FEATURE ENGINEERING TECHNIQUES TO CLASSIFY CAUSE OF DEATH FROM FORENSIC AUTOPSY REPORTS

GHULAM MUJTABA

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

FEATURE ENGINEERING TECHNIQUES TO CLASSIFY CAUSE OF DEATH FROM FORENSIC AUTOPSY REPORTS

GHULAM MUJTABA

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Ghulam Mujtaba

Matric No: WHA140037

Name of Degree: Doctor of Philosophy

Title of Thesis: FEATURE ENGINEERING TECHNIQUES TO CLASSIFY CAUSE OF DEATH FROM FORENSIC AUTOPSY REPORTS

Field of Study: Data Mining (Computer Science)

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

FEATURE ENGINEERING TECHNIQUES TO CLASSIFY CAUSE OF DEATH FROM FORENSIC AUTOPSY REPORTS

ABSTRACT

Forensic autopsy focuses on revealing the cause of death (CoD) by examining a dead body. This process is performed by medical pathologists during the investigation of criminal and civil law cases. In forensic autopsy, pathologists examine corpses externally and anatomically to collect autopsy findings. Moreover, these experts collect the history of the deceased and death scene-related information from the deceased's relatives and eyewitnesses. Afterward, the pathologists determine the CoD through their expert knowledge while correlating the current autopsy findings with previous autopsy reports. Therefore, determining the CoD from autopsy findings is laborious, time consuming, and subject to inconsistencies associated with any labor-intensive process. Hence, automated text classification (ATC) techniques must be employed to overcome the aforementioned issues in determining the CoD. This study aimed to employ ATC techniques to classify the CoD from forensic autopsy reports. In the ATC technique, feature engineering is a highly important step because the success or failure of any ATC model is heavily dependent on the quality of the features used in the classification task. In ATC, the traditional feature engineering techniques include bag of words (BoW) and *n*-gram. This study argues that BoW and its variant techniques are inadequate in determining the CoD from forensic autopsy reports because these techniques ignore word-order, word-context, and word-level synonymy and polysemy. To overcome the aforementioned issues of BoW and its variant techniques, this study aimed to achieve the following four main objectives. First, this work intended to investigate the existing feature engineering techniques to classify free-text clinical reports, including forensic autopsy reports. Second, this study aimed to develop semi-automated expert-driven feature engineering to overcome the issue of word-level synonymy and polysemy. Third, this research sought to propose a fully automated conceptual graph-based feature engineering technique to address issues in word-order and word-context. Finally, this work intended to evaluate the proposed techniques by comparing their performances with existing baseline techniques. For the experimental evaluation, forensic autopsy reports of 16 different CoDs were obtained from a very large hospital in Kuala Lumpur, Malaysia. These reports were preprocessed by applying various text preprocessing techniques. The discriminative features were then extracted from the preprocessed reports through the proposed feature engineering techniques and formed numeric master feature vectors. These master feature vectors were fed as input to six machine learning algorithms to construct and evaluate the classification models. Furthermore, to show the effectiveness of the proposed techniques, this study compared their performances with five state-of-the-art baseline feature engineering techniques. Experimental results showed that the proposed techniques outperformed the traditional BoW and its variant techniques. Moreover, support vector machines and random forest algorithms outperformed the four other algorithms. The proposed techniques are feasible and practical in determining the CoD from forensic autopsy reports and can assist pathologists to accurately and rapidly determine the CoD from autopsy findings. Finally, the proposed techniques are generally applicable to other kinds of free-text clinical reports.

Keywords: Automated Text Classification Techniques, Forensic Autopsy Reports, Supervised Machine Learning Algorithms, Feature Engineering Techniques, Free-Text Clinical Reports

TEKNIK-TEKNIK KEJURUTERAAN CIRI UNTUK MENGKLASIFIKASIKAN PENYEBAB KEMATIAN DARI LAPORAN AUTOPSI FORENSIK

ABSTRAK

Autopsi forensik memfokuskan dalam mendedahkan penyebab kematian (CoD) dengan pemeriksaan mayat. Ia dilakukan oleh ahli patologi perubatan semasa penyiasatan kes jenayah dan undang-undang sivil. Dalam autopsi forensik, pakar patologi memeriksa mayat secara luaran dan anatomik untuk mengumpul hasil autopsi. Mereka juga mengumpul maklumat sejarah si mati, dan maklumat berkaitan keadaan kematian daripada saudara-mara dan saksi saksi si mati. Seterusnya, pakar patologi menentukan CoD dengan menggunakan kepakaran mereka dan mengaitkan penemuan autopsi semasa dengan laporan autopsi terdahulu. Oleh itu, menentukan CoD daripada penemuan autopsi adalah sukar, memakan masa dan tertakluk kepada ketidaktetapan yang berkaitan dengan proses kerja intensif. Oleh itu, teknik klasifikasi teks automatik (ATC) perlu digunakan untuk mengatasi isu-isu tersebut dalam menentukan CoD. Kajian ini bertujuan untuk menggunakan teknik ATC dalam mugklasifikasikan laporan autopsi forensik CoD. Dalam teknik ATC, kejuruteraan ciri adalah langkah yang sangat penting kerana kejayaan atau kegagalan model ATC sangat bergantung kepada kualiti ciri yang digunakan dalam proses klasifikasi. Dalam teknik ATC, teknik kejuruteraan ciri tradisional adalah *Bag of Words* (BoW) dan n-gram. Kajian ini mendapati bahawa BoW dan teknik variannya tidak mencukupi untuk menentukan CoD dari laporan autopsi forensik kerana teknik teknik tersebut mengabaikan susunan, konteks, dan tahap "*synonymy*" dan "*polysemy*" perkataan didalam laporan. Oleh itu, bagi mengatasi isu-isu BoW dan teknik variannya, kajian ini bertujuan untuk mencapai empat objektif utama. Pertama, untuk mengkaji teknik kejuruteraan ciri yang sedia ada bagi mengklasifikasikan laporan klinikal bebas teks termasuklah laporan autopsi forensik. Kedua, untuk mencadangkan teknik kejuruteraan ciri separuh automatik yang didorong oleh pakar bagi mengatasi masalah perkataan "*synonymy*" dan "*polysemy*". Ketiga, untuk mencadangkan teknik kejuruteraan ciri berasaskan graf konseptual sepenuhnya bagi mengatasi masalah susunan dan konteks perkataan. Akhir sekali, kajian diteruskan dengan penilaian teknik yang dicadangkan dengan membandingkan pencapaian teknik tersebut dengan teknik kejuruteraan ciri yang sedia ada. Merujuk kepada eksperimen, laporan autopsi forensik terdiri dari enam belas perbezaan CoD yang diperolehi dari salah sebuah hospital terbesar di Kuala Lumpur, Malaysia. Laporan ini telah diproses terlebih dahulu dengan menggunakan pelbagai teknik pra-pemprosesan teks. Seterusnya, ciri-ciri diskriminatif telah diekstrak dari laporan pra-proses dengan menggunakan teknik kejuruteraan ciri yang telah dicadangkan dan telah membentuk vektor ciri induk numerik. Ciri utama vektor tersebut kemudiannya digunakan sebagai input kepada enam algoritma pembelajaran mesin untuk membina dan menilai model klasifikasi. Tambahan pula, untuk memastikan keberkesanan teknik-teknik yang telah dicadangkan, hasil keputusan teknik-teknik tersebut telah dibandingkan dengan lima teknik kejuruteraan ciri asas terkini. Keputusan eksperimen menunjukkan bahawa teknik-teknik yang telah dicadangkan telah mengatasi BoW tradisional dan teknik variannya. Selain itu, "*Support vector machine*", dan "*Random Forest*" telah mengatasi empat lagi algoritma. Teknik-teknik yang telah dicadangkan adalah boleh dicapai dan praktikal untuk menentukan CoD dari laporan autopsi forensik dan boleh membantu ahli patologi untuk menentukan CoD dengan tepat dan cepat dari penemuan autopsi. Kesimpulannya, teknik-teknik yang telah dicadangkan pada umumnya boleh digunakan untuk laporan klinikal bebas teks yang lain.

Kata Kunci: Teknik Klasifikasi Teks Automatik, Laporan Autopsy Forensik, Algoritma Pembelajaran Mesin yang Diselia, Teknik Kejuruteraan Ciri, Laporan Klinikal Bebas Teks

ACKNOWLEDGEMENTS

First, I would extend my thanks and immense gratitude to Allah for endowing me the strength, wisdom, and endless blessings to do my PhD. I would like also to sincerely express my deepest gratitude to my supervisors Dr. Nor Liyana Mohd Shuib and Dr. Ram Gopal Raj for their precious support, supervision, encouragement and inspirations to me during my PhD journey. Their continuous support and guidance helped me producing a valuable piece of research conveyed in this thesis. I would also like to extend my heartiest appreciation to my employer Sukkur IBA University, Sukkur, Pakistan, who has supported my PhD studies throughout the tenure.

As well, I would like to extend heartiest appreciation to my parents whom have sacrificed their time, efforts to train me to become a person of importance and value to the society. They are always offering me continuous help and support in every oppressive and repressive moment. My gratitude to my darling wife Dr. Khairunisa Shaikh, my lovely sons Ahmad Mujtaba, and Abbas Mujtaba, and Sister Kalsoom Rani Shaikh for their patience, prayers and love throughout my studies. Thank you for your inspirations and motivations.

This thesis is dedicated to my grandfather Late Haji Mohammad Urs Shaikh, my father Mr. Noorullah Shaikh, my mother Zaibulnisa Shaikh, my beloved wife, and my employer Sukkur IBA University for their endless support and motivation. Special thanks to my beloved wife for her emotional support and unconditional love throughout the journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

χ^2	:	Chi-Square
ANN	:	Artificial Neural Network
ATC	:	Automated Text Classification
AUC	:	Area Under the Curve
BNSS	:	Bi-Normal Separation Score
BoC	:	Bag of Concepts
BoP	:	Bag of Phrases
BoW	:	Bag of Words
BP	:	Basic Preprocessing Tasks
BR	:	Binary Representation
CBR	:	Case-based Reasoning
CGFE	:	Conceptual Graph-based Feature Engineering
CNN	:	Convolutional Neural Network
CoD	:	Cause of Death
DFS	:	Distinguishing Feature Selector
DT	:	Decision Tree
EBMC	:	Efficient Bayesian Multivariate Classification
ED	:	Expert-Driven
ESA	:	Entropy Scoring Algorithm
FN	:	False Negative
FP	:	False Positive
GA	:	Genetic Algorithm
GI	:	Gini-Index
GoW	:	Graph of Words
ICD-10	:	International Classification of Disease Tenth Edition

ICD-9	:	International Classification of Disease Ninth Edition
IG	:	Information Gain
*k*NN	:	*k*-Nearest Neighbour
LM	:	Lemmatization
LR	:	Linear Regression
LSFS	:	Local Semi-Supervised Feature Selection
MDA	:	Multiple Discriminant Analysis
MFV	:	Master Feature Vector
MI	:	Mutual Information
MoD	:	Manner of Death
NB	:	Naive Bayes
NLTK	:	Natural Language Toolkit
NM	:	Normalization
N-TFiDF	:	Normalized TFiDF
PC	:	Pearson Correlation
PCA	:	Principal Component Analysis
PoS	:	Parts of Speech
RB	:	Rule-based
RF	:	Random Forest
SC	:	Spell Correction
Short Forms		Full Forms
SM	:	Stemming
SML	:	Supervise Machine Learning
SNOMED CT	:	Systematized Nomenclature of Medicine -- Clinical Terms
ST	:	Sentence Tokenization
SVM	:	Support Vector Machine
TF	:	Term Frequency

TFiDF	:	Term Frequency with Inverse Document Frequency
TN	:	True Negative
TP	:	True Positive
WT	:	Word Tokenization

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

This chapter discusses the study background and underlying motivation. It also presents the problem statement, followed by the research aim, objectives, and research questions. This chapter briefly describes the general research design, scope, contribution, and significance of the study. Finally, it states the thesis organization.

