this process. The highest 500 terms are selected in this process since the remaining terms afterwards are observed and considered having low term weightages. As an example, through the constructed enhanced supervised TWS with TF.RF.Enhanced, 78 terms are filtered and finalized for the evaluation in this chapter, together with 90 terms for TF.IDEFEC-based.Enhanced, 40 terms for TF.RF, 39 terms for TF.IDFEC-based, and 81 terms for TF-IDF.

Universiti Utara Malaysia

6.2.2 Step 2: Calculate Distance for Syntactic Comparison

With the extracted terms from TF.RF.Enhanced, all of the terms are gathered to calculate the distance for syntactic comparison.

Step 2.1: Compute Euclidean Distance: As described in Chapter Three, the Euclidean distance is utilised in this step to compute the interval between two concepts. The computed distance is used as a reference to develop associations between concepts.

Step 2.2: Perform hierarchy of terms using hierarchical clustering method: From the computed Euclidean distance, the data is gathered and processed for the development of hierarchy of terms. In this way, the hierarchy can refer to for the next step in designing the learned ontology. As seen in Figure 6.1, terms with nearest Euclidean distance values are grouped, while others will be in other groups. As an example, for the term 'disaster' and 'management' are group together as the values are close, with the computed Euclidean distance of 0.39 while terms under these concepts are defined with 'is-a' relationship as parent and child linkage.

Universiti Utara Malaysia

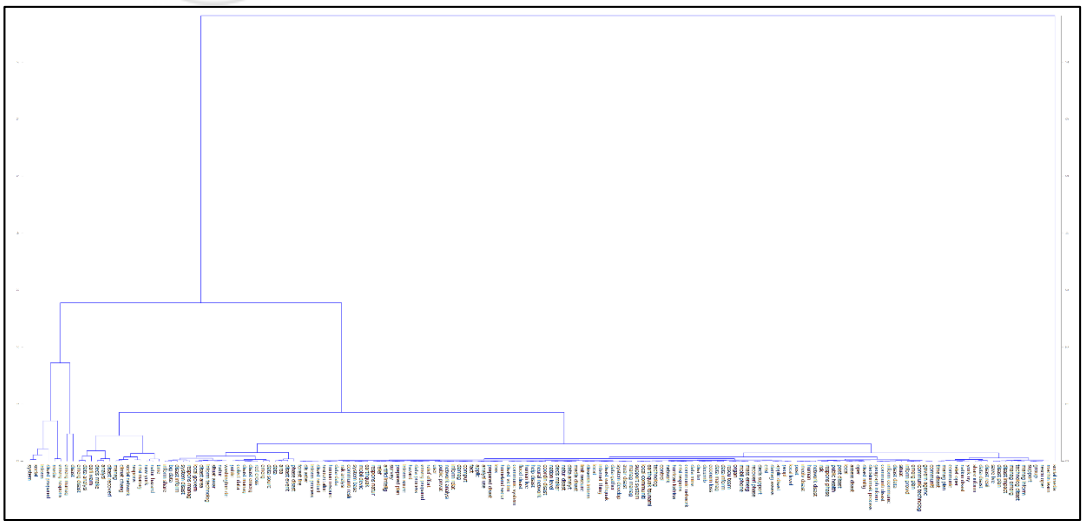

Figure 6.1. Hierarchy of terms by using hierarchical clustering method for TF.RF.Enhanced result.

6.2.3 Step 3: Design the Learned Ontology

A learned ontology is developed by using the concepts extracted from the text mining process as shown in Figure 6.2 with TF.RF.Enhanced and Figure 6.3 with TF.IDFEC-based.Enhanced. This includes referring to the hierarchy of terms data for the association. However, the grouping involves human or manual intervention to make certain the concepts are correctly mapped. Terms that contain closer distance are grouped together, while terms that contains high distance value is considered a parent node with 'is-a' relationship definition. As an example for TF.RF.Enhanced in Figure 6.2, the term 'Hurricane Katrina' and 'Earthquake' are grouped together, while the 'Natural Hazard' is assigned as the parent node for both of the terms. This process is crucial since the learned ontology will be used to be evaluated against the gold-standard or referenced ontology. The ontology for TF.RF, TF.IDFEC-based, and TF-IDF are not designed for this step since TF-RF.Enhanced and TF.IDFEC-based is the only schemes that obtained the best result, while TF.RF, TF.IDFEC-based, and TF-IDF schemes is assessed for this evaluation for the purpose of comparison.

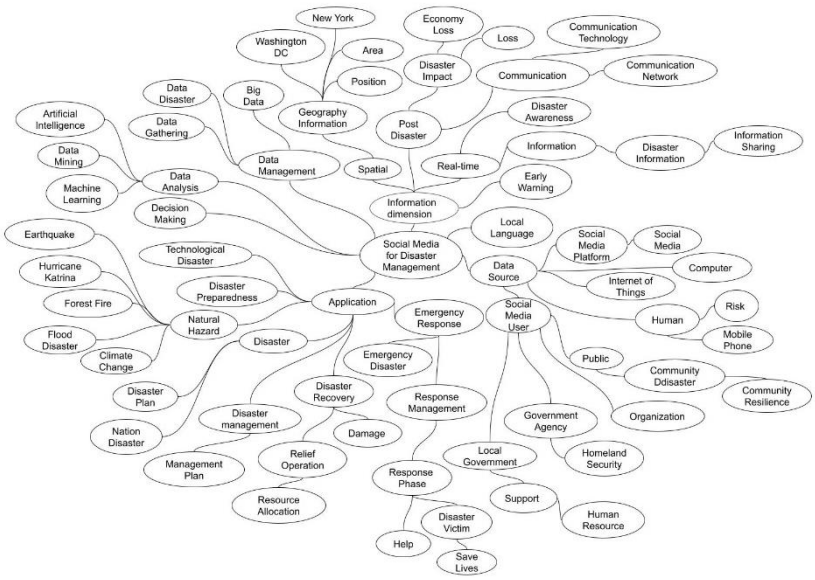

Figure 6.2. The designed learned ontology based on TF.RF.Enhanced result.

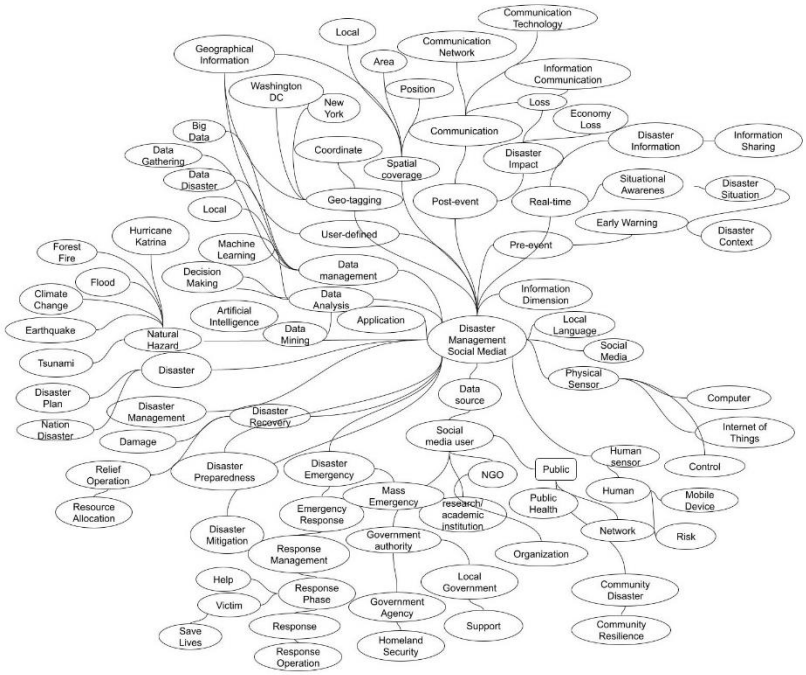

Figure 6.3. The designed learned ontology based on TF.IDFEC-based.Enhanced result.

6.2.4 Step 4: Evaluate the Learned Ontology through Gold-standard Based Evaluation Method

After the learned ontology has been fully designed, the evaluation is ready to be conducted against the gold-standard ontology. Three main steps are involved in this step, namely, 1) Extract gold-standard based ontology concepts, 2) Perform mapping between learned and referenced ontology concepts, and 3) Compute precision, recall, and F-measure.

<u>Step 4.1:</u> Extract gold-standard-based ontology concepts: To facilitate the mapping process, the gold-standard-based ontology concepts are extracted and transformed into a common representation. This will make the mapping process between learned and referenced ontology more effective. Table 6.6 shows the extracted gold-standard-based ontology concepts. Basically, there are five main segments for the gold-standard ontology, namely, Information Dimension, Methodology, Application, Data Source,

and Language. From the main segments, the sub-segments are also extracted for each segment.

Table 6.6

List of extracted terms from the gold-standard taxonomy

Extracted Terms from Phengsuwan et al. (2021) DMSM Taxonomy		
Data source	Social Media User	Government Authority
		Research/Academic Institution
		NGO
		Public
	Sensor	Human Sensor
		Physical Sensor
	Social Media Platform	
	Third Party	
Language	Mixed Script	
	Mixed Language	
	Local Language	
	Global Language	
Information Dimension	Temporal	Post-event
		Real-time
		Pre-event
	Spatial	Spatial Coverage
		User-defined
		Geo-tagging
Methodology	Data Management	
	Data Analysis	
Application	Disaster Management Type	
	Disaster Management Phase	

Universiti Utara Malaysia

Step 4.2: Perform mapping between learned and referenced ontology concepts: With both ontology concepts extracted to common representation, the concepts between the learned ontology are mapped to referenced ontology. If the concept exists in both ontologies, then the mapped value for the particular concept is set as 1. The mapped values are accumulated for the next step in measuring the precision, recall, and F-measure.

Step 4.3: Compute precision, recall, and F-measure: From the accumulated mapped values, the precision, recall, and F-measure of both ontologies are measured.

Table 6.7, Table 6.8, Table 6.9, Table 6.10, and Table 6.11 presents the computation including the mapped terms for gold-standard based and learned ontology, indicated by the greyed cells for TF.RF,Enhanced, TF.IDFEC-based.Enhanced, TF.RF, TF.IDFEC-based, and TF-IDF TWS. The results of each scheme are discussed and analysed in the next section.

6.3 Results and Discussion

Based on the developed DMSM ontology from the previous section, the outcome is discussed and analysed in this section. The discussion is categorised based on the enhanced and current supervised TWS, including benchmark technique, TF-IDF, the unsupervised TWS.

Table 6.7 shows the end result of the enhanced learned ontology generated based on the mapping with the gold-standard ontology. The first column group listed all of the gold standard extracted terms and the learned ontology terms that are mapped to the referenced terms are listed in the second column group. Within each column is the segmentation for the concepts and mapped terms, with the total number of extracted terms and mapped terms are labelled with (1) indicating 1 count for the aggregation. Subsequent columns present the total learned extracted terms, total gold extracted terms, and total mapped terms, in which all of the values are used in the precision, recall, and F-measure computation. The precision, recall, and F-measure rate are shown at the bottom of the table. Five evaluations are performed for two of the best-supervised TWS, TF.RF.Enhanced and TF.IDFEC-based.Enhanced, the existing supervised TWS, TF.RF and TF.IDFEC-based, and unsupervised TWS, TF-IDF.

The results show significant improvement by using supervised TWS compared to existing works by using unsupervised TWS such as the study conducted by Liu &

Zhang (2020), an average accuracy of 53.16% and 63.73% recall are obtained. Salatino et al. (2019) also developed classification of research papers according to the Computer Science Ontology (CSO), found the results for precision of 40.4%, recall of 24.1%, and 30.1% F-measure by using TF-IDF. Compared to these results, the enhanced supervised TWS, TF.IDFEC-based.Enhanced and TF.RF.Enhanced achieved better results with 93.33% and 91.03% for precision, 80.8% and 78.02% for recall, and 0.87 and 0.84 for F-measure, respectively.

6.3.1 DMSM Ontology Development based on TF.RF.Enhanced

Table 6.7 presents the result of the mapping between learned and gold-standard extracted terms based on the TF.RF.Enhanced supervised TWS. From the mapping, the parent segments for this ontology are identified as Data Source, Language, Information Dimension, Methodology, and Application.

For Data Source, it represents the various source of social media data used to facilitate social media data analysis in DM. Type of sources could be obtained from Social Media User, Sensor, Social Media Platform, and Third Party. The concepts such as Local Government, Support, Mobile Phone, and Computer are examples of concepts that are introduced from the learned ontology for the purpose of to better describe the real-time scenario for Data Source in DM.

Language parent concept consists of language used for writing posts on social media, namely, Global Language, Local Language, Mixed Language, and Mixed Script types. The learned ontology mapped a new concept called Local to represent the local language used in posts for social media. The next segment, Information Dimension, represents the methodologies for comparing social media content to identify Spatial and Temporal information. In this parent concept, interesting findings are gathered

from the mapped learned ontology such as Position, Early Warning, and Communication are added.

The methodology concept represents the methodologies and algorithms used to examine social media data, categorised by Data Management and Data Analysis. This concept is further improved as an example by adding Big Data, Artificial Intelligence, and Data Mining concepts to the Methodology. Finally, the Application concept which represents the current utilisation of social media data for disaster management, which are categorised into the two elements of Disaster Management Phases and Disaster Management Types. These concepts are also extended with Technological Disaster and Natural Hazard.

As shown in Table 6.7, as an example, the total extracted term from the learned ontology is 78, and the total mapped terms from the learned ontology, C_L is 71. Therefore, the total mapped terms divided by the total extracted terms, $C_G \cup C_L$ / C_L is 71 / 78 *100 = 91.03% obtained for the Probabilistic Cotopy Precision (PCP). Meanwhile, the Probabilistic Cotopy Recall (PCR), recall also has achieved 78.02% from all of the terms as seen in Table 6.7, with the total mapped terms divided by the total gold-standard extracted terms, $C_G \cup C_L$ / C_G. As shown in Table 6.7, overall, from the computation, it is found that the learned and gold-standard ontology are mapped well with 91.03% precision, 78.02% recall, and 0.84 F-measure values.

91.03% precision and 78.02% recall indicated that the text mining process has achieved high identified concepts for DMSM domain ontology. F-measure computation successfully obtained the value of 0.84 which falls under 0 to 1, indicating a good result. Overall, the constructed enhanced supervised TWS introduced in this study has achieved a very good outcome. The next section presents and calculates the

total mapped terms, precision, recall, and F-measure value for the second enhanced supervised TWS, TF.IDFEC-based.Enhanced.

Figure 6.4 depict the aligned ontology that correlates between the learned and gold ontology based on TF.RF.Enhanced result. Basically, the design is based on the gold ontology, with the identified concepts from TF.RF.Enhanced is added to the ontology. The process of including the new concepts is by using the mapping results in Table 6.7.

Concepts that are mapped with the gold ontology are included in the final design based on the gold ontology segments. The main components of the gold ontology which are Data Source, Language, Information Dimension, Methodology, and Application is maintained including the child nodes such as Social Media User and Sensor. From the ontology developed, it can be seen that new concepts are added for Government Authority, which are Government Agency and Homeland Security. This suggests that Homeland Security is responsible in supporting and coordinating the social media for the government agency. The term Hurricane Katrina, Forest Fire, and Flood Disaster is the most discussed natural disasters when using social media during the disaster event.