1.1 Background

The widespread implementation of electronic databases has improved the accessibility of plaintext clinical information for supplementary use. Numerous automated text classification (ATC) techniques have been employed to obtain useful information from free-text clinical data (Lin et al., 2013; Chomutare, 2014; de la Iglesia et al., 2014; Marafino, Davies, Bardach, Dean, & Dudley, 2014; Abacha et al., 2015; Iqbal et al., 2015; Koopman, S. Karimi, et al., 2015; Koopman, G. Zuccon, A. Nguyen, A. Bergheim, & N. Grayson, 2015; Sarker & Gonzalez, 2015). ATC refers to the task of automatically classifying text documents into one or more predefined categories (Meadow, 1992; Aggarwal & Zhai, 2012a). For instance, Koopman, G. Zuccon, et al. (2015) used ATC techniques to determine cancer-related causes of death (CoDs) from death certificates. Moreover, Jouhet et al. (2012) developed an intelligent tool by using ATC techniques to automatically categorize pathology reports. Schuemie et al. (2012) adopted ATC techniques to classify epidemiological studies. These aforementioned works provide concrete proof that ATC techniques are suitable for classifying free-text clinical reports, such as epidemiological reports, cancer-related reports, and pathology reports. However, these ATC techniques have been rarely adopted to classify the CoDs from forensic autopsy reports.

Forensic autopsy (also known as postmortem) is a surgical procedure that involves the external and internal examinations of a deceased body to determine the manner of death

(MoD) and CoD. Forensic autopsies are usually performed by a team of experts (known as pathologists) during criminal and civil law case investigations. During external examinations, experts collect external body information from head to toe. The experts also gather information on signs of postmortem changes, recent medical therapy, and injuries. During internal examinations, experts record information on the body's internal organs, such as the brain, heart, kidneys, and abdomen. Death scene information, eyewitness information, and deceased history-related data are also gathered to determine the MoD and CoD.

In general, a forensic autopsy examination usually lasts for 2–4 h depending on case complexity (James, Nordby, & Bell, 2002; Vij, 2014). After forensic autopsy examination, the team of experts correlate the autopsy findings with the medical history, premortem and postmortem laboratory studies, microscopic tissue findings, toxicology, other related medical procedures and documents, and similar relevant past cases to determine the MoD and CoD in accordance with the International Classification of Disease Tenth Edition (ICD-10) coding system (Organization, 1979). Hence, the main purpose of forensic autopsy examination is to determine the MoD and CoD. Finally, forensic autopsy examination culminates in generating forensic autopsy reports.

The initial version of the forensic autopsy report is prepared within 2–3 days. However, final reports become available between 30 and 45 days, and some complex cases may take up to 90 days (James et al., 2002; Vij, 2014). Therefore, determining the MoD and CoD is laborious and time consuming after forensic autopsy examination. Moreover, assigning CoDs by labor-intensive processes is subject to inconsistencies (Hoelz, Ralha, & Geeverghese, 2009; Vij, 2014). Hence, ATC techniques are needed to predict the MoD and CoD from complex autopsy reports to minimize time, labor, and irregularities.

In general, ATC can be commonly performed using five different techniques namely, ontology-based technique, rule-based technique, semi-supervised machine learning-based technique, unsupervised machine learning-based technique, and supervised machine learning-based (SML) technique (Manning, Raghavan, & Schütze, 2008; Aggarwal & Zhai, 2012a). However, of these techniques, the SML-based technique is the most widely used for classifying text documents (Aggarwal & Zhai, 2012a; Witten, Frank, Hall, & Pal, 2016).

The SML-based ATC technique comprises five main steps. First, a set of labeled text documents is prepared by labeling all the documents with their respective classes or categories. Second, several text preprocessing steps (such as replacing special characters and punctuation marks with spaces, normalizing case, removing duplicate characters, removing user-defined or built-in stop-words, and word stemming) are applied on collected text documents to clean the collected text documents. Third, the content of the text document is converted into useful word features by using state-of-the-art feature engineering techniques. The outcome of this step is the numeric master feature vector (MFV). In numeric MFV, rows and columns represent documents and features, respectively. Fourth, this MFV is fed to various state-of-the-art machine learning algorithms (such as support vector machine [SVM], random forest, naïve Bayes [NB], and decision trees) to construct the classification model. Finally, the performance of the constructed model is evaluated on unlabeled text documents (also known as test set).

Of these five steps, feature engineering is the key in the SML-based ATC technique (Wolpert & Macready, 1995; Aggarwal & Zhai, 2012a; Domingos, 2012; Witten et al., 2016). The success or failure of any text classification model heavily depends on the quality of features used in a classification task. If the extracted features correlate well with the class, classification becomes easy and accurate. By contrast, if the extracted features do not correlate well with the class, then the classification task becomes difficult and less accurate.

The traditional feature engineering technique is bag of words (BoW) (Meadow, 1992; Aggarwal & Zhai, 2012a). In this technique, the free-text clinical report is represented as the bag of its words, in which grammar and word order are disregarded but word frequency is maintained (Figueiredo et al., 2011; Papadakis, Giannakopoulos, & Paliouras, 2016; Passalis & Tefas, 2016). The BoW technique ignores the word context in a text when applied to free-text clinical reports (such as forensic autopsy reports). Therefore, effective (highly accurate and consuming minimal computational time and resources) feature engineering techniques are required to overcome the limitations of the traditional BoW technique for classifying free-text clinical reports (such as forensic autopsy reports).

1.2 Research Motivation

Many hospitals use electronic database systems, where autopsy findings and reports are stored in free-text format. Occasionally, experts utilize and correlate these previously stored autopsy reports to solve future cases. However, the greatest challenges in performing autopsy are lack of human resources and insufficient investigation time to determine the MoD and CoD (Hoelz et al., 2009). Therefore, ascertaining the MoD and CoD on the basis of the ICD-10 coding system is laborious and time consuming after autopsy examination. Assigning CoDs by labor-intensive processes is subject to inconsistencies. The automatic prediction of the MoD and CoD from complex autopsy reports by ATC techniques can minimize time, labor, and irregularities. These reports must be converted into actionable information that can be exploited by pathologists to accurately and rapidly determine the MoD and CoD.

Danso, Atwell, and Johnson (2013) adopted SML-based ATC techniques to classify the CoD using verbal autopsy reports. In the present study, the authors used the traditional BoW feature engineering technique to extract useful features from verbal autopsy reports. The authors used the SVM algorithm to classify CoDs and achieved 58.7% prediction accuracy. Yeow, Mahmud, and Raj (2014) adopted case-based reasoning (CBR) coupled with the NB algorithm and BoW feature engineering technique to support decision making in forensic autopsy reports to classify the CoD. Experimental results showed 80% prediction accuracy. Mujtaba, Shuib, Raj, Rajandram, and Shaikh (2016) applied the traditional BoW feature engineering technique to determine the MoD and CoD from forensic autopsy reports and achieved 78% classification accuracy.

In all three aforementioned studies (Danso et al., 2013; Yeow et al., 2014; Mujtaba et al., 2016), authors employed the traditional BoW feature engineering technique to extract useful features from autopsy reports. However, the obtained low prediction accuracy demonstrated that the traditional BoW feature engineering technique is inappropriate for excavating discriminative features from autopsy reports. The traditional BoW technique exhibits poor performance because it ignores the word context and order in free-text autopsy reports.

Several variants and extensions of the BoW technique have been proposed in the literature. Examples include the *n*-gram technique (Cavnar & Trenkle, 1994; Sebastiani, 2002), skip-gram technique (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), continuous BoW (CBoW) technique (Wang, 2014), and entropy-optimized BoW (EO-BoW) technique (Passalis & Tefas, 2016) to capture some word dependencies and word order. However, the aforementioned techniques involve word sequences and fail to capture word inversion and subset matching (e.g., "*information about deceased*" versus "*deceased information*") (Joachims, 1998b; Sebastiani, 2002; Papadakis et al., 2016) in classifying autopsy reports.

These aforementioned feature engineering techniques cannot handle complex semantic information, such as word-level synonymy and polysemy (Yadav, Sharan, & Joshi, 2014; Malliaros & Skianis, 2015; Dasondi, Pathak, & Singh, 2016; Jiang, Li, & Huang, 2016; Papadakis et al., 2016), when applied to autopsy reports to classify forensic autopsy reports. For instance, in many reports, pathologists use the word "*heart attack*" and "*myocardial infarction*" interchangeably. Hence, in this case, traditional BoW or variants of the BoW techniques fail to capture this word-level synonymy. Effective (highly accurate and consuming minimal computational time and resources) feature engineering techniques remain to be developed to capture the word order, word context, and word-level synonymy and polysemy for classifying free-text clinical reports (such as forensic autopsy reports). This strategy would enhance the accuracy of classifying CoDs from forensic autopsy reports.

1.3 Problem Statement

Danso et al. (2013), Yeow et al. (2014), and Mujtaba et al. (2016) applied the conventional BoW and *n*-gram (Cavnar & Trenkle, 1994) feature engineering techniques to discover result-oriented features from autopsy reports to determine CoD. However, the obtained low prediction accuracy demonstrated that these conventional feature engineering techniques are not suitable for discovering the discriminative features from autopsy reports because these techniques ignore the word context and order in free-text autopsy reports. To overcome the limitations of BoW and *n*-gram techniques, many researchers have proposed state-of-the-art feature engineering techniques as variants of the BoW and *n*-gram feature engineering techniques. Examples of such techniques include skip-gram (Mikolov et al., 2013), CBoW (Wang, 2014), and EO-BoW (Passalis

& Tefas, 2016), which can capture word dependency and word order. However, the aforementioned techniques involve word sequences and fail to capture word inversion and subset matching (e.g., “*information about deceased*” vs. “*deceased information*”) (Chakravarthy, Venkatachalam, & Telang, 2010; Danso et al., 2013; Yeow et al., 2014; Mujtaba et al., 2016; Papadakis et al., 2016) in classifying CoDs from forensic autopsy reports. Moreover, these aforementioned feature engineering techniques cannot consider word-level synonymy and polysemy (e.g., “*heart attack*” vs. “*myocardial infarction*”) in classifying CoDs from forensic autopsy reports (Koopman, G. Zuccon, et al., 2015; Mujtaba et al., 2016; Nam, Kim, Kim, Ngo, & Zong, 2016; Papadakis et al., 2016; Passalis & Tefas, 2016). Therefore, effective feature engineering techniques remain to be developed to overcome the limitations of existing feature engineering techniques and improve the performance in classifying free-text clinical reports (such as forensic autopsy reports for categorizing CoDs).

1.4 Aim and Objectives

This research primarily aimed to determine the feasibility of predicting MoDs and CoDs from free-text forensic autopsy reports by using SML-based ATC techniques. This research also proposes effective feature engineering techniques to overcome the limitations of the traditional BoW feature engineering technique and its variants to accurately and efficiently classify MoDs and CoDs from forensic autopsy reports and assist pathologists. In this thesis, the phrase “effective feature engineering techniques” refers to the techniques that achieve higher accuracy than existing feature engineering techniques and consume minimal computational time and resources. Thus, to accomplish the above objectives, this research has the following objectives:

1. To investigate the existing feature engineering techniques for classifying free-text clinical reports, including forensic autopsy reports.

2. To develop an effective semi-automated expert-driven feature engineering technique for addressing the issue of word-level synonymy and polysemy to classify CoDs from forensic autopsy reports.
3. To develop an effective fully automated conceptual graph-based feature engineering technique to address the issues of word order, word context, and word-level synonymy and polysemy in the text to classify CoDs from forensic autopsy reports.
4. To evaluate the performance of the proposed feature engineering techniques by using real-world forensic autopsy report datasets and by comparing the performance of the proposed techniques with those of baseline feature engineering techniques.

1.5 Research Questions (RQs)

The RQs of each objective are given below.

Objective 1: To investigate the existing feature engineering techniques for classifying free-text clinical reports, including forensic autopsy reports.

RQ1: What are the existing feature engineering techniques for classifying free-text clinical reports?

RQ2: How feasible are the existing feature engineering techniques in terms of their performance in classifying forensic autopsy reports and determining the MoDs and CoDs from free-text forensic autopsy reports?

RQ3: What are the limitations of the existing feature engineering techniques in determining the MoDs and CoDs from free-text forensic autopsy reports?

Objective 2: To develop an effective semi-automated expert-driven feature engineering technique for addressing the issue of word-level synonymy and polysemy to classify CoDs from forensic autopsy reports.

RQ4: How much of the classification performance of forensic autopsy reports can be enhanced through the proposed semi-automated expert-driven feature engineering technique?

RQ5: How important is the proposed semi-automated feature engineering technique in classifying forensic autopsy reports?

RQ6: What are the limitations of the proposed semi-automated expert-driven feature engineering technique?

Objective 3: To develop an effective fully automated conceptual graph-based feature engineering technique to address the issues of word order, word context, and word-level synonymy and polysemy in the text to classify CoDs from forensic autopsy reports.

RQ7: How much of the classification performance of the forensic autopsy reports can be enhanced through the fully automated feature engineering technique without human expert intervention?

RQ8: How can graph theory concepts be exploited in obtaining word order and word context from free-text forensic autopsy reports?

RQ9: How can existing medical or clinical ontologies be utilized to extract word-level synonymy and polysemy from free-text forensic autopsy reports?

Objective 4: To evaluate the performance of the proposed feature engineering techniques by using real-world forensic autopsy reports and by comparing the proposed techniques' performance with those of baseline feature engineering techniques.

RQ10: How can the performance of the proposed feature engineering techniques be evaluated?

RQ11: How much of the performance of the proposed feature engineering techniques is improved relative to those of the conventional and state-of-the-art feature engineering techniques?

RQ12: Which of the proposed semi-automated expert-driven and fully automated conceptual graph-based feature engineering techniques is effective?

1.6 Research Methodology and Design

The general research design of this study is shown in Figure 1.1. As presented, this research consists of five main phases. These phases are discussed briefly in the subsequent sections. The specific research design is found in Chapter 3.

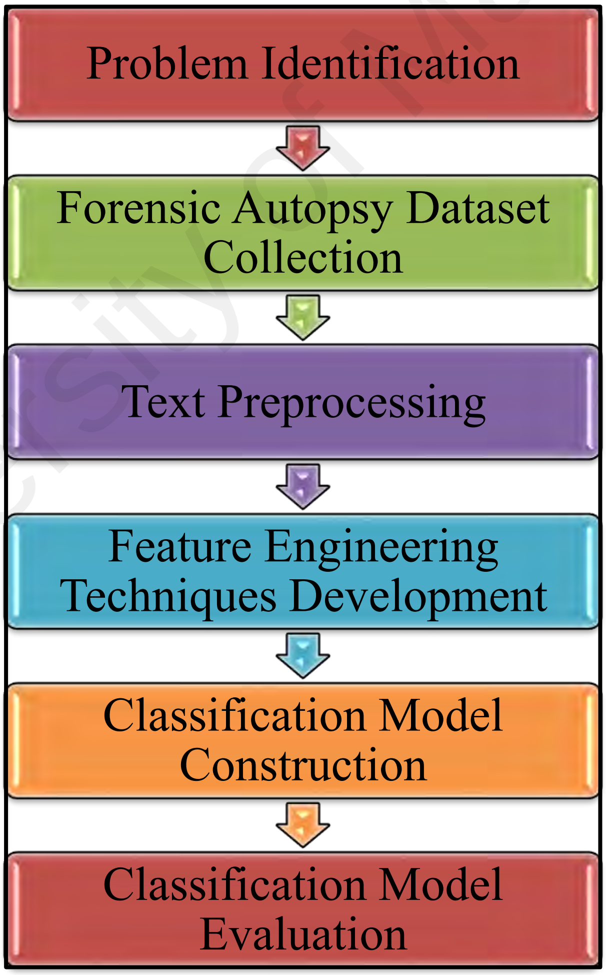

Problem Identification
Forensic Autopsy Dataset Collection
Text Preprocessing
Feature Engineering Techniques Development
Classification Model Construction
Classification Model Evaluation

Figure 1.1: Research methodology and design

1.6.1 Problem Identification

This step identifies the research problem by reviewing the existing literature on the classification of clinical text and forensic autopsy reports. This step also involves a comparative study of various existing SML-based ATC techniques to classify CoDs from the collected autopsy reports. The results of this comparative study are precisely discussed in Chapter 2.

1.6.2 Forensic Autopsy Dataset Collection

This step discusses the real-world forensic autopsy dataset used in the present experiments. This dataset was collected from a well-known emergency hospital situated at Kuala Lumpur, Malaysia. The collected dataset included the forensic autopsy reports involving four MoDs and 16 CoDs. The details of this step are presented in Chapter 3.

1.6.3 Text Preprocessing

This step considers the various text preprocessing tasks applied on the collected forensic autopsy dataset to remove the irrelevant features. These tasks include removing stop-words, eliminating punctuation marks, converting to lower case, tokenization, stemming, and lemmatization. This step is elaborated in Chapter 3.

1.6.4 Feature Engineering Technique Development

In this step, two effective feature engineering techniques are developed to extract discriminative features from the collected dataset. As discussed in Section 1.1, in the SML-based ATC technique, the key step is feature engineering. Therefore, the accuracy of any machine learning algorithm heavily depends on the quality of features extracted through the feature engineering technique. The traditional feature engineering technique BoW and its variants suffer from three major weaknesses, namely, ignoring word order, word context, and word-level synonymy and polysemy. To overcome these limitations, this thesis proposed and developed two effective feature engineering techniques, namely, semi-automated expert-driven feature engineering technique and fully automated conceptual graph-based feature engineering technique, for this step. The details of these techniques are given in Chapters 4 and 5. The proposal of the two techniques is justified in Chapter 2 (Sections 2.8.1 and 2.11.1). In addition, the difference between both the proposed techniques is also shown in Table 5.8, Chapter 5, Section 5.5.

1.6.5 Classification Model Construction

This step constructs an effective classification model through various machine learning algorithms (such as SVM and NB) coupled with proposed feature engineering techniques to classify CoDs from forensic autopsy reports. The specifics of this step are discussed in Chapters 4 and 5.

1.6.6 Classification Model Evaluation

This step evaluates the performance of the constructed classification models by using *k*-fold cross validation. Four evaluation metrics are used to measure the performances of the constructed classification models. These metrics are macro precision, macro recall, macro F-measure, and overall accuracy. This step also evaluates the performance of the proposed feature engineering techniques with the state-of-the-art baseline feature engineering techniques. Finally, this step identifies the best classification model via proposed feature engineering techniques and machine learning algorithms. The detailed results are discussed in Chapters 4 and 5.

1.7 Research Scope

This study was conducted on the basis of certain delimitations for increased focus. The boundaries were as follows:

- ❖ The forensic autopsy corpus collected for this research belonged to four MoDs and 16 CoDs. Despite the thousands of CoDs, the generalizability of the proposed techniques in theory should be applied to all other CoDs. In addition, it is believed that the proposed techniques have potential to show the similar classification performance on more CoDs.

- The proposed feature engineering techniques have been evaluated on the forensic autopsy dataset. However, the generalizability of the proposed techniques in theory and practice should classify other kinds of free-text clinical reports, such as verbal autopsy reports, pathology reports, radiology reports, and other related clinical report types.
- The prediction model requires complete forensic autopsy findings and previously stored forensic autopsy reports as input to determine the MoDs and CoDs.

1.8 Research Contribution

The contributions of this research to the current literature are as follows:

- **Literature Analysis:** The conducted literature review exposed the weaknesses of existing feature engineering techniques. ATC techniques that classify free-text clinical reports, including forensic autopsy reports, were comprehensively reviewed by exploiting the procedural decision analysis in six aspects, namely, the types and characteristics of clinical reports and datasets, preprocessing techniques, feature engineering techniques, machine learning algorithms, and performance metrics.
- **Free-text Forensic Autopsy Corpus:** This research built the corpus of complete forensic autopsy reports to be used as resource for clinical text classification research.
- **Comparative Study of SML-based ATC Techniques:** This research conducted a comparative study of SML-based ATC techniques to classify CoDs from forensic autopsy reports. Moreover, the study identified the suitable techniques at

various phases of the SML-based ATC process to classify CoDs from forensic autopsy reports.

- **Effective Semi-Automated Expert-driven Feature Engineering Technique:** This research proposed and developed an effective semi-automated expert-driven feature engineering technique to classify CoDs from forensic autopsy reports. This technique aims to overcome the limitation of world-level synonymy and polysemy in existing baseline feature engineering techniques. This technique also provides the benchmark classification performance for other fully automated feature engineering techniques that will be developed to classify forensic autopsy reports in the near future.
- **Effective Fully Automated Conceptual Graph-based Feature Engineering Technique:** This research developed an effective fully automated conceptual graph-based feature engineering technique for classifying CoDs from forensic autopsy reports. This work also aimed to overcome the limitations (such as word order, word context, and world-level synonymy and polysemy) of existing baseline feature engineering techniques. This technique was built to compete with the semi-automated expert-driven method and overcome the limitations of the semi-automated expert-driven technique.

All the proposed techniques in this thesis have been published in reputable ISI-indexed journals and A-rank conferences organized in the United States of America (Refer to page 202-203 for the list of publications).

1.9 Research Significance

The significance of this research was observed in two domains, namely, forensic autopsy research community and SML-based ATC research community.

1.9.1 Research Significance in Forensic Autopsy Research Community

As mentioned in Section 1.1, determining the MoD and CoD from forensic autopsy findings is laborious and time consuming. The massive effort and duration involved are due to the need for experts to correlate the autopsy findings with the death scene information, histopathology and toxicology reports, and similar past cases. Experts may also assign incorrect and inconsistent CoDs. Currently, no computer-aided expert system can process free-text forensic autopsy findings to ascertain the MoD and CoD. Hence, this domain requires a computer-aided expert system that assists pathologists in determining the MoD and CoD from forensic autopsy reports accurately and efficiently. To address this need, this study developed an effective classification model that can assist experts in determining the MoD and CoD. The developed classification models can also reduce the time and effort spent in ascertaining the MoD and CoD from forensic autopsy reports. Finally, this research introduced a new focus in this area that can be expanded further.

1.9.2 Research Significance in ATC and SML Research Community

This research applies SML-based ATC techniques to classify MoDs and CoDs from forensic autopsy reports. The ATC techniques were applied because forensic autopsy reports are usually prepared in an unstructured free-text format. As presented in Section 1.1, five popular means can be used to classify free-text clinical reports. These methods are ontology-based, rule-based, semi-supervised, unsupervised, and SML-based ATC techniques. This research classifies forensic autopsy reports via SML-based ATC techniques because of the methods' popularity and beneficial results (Manning et al., 2008; Aggarwal & Zhai, 2012a).

In SML-based ATC techniques, the most important step is feature engineering (Wolpert & Macready, 1995; Aggarwal & Zhai, 2012a; Domingos, 2012; Witten et al.,

2016) because the success or failure of any text classification model is heavily dependent on the quality of features used in the classification task (Domingos, 2012). The traditional feature engineering technique for ATC is BoW. In this technique, the textual document is represented as the bag of its words, in which grammar and word order are disregarded but word frequency is maintained. Hence, the BoW technique ignores the word context in the text (Figueiredo et al., 2011; Papadakis et al., 2016; Passalis & Tefas, 2016). Mujtaba et al. (2016) applied the traditional BoW feature engineering technique to classify CoD from forensic autopsy reports and achieved 78% classification accuracy. To improve the classification accuracy of forensic autopsy reports, this research proposed two effective feature engineering techniques (namely, semi-automated expert-driven technique and fully automated conceptual graph-based technique) to overcome the limitations of traditional BoW and state-of-the-art variants of traditional BoW techniques. The feasibility of these techniques was established with promising results and formed the basis for further ATC and SML research within the context of forensic autopsy reports.

1.10 Thesis Overview

The remaining structure of this thesis is organized as following:

Chapter 2: This chapter discusses the survey of autopsy reports. A brief overview of field is given, which describes the process involves in gathering the autopsy findings; purpose of autopsy; types of autopsy reports; issue in determining the MoD and CoD manually; and the need of automation for determining MoD and CoD from autopsy reports. Moreover, this chapter presents the literature review of various SML-based ATC techniques, and feature engineering techniques used for classifying free-text clinical reports. In addition, this chapter also presents a literature analysis on the various SML-based ATC techniques, and feature engineering techniques that currently exist in carrying out analysis and predicting the CoD from autopsy reports, which informs the formulation of this research.

Chapter 3: This chapter presents the methodology used in this thesis to develop the proposed feature engineering techniques for classifying MoD and CoD from forensic autopsy reports. Moreover, this chapter discusses in detail the dataset used for the experiments. Moreover, it also discusses the various text preprocessing steps which were applied on collected dataset to remove noisy features. Furthermore, it discusses briefly the proposed feature engineering techniques, machine learning algorithms that were used for constructing the classification models, and performance metrics that were used to measure the performance of classification models. Finally, it also discusses the baseline feature engineering techniques that were used as a benchmark to evaluate the performance of proposed feature engineering techniques.

Chapter 4: This chapter explains in detail the proposed semi-automated expert-driven feature engineering technique for classifying forensic autopsy reports. Moreover, it also discusses the experimental setup, results obtained through semi-automated expert-driven technique, and discusses the findings.