Universiti Utara Malaysia

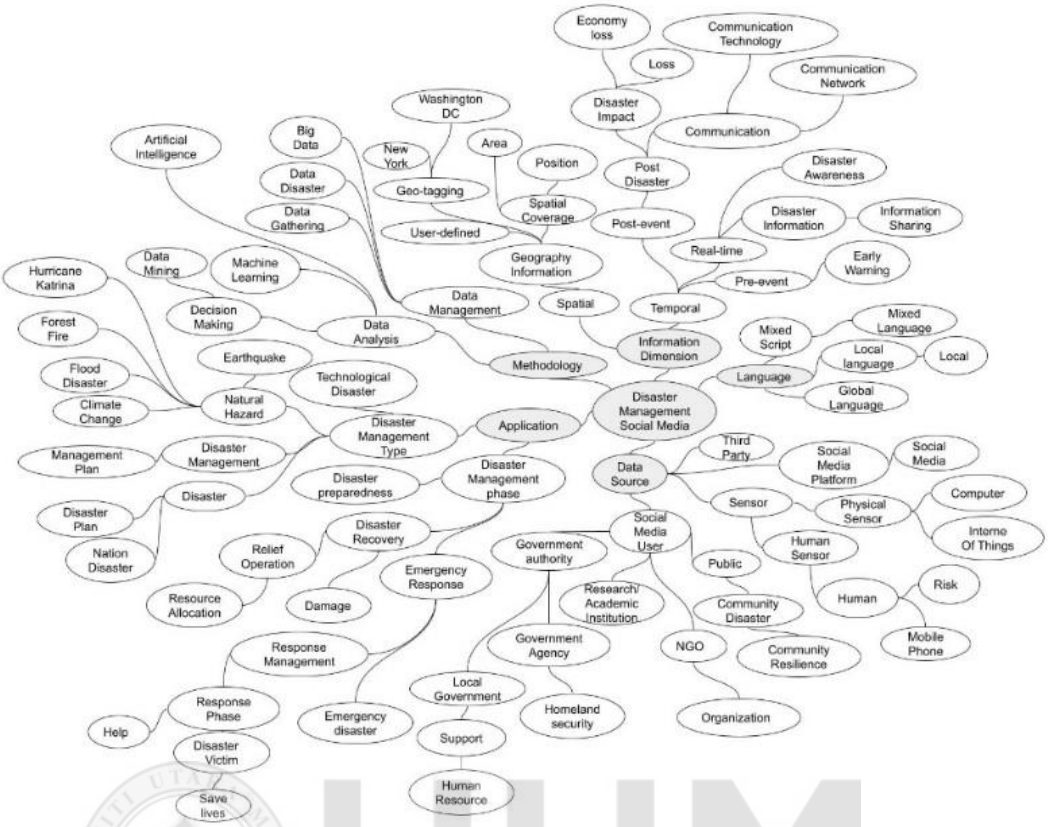

Figure 6.4. Ontology developed from TF.RF.Enhanced results.

Universiti Utara Malaysia

Table 6.7

Precision and recall for mapped learned and gold-standard ontologies for TF.RF.Enhanced supervised TWS.

Gold-standard Based Taxonomy of DMSM Concepts (Phengsuwan et al., 2021)					Learned Ontology Concepts			Total Learned Extracted Terms	Total Gold Extracted Terms	Total Mapped Terms
Data Source (1)	Social Media User (1)	Government Authority (1)	Local Government (1) Support (1) - Human Resource (1) Government Agency (1) - Homeland Security (1)	Data Source (1)	Social Media User (1)	Government Authority	Local Government (1) Support (1) - Human Resource (1) Government Agency (1) - Homeland Security (1)	78	25	17
		Research/Academic Institution (1)				Research/Academic Institution				
		NGO (1)	Organization (1)			NGO	Organization (1)			
		Public (1)	Community Disaster (1) - Community Resilience (1)			Public (1)	Community Disaster (1) - Community Resilience (1)			
	Sensor (1)	Human Sensor (1)	Human (1) - Risk (1) - Mobile Phone (1)		Sensor	Human Sensor	Human (1) - Risk (1) - Mobile Phone (1)			
		Physical Sensor (1)	Computer (1) Internet of Things (1)			Physical Sensor	Computer (1) Internet of Things (1)			
	Social Media Platform (1)	Social Media (1)			Social Media Platform	Social Media (1)				
	Third Party (1)				Third Party					
	Mixed Script (1)			Language	Mixed Script				5	1

Universiti Utara Malaysia

Language (1)	Mixed Language (1)				Mixed Language		78		
	Local Language (1)				Local Language (1)				
	Global Language (1)				Global Language				
Information Dimension (1)	Temporal (1)	Post-event (1)	Communication (1) - Communication Technology (1) - Communication Network (1) Disaster Impact (1) - Economy Loss (1) - Loss (1)	Information Dimension (1)	Temporal	Post-event (1)	Communication (1) - Communication Technology (1) - Communication Network (1) Disaster Impact (1) - Economy Loss (1) - Loss (1)	78	24	19
		Real-time (1)	Situational awareness (1) Information (1) - Disaster Information (1) -- Information Sharing (1)			Real-time (1)	Situational awareness (1) Information (1) - Disaster Information (1) -- Information Sharing (1)			
		Pre-event (1)	Early warning (1)			Pre-event	Early warning (1)			
	Spatial (1)	Spatial Coverage (1)	Area (1) Position (1)		Spatial (1)	Spatial Coverage	Area (1) Position (1)			
		User-defined (1)				User-defined				
		Geo-tagging (1)	New York (1) Washington DC (1)			Geo-tagging	New York (1) Washington DC (1)			
Methodology (1)	Data Management (1)		Big Data (1) Data Disaster (1) Data Gathering (1)	Methodology	Data Management (1)		Big Data (1) Data Disaster (1) Data Gathering (1)	78	10	9
	Data Analysis (1)		Machine Learning (1) Data Mining (1) Artificial Intelligence (1) Decision making (1)		Data Analysis (1)		Machine Learning (1) Data Mining (1) Artificial Intelligence (1) Decision making (1)			

Universiti Utara Malaysia

Application (1)	Disaster Management Type (1)	Natural hazard (1) - Climate change (1) - Earthquake (1) - Hurricane Katrina (1) - Forest Fire (1) - Flood Disaster (1) Disaster Management (1) - Management plan (1) Disaster (1) - Disaster plan (1) - Nation Disaster (1) Technological Disaster (1)	Application (1)	Disaster Management Type	Natural hazard (1) - Climate change (1) - Earthquake (1) - Hurricane Katrina (1) - Forest Fire (1) - Flood Disaster (1) Disaster Management (1) - Management plan (1) Disaster (1) - Disaster plan (1) - Nation Disaster (1) Technological Disaster (1)	78	27	25
	Disaster Management Phase (1)	Disaster preparedness (1) Disaster recovery (1) - Relief operation (1) -- Resource Allocation (1) - Damage (1) Disaster Emergency (1) - Emergency Response (1) - Response Management (1) -- Response Phase (1) --- Victim (1) -- Save lives (1) --- Help (1)		Disaster Management Phase	Disaster preparedness (1) Disaster recovery (1) - Relief operation (1) -- Resource Allocation (1) - Damage (1) Disaster Emergency (1) - Emergency Response (1) - Response Management (1) -- Response Phase (1) --- Victim (1) -- Save lives (1) --- Help (1)			
						Total	**91**	**71**
							Precision	**91.03%**
							Recall	**78.02%**
							F-measure	**0.84**

Universiti Utara Malaysia

6.3.2 DMSM Ontology Development based on TF.IDFEC-based.Enhanced

Similar with the TF.IDFEC-based.Enhanced scheme, good results have been obtained with the enhanced ontology from the gold-standard based as shown in Table 6.8. The parent nodes are maintained with Data Source, Language, Information Dimension, Methodology, and Data Source.

The TF.IDFEC-based.Enhanced scheme has introduced several new terms for improvement, as an example for Data Source segment, Government Authority is included with 'Local Government', 'Support', 'Government Agency', and 'Homeland Security'. This indicated that under government authority there is local government that provides support during the disaster event. Other than that, one of the government agencies involved in facilitating the DM operation is Homeland Security in which provides emergency response and promote relevant government officials to communicate with the affected areas. It is apparent for the Language segment; no new terms are introduced except for mapped terms 'Local language' between learned and gold-standard concepts.

Universiti Utara Malaysia

Identical newly added terms with the TF.IDFEC-based.Enhanced scheme are identified for Information Dimension segment, but the distinct terms added are Coordinate and Geographical Information. This finding signifies that the geo-tagging feature provided from social media capabilities allows coordinate and geographic information such as place name and features. The other segments, Methodology and Application consist of similar results as TF.IDFEC-based.Enhanced extended with Natural Hazard and DM phases such as Mitigation, Preparedness, Response, and Recovery.

Figure 6.5 shows the final design of the ontology by using TF.IDEFEC-based.Enhanced mapping results. The alignment helps in making the gold ontology

richer and more expressive in delivering information. As an example, an interesting finding is identified with the inclusion of coordinate and geographic information for geo-tagging. The association between these terms indicate that geo-tagging features enable geographic coordinates in social media application. The concept from the ontology also conveys through geo-tagging feature in social media enable user to share geographic information such as image, latitude, longitude, and name of places with other users. This information will be beneficial for the users especially the affected community in getting frequent updates and geographic information regarding the disaster areas.

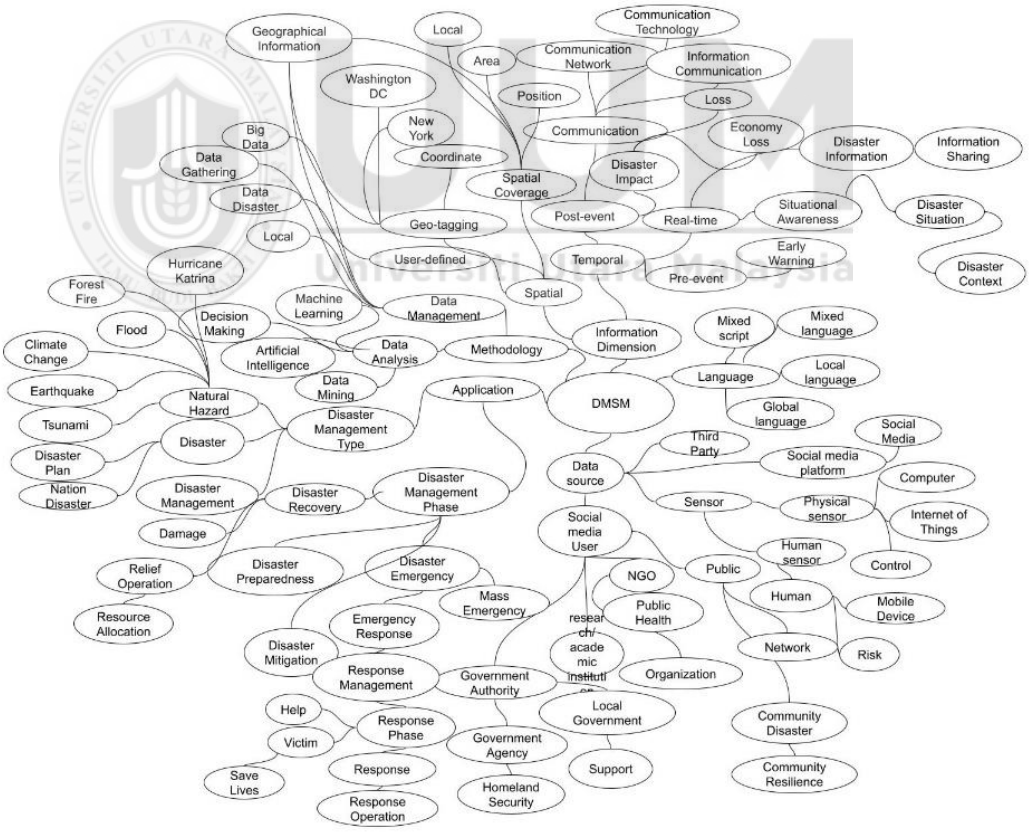

Figure 6.5. Ontology developed from TF.IDFEC-based.Enhanced results.

Table 6.8

Precision and recall for mapped learned and gold-standard ontologies for TF.IDFEC-based.Enhanced supervised TWS.

Gold-standard Based Taxonomy of DMSM Concepts (Phengsuwan et al., 2021)				Learned Ontology Concepts			Total Learned Extracted Terms	Total Gold Extracted Terms	Total Mapped Terms
Data Source (1)	Social Media User (1)	Government Authority (1)	Local Government (1) Support (1) - Human Resource (1) Government Agency (1) - Homeland Security (1)	Social Media User (1)	Government Authority	Local Government (1) Support (1) Government Agency (1) - Homeland Security (1)	90	26	19
		Research/Academic Institution (1)			Research/Academic Institution				
		NGO (1)	Organization (1)		NGO (1)	Organization (1)			
		Public (1)	Community Disaster (1) - Community Resilience (1) Network (1)		Public (1)	Community Disaster (1) - Community Resilience (1) Network (1)			
	Sensor (1)	Human Sensor (1)	Human (1) - Risk (1) - Mobile Phone (1)	Sensor	Human Sensor	Human (1) - Risk (1) - Mobile Phone (1)			
		Physical Sensor (1)	Internet of Things (1) Computer (1) Control (1)		Physical Sensor	Internet of Things (1) Computer (1) Control (1)			
	Social Media Platform (1)	Social Media (1)		Social Media Platform	Social Media (1)				
	Third Party (1)			Third Party					

Universiti Utara Malaysia

Language (1)	Mixed Script (1)			Language	Mixed Script			90	5	1
	Mixed Language (1)				Mixed Language					
	Local Language (1)				Local Language (1)					
	Global Language (1)				Global Language					
Information Dimension (1)	Temporal (1)	Post-event (1)	Communication (1) - Communication Network (1) - Communication Technology (1) - Information Communication (1) Disaster Impact (1) - Loss (1) - Economy Loss (1)	Information Dimension (1)	Temporal	Post-event (1)	Communication (1) - Communication Network (1) - Communication Technology (1) - Information Communication (1) Disaster Impact (1) - Loss (1) - Economy Loss (1)	90	32	26
		Real-time (1)	Situational Awareness (1) - Disaster Situation (1) - Disaster Context (1) Information (1) - Disaster Information (1) - Information Sharing (1)			Real-time (1)	Situational Awareness (1) - Disaster Situation (1) - Disaster Context (1) Information (1) - Disaster Information (1) - Information Sharing (1)			
		Pre-event (1)	Early warning (1)			Pre-event	Early warning (1)			
	Spatial (1)	Spatial Coverage (1)	Area (1) Geographical Information (1) Local (1) Position (1)		Spatial	Spatial Coverage	Area (1) Geographical Information (1) Local (1) Position (1)			
		User-defined (1)	Geographical Information (1)			User-defined	Geographical Information (1)			

Universiti Utara Malaysia

	Geo-tagging (1)	Coordinate (1) Geographical Information (1) New York (1) Washinton DC (1)		Geo-tagging	Coordinate (1) Geographical Information (1) New York (1) Washinton DC (1)			
Methodology (1)	Data Management (1)	Big Data (1) Data Gathering (1) Data Disaster (1) Local (1)	Methodology	Data Management (1)	Big Data (1) Data Gathering (1) Data Disaster (1) Local (1)	90	11	10
	Data Analysis (1)	Artificial Intelligence (1) Machine Learning (1) Data Mining (1) Decision Making (1)		Data Analysis (1)	Artificial Intelligence (1) Machine Learning (1) Data Mining (1) Decision Making (1)			
Application (1)	Disaster Management Type (1)	Natural Hazard (1) - Climate Change (1) - Earthquake (1) - Tsunami (1) - Hurricane Katrina (1) - Forest Fire (1) - Flood (1) Disaster (1) - Disaster Plan (1) - Nation Disaster (1) Disaster Management (1)	Application (1)	Disaster Management Type	Natural Hazard (1) - Climate Change (1) - Earthquake (1) - Tsunami (1) - Hurricane Katrina (1) - Forest Fire (1) - Flood (1) Disaster (1) - Disaster Plan (1) - Nation Disaster (1) Disaster Management (1)	90	30	28
	Disaster Management Phase (1)	Disaster Recovery (1) - Damage (1) - Relief Operation (1) -- Resource Allocation (1) Disaster Emergency (1) - Mass Emergency (1) - Emergency Response (1) -- Response Management (1) --- Response Phase (1) ---- Victim (1) ----- Save lives (1)		Disaster Management Phase	Disaster Recovery (1) - Damage (1) - Relief Operation (1) -- Resource Allocation (1) Disaster Emergency (1) - Mass Emergency (1) - Emergency Response (1) -- Response Management (1) --- Response Phase (1) ---- Victim (1) ----- Save lives (1) ---- Response (1) ----- Response Operation (1) ---- Help (1) Disaster Mitigation (1) Disaster Preparedness (1)			

Universiti Utara Malaysia

	---- Response (1) ----- Response Operation (1) ---- Help (1) Disaster Mitigation (1) Disaster Preparedness (1)			
		Total	**104**	**84**
			Precision	**93.33%**
			Recall	**80.80%**
			F-measure	**0.87**

Universiti Utara Malaysia

The precision, recall, and F-measure for the developed ontology based on the TF.IDFEC-based.Enhanced scheme is calculated and presented in Table 6.8. The precision for the overall mapped terms, $C_G \cup C_L$ / C_L obtained 93.33% with 84 terms are mapped from 90 total terms between the learned and gold-standard concepts. The calculated value expresses that the term extracted by using this scheme is relevant to the DMSM domain with a high precision rate. The recall, $C_G \cup C_L$ / C_G also has achieved 80.80% from all of the terms as seen in Table 6.9, which means that 84 terms are correctly identified and mapped to the 104 terms of the gold-standard concepts.