Chapter 5: This chapter explains in detail the proposed fully-automated conceptual graph-based feature engineering technique for classifying forensic autopsy reports. Moreover, it also discusses the experimental setup, results obtained through fully-automated conceptual graph-based technique, and discusses the findings.

Chapter 6: This chapter concludes the thesis by reappraising the research objective. The main contributions are summarized. It discusses the limitations of the research and proposes future directions.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter presents a review of existing related literature on clinical text classification, including the classification of forensic autopsy reports. The discussion begins with the description, purpose, and types of autopsy reports in Section 2.2. In Section 2.3, some fundamental automated text classification techniques are presented. In Section 2.3.1, the detailed process of the supervised machine learning-based automated text classification approach is discussed. Sections 2.4–2.10 present a comprehensive review of automated text classification techniques for free-text clinical reports, including forensic autopsy reports. Section 2.11 enumerates the potential limitations and challenges in classifying free-text clinical reports, including forensic autopsy reports. Section 2.12 identifies the research gap for this thesis. Finally, Section 2.13 concludes this chapter. The overall organization of this chapter is depicted in Figure 2.1.

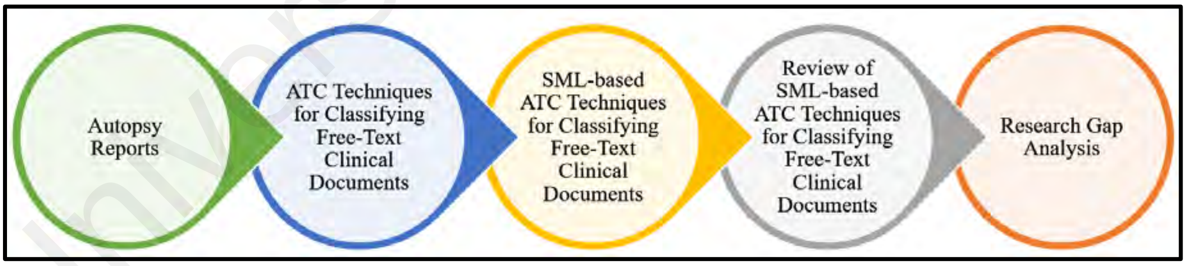

Autopsy Reports
ATC Techniques for Classifying Free-Text Clinical Documents
SML-based ATC Techniques for Classifying Free-Text Clinical Documents
Review of SML-based ATC Techniques for Classifying Free-Text Clinical Documents
Research Gap Analysis

Figure 2.1: Organization of Chapter 2

2.2 Autopsy Reports

An autopsy (also known as postmortem) is a surgical procedure that involves the external and internal examinations of deceased body to determine the Manner of Death (MoD), and Cause of Death (CoD). During external examinations, experts collect external body information from head to toe. Moreover, experts collect information on signs of postmortem changes, signs of recent medical therapy, and injuries. During internal examinations, experts collect information about the body's internal organs, such as the brain, heart, kidneys, and abdomen. In addition, death scene information, eye witness information, and deceased history related information is also gathered to determine the MoD and CoD. Autopsies are performed by pathologists during criminal and civil law case investigations. By and large, the autopsy examination usually takes 2 to 4 hours depending on case complexity (James et al., 2002; Vij, 2014). After the autopsy examination, the team of experts correlate the autopsy findings with medical history, premortem and postmortem laboratory studies, microscopic findings of tissues, toxicology, other related medical procedures and documents, and similar relevant past cases to determine the MoD, and CoD in accordance with the International Classification of Disease Tenth Edition (ICD-10) coding system (Organization, 1979). Finally, Autopsy examination culminates in the generation of autopsy reports.

There are three kinds of autopsy reports (as shown in Figure 2.2) namely, clinical autopsy reports, forensic autopsy reports, and verbal autopsy reports (Costache et al., 2014; Vij, 2014; Miasnikof et al., 2015a). Clinical autopsy is performed to discover the medical CoD. Clinical autopsy is usually conducted in situations of uncertain deaths. Thus, preventive actions should be carried out to avoid such incidents in future. Forensic autopsy is performed to discover the CoD in criminal matter. In verbal autopsy, an interview is conducted from the relatives or witnesses of the deceased person to discover the CoD. This method is common in low economical countries, where health facilities are insufficient. The initial version of autopsy report is prepared within 2 to 3 days. However, final reports become available between 30 to 45 days, and some complex cases may take up to 90 days (James et al., 2002; Vij, 2014). Therefore, it is laborious and time consuming to determine the MoD and CoD after autopsy examination. Moreover, it is subject to inconsistencies to assign CoDs with labor intensive processes (Hoelz et al., 2009). Hence, the automatic prediction of MoD and CoD through Automated Text

Classification (ATC) techniques is needed to minimize time, labor, and irregularities. The ATC process is discussed in subsequent section (Section 2.3).

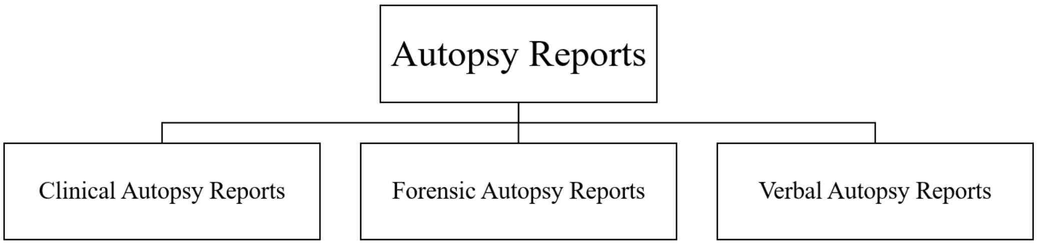


Figure 2.2: Types of autopsy reports (Costache et al., 2014)

2.3 Automated Text Classification: The Biomedical Domain

The extensive number of electronic health records contain useful information in free-text format. Thus, to be able to assist in medical decision-making process, this free-text is need to efficiently processed and transformed into useful and structured format. It has long been recognized that free-text clinical reports are beneficial for secondary use. A number of researchers across the globe employed clinical text mining to mine useful information (such as medical concepts or medical entity) from free-text clinical reports (Wang, E. Coiera, W. Runciman, & F. Magrabi, 2017; Wu & Wang, 2017; Yoon, Roberts, & Tourassi, 2017; Parlak & Uysal, 2018). There are various applications of clinical text mining including, clinical information extraction (Xu, Stenner, et al., 2010), clinical relation extraction (Porumb, Barbantan, Lemnaru, & Potolea, 2015; Barbantan, Porumb, Lemnaru, & Potolea, 2016), clinical document clustering (Ko & Seo, 2000; Renganathan, 2017), biomarkers identification (Vangay, Steingrimsson, Wiedmann, & Stasiewicz, 2014; Gutierrez-Sacristan et al., 2017), disease surveillance detection (Al-garadi, Khan, Varathan, Mujtaba, & Al-Kabsi, 2016), and automated text classification (ATC) of clinical documents (Meadow, 1992; Aggarwal & Zhai, 2012a).

ATC of clinical documents is one of the eminent research area inside clinical text mining domain (Kaurova, Alexandrov, & Blanco, 2011; Holzinger, Schantl, Schroettner, Seifert, & Verspoor, 2014; Spasić, Livsey, Keane, & Nenadić, 2014). It refers to the task of automatically classifying free-text clinical documents or reports into one or more than one predefined categories (Meadow, 1992; Aggarwal & Zhai, 2012a). In general, there are five different techniques to classify free-text clinical reports namely, supervised machine learning-based (SML-based) ATC (Hastie, Tibshirani, & Friedman, 2009), unsupervised machine learning-based ATC (Ko & Seo, 2000), semi-supervised machine learning-based ATC (Zhu & Goldberg, 2009; Settles, 2010), ontology-based ATC (Hotho, Maedche, & Staab, 2002), and rule-based ATC (Deng, Groll, & Denecke, 2015; MacRae et al., 2015). Of these, the most widely used approach for clinical text classification is the SML-based ATC (Sebastiani, 2002; Hastie et al., 2009; Witten et al., 2016). Moreover, these techniques obtained better classification results compared to other related techniques (Spasić et al., 2014; Al-garadi et al., 2016; Burger, Abu-Hanna, de Keizer, & Cornet, 2016; Kim & Delen, 2016). Section 2.3.1 discusses in detail the process of SML-based ATC technique.

2.3.1 Supervised Machine Learning-based ATC Techniques

In SML-based ATC technique, each clinical report $(d_1, d_2, d_3, ... d_n)$ belonging to certain free text clinical dataset (D) is labelled with pre-define class or category $(C_1, C_2, C_3, ... C_n)$. Several text preprocessing steps are applied on collected clinical dataset to remove non-discriminative or noisy content. Afterwards, various features are extracted from clinical reports and these extracted features are represented in numeric form to construct a numeric MFV $\begin{pmatrix} f_1, f_2, f_3, ... f_n \\ f_1, f_2, f_3, ... f_n \\ f_1, f_2, f_3, ... f_n \end{pmatrix}$ using feature representation techniques. Finally, this MFV is then fed as an input to a text classifier to develop a classification model. Often, all features in MFV do not contribute to better classification accuracy and also slowdown the classification process. Thus, to overcome these issues, the most powerful subset of features is selected from MFV by employing feature selection schemes. The steps of feature selection and classification model development are performed iteratively until the optimal learning curve is achieved with the help of suitable feature subset. Finally, the constructed model is evaluated on a separate unlabeled set of document (U_D). To summarize, the process of SML-based ATC technique mainly comprised of five steps namely, clinical reports dataset collection, clinical reports preprocessing, feature engineering, classification model development, and classification model evaluation. Figure 2.3 also depicts the process of SML-based ATC technique. Moreover, the detail of each step is presented in subsequent sections.

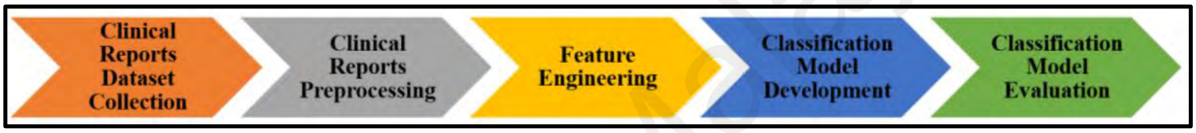

Clinical Reports Dataset Collection
Clinical Reports Preprocessing
Feature Engineering
Classification Model Development
Classification Model Evaluation

Figure 2.3: The Process of SML-based ATC technique

2.3.1.1 Clinical Reports Dataset Collection

As shown in Figure 2.3, in the first step the set of free text clinical reports is collected. Moreover, these reports are labelled with their respective classes or categories by domain experts. For instance, in the case of classifying autopsy reports according to MoD and CoD, these reports are first labelled by pathologists. The pathologists assign each autopsy report a unique MoD and CoD. The labelled set of clinical reports (more specifically autopsy reports) is known as training set. Thus, this training set is a pre-requisite for developing a classification model.

2.3.1.2 Clinical Reports Preprocessing

In SML-based ATC technique, preprocessing is one of the step that is responsible for removing the meaningless data from the collected dataset to improve the quality of clinical text classification models (Spasić et al., 2014; Witten et al., 2016). This step is followed after dataset collection. In general, the preprocessing techniques including, removal of stop-words, removal of punctuations or special symbols, removal of empty spaces, case conversion, spell correction, tokenization, stemming, lemmatization, and normalization are usually applied to remove the meaningless terms from the free-text clinical reports (Spasić et al., 2014; Witten et al., 2016). These techniques are briefly defined as follows;

- ***Removal of stop-words:*** This task removes the stop-words (such as *a*, *an*, *the*, etc.) from the clinical reports.
- ***Removal of punctuation or special symbol:*** This task removes the punctuation or any special symbol (such as hyphen, double quotes, single quotes, @, #, etc.) from the clinical reports.
- ***Removal of empty spaces:*** This task removed the white spaces (if any) from the clinical documents.
- ***Case conversion:*** This task converts all the words in q unique case (for instance, lower case).
- ***Spell correction:*** This task automatically corrects the spelling of texts available in the clinical reports.
- ***Tokenization:*** In tokenization the clinical text is tokenized into smaller units of text such as, sentences or words. In sentence tokenization, the input text is divided into unique sentences. Moreover, in word tokenization, input text is tokenized into unique word tokens.

Text Normalization techniques: Clinical reports may comprise of different forms of a word due to grammatical reason (such as, *organize*, *organized*, *organizes*). Thus, various text normalization techniques are applied to convert different forms of word into a common form. In clinical text classification, two most widely used text normalization techniques are stemming and lemmatization. The stemming technique converts different forms of a word into their stem or root form. Stemming is a word normalization technique that chops the ends of the words. For instance, the words *organizes*, and *organized* will be normalized into their root or stem word *organize*. For English language, variety of stemming algorithms are available such as, Porter stemmer, Snowball stemmer, Lovins stemmer, Dawson stemmer, and Husk stemmer (Paice, 1994; Jivani, 2011). In selected primary studies, most of the authors have employed Porters stemmer (Porter, 1980; Willett, 2006). Moreover, in literature Porters algorithm has constantly been revealed to effective empirically (Paice, 1994; Jivani, 2011). The lemmatization is also a word normalization technique that uses a vocabulary and morphological analysis of a given word and remove the inflectional endings of that word to convert it into a dictionary form. This process normalizes the words into basic forms. Lemmatization is similar to stemming, however, it does not require to produce the stem of the word but to replace the suffix of the input word with a (typically) different word suffix to produce its normalized form. For example, the word stemmer and word lemmatization will transform the words *worked*, *working*, *works* into word *work*, however, the words *computed*, *computing*, and *computes* will be normalized to *comput* by stemmer, and *compute* by lemmatization (Plisson, Lavrac, & Mladenić, 2004; Toman, Tesar, & Jezek, 2006). Besides, the stemming and lemmatization techniques, researchers also apply other types of text normalization techniques to clean clinical documents. For instance, if input text contains any number or dates then those numbers and dates will be converted to words *number* and *date* respectively.

2.3.1.3 Feature Engineering

In feature engineering, the content of clinical reports $\left(d_1, d_2, d_3, \ldots d_n\right)$ is converted into useful word features $\left(w_{f1}, w_{f2}, w_{f3}, \ldots w_{fn}\right)$ by using various feature engineering techniques. In SML-based ATC technique, the most important step is the feature engineering (Wolpert & Macready, 1995; Aggarwal & Zhai, 2012a; Domingos, 2012;

Witten et al., 2016). This is because, the success or failure of any clinical text classification model is heavily depending upon the quality of features used in the classification task. If the extracted features correlate well with the class, the classification will be easy and more accurate. In contrast, if the extracted features do not correlate well with the class, the classification task will be difficult and less accurate. Often, collected free-text clinical reports are not available in a form that is amenable to learning classification rules. Thus, to make these reports useful for clinical text classification task, various features are extracted from these reports and these extracted features are amenable to learning classification rules. In general, most of the clinical text classification effort is required by feature engineering step. Moreover, it is also one of the interesting step in clinical text classification process, where perception, innovation, intuition, creativity, and "black art" are equally important as the technical and subject knowledge. Often, the construction of classification model is the fastest step in clinical text classification, this is because feature engineering is responsible for extracting the discriminative features from free-text clinical reports and transforming those features into numeric master feature vector. This master feature vector is then used by classifier to learn the classification rules and develop a classification model quickly. Feature engineering is more difficult than classification because it is domain-specific and the classification task is general-purpose. Nonetheless, there is no sharp frontier between the two, and this is another reason the most useful learners are those that facilitate incorporating knowledge. Feature engineering steps is further sub-divided into three sub-steps namely, feature extraction, feature value representation, and feature selection. These are described in subsequent sub-sections.

(a) ***Feature Extraction***

Feature extraction is the process of extracting useful features from free-text clinical reports. In clinical text classification, a feature in an individual measureable characteristic of a phenomena being observed. Moreover, choosing an informative and discriminative features from free-text clinical reports is a crucial step for constructing of a classification model. Thus, most often researchers working in the field of clinical text classification paying more attention and efforts to discover the discriminative features that prove useful for classification model. In clinical text classification, the most commonly used features are Bag of Words (BoW), Bag of Phrases (BoP), *n*-gram, and Bag of Concepts (BoC).

- ***Bag of Words (BoW)***: In BoW model, the unique words are extracted from all clinical reports available in the dataset irrespective of their categories. All the extracted words are then sorted in ascending order and stored in a list called ‘bag of words (BoW)’. In BoW each available word represents an independent, and discriminative feature (Tong & Koller, 2001).
- ***n-gram***: An n-gram is the contiguous sequence of *n* items (such as words, or characters) from a given sequence of clinical text (Cavnar & Trenkle, 1994). They are typically a set of co-occurrence words within a given window. In *n*-gram, *n* maybe ‘1’, ‘2’, ‘3’, or any number. When the *n* = 1, it is called unigram, when *n* = 2, it is called bigram, and when *n* = 3, it is called as trigram.
- ***Bag of Phrases (BoP)***: In BoP, medical phrases are extracted from the clinical reports through the help of some tool such as, MetaMap (Aronson, 2001). For instance, suppose a sentence ‘*multiple grazed abrasions over the back of right hand*’ is available in a clinical report. The MetaMap tool will extract following phrases from the sentence (‘*multiple*’, ‘*grazed*’, ‘*abrasion*’, and ‘*Back of right hand*’).
- ***Bag of Concepts (BoC)***: Medical experts (such as physician or surgeon) may use different terms to describe same condition in free-text clinical reports. For instance, experts may use the term ‘heart attack’ or ‘Myocardial infarction’ interchangeably. Though, both the terms belong to same medical concept however, the feature extraction techniques cannot identify the relationship between these two terms.

Therefore, to overcome this issue, specialized medical ontologies are developed such as, SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms) (Stearns, Price, Spackman, & Wang, 2001; Donnelly, 2006). It is a standardized medical ontology where medical related terms are categorized into medical hierarchical concepts. The root concept is SNOMED CT concept and this concept has many child concepts such as medical conditions, body structures, procedures, etc. In SNOMECTD CT, each medical term has a unique concept id and similar medical terms share the same concept id. Moreover, each medical terms can belong to one or more concepts. For instance, the words 'heart attack' and 'Myocardial infarction' have the same concept id (22298006) and both belong to same parent concept '*Ischemic heart disease*' (concept id- 414545008). Thus, BoC includes the SNOMEDT CT concepts as a features.

(b) ***Feature Representation***

Feature representation, or term-weighing is vital in automatic text classification (Salton & Buckley, 1988). An important step after extracting features from clinical reports is transforming extracted features into numeric vectors for linear algebraic methods. Transforming clinical reports into numeric vectors is known as feature value representation or term-weighing (Debole & Sebastiani, 2004). In general, to classify free-text clinical reports, four types of feature representation techniques are used namely, binary representation (BR), term frequency (TF), term frequency with inverse document frequency (TFiDF),and normalized TFiDF (N-TFiDF) (Debole & Sebastiani, 2004). All these four techniques are briefly discussed in subsequent paragraphs.

- ***Binary Representation:*** In BR, the value of a feature can be '0' or '1', where '1' represents the occurrence of a feature in the document and '0' represents its non-occurrence (Salton & Buckley, 1988).

- ***Term Frequency:*** In TF, the frequency of a term is referred to as the occurrence of term in a document (Ramos, 2003). However, if all documents contain the same term with more or less the same frequency, then that term is not a discriminative feature (Aizawa, 2003). Therefore, a new TFiDF feature representation scheme was introduced to address this issue.
- ***Term Frequency with Inverse Document Frequency:*** The crux of TFiDF is that the term *t* is a discriminative feature if the term *t* frequently occurs in a particular document. Moreover, if the same term *t* appears frequently in many other documents, then term *t* is not a powerful feature for a given set of documents (Ramos, 2003).
- ***Normalized Term Frequency with Inverse Document Frequency:*** In N-TFiDF term frequency and document frequency are combined with a normalized factor such as, the length of the clinical reports to ensure features found in long and short clinical reports are equally important (Debole & Sebastiani, 2004).

(c) ***Feature Selection***

High dimensional data often contain irrelevant or redundant features, which causes some of the limitations in clinical text classification tasks. These limitations reduce the accuracy of the text classification algorithms, slow down the classification process, produce problems in storing and retrieving the information and make it harder to interpret the classification results. To overcome these limitations, feature selection techniques are often used. These techniques are responsible for selecting the most relevant subset of features for classification task according to some selection criteria. For reasons of both efficiency and efficacy, feature selection is widely used when applying SML-based ATC technique for clinical text classification. The most commonly used feature selection techniques are Information Gain (IG), the Chi-Square $\left(\chi^2\right)$ and Pearson correlation (PC). The brief introduction of these techniques is given below.

- **Information Gain (IG):** It identifies the importance of a given attribute in a feature vector, is the expected reduction in entropy caused by partitioning the examples according to a given attribute (Yang & Pedersen, 1997). If we have a set with k different values in it, we can calculate the entropy using Equation 2.1.

$$Entropy_{(Set)} = I_{(Set)} = -\sum_{i=i}^{k} P(Value_i).\log_2(P(Value_i)) \qquad (2.1)$$

Where *P*(Value*ᵢ*) is the probability of getting the ith value when randomly selecting one from the set.

So, for the set S = {s1, s1, s1, s2, s2, s2}

$$Entropy_{(R)} = I_{(R)} = -\left[\left(\frac{3}{8}\right)\log 2\left(\frac{3}{8}\right) + \left(\frac{5}{8}\right)\log 2\left(\frac{5}{8}\right)\right]$$

Let AT be the set of all attributes, TS the set of all training examples, V (X, A) is value of specific example X of attribute A in AT and H specifies the entropy. The information gain of an attribute A in AT set is defined as shown in Equation 2.2.

$$IG(TS, A) = H(TS) - \sum_{\upsilon \in values(A)} \left(\frac{\left|\{X \in TS \middle| Value(X, A) = \upsilon\}\right|}{|TS|} .H(\{X \in TS \mid Value(X, A) = \upsilon\}|) \right) \qquad (2.2)$$

- **Chi-Square Test (χ^2):** It is the statistical test that measures the relevance of term *t* with class *c* (Yang & Pedersen, 1997) by applying equation 2.3. Here, *tc* is the total number of times *t* and *c* appear together. The *tNc* is the number of times the *t* occurs without *c*. The *cNt* is the number of times *c* occurs without *t*. The *NtNc* is the number of times neither *t* nor *c* occurs together and *N* is the total number of documents. The χ^2 test has a natural value of zero if *t* and *c* are independent (Yang & Pedersen, 1997). The mathematical definition of χ^2 is shown in Equation 2.3.

$$Chi-Square(W, C) = \frac{N \times ((tc \times NtNc) - (cNt \times tNc))\,\hat{}\,2}{(tc + cNt) \times (tNc + NtNc) \times (tc + tNc) \times (cNw + NtNc)} \qquad (2.3)$$

- **Pearson correlation (PC):** Correlation-based feature selection is a commonly used method for reducing feature dimensionality and evaluating the discrimination power of a feature in classification methods. It is also a straightforward method for choosing significant features. Pearson correlation measures the relevance of a feature by computing the Pearson correlation between it and a class. Pearson correlation coefficient measures the linear correlation between two attributes (Benesty, Chen, Huang, & Cohen, 2009). The subsequent value lies between -1 and +1, with -1 implying absolute negative correlation (as one attribute increases, the other decreases), +1 denoting absolute positive correlation (as one attribute increases, the other also increases), and 0 meaning no linear correlation between the two attributes. For two attributes or features X and Y, Pearson correlation coefficient measures the correlation (Hall, 1999) as shown in Equation 2.4.

$$r_{xy} = \frac{\sum\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\left(n-1\right)S_x S_y} \tag{2.4}$$

where $\bar{x}$ and $\bar{y}$ are the sample means for X and Y, respectively, S_x and S_y are the sample standard deviations for X and Y, respectively, and n is the size of the sample used to compute the correlation coefficient (Hall, 1999).

2.3.1.4 Classification Model Development

The fourth step in SML-based ATC process is the construction of a free-text clinical reports classification model using machine learning algorithms. Numerous machine learning algorithms (such as, supervised, semi-supervised, unsupervised, and rule-based) have been applied to classify free-text clinical reports. Of these, most widely used machine learning algorithms are either rule-based, or SML-based algorithms (as shown in Figure 2.4) and are briefly discussed in subsequent paragraphs.