From these two measures, both precision and recall values, F-measure is calculated with the achieved rate of 0.87, whereby the F-measure value falls under the range of 0 to 1. The F-measure rate is closed to the value 1 indicating that the precision and recall obtained a good outcome. The total mapped terms, precision, recall, and F-measure value for the existing supervised TWS, TF.RF, is presented and calculated in the next section.

Universiti Utara Malaysia

6.3.3 DMSM Ontology Development based on TF.RF

In this evaluation, TF.RF scheme extracted terms are selected for the DMSM ontology development as a comparison between the existing and enhanced RF scheme based on the proposed formula in this study. As shown in Table 6.9, for Data Source segment, several terms are introduced from this scheme's extracted terms such as 'Support', 'Organization', 'Community', 'Network', 'Risk', 'Computer', and 'Social Media', including mapped terms such as 'Government Authority' and 'Public' terms, similar with TF.RF.Enhanced and TF.IDFEC-based.Enhanced schemes. The Language

segment contains one mapped term with the learned ontology which is ‘Local Language’ similar with previous evaluated schemes.

New terms are seen added to the Information Dimension, which are ‘Communication’, ‘Information’, and ‘Area’ with mapped terms which are ‘Post-event’ and ‘Real-time’. The Methodology segment is observed did not include any new terms from the mapping, but several new terms are added to the Application segment, mainly on ‘Disaster Management Type’ and ‘Disaster Management Phase’.

The evaluation results of this scheme are measured by the precision, recall, and F-measure rate shown in Table 6.9. The precision for the TF.RF supervised term weighting scheme, $C_G \cup C_L$ / C_L has achieved 65% while the recall, $C_G \cup C_L$ / C_G has achieved 50% rate. The F-measure has also been computed with the obtained rate of 0.57. The outcome is found to be lower as opposed to TF.RF.Enhanced and TF.IDFEC-based.Enhanced, giving the meaning that from these schemes, the extracted terms are less relevant to the DMSM domain for the ontology development. The next section presents the evaluation performed on the TF.IDEFEC-based supervised TWS in order to measure the precision, recall, and F-measure rate.

Universiti Utara Malaysia

Table 6.9

Precision and recall for mapped learned and gold-standard ontologies for TF.RF supervised TWS.

Gold-standard Based Taxonomy of DMSM Concepts (Phengsuwan et al., 2021)				Learned Ontology Concepts				Total Learned Extracted Terms	Total Gold Extracted Terms	Total Mapped Terms
Data Source (1)	Social Media User (1)	Government Authority (1)	Support (1)	Data Source	Social Media User	Government Authority (1)	Support (1)	40	18	9
		Research/Academic Institution (1)				Research/Academic Institution				
		NGO (1)	Organization (1)			NGO	Organization (1)			
		Public (1)	Community (1) Network (1)			Public (1)	Community (1) Network (1)			
	Sensor (1)	Human Sensor (1)	Risk (1)		Sensor	Human Sensor	Risk (1)			
		Physical Sensor (1)	Computer (1)			Physical Sensor	Computer (1)			
	Social Media Platform (1)	Social media (1)			Social Media Platform	Social Media (1)				
	Third Party (1)				Third Party					
Language (1)	Mixed Script (1)			Language	Mixed Script			40	5	1
	Mixed Language (1)				Mixed Language					
	Local Language (1)				Local Language (1)					
	Global Language (1)				Global Language					

Universiti Utara Malaysia

Information Dimension (1)	Temporal (1)	Post-event (1)	Communication (1)	Information Dimension	Temporal	Post-event (1)	Communication (1)			
		Real-time (1)	Information (1)			Real-time (1)	Information (1)	40	12	5
		Pre-event (1)				Pre-event				
	Spatial (1)	Spatial Coverage (1)	Area (1)		Spatial	Spatial Coverage	Area (1)			
		User-defined (1)				User-defined				
		Geo-tagging (1)				Geo-tagging				
Methodology (1)	Data Management (1)			Methodology	Data Management			40	3	0
	Data Analysis (1)				Data Analysis					
Application (1)	Disaster Management Type (1)		Disaster (1) - Nation disaster (1) - Plan (1) Natural hazard (1) - Earthquake (1) - Flood (1) Disaster Management (1)	Application (1)	Disaster Management Type (1)		Disaster (1) - Nation disaster (1) - Plan (1) Natural hazard (1) - Earthquake (1) - Flood (1) Disaster Management (1)	40	14	11
	Disaster Management Phase (1)		Disaster Preparedness (1) Disaster Recovery (1) - Relief (1) Disaster Response (1)		Disaster Management Phase (1)		Disaster Preparedness (1) Disaster Recovery (1) - Relief (1) Disaster Response (1)			
								Total	**52**	**26**
									Precision	**65.0%**
									Recall	**50.0%**
									F-measure	**0.57**

Universiti Utara Malaysia

6.3.3 DMSM Ontology Development based on TF.IDFEC-based

In Table 6.10, the mapping between learned and the gold-standard extracted terms by using the TF.IDFEC-based supervised TWS is evaluated. Overall the outcome is identical to the TF.RF results for the mapping with slight difference total number of extracted terms, whereby TF.RF consists of 40 while TF.IDFEC-based consists of 39 terms.

However, the precision, $C_G \cup C_L / C_L$ and recall, $C_G \cup C_L / C_G$ for this scheme obtained 66.67% and 50% rate, respectively. The F-measure is computed based on these values, which achieved 0.57 rate, expressing a low mapping result for the extracted terms. This outcome signifies that the TF.IDFEC-based supervised TWS is not suitable in performing text mining process for binary classification ontology development, similar with the TF.RF scheme. The next section exhibits the evaluation of TF-IDF extracted terms for ontology development.

Universiti Utara Malaysia

Table 6.10

Precision and recall for mapped learned and gold-standard ontologies for TF.IDFEC-based supervised TWS.

Gold-standard Based Taxonomy of DMSM Concepts (Phengsuwan et al., 2021)				Learned Ontology Concepts				Total Learned Extracted Terms	Total Gold Extracted Terms	Total Mapped Terms
Data Source (1)	Social Media User (1)	Government Authority (1)	Support (1)	Data Source	Social Media User	Government Authority (1)	Support (1)	39	18	9
		Research/Academic Institution (1)				Research/Academic Institution				
		NGO (1)	Organization (1)			NGO	Organization (1)			
		Public (1)	Community (1) Network (1)			Public (1)	Community (1) Network (1)			
	Sensor (1)	Human Sensor (1)	Risk (1)		Sensor	Human Sensor	Risk (1)			
		Physical Sensor (1)	Computer (1)			Physical Sensor	Computer (1)			
	Social Media Platform (1)	Social media (1)			Social Media Platform	Social Media (1)				
	Third Party (1)				Third Party					
Language (1)	Mixed Script (1)			Language	Mixed Script			39	5	1
	Mixed Language (1)				Mixed Language					
	Local Language (1)				Local Language (1)					
	Global Language (1)				Global Language					
Information Dimension (1)	Temporal (1)	Post-event (1)	Communication (1)	Information Dimension	Temporal	Post-event (1)	Communication (1)	39	12	5
		Real-time (1)	Information (1)			Real-time (1)	Information (1)			
		Pre-event (1)				Pre-event				
	Spatial (1)	Spatial Coverage (1)			Spatial	Spatial Coverage	Area (1)			

Universiti Utara Malaysia

	User-defined (1) Geo-tagging (1)			User-defined Geo-tagging				
Methodology (1)	Data Management (1) Data Analysis (1)		Methodology	Data Management Data Analysis		39	3	0
Application (1)	Disaster Management Type (1)	Disaster (1) - Nation disaster (1) - Plan (1) Natural hazard (1) - Earthquake (1) - Flood (1) Disaster Management (1)	Application	Disaster Management Type	Disaster (1) - Nation disaster (1) - Plan (1) Natural hazard (1) - Earthquake (1) - Flood (1) Disaster Management (1)	39	14	11
	Disaster Management Phase (1)	Disaster Preparedness (1) Disaster Recovery (1) - Relief (1) Disaster Response (1)		Disaster Management Phase	Disaster Preparedness (1) Disaster Recovery (1) - Relief (1) Disaster Response (1)			
						Total	52	26
							Precision	66.67%
							Recall	50.0%
							F-measure	0.57

Universiti Utara Malaysia

6.3.4 DMSM Ontology Development based on TF-IDF

The final evaluation is conducted as an assessment of the benchmark scheme, TF-IDF unsupervised TWS. As explained in previous chapters, TF-IDF ignores classification information of the document distribution, as compared to the supervised TWS. Therefore, it is necessary to perform this evaluation in order to measure the impact of the results based on the term weighting process performed through ontology development for the DMSM domain.

Table 6.11 presents the outcome of the evaluation with five main segments from the gold-standard extracted terms, which are Data Source, Language, Information Dimension, Methodology, and Application. In the Data Source segment, two concepts are mapped to the gold-standard terms from the learned ontology, which are Sensor and Government Authority. Several terms are added to this segment from the learned extracted terms which are 'Organization', 'Community', 'Computer', 'Internet of Things', 'Radar', and 'Wireless'.

Universiti Utara Malaysia

Information Dimension segment obtained only one mapping which is 'Spatial', while application includes new terms from the mapped learned extracted terms for the Disaster Management Phase segment. No mapping is found for the Methodology segment, while only one mapping is found for the Language segment, which is 'Local Language'.

Table 6.11

Precision and recall for mapped learned and gold-standard ontologies for TF.IDF supervised TWS.

Gold-standard Based Taxonomy of DMSM Concepts (Phengsuwan et al., 2021)				Learned Ontology Concepts				Total Learned Extracted Terms	Total Gold Extracted Terms	Total Mapped Terms
Data Source (1)	Social Media User (1)	Government Authority (1)		Data Source	Social Media User	Government Authority (1)		81	18	9
		Research/Academic Institution (1)				Research/Academic Institution				
		NGO (1)	Organization (1)			NGO	Organization (1)			
		Public (1)	Community (1)			Public	Community (1)			
	Sensor (1)	Human Sensor (1)			Sensor (1)	Human Sensor				
		Physical Sensor (1)	Computer (1) Internet of things (1) Radar (1) Wireless (1)			Physical Sensor	Computer (1) Internet of things (1) Radar (1) Wireless (1)			
	Social Media Platform (1)	Social media (1)			Social Media Platform	Social media (1)				
	Third Party (1)				Third Party					
Language (1)	Mixed Script (1)			Language	Mixed Script			81	5	1
	Mixed Language (1)				Mixed Language					
	Local Language (1)				Local Language (1)					
	Global Language (1)				Global Language					
	Temporal (1)	Post-event (1)	Communication (1)	Information Dimension	Temporal	Post-event	Communication	81	9	1

Universiti Utara Malaysia

Information Dimension (1)		Real-time (1)	Information (1)			Real-time	Information			
		Pre-event (1)				Pre-event				
	Spatial (1)	Spatial Coverage (1)			Spatial (1)	Spatial Coverage				
		User-defined (1)				User-defined				
		Geo-tagging (1)				Geo-tagging				
Methodology (1)	Data Management (1)			Methodology	Data Management			81	3	0
	Data Analysis (1)				Data Analysis					
Application (1)	Disaster Management Type (1)			Application	Disaster Management Type					
	Disaster Management Phase (1)		Disaster Relief (1) Disaster Preparedness (1) Disaster Resilience (1)		Disaster Management Phase		Disaster Relief (1) Disaster Preparedness (1) Disaster Resilience (1)	81	6	3
								Total	41	14
									Precision	17.28%
									Recall	34.15%
									F-measure	0.23

Universiti Utara Malaysia

Table 6.11 also presents the calculated precision, recall, and F-measure values for the TF-IDF extracted terms. The precision for this evaluation, $C_G \cup C_L$ / C_L, is achieved with 17.28% and 34.15% rate for recall, $C_G \cup C_L$ / C_G. The values obtained are the lowest compared to other supervised TWS evaluated in this section, TF.RF.Enhanced, TF.IDFEC-based.Enhanced, TF.RF, and TF.IDFEC-based.The F-measure is also calculated with the obtained rate of 0.23, whereby also considered as the lowest compared to the other schemes. The summary of all of the results is presented in the next section.

6.3.5 Results Comparison between All Evaluations

The comparison between all evaluations presented in the previous section is presented in Table 6.12. In Table 6.12, the evaluation conducted by using five types of TWS are listed, whereby the supervised TWS are TF.RF.Enhanced, TF.IDFEC-based.Enhanced, TF.RF, TF.IDFEC-based, and an unsupervised TWS, TF-IDF. In the evaluation, the precision, recall, and F-measure rate are calculated for each scheme. The precision is displayed visually in Figure 6.6 and the recall is displayed in Figure 6.7 for comparison.

Universiti Utara Malaysia

Table 6.12

Summary of the evaluation results

Term Weighting Schemes	Type of Weighting Scheme	Precision (%)	Recall (%)	F-measure
TF.RF.Enhanced	Supervised	91.03	78.02	0.84
TF.IDFEC-based.Enhanced	Supervised	93.33	80.80	0.87
TF.RF	Supervised	65.00	50.00	0.57
TF.IDFEC-based	Supervised	66.67	50.00	0.57
TF.IDF	Unsupervised	17.28	34.15	0.23