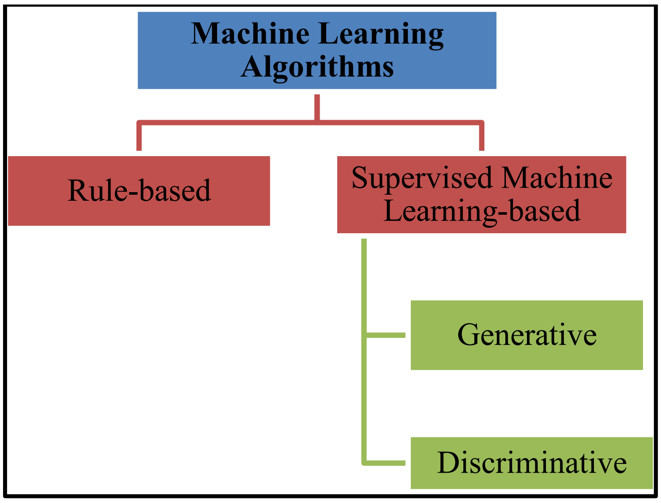

Machine Learning Algorithms
Rule-based
Supervised Machine Learning-based
Generative
Discriminative

Figure 2.4: Machine learning algorithms used in related literature

- ***Rule-based Algorithms:*** In rule-based (RB) algorithms, rules are either written manually, or generated automatically and then verified manually to save time (Sebastiani, 2002; Witten et al., 2016). The rule-based approach is simple and flexible where rules can be understood and can be improved over time unlike the supervised machine learning algorithms that work like a black-box system.
- ***SML-based Algorithms:*** These algorithms learn the classification rules from the features that were extracted from labeled datasets or training set. After learning the classification rules, these algorithms are capable of predicting the category of unlabeled clinical reports using test set (Sebastiani, 2002; Hastie et al., 2009; Witten et al., 2016). There are further two kinds of SML-based algorithms namely, generative and discriminative (Aggarwal & Zhai, 2012b; Alabbas, Al-Khateeb, & Mansour, 2016; Witten et al., 2016).
- ***Generative SML-based Algorithms:*** These algorithms learn the joint probability distribution $p(x, y)$. These models model the distribution of individual class. The algorithms do not focus on differences between the classes, however, such algorithms try to build a model that is representative of particular class. Naïve Bayes (NB) is a

good example of generative SML-based algorithm (Sebastiani, 2002; Aggarwal & Zhai, 2012b; Witten et al., 2016).

- ***Discriminative SML-based Algorithms:*** The discriminative algorithms learn the conditional probability distribution $p(y|x)$. These algorithms learn the hard and soft boundary between class. The discriminative algorithms highlight the differences between two classes. The examples of discriminative algorithms include, support vector machine, linear regression, decision trees, and neural networks (Sebastiani, 2002; Aggarwal & Zhai, 2012b; Witten et al., 2016).

This section briefly explains some common SML-based algorithms, namely, Naive Bayes (NB), Decision Tree (DT), Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF) and Genetic Algorithm (GA), *k*-Nearest Neighbour (*k*NN), and Ensemble voted Classifier. Good overviews of the text classification work and different text classification techniques can be found in Hayes and Weinstein (1990) and Yang (1999).

(a) ***Naive Bayes***

It is a simple probabilistic classifier. The NB method is responsible for making a probabilistic model of data within each class. It is a statistical analysis algorithm that works on numeric data (Sahami, Dumais, Heckerman, & Horvitz, 1998). It requires a small amount of training data to predict the parameters essential for classification. It is a simple and fast classification algorithm. It works well with text representations, such as BoW. The detail discussion on Naïve Bayes classifier is found in (Lewis, 1998a).

(b) ***Decision Tree***

It is the most commonly used algorithm for the task of classification and prediction. The DT represents rules that can be easily understandable by humans, and constructs the classifier in hierarchical form. The tree has a decision node, leaf node, edge and path

(Quinlan, 1986). There are variety of DT classifiers such as ID3, C4.5, C5.0, etc. One of the key drawback of DT classifier is that they are prone to overfitting. This is because, the trees if grown deeper, are able to fit all kinds of variations in the data, including the noise. Moreover, small changes in the data can drastically affect the structure of a DT. Therefore, to improve the performance of individual trees, ensemble methods such as random forest were proposed, in which many trees are constructed and trained by splitting the training set and final predictions are aggregated across the trees. The details of DT classifiers is discussed in (Safavian & Landgrebe, 1990).

(c) ***Random Forest***

It was originally developed by UC Berkeley visionary Leo Breiman in a paper he published in 1999 (Breiman, 2001). The RF algorithm works as a large collection of decorrelated decision trees. This is based on the bagging technique. In a RF, from a training set, different sub-training sets are created. From each sub-training set a DT classifier is constructed. For the test dataset, each input vector will be classified by all the decision trees in a forest and the forest chooses the classifier having the most votes (Liaw & Wiener, 2002). RF shows significant performance over a single DT. This classifier also overcomes the issue of overfitting. Fernández-Delgado, Cernadas, Barro, and Amorim (2014) compared 179 classifiers on 121 different datasets and they found that the RF is the best classifiers compared to other classifiers used in the study.

(d) ***Support Vector Machines***

The SVM classifier is highly influenced by advances in statistical learning theory. SVMs play a vital role in the application of image classification, handwriting recognition and bioinformatics. SVMs learn by example. Each example consists of a *m* number of data points (*x1,……xm*) followed by a class label *+1* or *-1*. *-1* represents one class label and *+1* represents another class label. An optimum hyperplane separates the two classes by minimizing the distance between *+1* and *-1* class labels. Such hyperplanes are termed as support vectors. The right side of the hyperplane contains the *+1* class and the left one contains the *-1* class. This separation of classes is performed with the help of training examples (Cristianini & Shawe-Taylor, 1999).

(e) ***Artificial Neural Network***

ANN comprised of three main layers. These layers are input layer, hidden layer and output layer. The input layer and hidden layer comprise many nodes, while the output layer has one node. The nodes in a neural network contain an activation function. With the help of the input layer, patterns are provided to the neural network, which interacts with hidden layers. The actual processing is done in hidden layers by allocating random weights to edges. The hidden layers are further connected to an output layer where the final answer is computed. Mostly, ANNs use learning rules for modifying the weights of the connections according to the input patterns. One of the popular learning rules is 'Delta rule', which is often utilized by 'backpropagation neural networks' (BPNNs). Delta rule is a supervised learning rule that occurs with each cycle (i.e. each time the network is presented with a new input pattern). The initial pattern is determined by a random 'guess' (Zurada, 1992).

(f) ***Genetic Algorithm***

The GA was first proposed by John Holland in the early 1970s (Mitchell, 1998). The idea behind the implementation of GA is to use the process like natural evolution to resolve the issue of optimization. In GA, a gene is comprised of a string of bits. Generally, the preliminary genes population is generated randomly. The bit string length depends on the nature of problem to be resolved. From the initial population a subset of genes is extracted based upon some quality fitness measurement. After the selection, the next step is mating and crossover of which there are different types. Random mating with crossover is one of the easy type, where, the genes in selected population are randomly selected and mated in pairs. Usually, a point for crossover is chosen for each selected pair. After crossover, the information is swapped between two pairs. In final mutation step, each gene bit has a definite Probability *P* to get inverted. The more detail on GA can be found in (Mitchell, 1998).

(g) ***k-Nearest Neighbour***

*k*NN employs instance-based learning. *k*NN is also termed as lazy learning classifier because it is the simplest classification algorithm that stores all the instances and classifies new instance using a similarity measure, such as the Euclidean distance shown in Equation 2.5 (Bao, Ishii, & Du, 2004; Fukunaga, 2013).

$$\sqrt{\sum_{i=1}^{k}\left(x_i - y_i\right)^2} \tag{2.5}$$

(h) ***Ensemble Voted Classifier***

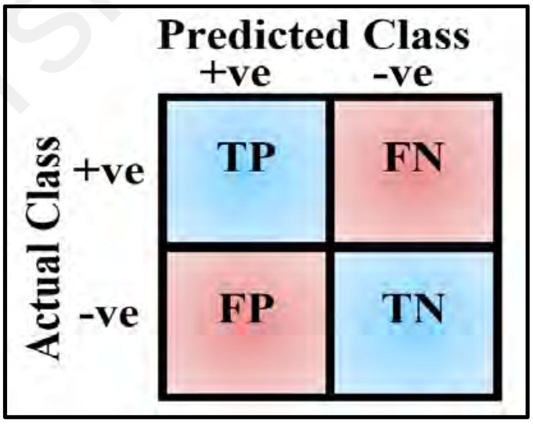
In ensemble voted classifier is a meta-classifier that constructs a set of classifiers and then classify new instances by taking majority vote of their prediction (Joachims, 1998a; Lewis, 1998a; Xu, Guo, Ye, & Cheng, 2012; Danso, Atwell, & Johnson, 2014; Yeow et al., 2014).

2.3.1.5 Classification Model Evaluation and Performance Metrics

The last step in the SML-based ATC technique is the evaluation of a constructed classifier. In this step, constructed classifier predicts the class of unlabelled clinical report (for instance, cancer or no cancer) using test dataset.

During the classification construction and evaluation phase, text classifier accuracy can be determined from confusion matrix (as shown in Figure 2.5 for binary classification problems). This matrix contains four different kind of cases namely true positive (TP), false positive (FP), false negative (FN), and true negative (TN). For instance, in case of classifying cancer related pathology reports, the TP is the number of correctly classified reports that actually belong to class “Cancer” and also predicted as “Cancer”. TN is the number of correctly classified reports that actually belong to “No Cancer” class and also predicted as “No Cancer”. FP is the number of misclassified reports that actually belong to “No Cancer” class and were predicted as “Cancer” class. FN is the number of misclassified reports that actually belong to “Cancer” class and were predicted as “No Cancer”.

In clinical text classification, several studies employed different types of performance metrics to evaluate the classification performance. The commonly used performance metrics for binary class problem are precision, recall, F-measure, accuracy, area under curve, sensitivity, and specificity. However, in multi-class problems the commonly used performance metrics are micro or macro-averaging of precision, recall, and F-measure are used. These performance metrics are briefly described in subsequent paragraphs. However, the detailed discussion on these performance metrics can be found in (Sokolova & Lapalme, 2009).



		Predicted Class	
		+ve	-ve
Actual Class	+ve	TP	FN
	-ve	FP	TN

Figure 2.5: Confusion matrix for binary class problem

(a) ***Precision***

It is the ratio of correctly predicted positive clinical reports to the total positively predicted clinical reports. It is also known as positive predictive value (PPV). It is formally defined as following:

$$precision = \frac{TP}{TP + FP} \tag{2.6}$$

(b) ***Recall***

It is the ratio of correctly predicted positive clinical reports to the all clinical reports in actual positive class. It is also known as true positive rate (TPR) or sensitivity. It is formally defined as following:

$$recall = \frac{TP}{TP + FN} \tag{2.7}$$

(c) ***F-measure***

It is the weighted average of precision and recall. It is formally defined as following:

$$F - measure = \frac{2 \times (precision \times recall)}{(precision + recall)} \tag{2.8}$$

(d) ***Accuracy***

It is the most widely used performance metric. It is the ratio of correctly predicted clinical reports to the total clinical reports. It is formally defined as following:

$$accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{2.9}$$

(e) ***Area Under the Curve***

AUC stands for area under the curve. It is used to compute the goodness of clinical report classifier by plotting a particular curve and computing area under that curve. The value of 1 for AUC shows the classifier performance is good. Conversely, when AUC value is 0.5 or lower than that shows the poor performance of clinical report classifier (Provost, Fawcett, & Kohavi, 1998; Hand & Till, 2001; Fawcett, 2006). It is formally defined as following:

$$AUC = \frac{S_0 - n_0\left(n_0 + 1\right)/2}{n_0 n_1} \qquad (2.10)$$

Here, n_0 and n_1 denote the count of positive and negative clinical reports respectively, and $S_0 = \sum r_i$ where r_i is the rank of i_{th} positive sample in ranked list.

(f) ***Specificity***

It measures the proportion of negative clinical reports that are correctly predicted a negative. It is formally defined as following:

$$specificit\,y = \frac{TN}{TN + FP} \qquad (2.11)$$

(g) ***Micro- and Macro-average of Precision, Recall and F-measure***

In micro averaging of precision, recall, and F-measure, individual TP, FP, and FN of the system for different sets are summed up and then apply them to get the statistics. Conversely, in macro averaging of precision, recall, and F-measure, simply the average of precision, recall or F-measure of the system on different sets is taken. The formal definitions of micro and macro averaging of precision, recall, and F-measure can be found in (Sokolova & Lapalme, 2009).

2.4 Review of ATC Techniques for Classification of Free-Text Clinical Reports

The pervasive use of electronic health databases has increased the accessibility of free-text clinical reports for supplementary use. Several ATC techniques such as SML-based ATC techniques or rule-based techniques have been utilized to discover useful information from free-text clinical reports. In recent years, many researchers have worked on clinical text classification and contributed relevant results to the academic literature. Thus, this section recapitulates the existing related works on ATC approaches for classifying free-text clinical reports, including autopsy reports, by exploiting the procedural decision analysis in six aspects, namely, types of clinical reports, datasets and their characteristics, preprocessing techniques, feature engineering techniques, SML-based algorithms, and performance metrics (shown in Figure 2.6). Moreover, the subsequent sections (Sections 2.5–2.10) present a detailed review of the six aforementioned aspects.

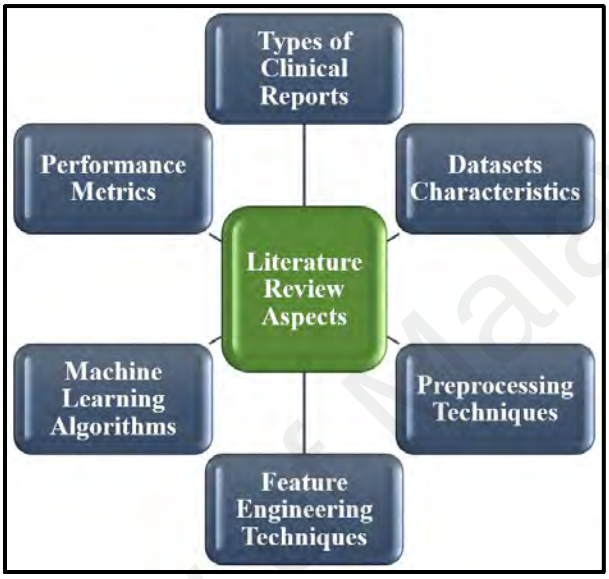

Types of Clinical Reports
Performance Metrics
Datasets Characteristics
Literature Review Aspects
Machine Learning Algorithms
Preprocessing Techniques
Feature Engineering Techniques

Figure 2.6: Literature review aspects

2.5 Types of Clinical Reports used in the Related Literature

ATC techniques have been employed in several types of free-text clinical reports, such as pathology reports, radiology reports, autopsy reports, death certificates, and biomedical documents. Overall, nine different types of clinical reports were identified from the literature as shown in Table 2.1. Moreover, the table shows the detailed kinds of clinical reports in each category with related references.

As shown in Table 2.1, majority of the studies employed pathology reports, followed by biomedical documents, radiology reports, and autopsy reports. Most of the pathology reports were used to detect breast cancer or other related cancers via text classification techniques. For instance, Rani, Gladis, and Mammen (2015), Napolitano, Marshall, Hamilton, and Gavin (2016), and Yoon et al. (2017) used pathology reports to detect cancer stages via text classification techniques. Moreover, Kasthurirathne et al. (2016) and Kasthurirathne et al. (2017) investigated the use of non-dictionary-based and dictionary-based text classification approaches to detect cancer from pathology reports. Saqlain, Hussain, Saqib, and Khan (2016) used pathology reports to predict the survival of patients diagnosed with heart failure. Sedghi, Weber, Thomo, Bibok, and Penn (2016) developed a migraine detection model using pathology reports to classify patients with migraine or no migraine.

Table 2.1: Types of clinical reports used in related literature

Report Types	Description	Studies
Influenza Related Reports	This includes emergency Department reports related to Influenza	(Pineda, Tsui, Visweswaran, & Cooper, 2013; Ye, Tsui, Wagner, Espino, & Li, 2014; MacRae et al., 2015; Pineda et al., 2015)
Radiology Reports	This includes the radiology reports related to CT Abdomen, CT Neuro, limb fractures, Cancer, Retrospective Study, Invasive Fungal (IFD) Disease, HIV, Audiologic Data, imaging, and Head CT Reports	(Mabotuwana, Lee, & Cohen-Solal, 2013; Wagholikar et al., 2013; Zuccon et al., 2013; Nguyen & Patrick, 2014; Y. Zhou et al., 2014; Bates, Fodeh, Brandt, & Womack, 2015; Martinez et al., 2015; Hassanpour & Langlotz, 2016; Masino, Grundmeier, Pennington, Germiller, & Crenshaw, 2016; K. Yadav et al., 2016; Shin, Chokshi, Lee, & Choi, 2017)
Bio-Medical Documents	This includes Medline abstracts and medical news articles	(Farshchi & Yaghoobi, 2013; Jo, 2013; Adeva, Atxa, Carrillo, & Zengotitabengoa, 2014; Alghoson, 2014; Jindal & Taneja, 2015; Parlak & Uysal, 2015; Rios & Kavuluru, 2015; H. Y. Zhou, Q. R. Zhang, H. X. Wang, & D. Zhang, 2015; Fragos & Skourlas, 2016; Mouriño-García, Pérez-Rodríguez, Anido-Rifón, & Gómez-Carballa, 2016; Parlak & Uysal, 2016b, 2018)
Tweets related to Healthcare	This includes tweets related to Influence like illness (ILI) disease and user comments about hospital services	(Greaves, Ramirez-Cano, Millett, Darzi, & Donaldson, 2013; X. Dai & M. Bikdash, 2015; Zuccon et al., 2015)
Death Certificates	This includes the death certificate written in	(Butt, Zuccon, Nguyen, Bergheim, & Grayson, 2013; B. Koopman, S. Karimi, et

Table 2.1: continued

Report Types	Description	Studies
	English and French and are related to cancer and other diseases	al., 2015; B. Koopman, G. Zuccon, A. Nguyen, A. Bergheim, & N. Grayson, 2015; Imane & Mohamed, 2017; Wu & Wang, 2017)
Autopsy Reports	This includes the verbal autopsy reports and forensic autopsy reports	(Danso et al., 2013, 2014; Yeow et al., 2014; Miasnikof et al., 2015b; Kalter, Perin, & Black, 2016; Mujtaba et al., 2016)
Pathology Reports	This includes pathology reports related to lymphoma, cancer, heart failure patients, stroke and migraine case reports, arthroplasty reports related to hip surgery and reports related to Cervical Spine	(Garla, Taylor, & Brandt, 2013; Luo, Sohani, Hochberg, & Szolovits, 2014; Deng et al., 2015; Kasthurirathne, Dixon, & Grannis, 2015; Rani et al., 2015; Kasthurirathne et al., 2016; Napolitano et al., 2016; Saqlain et al., 2016; Sedghi et al., 2016; Kasthurirathne et al., 2017; Lauren, Qu, Zhang, & Lendasse, 2017; Oleynik, Patrão, & Finger, 2017; Yoon et al., 2017)
Other Clinical Reports	This includes discharge summaries of patients, nursing care records, patient history reports suffering from diabetes, child abuse consultation reports, and radiotherapy reports	(Afzal et al., 2013; Wei, Ju, Chun, Hua, & Jin, 2013; Gatta, Vallati, De Bari, & Ozsahin, 2014; L. Zhou et al., 2015; Lopprich et al., 2016; Amrit, Paauw, Aly, & Lavric, 2017; Barak-Corren et al., 2017; Buchan, Filannino, & Uzuner, 2017; Clark, Wellner, Davis, Aberdeen, & Hirschman, 2017; Hassanpour, Langlotz, Amrhein, Befera, & Lungren, 2017; Lucini et al., 2017; Y. Wang, E. Coiera, W. Runciman, & F. Magrabi, 2017)
Combination of various Reports	This includes the multi-modality reports where different set of reports belong to same disease were combined for classification task.	(Kavuluru, Rios, & Lu, 2015; Sarker & Gonzalez, 2015; Kocbek et al., 2016)

Radiology reports were also used extensively in the field of medical text classification. Zuccon et al. (2013) and Wagholikar et al. (2013) used radiology reports to identify limb fractures via text classification techniques. Shin et al. (2017) and Yadav et al. (2016) employed radiology reports relating to brain computed tomography (brain or head CT reports) to identify pediatric traumatic brain injury (TBI). Bates et al. (2015) used radiology reports to detect the human immunodeficiency virus (HIV) by automated text classification techniques. In addition, researchers have also classified influenza-related clinical reports to detect influenza-like illnesses using SML-based ATC techniques (Pineda et al., 2013; Ye et al., 2014; MacRae et al., 2015; Pineda et al., 2015). Researchers have also used Twitter tweets to predict influenza-related tweets (X. F. Dai & M. Bikdash, 2015; Zuccon et al., 2015).

Verbal autopsy reports (Danso et al., 2013, 2014; Miasnikof et al., 2015b; Kalter et al., 2016) and forensic autopsy reports (Yeow et al., 2014; Mujtaba et al., 2016) have also been utilized to predict cause of death (CoD) from autopsy findings using SML-based ATC techniques. Death certificates have also been used to automatically assign ICD-10 codes to such forms (Butt et al., 2013; Koopman, S. Karimi, et al., 2015; Koopman, G. Zuccon, et al., 2015; Imane & Mohamed, 2017; Wu & Wang, 2017). Finally, recent studies collected various clinical reports from different sources, combined those reports, and used those combined reports for developing the classification model (Kavuluru et al., 2015; Sarker & Gonzalez, 2015; Kocbek et al., 2016). For instance, Kavuluru et al. (2015) combined pathology and radiology reports to develop a text classification model to automatically assign ICD-9 codes to electronic medical reports. Kocbek et al. (2016) combined the pathology reports, radiology reports, and patients' admission-related meta-data to predict the rate of admissions against disease. In all these three aforementioned studies, authors reported that combining data from various sources or combining features of different reports can produce highly reliable and accurate predictions.

2.6 Review of Dataset and their Characteristics

The free-text clinical report dataset is the essential ingredient of ATC techniques. Nonetheless, such a dataset is useless on its own until some useful knowledge or patterns are extracted from it. The literature related to ATC of free-text clinical reports shows that authors mostly collected customized datasets of free-text clinical reports from their country. Moreover, the datasets were divided into training and test sets. For instance, Ye et al. (2014) collected the corpus of influenza-related clinical reports to detect influenza. The collected corpus comprised 592 influenza-related reports and 29,092 non-influenza-related reports. For the training set, authors used 468 influenza-related reports and 29,004 non-influenza-related reports to construct the classification model. To test the performance of the constructed classification model, authors used the test set that comprised 124 influenza-related reports and 87 non-influenza-related reports. The datasets (related to free-text clinical reports) used in the literature can be categorized into two major categories: homogenous datasets and heterogeneous datasets (as shown in Figure 2.7). The data sources (from where the dataset is collected) can also be categorized into homogenous or heterogeneous sources. This relationship is shown in Figure 2.8 and described in subsequent paragraphs.

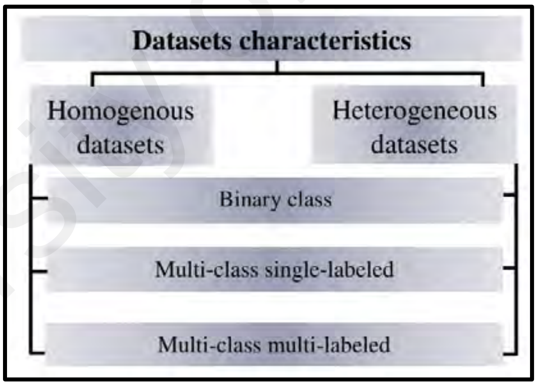


Figure 2.7: Types of datasets used in related literature

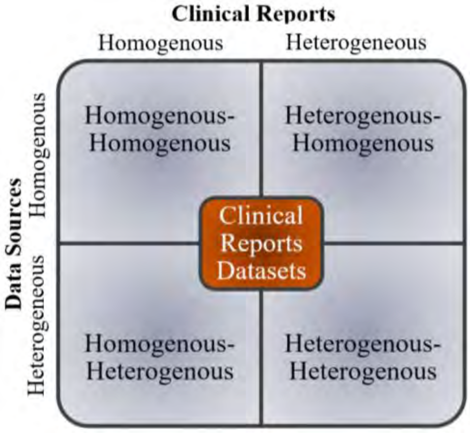


Figure 2.8: The dataset and data source matrix

Homogenous–Homogenous**:** Here, the dataset consists of one type of clinical report (such as pathology report), and the dataset is usually collected from one data source or hospital. In previous studies (Pineda et al., 2013; Ye et al., 2014; MacRae et al., 2015; Pineda et al., 2015), authors collected influenza-related emergency department reports from one hospital to develop a classification model for detecting influenza-like illnesses. In these studies, authors mentioned that their results may not be generalized because the constructed classification model was trained on emergency department reports of one hospital.