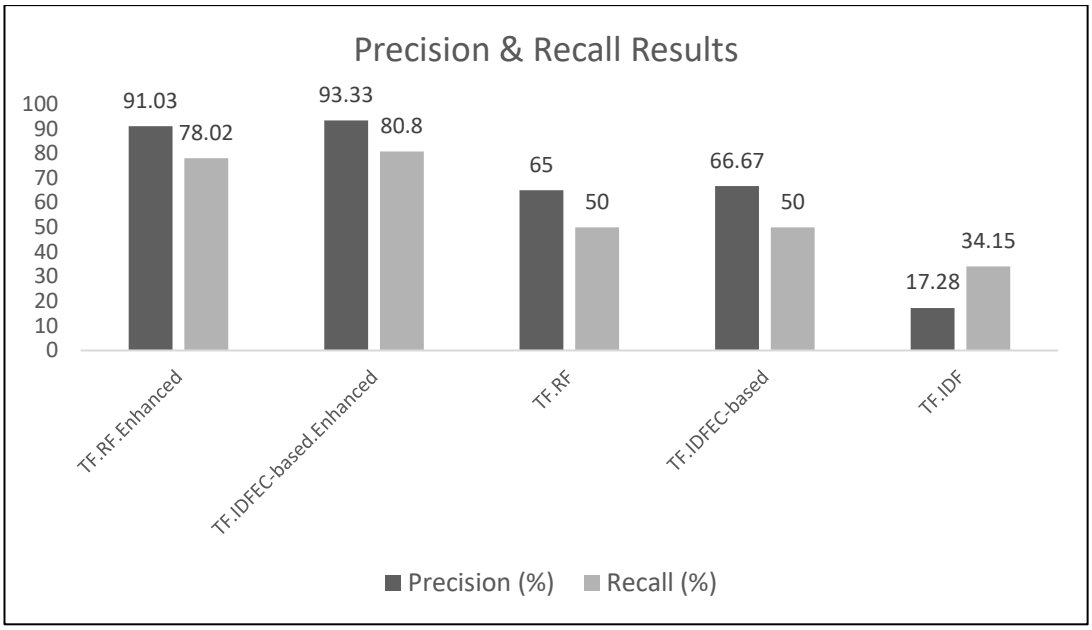

Figure 6.6. Summary of the evaluation precision & recall results

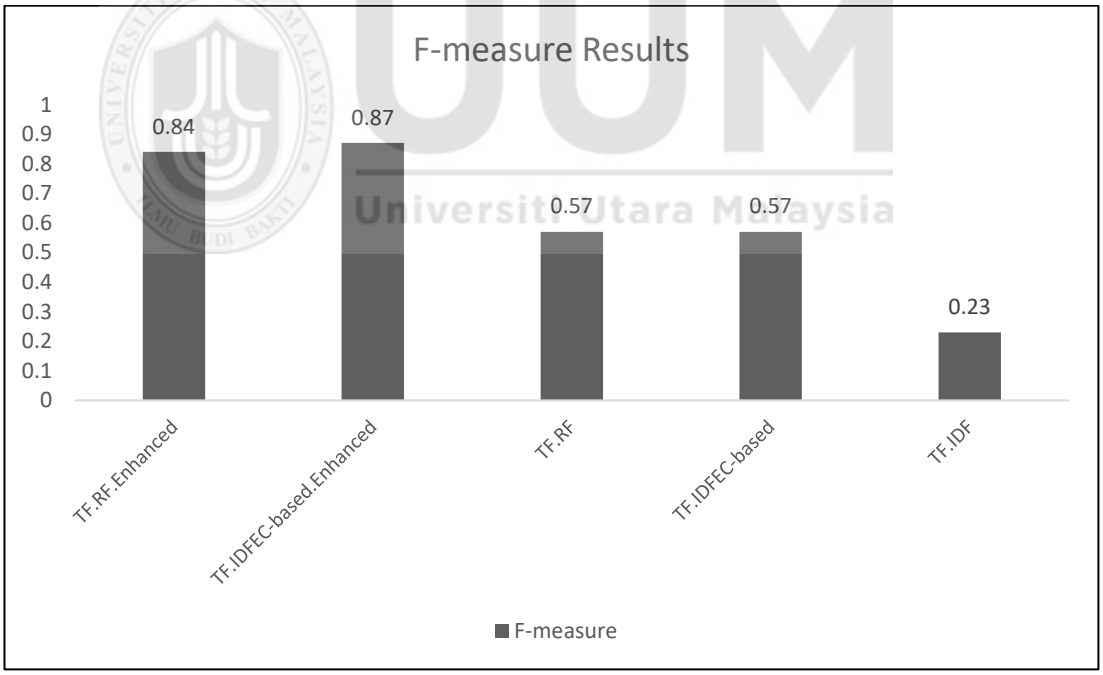

Figure 6.7. Summary of the evaluation F-measure results

In Figure 6.6, the precision and recall for each evaluation result are shown with the highest precision is achieved by TF.IDFEC-based.Enhanced, followed by

TF.RF.Enhanced, TF.IDFEC-based, TF.RF, and finally TF-IDF. Following the results, since TF.IDFEC-based.Enhanced and TF.RF.Enhanced obtained the highest precision, 93.33% and 91.03%, respectively, it is proven that the introduced enhanced supervised TWS in this study shows a good result, by extracting the most relevant results for DMSM ontology, compared to its predecessor, TF.RF and TF.IDFEC-based. TF.RF and TF.IDFEC-based TWS acquire poor precision, with 66.67% and 65%, respectively. The lowest precision rate is obtained by TF-IDF, with 17.28%, expressing that unsupervised TWS performed poorly in carrying out text mining to ontology development operation with binary classification.

A similar result is obtained in Figure 6.6 for recall, whereby TF.IDFEC-based.Enhanced and TF.RF achieved 80.8% and 78.02% as the highest recall rate. TF.RF and TF.IDFEC-based obtained both 50% and TF-IDF obtained the lowest result with a 34.15% recall rate. The results demonstrated that extracted terms by TF.IDFEC-based.Enhanced and TF.RF supervised TWS have correctly identified the terms for DMSM ontology while TF.IDFEC-based, TF.RF, and TF-IDF performed poorly on this.

Universiti Utara Malaysia

Figure 6.7 displays the F-measure results for all evaluations, in order to combine both precision and recall into a single measure that captures both properties. The lowest score for F-measure is zero while the highest will be one. The result observed that TF.IDFEC-based.Enhanced and TF.RF.Enhanced acquired the highest F-measure score, 0.87 and 0.84, respectively, while, TF.IDFEC-based and TF.RF both obtains 0.57, and the lowest score is TF-IDF with 0.23 score.

To sum up, by using the enhanced supervised TWS introduced in this study, higher weightage is obtained for positive category compared to the existing scheme. Therefore extracted terms with higher precision is obtained for the DMSM ontology

development. The final results also implies that supervised TWS deliver a better performance in term of precision, recall, and F-measure compared to unsupervised TWS.

6.4 Summary

This chapter presents the evaluation of the constructed enhanced supervised TWS. The verification stage was conducted through the gold-standard-based evaluation method by calculating the distance for syntactic comparison, designing the ontology, and evaluating the learned ontology through the gold-standard-based evaluation method. From the results, it is found that the precision, recall, and F-measure obtained a good score for TF.IDFEC-based.Enhanced and TF.RF.Enhanced with 93.33% and 91.03% for precision, 80.8% and 78.02% for recall, and 0.87 and 0.84 for F-measure, respectively.

Universiti Utara Malaysia

CHAPTER 7

CONCLUSION AND FUTURE WORKS

7.1 Introduction

The research accomplishments achieved though the experimental work and results analysis are outlined in this final chapter, as well as future directions especially for the present study's field, text mining. Identified contribution for this research are sums up and explained in Section 7.2 while Section 7.3 presents respective suggestions and recommendations relevant for this study future works

7.2 Study Recapitulation

Three objectives are defined for this study, particularly to identify existing supervised TWS for ontology development, to construct an enhanced supervised TWS for DMSM ontology development based on binary classification, and to evaluate the enhanced supervised TWS through the developed ontology. The research is summarised in reference to these objectives respectively.

7.2.1 Objective 1: To Identify Existing Supervised Term Weighting Schemes for Ontology Development

The first objective was answered on account of the experimental work as described in detail in Chapter Four. In order to address this objective, the existing supervised TWS are identified. The text mining techniques are utilised with knowledge base consolidated from five digital libraries to achieve higher reliability, which are ACM DL, Google Scholar, Science Direct, Scopus, and Wiley Online Library. The document

weightage scheme for the text mining process is utilising several existing supervised TWS to express how crucial the terminology is to a document in a corpus. The selected schemes which are x^2, IG, OR, GR, RF, MI, IDFEC-based, and SW are evaluated in this study. The criteria in selecting the schemes is based on the reviews by Maisonnave, Delbianco, Tohmé, & Maguitman (2019) and Mazyad, Teytaud, & Fonlupt (2018), whereby these schemes are the well-established supervised TWS and most commonly used in experimental works for evaluation and comparison. Maisonnave, Delbianco, Tohmé, & Maguitman (2019 and Mazyad, Teytaud, & Fonlupt (2018) also stated that among the latest supervised TWS developed, IDFEC-based and SW has received positive results and usually evaluated as benchmark comparison among other supervised TWS (Maisonnave, Delbianco, Tohmé, & Maguitman, 2019; Mazyad, Teytaud, & Fonlupt, 2018).

The outcome from the study revealed that the existing supervised TWS did not achieve the main objective of this study, in order to retrieve the concepts that fully represent the DMSM domain. This is because all of the TWS present results that emphasized more on the negative category. Based on this observation, the existing TWS must be further improved to answer the objective of this research.

Universiti Utara Malaysia

7.2.2 Objective 2: To Construct an Enhanced Supervised Term Weighting Scheme for Disaster Management Social Media Ontology Development based on Binary Classification

This study has successfully constructed an enhanced supervised TWS for DMSM ontology development based on binary classification as discussed in Chapter Five. The experimental results showed that the existing supervised TWS obtained ambiguous performance in terms of various categories of distribution data sets and different

learning algorithms. Most of the schemes which are based on Chi-square (x^2), Information Gain (IG), Odds Ratio (OR), Gain Ratio (GR), Relevance Frequency (RF), Mutual Information (MI), Inverse Document Frequency Excluding Category (IDFEC-based), and Short-text Term Weighting (SW) have not shown satisfactory performance since irrelevant results are extracted from term weightages. Therefore, this experiment aims to construct an enhanced supervised TWS for DMSM ontology development based on binary classification.

The enhanced supervised TWS are constructed based on the experiments conducted to identify the existing schemes, from the limitations specified. A new equation has been introduced by incorporating a new term frequency based on category information and positive category ratio, factored with the existing supervised TWS. By using this equation, another experiment has been conducted to compute the term weightages. At the end of the process, the new term weightages are analysed with the existing supervised TWS results. The outcome demonstrated that the enhanced formula generates improved results compared to existing schemes in which higher weightage is obtained for positive classification, instead of negative class. The detail explanation on the constructed enhancement is elaborated in Chapter Five.

Universiti Utara Malaysia

7.2.3 Objective 3: To Evaluate the Enhanced Supervised Term Weighting Schemes through the Developed Ontology

The final objective was answered by conducting the evaluation through the gold-standard-based evaluation method. A referenced ontology has been selected and extracted all of the ontology concepts for mapping evaluation. Learning ontology is generated by using the new text-mining results and mapped between the learned and referenced ontology.

The completion of the evaluation revealed that the enhanced TWS, TF.IDFEC-based.Enhanced and TF.RF.Enhanced has been confirmed to successfully obtain high precision of 93.33% and 91.03%, 80.8% and 78.02% for recall, and 0.87 and 0.84 for F-measure, respectively. Chapter Six elaborated in detail the assessment to fulfil the final objective of this thesis.

7.3 Contributions

The contributions for this study are highlighted into two main categories which are text mining to ontology development for binary classification and DMSM ontology development.

7.3.1 Text Mining to Ontology Development for Binary Classification

Current text mining to ontology development is utilising unsupervised TWS, TF-IDF in order to carry out the operation. It is known that TF-IDF experiences the main limitation, which ignores the classification information of a corpus document distribution. By emphasizing the classification information of a corpus, features dimensionality can be reduced and boosts the importance of words that are distributed between the positive and the negative class.

To reduce the dimensionality of a corpus, this study applies binary classification to distribute the corpus into positive and negative categories. However, the current technique for text mining operation must be enhanced to use supervised instead of unsupervised TWS. In this study, eight supervised TWS are selected, in order to identify which are the best schemes in order to perform the text mining process for binary classification.

A notable contribution is achieved to the text mining field, especially for binary classification data set, by performing analysis on the eight selected supervised TWS, Chi-square (x^2), Information Gain (IG), Odds Ratio (OR), Gain Ratio (GR), Relevance Frequency (RF), Mutual Information (MI), Inverse Document Frequency Excluding Category (IDFEC-based), and Short-text Term Weighting (SW). The performance for each of the schemes is analysed and discussed in the experimental work of this study.

According to the results of the analysis, this research introduced enhancement on the existing scheme, in order to improve the outcome of the experimental work. This is another significant contribution of this study, by introducing integration of positive term frequency ratio and positive category ratio to the existing supervised TWS. The experimental work has shown an improvement to the existing schemes, especially for RF and IDFEC-based, whereby the schemes obtained higher weightage terms on positive category while others on negative category, in acquiring higher appropriate terms for the DMSM ontology development. Based on both schemes the extracted terms are utilised in the next step of the experimental work in which to develop DMSM ontology.

Universiti Utara Malaysia

7.3.2 DMSM Ontology Development

Based on the previous step, the supervised TWS have been analysed for its limitations and enhanced. The enhanced equation is further evaluated to ensure the reliability of the equation by using the gold-standard-based evaluation method.

The evaluation process provides another important study's contribution, in which a method to evaluate the generated learned ontology is detailed out to be assessed with existing the gold-standard-based representation, in the aspect of precision, recall, and F-measure. Five evaluations are carried out, whereby evaluation of the enhanced

supervised TWS, TF.RF.Enhanced and TF.IDFEC-based.Enhanced are performed, together with the existing schemes, TF.RF and TF.IDFEC-based. An evaluation of the unsupervised TWS is also done as a comparison with the TF-IDF scheme.

Benchmark results are presented at the end of the evaluation, with the enhanced supervised TWS, TF.IDFEC-based.Enhanced and TF.RF.Enhanced achieved a precision of 93.33% and 91.03%, 80.8% and 78.02% for recall, and 0.87 and 0.84 for F-measure. This work provides the facilitation of the text mining process for ontology development by using binary classification. To the ontology engineers, a text mining process has properly exhibited a comprehensive evaluation process that unifies both the learned and gold-standard based ontology.

7.4 Limitations and Future Works

The limitations and future works identified is summarised in this section. Several recommendations are identified and listed below:

i) The concepts or terminologies extraction in this research is accomplished through the tokenization technique. Alternatively, several different text mining techniques has the potential to be employed, namely rule application and selection. Rule applications verify the state of the chosen rules and fulfil the form based on the findings of the corresponding rules. This is useful in reducing the involvement of human or manual effort in extracting and verifying the identified terms from the primary source. Besides that, rule selection refers to extraction rules which are using triggers for instance keywords. The terms is checked to validate the triggering items and ensuring the rules assigned.

ii) Manual mapping process is required to perform the mapping process of the learned to gold-standard ontology. Consequently, this process is time and

resource consuming to be completed. Possibility of errors are also high that might cause generating an incorrect ontology. Improvement can be explored to seek the ideal approach which allows automated transformation, eliminating the necessity of mapping the ontology manually.

iii) The research scope if limited to the concept stage of the ontology development. Thereby, identifying relationship level in mapping the generated ontology can be further explored. Accordingly, valuable, and significant outcome can be accomplished to semantically analyse the DMSM domain.

7.5 Conclusions

The research mainly contributes to the constructed enhanced supervised TWS for DMSM ontology development based on binary classification. The enhancement to done since the existing supervised TWS experienced several limitations which emphasize on negative category information. Therefore, the existing schemes are further improved to increase the weightages on the positive document distribution.

To enhance the existing supervised schemes, the enhanced supervised scheme is constructed by considering the category information as the term frequency factor and introducing a positive category ratio to emphasize the positive classes. From the text mining outcome, the equation is evaluated by developing a DMSM ontology as an evaluation, whereby positive results are gained for TF.IDFEC-based.Enhanced and TF.RF.Enhanced. The results show high precision rates of 93.33% and 91.03%, 80.8% and 78.02% for recall, and 0.87 and 0.84 for F-measure. Thus, this enhanced technique can also be used in other domains for ontology development with binary classification problem.

Universiti Utara Malaysia

REFERENCES

Abdalla, R., & Esmail, M. (2018). *Basic Concept of Disaster Management and Emergency Response*. 11–22.

Abualigah, L. M., Khader, A. T., Al-Betar, M. A., & Alomari, O. A. (2017). Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. *Expert Systems with Applications*, *84*, 24–36.

Ahmad, T., Iqbal, J., Ashraf, A., Truscan, D., & Porres, I. (2019). Model-based testing using UML activity diagrams: A systematic mapping study. *Computer Science Review*, *33*, 98–112.

Ajami, H., & Mcheick, H. (2018). *Ontology-Based Model to Support Ubiquitous Healthcare Systems for COPD Patients*. https://doi.org/10.3390/electronics7120371

Al-janabi, M. I., Qutqut, M. H., & Hijjawi, M. (2018). *Machine Learning Classification Techniques for Heart Disease Prediction : A Machine Learning Classification Techniques for Heart Disease Prediction : A Review*. *October*. https://doi.org/10.14419/ijet.v7i4.28646

Albitar, S., Fournier, S., & Espinasse, B. (2014). An effective TF/IDF-based text-to-text semantic similarity measure for text classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8786*, 105–114.

Aldrraji, M. (2022). *Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets*. *August*, 375–383. https://doi.org/10.11591/ijeecs.v28.i1.pp375-383

Alexander, D. (2016). *Principles of Emergency Planning and Management Reviewed by Paul Shrivastava*. *December*. https://doi.org/10.1057/palgrave.rm.8240152

Alhaj, T. A., Siraj, M. M., Zainal, A., Elshoush, H. T., & Elhaj, F. (2016). Feature selection using information gain for improved structural-based alert correlation. *PLoS ONE*, *11*(11), 1–18.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*.

Allaire, M. C. (2016). *Disaster loss and social media: Can online information increase flood resilience?* 7408–7423.

Alobaidi, M., Malik, K. M., & Sabra, S. (2018). Linked open data-based framework for automatic biomedical ontology generation. *BMC Bioinformatics*, *19*(1), 319.

Alsaeedi, A. (2020). A survey of term weighting schemes for text Classification. *International Journal of Data Mining, Modelling and Management*, *12*(2), 237–254.

Alshaer, H. N., Otair, M. A., Abualigah, L., Alshinwan, M., & Khasawneh, A. M. (2021). Feature selection method using improved CHI Square on Arabic text classifiers: analysis and application. *Multimedia Tools and Applications*, *80*(7), 10373–10390.

Alsmadi, I., & Hoon, G. K. (2019). Term weighting scheme for short-text classification: Twitter corpuses. *Neural Computing and Applications*, *31*(8), 3819–3831.

Amith, M., He, Z., Bian, J., Lossio-Ventura, J. A., & Tao, C. (2018). Assessing the practice of biomedical ontology evaluation: Gaps and opportunities. *Journal of*

Biomedical Informatics, *80*(February), 1–13.

Andreasen, M. M., Wognum, N., & McAloone, T. (2002). Design Typology and Design Organisation. *DESIGN 2002 - 7th International Design Conference*, 1–6.

Aphinyanaphongs, Y., Fu, L. D., Li, Z., Peskin, E. R., Efstathiadis, E., Aliferis, C. F., & Statnikov, A. (2014). A Comprehensive Empirical Comparison of Modern Supervised Classification and Feature Selection Methods for Text Categorization. *Journal of the Association for Information Science and Technology*.

Artur, M. (2021). ScienceDirect ScienceDirect Review the performance of the Bernoulli Na ï ve Bayes Classifier in Intrusion Detection Systems using Recursive Feature Elimination with Cross-validated selection of the best number of features. *Procedia Computer Science*, *190*(2019), 564–570. https://doi.org/10.1016/j.procs.2021.06.066

Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., & Abbasi, H. M. (2018a). A survey of ontology learning techniques and applications. *Database*, *2018*(2018), 1–24. https://doi.org/10.1093/database/bay101

Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., & Abbasi, H. M. (2018b). A survey of ontology learning techniques and applications. *Database*, *2018*(2018), 1–24.

Attigeri, G., Manohara Pai, M. M., Pai, R. M., & Kulkarni, R. (2018). Knowledge Base Ontology Building for Fraud Detection Using Topic Modeling. *Procedia Computer Science*, *135*, 369–376. https://doi.org/10.1016/j.procs.2018.08.186

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135–160.

Benabdallah, A., Abderrahim, M. A. E., & Abderrahim, M. E. A. (2017). Extraction of terms and semantic relationships from Arabic texts for automatic construction of an ontology. *International Journal of Speech Technology*, *20*(2), 289–296.

Benaben, F., Lauras, M., Truptil, S., & Salatge, N. (2016). A metamodel for knowledge management in crisis management. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2016-March*, 126–135. https://doi.org/10.1109/HICSS.2016.24

Berquand, A., Moshfeghi, Y., & Riccardi, A. (2020). Space mission design ontology: Extraction of domain-specific entities and concepts similarity analysis. *AIAA Scitech 2020 Forum*, *1*, 2253.

Biagetti, M. T. (2020). Ontologies (as knowledge organization systems). In *ISKO Encyclopedia of Knowledge Organization*. https://www.researchgate.net/publication/344070697

Blanco, V., Japón, A., & Puerto, J. (2020). Optimal arrangements of hyperplanes for SVM-based multiclass classification. *Advances in Data Analysis and Classification*, *14*(1), 175–199. https://doi.org/10.1007/s11634-019-00367-6

Borgo, S., Ferrario, R., Masolo, C., & Vieu, L. (2019). *Ontology Makes Sense: Essays in Honor of Nicola Guarino* (Vol. 316). IOS Press.

C, C., Sundaram, S. M., N S, K., S, M., & Hegde M, N. (2018). Ontology Based Text Mining Method Using Cluster Approach. *AIJR Proceedings*, 1–5.

Canbek, G., Temizel, T. T., Sagiroglu, S., & Baykal, N. (2017). Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. *2nd International Conference on Computer Science and Engineering, UBMK 2017*, 821–826.

Carvalho, F., & Guedes, G. P. (2020). *TF-IDFC-RF: A Novel Supervised Term Weighting Scheme*. 1–28.

Centers for Disease Control and Prevention. (2012). Crisis and Emergency Risk Communication. https://www.cdc.gov/

Chaawa, M., Thabet, I., Hanachi, C., & Ben Said, L. (2017). Modelling and simulating a crisis management system: an organisational perspective. *Enterprise Information Systems*, *11*(4), 534–550.

Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems and Their Applications*, *14*, 20–26.

Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V, & Rong, X. (2015). Data Mining for the Internet of Things: Literature Review and Challenges. *International Journal of Distributed Sensor Networks*, *11*(8).

Chen, K., Zhang, Z., Long, J., & Zhang, H. (2016). Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Systems with Applications*, *66*, 1339–1351.

Chowdhury, G. (2003). Natural language processing. *Annual Review of Information Science and Technology*, *37*, 51–89.

Chughtai, G. R., Lee, J., Kabir, A., Abbasi, R., & Shehzad Hassan, M. A. (2019). Ontology-Based Academic Article Recommendation. *2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2018*, 93–96.

Conneau, A., Schwenk, H., Yann, L. C., & Barrault, L. (2017). Very deep convolutional networks for text classification. *European Chapter of the*

Universiti Utara Malaysia

Association for Computational Linguistics (EACL).

Correa, J. C., Garzón, W., Brooker, P., Sakarkar, G., Carranza, S. A., Yunado, L., & Rincón, A. (2019). Evaluation of collaborative consumption of food delivery services through web mining techniques. *Journal of Retailing and Consumer Services*, *46*(March 2018), 45–50.

Cowie, J., & Lehnert, W. (1996). Information Extraction. *Communications of the ACM*, *39*(1), 80–91.

Cozzolino, A. (2012). Humanitarian logistics: cross-sector cooperation in disaster relief management. *Springer Science & Business Media*.

Cross, N. (2007). Forty years of design research. *Design Studies*, *28*(1), 1–4.

Cutter, S. L. (2003). *Review Article GI Science , Disasters , and Emergency Management*. *7*(4), 439–445.

Cyganiak, R. (2005). *A relational algebra for SPARQL query developers with a powerful tool to extract information from large A relational algebra for SPARQL*.

Debole, F., & Sebastiani, F. (2003). Supervised term weighting for automated text categorization. *Proceedings of the ACM Symposium on Applied Computing*, *January*, 784–788.

Deen, S. (2015). Pakistan 2010 floods. Policy gaps in disaster preparedness and response. *International Journal of Disaster Risk Reduction*, *12*, 341–349.

Domeniconi G., Moro G., Pasolini R., S. C. (2016). A comparison of term weighting schemes for text classification and sentiment analysis with a supervised variant of tf. idf. *International Conference on Data Management Technologies and Applications*, 39–58.

Universiti Utara Malaysia

Domeniconi, G., Moro, G., Pasolini, R., & Sartori, C. (2015). A study on term weighting for text categorization: A novel supervised variant of tf.idf. *DATA 2015 - 4th International Conference on Data Management Technologies and Applications, Proceedings, September*, 26–37.

Edson, C. T. J., & Takahashi, B. (2016). Log in if you survived: Collective coping on social media in the aftermath of Typhoon Haiyan in the Philippines. *New Media & Society*, 1–16.

Ehnis, C. F. (2018). *Social Media within Emergency Management Organisations - A case study exploring Social Media utilisation for Emergency and Disaster Management Social Media within Emergency Management Organisations A case study exploring Social Media utilisation for Emerg*. *March*.

Ehsani, R., & Drabløs, F. (2020). Robust Distance Measures for kNN Classification of Cancer Data. *Cancer Informatics*, *19*. https://doi.org/10.1177/1176935120965542

Eiden, A., Eickhoff, T., Gries, J., Göbel, J. C., & Psota, T. (2021). Supporting semantic PLM by using a lightweight engineering metadata mapping engine. *Procedia CIRP*, *100*, 690–695. https://doi.org/10.1016/j.procir.2021.05.146

Ertek, G., Chi, X., Zhang, A. N., & Asian, S. (2017). Text mining analysis of wind turbine accidents: An ontology-based framework. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-Janua*(Figure 1), 3233–3241.

Fauzi, M. A., Arifin, A. Z., & Yuniarti, A. (2017). Arabic book retrieval using class and book index based term weighting. *International Journal of Electrical and Computer Engineering*, *7*(6), 3705–3710.

Universiti Utara Malaysia

Fettke, P. (2020). Conceptual modelling and artificial intelligence overview and research challenges from the perspective of predictive business process management. *CEUR Workshop Proceedings*, *2542*, 157–164.

Gao, W., Hu, L., & Zhang, P. (2018). Feature selection by maximizing part mutual information. *ACM International Conference Proceeding Series*, 120–127.

Gayathri, R., & Uma, V. (2018). Ontology based knowledge representation technique, domain modeling languages and planners for robotic path planning: A survey. *ICT Express*, *4*(2), 69–74.

Genesereth, M. R., & Nilsson, N. J. (1987). nonmonotonic.pdf. In *Logical Foundations of Artificial Intelligence*. Morgan Kaufmann.

George, A., Stead, T. S., & Ganti, L. (2020). What's the Risk: Differentiating Risk Ratios, Odds Ratios, and Hazard Ratios? *Cureus*, *12*(8).

Gómez-pérez, A., Corcho, O., & Madrid, U. P. De. (2002). Ontology Languages for the Semantic Web. *IEEE Intelligent Systems*, *17*(1), 54–60.

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. 1–17.

Guizzardi, G., Fonseca, C. M., Almeida, J. P. A., Sales, T. P., Benevides, A. B., & Porello, D. (2021). Types and taxonomic structures in conceptual modeling: A novel ontological theory and engineering support. *Data and Knowledge Engineering*, *134*. https://doi.org/10.1016/j.datak.2021.101891

Guo, A., & Yang, T. (2016). Research and improvement of feature words weight based on TFIDF algorithm. *Proceedings of 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2016*, 415–419.

Universiti Utara Malaysia

Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., Zamani, H., Wu, C., Croft, W. B., & Cheng, X. (2020). A Deep Look into neural ranking models for information retrieval. *Information Processing and Management*, *57*(6), 102067.

Gupta, G., & Malhotra, S. (2015). Text Document Tokenization for Word Frequency Count using Rapid Miner (Taking Resume as an Example). *International Conference on Advancements in Engineering and Technology*, 24–26.

Gupta, S., Starr, M. K., Farahani, R. Z., & Matinrad, N. (2016). Disaster Management from a POM Perspective: Mapping a New Domain. *Production and Operations Management*, *25*(10), 1611–1637.

Gurusamy, V., & Kannan, S. (2015). Preprocessing Techniques for Text Mining. *RTRICS*.

Ha, K. M. (2017). Three-Domain Model for Disaster Management Framework. *Global Journal of Flexible Systems Management*, *18*(4), 321–329.

Haddoud, M., Mokhtari, A., Lecroq, T., & Abdeddaïm, S. (2016). Supervised term weights for biomedical text classification: Improvements in nearest centroid computation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9874 LNCS*, 98–113.

Hadj-Kacem, M., & Bouassida, N. (2021). A Multi-Label Classification Approach for Detecting Test Smells over Java Projects. *Journal of King Saud University - Computer and Information Sciences*, *xxxx*.

Hamdan, H., Bellot, P., & Bechet, F. (2016). *Supervised Term Weighting Metrics for Sentiment Analysis in Short Text*.

Hasan, S. Al, Protim, J., & Fahim, N. (2022). *Classification of Multi-Labeled Text*

Articles with Reuters Dataset using SVM. *February*. https://doi.org/10.13140/RG.2.2.33450.49602

Hawalah, A. (2019). Semantic ontology-based approach to enhance arabic text classification. *Big Data and Cognitive Computing*, *3*(4), 1–14.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105.

Horita, F. E. A., de Albuquerque, J. P., Marchezini, V., & Mendiondo, E. M. (2017). Bridging the gap between decision-making and emerging big data sources: An application of a model-based framework to disaster management in Brazil. *Decision Support Systems*, *97*, 12–22.

Horney, J., Nguyen, M., Salvesen, D., Tomasco, O., & Berke, P. (2016). Engaging the public in planning for disaster recovery. *International Journal of Disaster Risk Reduction*, *17*, 33–37.

Houston, J. B., Hawthorne, J., Perreault, M., & Halliwell, M. R. (2014). Social media and disasters : a functional framework for social media use in disaster planning , response , and. *Disasters*.

Huang, G., & Cai, Z. (2021). Understanding Social Security Development: Lessons From the Chinese Case. *SAGE Open*, *11*(3).

Huang, Q., & Xiao, Y. (2015). Geo-Information Geographic Situational Awareness: Mining Tweets for Disaster. *International Journal of Geo-Information*, *4*(3), 1549–1568.

Iroju, O., State, O., Soriyan, A., Gambo, I., Science, C., & State, O. (2012). *State-of-the Art : A Comparative Analysis of Ontology Matching Systems*. *5*(4).

Jalilifard, A., Caridá, V. F., Mansano, A. F., Cristo, R. S., & da Fonseca, F. P. C.

(2021). Semantic Sensitive TF-IDF to Determine Word Relevance in Documents. *Advances in Computing and Network Communications*, *736 LNEE*, 327–337.

Jin, Y. (2010). Making Sense Sensibly in Crisis Communication : How Publics ' Crisis Appraisals Influence Their Negative Emotions , Coping Strategy Preferences , and Crisis Response Acceptance. *Communication Research*, 1–31.

Jonnalagadda, S., Cohen, T., Wu, S., & Gonzalez, G. (2012). Enhancing clinical concept extraction with distributional semantics. *Journal of Biomedical Informatics*, *45*(1), 129–140.

Jurafsky, D., & Martin, J. H. (2019). N-gram Language Models. In *Speech and Language Processing* (pp. 30–55).

Jurgens, M., & Helsloot, I. (2018). The effect of social media on the dynamics of (self) resilience during disasters: A literature review. *J Contingencies and Crisis Management*, *26*, 79–88.

Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, *52*(1), 273–292. https://doi.org/10.1007/s10462-018-09677-1

Karagiannis, D., & Buchmann, R. A. (2018). A proposal for deploying hybrid knowledge bases: The Adoxx-to-GraphDB interoperability case. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2018-Janua*, 4055–4064.

Kastolani, W., & Mainaki, R. (2018). Does Educational Disaster Mitigation Need To Be Introduced In School? *SHS Web of Conferences*, *42*.

Katutu, R., Nyamweha, B. R., Kabaseke, C., Koojo, M. S., Sekajugo, J., Martens, K., Jacobs, L., Maes, J., & Kervyn, M. (2019). Study Of Natural Disasters and Of

Their Impact on the Environmental Condition Rwenzori Mountain Region. *Scientific and Technical Journal*, *5*, 31–37.

Kaya, E., Agca, M., Adiguzel, F., & Cetin, M. (2019). Spatial data analysis with R programming for environment. *Human and Ecological Risk Assessment*, *25*(6), 1521–1530.

Khzam, N. B., & Mussbacher, G. (2018). Domain-specific software language for crisis management systems. *Proceedings - 2018 8th International Model-Driven Requirements Engineering Workshop, MoDRE 2018*, 36–45.

Konys, A. (2019). Knowledge repository of ontology learning tools from text. *Procedia Computer Science*, *159*, 1614–1628.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, *2*(1), 1–15.

Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J., & Valencia, A. (2017). Information retrieval and text mining technologies for chemistry. *Chemical Reviews*, *117*(12), 7673–7761.

Kuntarto, G. P., Moechtar, F. L., Gunawan, I. P., Santoso, B. I., & Ahmadin, Y. (2017). Dwipa ontology II: A semi-automatic ontology population process for Bali Tourism based on the ontology population methodology. *Proceeding of 2017 International Conference on Smart Cities, Automation and Intelligent Computing Systems, ICON-SONICS 2017*, *2018-Janua*, 42–47.

Kunz, N., Reiner, G., & Gold, S. (2014). Investing in disaster management capabilities versus pre-positioning. *International Journal of Production Economics*, *157*(November), 261–272.

Le, C. C., Prasad, P. W. C., Alsadoon, A., Pham, L., & Elchouemi, A. (2019). Text

classification: Naïve bayes classifier with sentiment Lexicon. *IAENG International Journal of Computer Science*, *46*(2), 141–148.

Liu, G., & Zhang, H. (2020). An ontology constructing technology oriented on massive social security policy documents. *Cognitive Systems Research*, *60*, 97–105.

Liu, H., Jiang, L., & Xu, Z. (2018). Entropy measures of probabilistic linguistic term sets. *International Journal of Computational Intelligence Systems*, *11*(1), 45–57.

Liu, Y., Huang, Y., Zhang, S., Zhang, D., & Ling, N. (2018). Integrating object ontology and region semantic template for crime scene investigation image retrieval. *Proceedings of the 2017 12th IEEE Conference on Industrial Electronics and Applications, ICIEA 2017*, 149–153.

Llesol, M., & Intal, G. L. (2019). Development and Validation of a Disaster Response Decision Metamodel in the Philippine Context. *2019 IEEE 6th International Conference on Industrial Engineering and Applications, ICIEA 2019*, 430–434.

Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems with Applications*, *38*(10), 12708–12716.

Maisonnave, M., Delbianco, F., Tohmé, F., & Maguitman, A. (2019). A flexible supervised term-weighting technique and its application to variable extraction and information retrieval. *Inteligencia Artificial*, *22*(63), 61–80.

Mazyad, A., Teytaud, F., & Fonlupt, C. (2018a). A comparative study on term weighting schemes for text classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10710 LNCS*, 100–108.

Mazyad, A., Teytaud, F., & Fonlupt, C. (2018b). Information gain based term weighting method for multi-label text classification task. *Advances in Intelligent*

Universiti Utara Malaysia

Systems and Computing, *868*, 607–615.

Mazyad, A., Teytaud, F., & Fonlupt, C. (2018c). Information gain based term weighting method for multi-label text classification task. *Advances in Intelligent Systems and Computing*, *868*(April), 607–615.

Mikulsen, M., & Diduck, A. P. (2016). Towards an integrated approach to disaster management and food safety governance. *International Journal of Disaster Risk Reduction*, *15*, 116–124.

Mileti, D., & Noji, E. K. (1999). *Disasters by Design : A Reassessment of Natural Hazards in the United States* (Issue July).

Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. In *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications* (Issue January, pp. 29–41).

Mladenic, D., & Grobelnik, M. (1998). Feature Selection for Classification Based on Text Hierarchy. *Text and the Web, Conference on Automated Learning and Discovery CONALD-98*, *September*.

Mohamed, K. A., Ellatif, M. A., & Farhan, M. S. (2018). Using ontology-based concept maps for requirements engineering: A case study. *ICENCO 2017 - 13th International Computer Engineering Conference: Boundless Smart Societies*, *2018-Janua*, 366–371.

Muskat, B., & Nakanishi, H. (2014). Integrating Tourism into Disaster Recovery Management. In B. Ritchie & K. Campiranon (Eds.), *Tourism Crisis and Disaster Management in Asia-Pacific* (pp. 97–115). CABI.

Muzamil, S. A. H. S., Zainun, N. Y., Ajman, N. N., Sulaiman, N., Khahro, S. H.,

Universiti Utara Malaysia

Rohani, M. M., & Ahmad, H. (2022). Proposed Framework for the Flood Disaster Management Cycle in Malaysia. *Sustainability*, *14*(7), 4088.

Nasir, N. S., Abdul Khanan, M. F., Othman, S. H., Abdul Rahman, M. Z., Razak, K. A., Mohd Salleh, M. R., Umar, H. A., & Abdul Razak, A. N. (2018). Assimilating geospatial metamodel and inventory mapping for non-structural mitigation of landslide. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *42*(4/W9), 217–228.

Nisbet, R., Elder, J., & Miner, G. D. (2018). *Handbook of statistical analysis and data mining applications*. Academic press.

Nishida, H., Yokoyama, J., Wagstaff, S. J., & Callomon, P. (2017). Disaster and Biodiversity. *Biology International*, *36*.

Noviantho, Isa, S. M., & Ashianti, L. (2017). Cyberbullying classification using text mining. *Proceedings - 2017 1st International Conference on Informatics and Computational Sciences, ICICoS 2017*, *2018-Janua*, 241–245.

Noy, N., & McGuinness, D. (2001). Ontology development 101: A guide to creating your first ontology. *Development*, *32*, 1–25.

Olteanu, A., Vieweg, S., & Castillo, C. (2015). What to Expect When the Unexpected Happens: Social Media Communications Across Crises. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, 994–1009.

Onan, A. (2021). Ensemble of Classifiers and Term Weighting Schemes for Sentiment Analysis in Turkish. *Scientific Research Communications*, *1*(1), 1–12.

Onan, A., & Tocoglu, M. A. (2021). A Term Weighted Neural Language Model and Stacked Bidirectional LSTM Based Framework for Sarcasm Identification. *IEEE*

Access, *9*, 7701–7722.

Oscar, R.-E., Albores, P., & Brewster, C. (2018). Disaster preparedness in humanitarian logistics: A collaborative approach for resource management in floods. *European Journal of Operational Research*, *264*(3), 978–993.

Othman, S. H. (2013). Supporting Domain Ontology through a Metamodel: A Disaster Management Case Study. In R. M. Colomb (Ed.), *Ontology-Based Applications for Enterprise Systems and Knowledge Management* (1st ed., pp. 191–209). IGI Global. https://doi.org/011

Othman, S. H., & Beydoun, G. (2016). A metamodel-based knowledge sharing system for disaster management. *Expert Systems with Applications*, *63*, 49–65.

Patel, S. P., & Upadhyay, S. H. (2020). Euclidean distance based feature ranking and subset selection for bearing fault diagnosis. *Expert Systems with Applications*, *154*.

Perrin, A. (2015). *Social Media Usage: 2005-2015*. http://www.pewinternet.org/2015/10/08/2015/Social-Networking-Usage-2005-2015/

Phengsuwan, J., Shah, T., Thekkummal, N. B., Wen, Z., Sun, R., Pullarkatt, D., Thirugnanam, H., Ramesh, M. V., Morgan, G., James, P., & Ranjan, R. (2021). Use of social media data in disaster management: A survey. *Future Internet*, *13*(2), 1–24.

Pintas, J. T., Fernandes, L. A. F., & Garcia, A. C. B. (2021). Feature selection methods for text classification: a systematic literature review. In *Artificial Intelligence Review* (Issue 0123456789). Springer Netherlands.

Procopio, C. H., & Procopio, S. T. (2007). Do You Know What It Means to Miss New

Universiti Utara Malaysia

Orleans ? Internet Communication , Geographic Community, and Social Capital in Crisis. *Journal of Applied Communication Social Capital in Crisis*, *35*(1), 67–87.

Qiu, J., Qi, L., Wang, J., & Zhang, G. (2018). A hybrid-based method for Chinese domain lightweight ontology construction. *International Journal of Machine Learning and Cybernetics*, *9*(9), 1519–1531.

Ramasubramanian, C., & Ramya, R. (2013). Effective Pre-Processing Activities in Text Mining using Improved Porter' s Stemming Algorithm. *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(12), 4536–4538.

Ritchie, H., & Roser, M. (2019). Natural Disasters. In *OFDA/CRED International Disaster Database*. https://doi.org/57

Rubin, D. L., Noy, N. F., & Musen, M. A. (2007). Protégé: a tool for managing and using terminology in radiology applications. *Journal of Digital Imaging*, *20 Suppl 1*(August), 34–46.

Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, P., & Mononen, J. (2018). Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. *Behavioural Processes*, *148*(January), 56–62.

Sabbah, T., Selamat, A., Selamat, M. H., Al-Anzi, F. S., Viedma, E. H., Krejcar, O., & Fujita, H. (2017). Modified frequency-based term weighting schemes for text classification. *Applied Soft Computing*, *58*, 193–206.

Şahin, D. Ö., & Kılıç, E. (2019). Two new feature selection metrics for text classification. *Automatika*, *60*(2), 162–171.

Salatino, A. A., Osborne, F., Thanapalasingam, T., & Motta, E. (2019). The CSO Classifier: Ontology-Driven Detection of Research Topics in Scholarly Articles. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11799 LNCS*(August 2020), 296–311.

Salih, A. M., Omar, M., & Yasin, A. (2018). Understanding Uncertainty of Software Requirements Engineering : A Systematic Literature Review Protocol. *Asia Pacific Requirements Engeneering Conference*, 164–171.

Scanlon, J., Helsloot, I., & Groenendaal, J. (2014). Putting It All Together: Integrating Ordinary People Into Emergency Response. *International Journal of Mass Emergencies and Disasters*, *32*(1), 43–63.

Schonlau, M., & Guenther, N. (2016). Text mining using N-grams. *Social Science Research Network*, *17*(4), 866–881. https://doi.org/10.2139/ssrn.2759033

Scopus. (2017). *Scopus: Content Coverage Guide*. Elsevier.

Seeger, M. W., Pechta, L. E., Price, S. M., Lubell, K. M., Rose, D. A., Sapru, S., Chansky, M. C., & Smith, B. J. (2018). A Conceptual Model for Evaluating Emergency Risk Communication in Public Health. *Health Security*, *16*(3), 193–203. https://doi.org/10.1089/hs.2018.0020

Seneviratne, K., Baldry, D., & Pathirage, C. (2010). Disaster Knowledge Factors In Managing Disasters Successfully. *International Journal of Strategic Property Management*, *14*, 376–390.

Seo, P. H., Lin, Z., Cohen, S., Shen, X., & Han, B. (2016). Hierarchical Attention Networks. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Universiti Utara Malaysia

Technologies, 1480–1489.

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The semantic web revisited. *IEEE Intelligent Systems*, *21*(3), 96–101.

Shalaby, W., & Zadrozny, W. (2018). Toward an interactive patent retrieval framework based on distributed representations. *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, *10*, 957–960.

Shaluf, I. M., & Ahmadun, F.-R. (2006). Disaster types in Malaysia: An overview. *Disaster Prevention and Management: An International Journal*, *15*(2), 286–298.

Shanavas, N., Wang, H., Lin, Z., & Hawe, G. (2020). Ontology-based enriched concept graphs for medical document classification. *Information Sciences*, *525*, 172–181. https://doi.org/10.1016/j.ins.2020.03.006

Shoeleh, F., Majd, M., Hamzeh, A., & Hashemi, S. (2015). *Knowledge Representation in Learning Classifier Systems : A Survey*.

Simon, H. A. (1996). The Sciences of the Artificial. In *Cambridge, MA* (Vol. 1). https://doi.org/10.1016/S0898-1221(97)82941-0

Simon, T., Goldberg, A., & Adini, B. (2015). Socializing in emergencies - A review of the use of social media in emergency situations. *International Journal of Information Management*, *35*(5), 609–619. https://doi.org/10.1016/j.ijinfomgt.2015.07.001

Singh, M. (2019). *Social Media Evolution*. Zenesys. https://www.zenesys.com/infographics/social-media-evolution

Singh, V. K., & Singh, V. K. (2015). Vector Space Model: an Information Retrieval System. *International Journal of Advanced Engineering Research and Studies E*

Universiti Utara Malaysia

Int. J. Adv. Engg. Res. Studies/IV/II, 141–143.

Sinyak, N., Tajinder, S., Madhu Kumari, J., & Kozlovskiy, V. (2021). Predicting real estate market trends and value using pre-processing and sentiment text mining analysis. *Real Estate: Economics, Management*, *1*, 35–43.

Slamet, C., Atmadja, A. R., Maylawati, D. S., Lestari, R. S., Darmalaksana, W., & Ramdhani, M. A. (2018). Automated Text Summarization for Indonesian Article Using Vector Space Model. *IOP Conference Series: Materials Science and Engineering*, *288*(1).

Slamet, C., Rahman, A., Sutedi, A., Darmalaksana, W., Ramdhani, M. A., & Maylawati, D. S. (2018). Social Media-Based Identifier for Natural Disaster. *IOP Conference Series: Materials Science and Engineering Conference*, *288*(1).

Somodevilla García, M., Vilariño Ayala, D., & Pineda, I. (2018). An overview of ontology learning tasks. *Computacion y Sistemas*, *22*(1), 137–146. https://doi.org/10.13053/CyS-22-1-2790

Steigenberger, N. (2016). Organizing for the Big One: A Review of Case Studies and a Research Agenda for Multi-Agency Disaster Response. *Journal of Contingencies and Crisis Management*, *24*(2), 60–72.

Steinbach, M., Karypis, G., & Kumar, V. (2000). *A Comparison of Document Clustering Techniques*.

Stephenson, J., Vaganay, M., Coon, D., Cameron, R., Hewitt, N., Campus, J., Road, S., Antrim, C., & Ireland, N. (2018). The role of Facebook and Twitter as organisational communication platforms in relation to flood events in Northern Ireland. *Journal of Flood Risk Management*, *11*(3), 339–350.

Sugumaran, V., & Storey, V. C. (2006). The role of domain ontologies in database

Universiti Utara Malaysia

design. *ACM Transactions on Database Systems*, *31*(3), 1064–1094.

Talib, R., Kashif, M., Ayesha, S., & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications*, *7*(11), 414–418. https://doi.org/10.14569/ijacsa.2016.071153

Tang, Z., Li, W., Li, Y., Zhao, W., & Li, S. (2020). Several alternative term weighting methods for text representation and classification. *Knowledge-Based Systems*, *207*, 106399.

Tankelevičienė, L. (2008). *Ontology and Ontology Engineering: Analysis of Concepts, Classifications and Potential Use in E-Learning Context* (Issue February).

Tulkens, S., Šuster, S., & Daelemans, W. (2019). Unsupervised concept extraction from clinical text through semantic composition. *Journal of Biomedical Informatics*, *91*(February), 103120.

Utomo, F. S., Suryana, N., & Azmi, M. S. (2020). Stemming impact analysis on Indonesian Quran translation and their exegesis classification for ontology instances. *IIUM Engineering Journal*, *21*(1), 33–50.

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, *50*(1), 104–112.

Verdonck, M., Gailly, F., Pergl, R., Guizzardi, G., Martins, B., & Pastor, O. (2019). Comparing traditional conceptual modeling with ontology-driven conceptual modeling: An empirical study. *Information Systems*, *81*(February 2021), 92–103.

Verma, T., Renu, & Gaur, D. (2014). Tokenization and Filtering Process in RapidMiner. *International Journal of Applied Information Systems*, *7*(2), 16–18.

Vijay Gaikwad, S., Chaugule, A., & Patil, P. (2014). Text Mining Methods and Techniques. *International Journal of Computer Applications*, *85*(17), 42–45.

Universiti Utara Malaysia

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing Techniques for Text Mining - An Overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.

Vijayarani, S., & Janani, R. (2016). Text Mining: Open Source Tokenization Tools – An Analysis. *Advanced Computational Intelligence: An International Journal*, *3*(1), 37–47.

Virtual Social Media Working Group and DHS First Responder Group. (2014). Using Social Media for Enhanced Situational Awareness and Decision Support. In *Department of Homeland Security, United States Government*.

Wallemacq, P., Delbiso, T., Below, R., Wahlstrom, M., McClean, D., Landelle, S., Leoni, B., House, R., & Mardi. (2015). *The Human Cost of Weather Related Disasters - 1995 - 2015*.

Wang, J., Xu, W., Yan, W., & Li, C. (2019). Text similarity calculation method based on hybrid model of LDA and TF-IDF. *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, 1–8.

Wang, T., Gu, H., Wu, Z., & Gao, J. (2020). Multi-source knowledge integration based on machine learning algorithms for domain ontology. *Neural Computing and Applications*, *32*(1), 235–245.

Weggenmann, B., & Kerschbaum, F. (2018). *SynTF: Synthetic and Differentially Private Term Frequency Vectors for PriWeggenmann, B., & Kerschbaum, F. (2018). SynTF: Synthetic and Differentially Private Term Frequency Vectors for Privacy-Preserving Text Mining, 305–314. Retrieved from https://arxiv*. 305–314.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2015). Fundamentals of Predictive Text Mining. In D. Gries & F. B. Schneider (Eds.), *Media* (2nd ed., Vol. 42). Springer-

Verlag.

Win, T. Z., Saing, N., & Kham, M. (2019). *Information Gain Measured Feature Selection to Reduce High Dimensional Data*. 68–73.

Wu, H., & Yuan, N. (2018). An Improved TF-IDF algorithm based on word frequency distribution information and category distribution information. *ACM International Conference Proceeding Series*, 211–215.

Xu, X., Zhang, L., Sotiriadis, S., Asimakopoulou, E., Li, M., & Bessis, N. (2018). CLOTHO: A Large-Scale Internet of Things-Based Crowd Evacuation Planning System for Disaster Management. *IEEE Internet of Things Journal*, *5*(5), 3559–3568.

Ye, Z. (2020). *A Label-based Conditional Mutual Information Estimator using A Label-based Conditional Mutual Information Estimator using Contrastive Loss Functions*.

Yiming, Y., & Jan O., P. (1997). A Comparative Study on Feature Selection in Text Categorization. *Proceeding ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning*, *53*(9), 412–420.

Yin, J., Karimi, S., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2015). Using Social Media to Enhance Emergency Situation Awareness. *International Joint Conference on Artificial Intelligence*, 4234–4238.

Zar Phyu, T., & Oo, N. N. (2016). Performance comparison of feature selection methods. *MATEC Web of Conferences*, *42*, 2–5.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality Assessment for Linked Data : A Survey. *Semantic Web*, *7*(1), 63–93.

Zavitsanos, E., Paliouras, G., & Vouros, G. A. (2011). Gold standard evaluation of

ontology learning methods through ontology transformation and alignment. *IEEE Transactions on Knowledge and Data Engineering*, *23*(11), 1635–1648.

Zaw, H. T., Maneerat, N., & Win, K. Y. (2019). Brain tumor detection based on Naïve Bayes classification. *Proceeding - 5th International Conference on Engineering, Applied Sciences and Technology, ICEAST 2019*. https://doi.org/10.1109/ICEAST.2019.8802562

Zhang, H., Wei, H., Tang, Y., & Pu, Q. (2019). *Research on Classification of Scientific and Technological Documents Based on Naive Bayes*. 327–331. https://doi.org/10.1145/3318299.3318330

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, *38*(3), 2758–2765.

Zhang, X., Gao, Y., Yan, X., Ordóñez, P., Pablos, D., Sun, Y., & Cao, X. (2015). From e-learning to social-learning : Mapping development of studies on social media-supported knowledge management. *Computers in Human Behavior*, *51*, 803–811.

APPENDIX A

EXAMPLE LIST OF EXTRACTED TERMS

TF.IDFEC-based.Enhanced

Terms	TP	TN	FP	FN	N	IDFEC-based	TF+	TF-	TF.IDFEC+	TF.IDFEC-	TF	TF.IDFEC-based	TF.IDFEC-based. Enhanced
social media	112	5	22	253	392	2.079181246	0.071837777	0.005124	0.149363759	0.010654403	0.076962103	0.160018161	956.3987653
disast	115	2	7	268	392	2.079181246	0.399546228	0.569046	0.830729024	1.183149861	0.968592271	2.013878885	345.2107169
can use	81	36	110	165	392	2.079181246	0.007827567	0.004557	0.016274931	0.00947495	0.012384625	0.02574988	325.2456413
use	116	1	2	273	392	2.079181246	0.168434487	0.130566	0.350205827	0.27147105	0.29900081	0.621676877	271.7308751
case studi	78	39	161	114	392	2.079181246	0.010436756	0.015278	0.021699907	0.031766633	0.02571519	0.053466541	271.7215064
emerg manag	80	37	139	136	392	2.079181246	0.012081679	0.016148	0.02512	0.03357513	0.028229925	0.058695131	267.431759
data collect	73	44	115	160	392	2.079181246	0.009047079	0.006429	0.018810517	0.013367148	0.015476123	0.032177665	231.9031155
media	116	1	1	274	392	2.079181246	0.086103233	0.009757	0.179024227	0.020286613	0.095860253	0.19931084	216.6362983
emerg respons	72	45	133	142	392	2.079181246	0.007714124	0.008282	0.016039062	0.017220033	0.015996246	0.033259095	213.3695858
intern journal	64	53	152	123	392	2.079181246	0.016080545	0.019306	0.033434368	0.040140758	0.035386586	0.073575126	173.4211813
inform	116	1	1	274	392	2.079181246	0.161911514	0.080155	0.336643383	0.166656883	0.242066567	0.503300266	161.3218751
social	116	1	1	274	392	2.079181246	0.121128758	0.07157	0.251848642	0.148807811	0.192699147	0.400656453	151.6064961
system	113	4	5	270	392	2.079181246	0.10380034	0.101787	0.21581972	0.211633405	0.205587236	0.427453126	148.2802925
intern confer	63	54	82	193	392	2.079181246	0.012365286	0.004803	0.025709671	0.009986046	0.01716816	0.035695716	143.2627901
need	114	3	4	271	392	2.079181246	0.043335224	0.055157	0.090101785	0.114682267	0.098492641	0.204784052	139.0507066

disast prepared	63	54	120	155	392	2.079181246	0.008876914	0.010154	0.018456713	0.021112231	0.019031022	0.039568944	135.7751484
can	116	1	1	274	392	2.079181246	0.104140669	0.099045	0.216527326	0.20593271	0.203185767	0.422460036	123.6167778
can also	58	59	103	172	392	2.079181246	0.004027226	0.003347	0.008373333	0.006958781	0.007374111	0.015332113	114.9745947
new york	56	61	128	147	392	2.079181246	0.004736245	0.00554	0.009847512	0.011519336	0.010276568	0.021366847	112.6021208
natur hazard	59	58	121	154	392	2.079181246	0.008026092	0.010305	0.0166877	0.021426753	0.018331472	0.038114453	112.049141
also	116	1	1	274	392	2.079181246	0.056097561	0.065293	0.116636997	0.135755183	0.121390177	0.252392179	111.4578787
relat	116	1	1	274	392	2.079181246	0.038088486	0.045003	0.079192866	0.093570036	0.083091795	0.172762902	110.556914
climat chang	35	82	120	155	392	2.079181246	0.0000284	0	5.90487E-05	0	0.0000284	5.90487E-05	106.4946492
time	116	1	1	274	392	2.079181246	0.049290981	0.062891	0.102484883	0.13076216	0.11218216	0.233247043	105.9727007
crisi manag	63	54	84	191	392	2.079181246	0.007515598	0.007129	0.01562629	0.014821807	0.014644273	0.030448098	104.5716957
keyword	95	22	20	255	392	2.079181246	0.011542825	0.008604	0.023999625	0.01788839	0.020146399	0.041888015	102.8816569
respons	115	2	2	273	392	2.079181246	0.082643222	0.10971	0.171830237	0.228106451	0.19235297	0.399936688	102.7302947
social network	67	50	52	223	392	2.079181246	0.008650028	0.003952	0.017984976	0.008216864	0.012601999	0.02620184	99.44399514
manag	116	1	1	274	392	2.079181246	0.227084515	0.337827	0.472149865	0.702404313	0.564911876	1.174554178	96.95208518
process	112	5	4	271	392	2.079181246	0.050482133	0.057143	0.104961504	0.118810357	0.10762499	0.223771861	87.38259372
earli warn	58	59	83	192	392	2.079181246	0.006097561	0.006429	0.012677934	0.013367148	0.012526605	0.026045082	82.57891913
can help	50	67	73	202	392	2.079181246	0.005275099	0.002307	0.010967887	0.004796447	0.007581991	0.015764333	78.80572359
data analysi	53	64	62	213	392	2.079181246	0.004764606	0.002099	0.009906479	0.004363981	0.0068635	0.01427046	74.10735102
emerg disast	49	68	77	198	392	2.079181246	0.004594441	0.002685	0.009552676	0.005582749	0.007279512	0.015135425	72.81161337
disast situat	50	67	74	201	392	2.079181246	0.005501985	0.00346	0.011439624	0.00719467	0.008962323	0.018634294	70.4884695
howev	110	7	5	270	392	2.079181246	0.022546795	0.033563	0.046878873	0.069784375	0.056110187	0.116663249	65.64476513
decis make	49	68	87	188	392	2.079181246	0.005275099	0.005446	0.010967887	0.011322762	0.010720878	0.022290648	64.13558049
differ type	52	65	62	213	392	2.079181246	0.002836075	0.001929	0.005896714	0.004010144	0.004764788	0.009906858	61.38300635
analysi	113	4	2	273	392	2.079181246	0.052155417	0.048483	0.108440565	0.100804021	0.100637973	0.209244586	60.88051791
disast recoveri	49	68	90	185	392	2.079181246	0.004310834	0.00537	0.008963005	0.011165501	0.009680977	0.020128506	60.04324248

provid	115	2	1	274	392	2.079181246	0.055331821	0.059072	0.115044885	0.122820501	0.114403391	0.237865385	57.82241945
follow	110	7	4	271	392	2.079181246	0.025212706	0.033809	0.052421785	0.070295473	0.059021915	0.122717259	55.82813873
acquir	48	69	66	209	392	2.079181246	0.003318208	0.00242	0.006899156	0.005032338	0.005738554	0.011931494	55.19879089
includ	115	2	1	274	392	2.079181246	0.042115712	0.052113	0.087566199	0.10835253	0.094228788	0.195918729	53.43437525
literatur review	44	73	103	172	392	2.079181246	0.003431651	0.005049	0.007135024	0.010497144	0.008480342	0.017632168	52.23350111
use social	61	56	25	250	392	2.079181246	0.010153148	0.001097	0.021110235	0.002280278	0.011249867	0.023390512	51.10077593
work	115	2	1	274	392	2.079181246	0.037436188	0.051073	0.07783662	0.106190196	0.088509271	0.184026816	50.56651806
correspond author	43	74	128	147	392	2.079181246	0.001276234	0.002647	0.002653522	0.005504119	0.003923487	0.008157641	50.30337767
use disast	50	67	50	225	392	2.079181246	0.002722632	0.001569	0.005660845	0.003263156	0.004292075	0.008924002	49.21292718
result	115	2	1	274	392	2.079181246	0.037039138	0.053796	0.077011081	0.111851576	0.09083511	0.188862657	48.74918039
report	101	16	8	267	392	2.079181246	0.025155984	0.029063	0.05230385	0.060427371	0.054219045	0.112731222	48.71617405
disast event	50	67	63	212	392	2.079181246	0.003630176	0.003933	0.007547794	0.008177549	0.007563238	0.015725343	46.91892084
relat disast	46	71	78	197	392	2.079181246	0.002552467	0.003177	0.005307042	0.006604943	0.005729171	0.011911985	46.81174018
futur research	44	73	64	211	392	2.079181246	0.003885423	0.002799	0.008078499	0.005818641	0.006683948	0.013897139	46.62376958
intern	112	5	1	274	392	2.079181246	0.0000284	0	5.90487E-05	0	0.0000284	5.90487E-05	46.57365991
can provid	46	71	52	223	392	2.079181246	0.002552467	0.001475	0.005307042	0.00306658	0.004027365	0.008373622	44.39502657
access	110	7	3	272	392	2.079181246	0.029239932	0.036551	0.060795118	0.075996168	0.065790939	0.136791287	43.56307446
phase disast	44	73	68	207	392	2.079181246	0.003119682	0.003139	0.006486384	0.006526313	0.006258568	0.013012697	42.47823177
area	110	7	3	272	392	2.079181246	0.046937039	0.063326	0.097590611	0.131666408	0.110263124	0.22925702	41.72474057
data	113	4	1	274	392	2.079181246	0.20833806	0.091274	0.433172587	0.189774187	0.299611578	0.622946774	40.84330008
import	110	7	3	272	392	2.079181246	0.025666478	0.03623	0.05336526	0.075327811	0.061896033	0.128693071	40.645429
data sourc	45	72	34	241	392	2.079181246	0.007742484	0.001153	0.016098028	0.002398223	0.00889593	0.018496251	38.45388678
knowledg	104	13	4	271	392	2.079181246	0.042682927	0.032032	0.088745541	0.066599849	0.074714694	0.155345391	38.00935493
red cross	42	75	60	215	392	2.079181246	0.003743619	0.003233	0.007783662	0.006722889	0.00697705	0.014506552	37.48447513
disast emerg	39	78	75	200	392	2.079181246	0.002552467	0.00278	0.005307042	0.005779325	0.005332083	0.011086367	37.32388579

also use	42	75	54	221	392	2.079181246	0.001843449	0.001399	0.003832865	0.002909321	0.003242712	0.006742186	35.74348421
collect data	40	77	56	219	392	2.079181246	0.002949518	0.002042	0.006132583	0.004246035	0.004991685	0.010378618	35.73991554
use data	43	74	45	230	392	2.079181246	0.002552467	0.001343	0.005307042	0.002791376	0.003895003	0.008098417	35.62820465
emerg	114	3	1	274	392	2.079181246	0.073624504	0.092578	0.153078688	0.192486932	0.16620274	0.34556562	34.99936369
journal disast	37	80	99	176	392	2.079181246	0.00794101	0.013709	0.016510799	0.028503477	0.021650001	0.045014276	34.91862251
studi	114	3	1	274	392	2.079181246	0.076290414	0.096908	0.158621598	0.201490099	0.1731988	0.360111697	34.80174646
avail onlin	35	82	95	180	392	2.079181246	0.001871809	0.002685	0.00389183	0.005582749	0.00455688	0.009474579	34.6309245
use inform	43	74	38	237	392	2.079181246	0.002750993	0.000908	0.005719813	0.001887127	0.003658623	0.00760694	34.52109053
requir	108	9	3	272	392	2.079181246	0.027793534	0.032731	0.057787795	0.068054509	0.060524932	0.125842304	34.37196109
big data	48	69	25	250	392	2.079181246	0.03142371	0.002042	0.065335589	0.004246035	0.033465877	0.069581624	33.95312499
inform share	46	71	34	241	392	2.079181246	0.004792967	0.001721	0.009965447	0.003577678	0.006513682	0.013543125	33.7014538
disast manag	114	3	1	274	392	2.079181246	0.091463415	0.126142	0.190169017	0.262271307	0.217605043	0.452440324	33.20889329
help	113	4	1	274	392	2.079181246	0.030204197	0.023277	0.0628	0.04839694	0.053481119	0.11119694	33.17245473
disast inform	44	73	40	235	392	2.079181246	0.005104935	0.002647	0.010614085	0.005504119	0.007752188	0.016118204	33.01019404
public	108	9	3	272	392	2.079181246	0.044100964	0.057483	0.091693897	0.119518029	0.101584182	0.211211926	32.49502272
method	107	10	3	272	392	2.079181246	0.033408962	0.035795	0.069463287	0.074423563	0.069203611	0.14388685	32.22045047
high	103	14	5	270	392	2.079181246	0.019171866	0.026832	0.039861784	0.055788186	0.046003671	0.09564997	31.87450423
inform can	45	72	33	242	392	2.079181246	0.002013613	0.000719	0.004186666	0.001493975	0.002732153	0.005680641	31.60510931
respons manag	47	70	40	235	392	2.079181246	0.002353942	0.001834	0.004894272	0.003813568	0.00418811	0.00870784	31.38555671
system disast	36	81	65	210	392	2.079181246	0.002353942	0.002156	0.004894272	0.004481927	0.004509563	0.009376199	31.35338873
disast occur	40	77	82	193	392	2.079181246	0.001843449	0.003423	0.003832865	0.007116041	0.00526597	0.010948906	31.0047486
follow section	36	81	63	212	392	2.079181246	0.001418037	0.001286	0.002948356	0.00267343	0.002703846	0.005621786	30.53205183
system use	37	80	46	229	392	2.079181246	0.003091322	0.001399	0.006427419	0.002909321	0.004490585	0.00933674	30.45111794
inform technolog	41	76	37	238	392	2.079181246	0.003148043	0.001248	0.006545352	0.002594799	0.004396034	0.009140151	29.71966782
base	113	4	1	274	392	2.079181246	0.050623936	0.049579	0.105256338	0.103084301	0.100203212	0.208340639	29.6746132

result show	36	81	64	211	392	2.079181246	0.002070335	0.002061	0.004304602	0.004285351	0.004131411	0.008589952	29.63684902
disast area	36	81	52	223	392	2.079181246	0.002750993	0.001815	0.005719813	0.003774255	0.004566253	0.009494068	28.94960814
type disast	36	81	64	211	392	2.079181246	0.002977879	0.003177	0.00619155	0.006604943	0.006154583	0.012796494	28.61529446
research	113	4	1	274	392	2.079181246	0.082756665	0.08891	0.172066106	0.184859795	0.171666564	0.3569259	28.31574989
number	110	7	2	273	392	2.079181246	0.031792399	0.041694	0.06610216	0.086689886	0.073486641	0.152792046	28.27039611
local govern	35	82	108	167	392	2.079181246	0.006352808	0.015392	0.013208639	0.032002525	0.021744696	0.045211164	28.0015896
situat awar	47	70	24	251	392	2.079181246	0.007146909	0.001513	0.014859719	0.003145211	0.008659625	0.01800493	27.65174707
washington dc	35	82	71	204	392	2.079181246	0.00190017	0.002458	0.003950798	0.005110968	0.004358334	0.009061766	27.47114984
build	99	18	6	269	392	2.079181246	0.022149745	0.033355	0.046053334	0.06935191	0.055505139	0.115405244	27.38052841
tabl show	34	83	61	214	392	2.079181246	0.001956892	0.001777	0.004068733	0.003695624	0.003734334	0.007764357	27.22555288

TF.RF.Enhanced

Terms	TP	TN	FP	FN	N	RF	TF+	TF-	TF.RF+	TF.RF-	TF	TF.RF	TF.RF. Enhanced
social media	112	5	22	253	392	0.387867954	0.071837777	0.005124	0.027863572	0.001987562	0.076962103	0.029851133	178.4146682
disast	115	2	7	268	392	0.385446195	0.399546228	0.569046	0.154003573	0.219336632	0.968592271	0.373340205	63.99642041
can use	81	36	110	165	392	0.396357878	0.007827567	0.004557	0.003102518	0.001806226	0.012384625	0.004908744	62.00213299
case studi	78	39	161	114	392	0.428816575	0.010436756	0.015278	0.004475454	0.006551646	0.02571519	0.0110271	56.04065831
emerg manag	80	37	139	136	392	0.413003755	0.012081679	0.016148	0.004989779	0.006669286	0.028229925	0.011659065	53.12202623
use	116	1	2	273	392	0.384695342	0.168434487	0.130566	0.064795963	0.050228256	0.29900081	0.115024219	50.2763298
data collect	73	44	115	160	392	0.390272568	0.009047079	0.006429	0.003530827	0.00250908	0.015476123	0.006039906	43.52935777
emerg respons	72	45	133	142	392	0.399161654	0.007714124	0.008282	0.003079182	0.003305906	0.015996246	0.006385088	40.96273802
media	116	1	1	274	392	0.384417517	0.086103233	0.009757	0.033099591	0.003750769	0.095860253	0.03685036	40.05364513
intern journal	64	53	152	123	392	0.401456582	0.016080545	0.019306	0.006455641	0.007750537	0.035386586	0.014206178	33.48485125
inform	116	1	1	274	392	0.384417517	0.161911514	0.080155	0.062241622	0.030813006	0.242066567	0.093054629	29.82662271

social	116	1	1	274	392	0.384417517	0.121128758	0.07157	0.046564016	0.027512911	0.192699147	0.074076928	28.03035706
system	113	4	5	270	392	0.383549417	0.10380034	0.101787	0.03981256	0.039040305	0.205587236	0.078852865	27.353469
need	114	3	4	271	392	0.383934549	0.043335224	0.055157	0.01663789	0.021176838	0.098492641	0.037814728	25.6766313
intern confer	63	54	82	193	392	0.366689032	0.012365286	0.004803	0.004534215	0.001761161	0.01716816	0.006295376	25.2661445
disast prepared	63	54	120	155	392	0.381377134	0.008876914	0.010154	0.003385452	0.003872545	0.019031022	0.007257997	24.90477297
can	116	1	1	274	392	0.384417517	0.104140669	0.099045	0.040033497	0.038074671	0.203185767	0.078108168	22.85536906
also	116	1	1	274	392	0.384417517	0.056097561	0.065293	0.021564885	0.025099625	0.121390177	0.04666451	20.60732369
relat	116	1	1	274	392	0.384417517	0.038088486	0.045003	0.014641881	0.01730006	0.083091795	0.031941941	20.44074531
new york	56	61	128	147	392	0.37675071	0.004736245	0.00554	0.001784384	0.002087321	0.010276568	0.003871704	20.40367044
can also	58	59	103	172	392	0.368697606	0.004027226	0.003347	0.001484829	0.001233988	0.007374111	0.002718817	20.38824557
natur hazard	59	58	121	154	392	0.377145343	0.008026092	0.010305	0.003027003	0.003886626	0.018331472	0.006913629	20.32473688
time	116	1	1	274	392	0.384417517	0.049290981	0.062891	0.018948317	0.024176471	0.11218216	0.043124787	19.59317519
respons	115	2	2	273	392	0.384038812	0.082643222	0.10971	0.031738205	0.042132801	0.19235297	0.073871006	18.97497895
keyword	95	22	20	255	392	0.375215194	0.011542825	0.008604	0.004331043	0.003228192	0.020146399	0.007559235	18.56632794
crisi manag	63	54	84	191	392	0.367326644	0.007515598	0.007129	0.002760679	0.002618552	0.014644273	0.005379232	18.47456546
manag	116	1	1	274	392	0.384417517	0.227084515	0.337827	0.087295265	0.129866755	0.564911876	0.21716202	17.92536359
climat chang	35	82	120	155	392	0.347487397	0.0000284	0	9.86864E-06	0	0.0000284	9.86864E-06	17.79813496
social network	67	50	52	223	392	0.361812502	0.008650028	0.003952	0.003129688	0.001429873	0.012601999	0.004559561	17.30492749
process	112	5	4	271	392	0.382608457	0.050482133	0.057143	0.019314891	0.02186334	0.10762499	0.041178231	16.08004086
earli warn	58	59	83	192	392	0.362121041	0.006097561	0.006429	0.002208055	0.002328092	0.012526605	0.004536147	14.38237488
can help	50	67	73	202	392	0.351704483	0.005275099	0.002307	0.001855276	0.000811344	0.007581991	0.00266662	13.33040415
data analysi	53	64	62	213	392	0.35195591	0.004764606	0.002099	0.001676931	0.000738718	0.0068635	0.002415649	12.54461111
emerg disast	49	68	77	198	392	0.351694821	0.004594441	0.002685	0.001615841	0.000944326	0.007279512	0.002560167	12.31613038
howev	110	7	5	270	392	0.381549592	0.022546795	0.033563	0.00860272	0.012806099	0.056110187	0.021408819	12.0464406
disast situat	50	67	74	201	392	0.351942377	0.005501985	0.00346	0.001936382	0.00121784	0.008962323	0.003154221	11.9315618

Universiti Utara Malaysia

analysi	113	4	2	273	392	0.382722768	0.052155417	0.048483	0.019961066	0.018555378	0.100637973	0.038516444	11.20650754
decis make	49	68	87	188	392	0.354231081	0.005275099	0.005446	0.001868604	0.001929064	0.010720878	0.003797668	10.92680883
provid	115	2	1	274	392	0.383762966	0.055331821	0.059072	0.021234304	0.022669481	0.114403391	0.043903785	10.67251987
differ type	52	65	62	213	392	0.351048293	0.002836075	0.001929	0.000995599	0.000677071	0.004764788	0.001672671	10.36388706
disast recoveri	49	68	90	185	392	0.355042295	0.004310834	0.00537	0.001530528	0.001906628	0.009680977	0.003437156	10.25302177
follow	110	7	4	271	392	0.381278305	0.025212706	0.033809	0.009613058	0.012890718	0.059021915	0.022503776	10.23771167
includ	115	2	1	274	392	0.383762966	0.042115712	0.052113	0.016162451	0.019999069	0.094228788	0.036161519	9.86260065
work	115	2	1	274	392	0.383762966	0.037436188	0.051073	0.014366623	0.019599958	0.088509271	0.03396658	9.333268549
acquir	48	69	66	209	392	0.348239631	0.003318208	0.00242	0.001155532	0.00084286	0.005738554	0.001998392	9.245180805
result	115	2	1	274	392	0.383762966	0.037039138	0.053796	0.014214249	0.020644902	0.09083511	0.034859151	8.99783512
literatur review	44	73	103	172	392	0.353303279	0.003431651	0.005049	0.001212414	0.001783719	0.008480342	0.002996133	8.875737618
report	101	16	8	267	392	0.376262464	0.025155984	0.029063	0.009465253	0.010935339	0.054219045	0.020400591	8.816002797
correspond author	43	74	128	147	392	0.360312566	0.001276234	0.002647	0.000459843	0.000953839	0.003923487	0.001413682	8.717344449
use social	61	56	25	250	392	0.351022853	0.010153148	0.001097	0.003563987	0.000384973	0.011249867	0.00394896	8.627213318
intern	112	5	1	274	392	0.381793373	0.0000284	0	1.08429E-05	0	0.0000284	1.08429E-05	8.552171549
use disast	50	67	50	225	392	0.346787486	0.002722632	0.001569	0.000944175	0.000544263	0.004292075	0.001488438	8.208244153
access	110	7	3	272	392	0.381008844	0.029239932	0.036551	0.011140673	0.013926257	0.065790939	0.02506693	7.982909949
disast event	50	67	63	212	392	0.349442481	0.003630176	0.003933	0.001268538	0.001374379	0.007563238	0.002642917	7.885538658
relat disast	46	71	78	197	392	0.34898645	0.002552467	0.003177	0.000890776	0.001108627	0.005729171	0.001999403	7.857257788
futur research	44	73	64	211	392	0.344103461	0.003885423	0.002799	0.001336988	0.000962982	0.006683948	0.00229997	7.71621066
area	110	7	3	272	392	0.381008844	0.046937039	0.063326	0.017883427	0.024127798	0.110263124	0.042011225	7.646036253
data	113	4	1	274	392	0.382450897	0.20833806	0.091274	0.079679078	0.034907639	0.299611578	0.114586717	7.51284034
import	110	7	3	272	392	0.381008844	0.025666478	0.03623	0.009779155	0.013803781	0.061896033	0.023582936	7.448252989
can provid	46	71	52	223	392	0.34366024	0.002552467	0.001475	0.000877181	0.000506864	0.004027365	0.001384045	7.337891059
phase disast	44	73	68	207	392	0.344895133	0.003119682	0.003139	0.001075963	0.001082587	0.006258568	0.00215855	7.046300271
knowledg	104	13	4	271	392	0.377263227	0.042682927	0.032032	0.016102699	0.012084408	0.074714694	0.028187107	6.896720491

emerg	114	3	1	274	392	0.383107427	0.073624504	0.092578	0.028206094	0.03546741	0.16620274	0.063673504	6.448940505
studi	114	3	1	274	392	0.383107427	0.076290414	0.096908	0.029227424	0.037126322	0.1731988	0.066353747	6.412527794
data sourc	45	72	34	241	392	0.339793573	0.007742484	0.001153	0.002630846	0.000391934	0.00889593	0.00302278	6.284388913
requir	108	9	3	272	392	0.379678692	0.027793534	0.032731	0.010552613	0.012427414	0.060524932	0.022980027	6.276653967
red cross	42	75	60	215	392	0.341503539	0.003743619	0.003233	0.001278459	0.001104228	0.00697705	0.002382687	6.156789327
disast emerg	39	78	75	200	392	0.341434525	0.002552467	0.00278	0.0008715	0.000949057	0.005332083	0.001820557	6.129173791
disast manag	114	3	1	274	392	0.383107427	0.091463415	0.126142	0.035040314	0.048325794	0.217605043	0.083366108	6.119030591
help	113	4	1	274	392	0.382450897	0.030204197	0.023277	0.011551622	0.00890228	0.053481119	0.020453902	6.101841808
public	108	9	3	272	392	0.379678692	0.044100964	0.057483	0.016744196	0.021825153	0.101584182	0.038569349	5.93390679
method	107	10	3	272	392	0.379012085	0.033408962	0.035795	0.0126624	0.013566605	0.069203611	0.026229005	5.873437017
also use	42	75	54	221	392	0.340453088	0.001843449	0.001399	0.000627608	0.000476383	0.003242712	0.001103991	5.852774786
collect data	40	77	56	219	392	0.338983782	0.002949518	0.002042	0.000999839	0.000692261	0.004991685	0.0016921	5.826933921
use data	43	74	45	230	392	0.339840149	0.002552467	0.001343	0.000867431	0.000456248	0.003895003	0.001323678	5.82339534
journal disast	37	80	99	176	392	0.344436934	0.00794101	0.013709	0.002735177	0.004721883	0.021650001	0.00745706	5.784615114
high	103	14	5	270	392	0.376847209	0.019171866	0.026832	0.007224864	0.010111491	0.046003671	0.017336355	5.77718656
avail onlin	35	82	95	180	392	0.341324591	0.001871809	0.002685	0.000638894	0.000916481	0.00455688	0.001555375	5.685115786
use inform	43	74	38	237	392	0.338742197	0.002750993	0.000908	0.000931877	0.000307453	0.003658623	0.00123933	5.624209085
big data	48	69	25	250	392	0.34084055	0.03142371	0.002042	0.010710475	0.000696053	0.033465877	0.011406528	5.565941791
inform share	46	71	34	241	392	0.34061688	0.004792967	0.001721	0.001632565	0.000586105	0.006513682	0.00221867	5.521059824
base	113	4	1	274	392	0.382450897	0.050623936	0.049579	0.01936117	0.018961639	0.100203212	0.038322808	5.458438243
disast inform	44	73	40	235	392	0.339895257	0.005104935	0.002647	0.001735143	0.000899789	0.007752188	0.002634932	5.396358976
research	113	4	1	274	392	0.382450897	0.082756665	0.08891	0.031650361	0.034003671	0.171666564	0.065654031	5.208484809
number	110	7	2	273	392	0.380741192	0.031792399	0.041694	0.012104676	0.015874715	0.073486641	0.027979391	5.176895638
respons manag	47	70	40	235	392	0.342422681	0.002353942	0.001834	0.000806043	0.000628061	0.00418811	0.001434104	5.168922377
inform can	45	72	33	242	392	0.339640306	0.002013613	0.000719	0.000683904	0.000244045	0.002732153	0.000927949	5.162786563
disast occur	40	77	82	193	392	0.34385229	0.001843449	0.003423	0.000633874	0.001176842	0.00526597	0.001810716	5.127524996

system disast	36	81	65	210	392	0.336745548	0.002353942	0.002156	0.000792679	0.000725896	0.004509563	0.001518575	5.078015247
follow section	36	81	63	212	392	0.336421971	0.001418037	0.001286	0.000477059	0.000432574	0.002703846	0.000909633	4.94023937
build	99	18	6	269	392	0.374387152	0.022149745	0.033355	0.00829258	0.012487831	0.055505139	0.020780411	4.930266701
system use	37	80	46	229	392	0.334769717	0.003091322	0.001399	0.001034881	0.000468431	0.004490585	0.001503312	4.902945398
inform technolog	41	76	37	238	392	0.336913586	0.003148043	0.001248	0.001060618	0.000420465	0.004396034	0.001481084	4.815818669
result show	36	81	64	211	392	0.336583023	0.002070335	0.002061	0.00069684	0.000693723	0.004131411	0.001390563	4.797686708
one	113	4	1	274	392	0.382450897	0.036954056	0.047859	0.014133112	0.01830355	0.084812617	0.032436661	4.707559134
disast area	36	81	52	223	392	0.334742175	0.002750993	0.001815	0.000920873	0.000607644	0.004566253	0.001528517	4.660803294
local govern	35	82	108	167	392	0.344309895	0.006352808	0.015392	0.002187335	0.005299579	0.021744696	0.007486914	4.637029308
type disast	36	81	64	211	392	0.336583023	0.002977879	0.003177	0.001002304	0.001069225	0.006154583	0.002071528	4.632314919
natur	113	4	1	274	392	0.382450897	0.047844583	0.065255	0.018298204	0.024956756	0.113099381	0.04325496	4.570531236
situat awar	47	70	24	251	392	0.339898623	0.007146909	0.001513	0.002429225	0.00051417	0.008659625	0.002943395	4.520428783

Universiti Utara Malaysia