Homogenous–Heterogeneous: Here, the dataset consists of one type of clinical report (such as pathology report), and the dataset is usually collected from different data sources or hospitals. For instance, in studies (Kasthurirathne et al., 2016, 2017), authors collected 7,000 cancer-related pathology reports from seven different healthcare systems and thirty different hospitals. The experimental findings showed that combining cancer-related pathology reports from various data sources improves generalization, reliability, and classification performance. Wang et al. (2017) developed a classification model to automate the identification of patients' safety incidents using incident reports. Here, the authors collected 6,000 incident reports from one hospital for training purposes and 5,950 incident reports from another hospital for testing purposes. The experimental findings showed the robustness of using incident reports from different data sources. Hassanpour et al. (2017) developed a classification model to automatically classify knee magnetic resonance imaging (MRI) reports into positive or negative class. For experiments, 706 reports were collected from Duke and 1748 reports were collected from Stanford healthcare organizations. Authors reported that combining knee MRI reports from two different organizations demonstrates improved classification performance. Barak-Corren et al. (2017) developed a prediction model to predict the risk of suicidal behavior of patients. For the experiments, authors collected the narrative clinical notes from a variety of hospitals situated in Boston, USA, to predict the risk of suicidal behavior of the patients. The homogenous datasets and heterogeneous data sources were collected and used to construct the generalized, accurate, and reliable classification models.

Heterogeneous–Homogenous: Here, the dataset consists of different types of clinical reports (such as pathology and radiology reports), and the dataset is usually collected from one data source or one hospital. Different reports are used for classification because a variety of reports can be prepared by a hospital for reporting the same disease; for instance, cancer cases can be reported into pathology and radiology reports. Thus, combining both of these reports in the auto-prediction model can enhance the prediction accuracy and credibility. For instance, Kavuluru et al. (2015) combined pathology and radiology reports to develop a text classification model to automatically assign ICD-9 codes to electronic medical reports. The authors collected the pathology and radiology reports from one hospital situated at the United Kingdom. Kocbek et al. (2016) combined three different clinical reports, namely, pathology, radiology, and patients' admission-related meta-data reports, to predict the rate of admissions against disease. The authors collected these reports from one hospital situated in Australia. The abovementioned studies reported that combining features of different clinical reports produce highly reliable and accurate predictions.

Heterogeneous–Heterogeneous: Here, the dataset consists of different types of clinical reports (such as pathology and radiology reports), and the dataset is usually collected from different data sources or different hospitals. This type of dataset is the most robust dataset for the development of the classification model. Moreover, the results generated from such datasets can be generalized on a wide scale. For instance, Sarker and Gonzalez (2015) collected Twitter tweets and daily strength instances related to adverse drug reaction events. Authors also collected adverse drug events reports from one hospital. Authors combined all these three datasets in the training set and developed a classification model for identifying adverse drug reactions. The experimental results showed that the classification performance significantly benefits from multi corpus training collected from different data sources (such as Twitter, daily strength, and hospital).

Both homogenous and heterogeneous datasets can be further divided into three subtypes, namely, binary class datasets, multi-class single labeled datasets, and multi-class multi-labeled datasets (as shown in Figure 2.7).

Binary Class Datasets: In binary class datasets, reports can be labeled in either of two classes (such as cancer positive or cancer negative). For instance, in studies (Kasthurirathne et al., 2016, 2017), authors collected 7,000 cancer-related pathology reports. Each report was labeled as cancer positive or cancer negative by three clinicians. Of these 7,000 reports, 1950 reports were cancer positive, and 5,050 reports were cancer negative. Of these 7,000 reports, 6,300 reports were used for training purpose and the remaining 700 reports were used for testing. Wagholikar et al. (2013) and Zuccon et al. (2013) collected radiology reports corpus related to limb fracture. This corpus comprised 99 radiology reports. Each report was labeled with “normal” or “abnormal” class. Of these reports, 90% were used as the training set, and the remaining 10% was used as the test set.

Multi-Class Single-Labeled Datasets: In multi-class single labeled datasets, clinical reports were composed of more than two categories; however, each report was categorized into one label. For instance, Mujtaba et al. (2016) developed a classification model to predict CoDs from forensic autopsy reports. Authors collected the dataset from one of the biggest hospitals situated in Kuala Lumpur, Malaysia. The dataset comprised 400 forensic autopsy reports. These 400 reports were labeled into eight different CODs.

Authors used tenfold cross validation (Kohavi, 1995; Refaeilzadeh, Tang, & Liu, 2009) to evaluate the classification model's performance. Danso et al. (2013) developed a prediction model to predict CoDs from verbal autopsy reports. The collected dataset comprised 6,407 verbal autopsy reports. The collected reports belonged to 16 different CoDs. Authors used tenfold cross validation to evaluate the classification model's performance. In studies (Jo, 2013; Adeva et al., 2014; Alghoson, 2014; Jindal & Taneja, 2015; Parlak & Uysal, 2015; Rios & Kavuluru, 2015; Zhou, Q. R. Zhang, H. X. Wang, & D. Zhang, 2015; Fragos & Skourlas, 2016; Mouriño-García et al., 2016; Parlak & Uysal, 2016a, 2018), authors used the subset of the OHSUMED dataset to classify medical abstracts into 23 cardiovascular diseases. This subset of the OHSUMED dataset contains 13,929 Medline abstracts. Nonetheless, each abstract may fall into more than one category, but the authors only considered those Medline abstracts, which fell under one category only. Of these 13,929 Medline abstracts, 6,286 abstracts were used in the training set and the remaining abstracts were used in the test set.

Multi-Class Multi-Labeled Datasets: In multi-class multi-labeled datasets, clinical reports comprised more than two categories, and each report was categorized into more than one class label. For instance, Imane and Mohamed (2017) collected the French Center for Epidemiology and Medical Causes of Death (CépiDC) dataset, which includes death certificates. The dataset contained 65,843 death certificates labeled by 3232 ICD-10 (the international classification of diseases code-version 10) codes. Each certificate maybe assigned one or more ICD-10 codes. Of these 65,843 certificates, 52,675 death certificates were used for training purposes, and 13,168 death certificates were used for testing purposes.

Table 2.2 shows the distribution of related literature based on datasets and data source matrix (shown in Figure 2.8). The majority of studies (57 out of 69) used one type of clinical report collected from one hospital (homogenous–homogenous). Of these 57 studies, 36 studies used multi-class single-labeled datasets, 20 used binary class datasets, and only one study used multi-class multi-label datasets. Moreover, five studies used one type of clinical reports that were collected from more than one data sources (please see homogenous–heterogeneous row in Table 2.2). Of these five studies, two used multi-class single-labeled datasets, and three employed binary class datasets. Furthermore, five studies used different types of clinical reports, which were collected from one data source (heterogeneous–homogenous). Of these five studies, four studies used multi-class single-labeled datasets, and one utilized a binary class dataset. Finally, only one study used different types of clinical reports that were collected from more than one different data source (heterogeneous– heterogeneous).

Table 2.2: Related literature based on dataset and data source matrix

Type of Dataset		References	Count
Homogenous - Homogenous	Binary class	(Afzal et al., 2013; Butt et al., 2013; Mabotuwana et al., 2013; Pineda et al., 2013; Wagholikar et al., 2013; Wei et al., 2013; Zuccon et al., 2013; Nguyen & Patrick, 2014; Ye et al., 2014; Bates et al., 2015; X. Dai & M. Bikdash, 2015; Deng et al., 2015; MacRae et al., 2015; Pineda et al., 2015; Zuccon et al., 2015; Lopprich et al., 2016; Saqlain et al., 2016; K. Yadav et al., 2016; Amrit et al., 2017; Lucini et al., 2017)	20
	Multi-class single-label	(Danso et al., 2013; Farshchi & Yaghoobi, 2013; Garla et al., 2013; Greaves et al., 2013; Jo, 2013; Adeva et al., 2014; Alghoson, 2014; Danso et al., 2014; Gatta et al., 2014; Luo et al., 2014; Yeow et al., 2014; Y. Zhou et al., 2014; Jindal & Taneja, 2015; Kasthurirathne et al., 2015; B. Koopman, S. Karimi, et al., 2015; B. Koopman, G. Zuccon, et al., 2015; Miasnikof et al., 2015b; Parlak & Uysal, 2015; Rani et al., 2015; Rios & Kavuluru, 2015; H. Y. Zhou et al., 2015; L. Zhou et al., 2015; Fragos & Skourlas, 2016; Hassanpour & Langlotz, 2016; Kalter et al., 2016; Masino et al., 2016; Mouriño-García et al., 2016; Napolitano et al., 2016; Parlak &	37

Table 2.2: continued

Type of Dataset		References	Count
		Uysal, 2016b; Buchan et al., 2017; Clark et al., 2017; Lauren et al., 2017; Oleynik et al., 2017; Shin et al., 2017; Wu & Wang, 2017; Yoon et al., 2017; Parlak & Uysal, 2018)	
	Multi-class multi-label	(Imane & Mohamed, 2017)	1
Homogenous - Heterogeneous	Binary class	(Martinez et al., 2015; Kasthurirathne et al., 2016, 2017)	3
	Multi-class single-label	(Sedghi et al., 2016; Barak-Corren et al., 2017)	2
	Multi-class multi-label	-	0
Heterogeneous - Homogenous	Binary class	(Hassanpour et al., 2017)	1
	Multi-class single-label	(Kavuluru et al., 2015; Kocbek et al., 2016; Mujtaba et al., 2016; Y. Wang et al., 2017)	4
	Multi-class multi-label	-	0
Heterogeneous - Heterogeneous	Binary class	(Sarker & Gonzalez, 2015)	1
	Multi-class single-label	-	0
	Multi-class multi-label	-	0

As can be seen from Table 2.2, very few studies have used heterogeneous-homogeneous and hetrogenous-hetrogenous clinical reports corpus and mostly have used either homogenous-homogenous and homogenous-heterogeneous dataset. However, it should be noted that several hospitals may have different medical documentation systems and patterns or styles, thereby possibly producing hurdles in generalizing constructed classifier to multiple hospitals. Moreover, one disease can be reported in a variety of reports. For example, cancer patients' findings can be reported in pathology and radiology reports. Hence, the practitioners in clinical text classification are suggested to use the heterogeneous–heterogeneous reports to develop a classification model.

2.7 Review of Preprocessing Techniques

In clinical text classification, preprocessing involves removing meaningless data from the collected dataset to improve the quality of clinical text classification models. In the related literature, preprocessing techniques such as removal of stop words, removal of punctuations or special symbols, removal of empty spaces, case conversion, spell correction, tokenization, stemming, lemmatization, and normalization were applied (as shown in Table 2.3). Table 2.3 shows the related literature based on applied preprocessing tasks. Majority of the studies employed basic pre-processing tasks (including stop word removal, removal of punctuation and white spaces, and case conversion) and word tokenization. In addition, these studies reported the effectiveness of these preprocessing techniques on clinical text classification. Nonetheless, few studies (Danso et al., 2013, 2014; Sarker & Gonzalez, 2015; Lauren et al., 2017) empirically investigated the presence and absence of stop words and reported that the presence of stop words produces better classification accuracy than their absence. In some studies, researchers demonstrated that applying stemming task with basic preprocessing tasks and word tokenization enhances classification performance (Jo, 2013; Adeva et al., 2014; Koopman, S. Karimi, et al., 2015; Koopman, G. Zuccon, et al., 2015; Sarker & Gonzalez, 2015). Buchan et al. (2017) and Wang et al. (2017) applied stemming and lemmatization for clinical text normalization with basic preprocessing tasks and word tokenization. They also reported the effectiveness of using stemming and lemmatization techniques.

Nonetheless, Lauren et al. (2017) applied text classification techniques to classify arthroplasty reports and empirically investigated the effectiveness of stemming and lemmatization to preprocess the arthroplasty reports. Experimental results showed the unsuitability of stemming and lemmatization when applied on psychiatric evaluation reports. Clark et al. (2017) applied text classification techniques for classifying psychiatric evaluation reports to detect the severity of mental disorders and reported the unsuitability of stemming and lemmatization when applied on psychiatric evaluation reports. Martinez et al. (2015) and Masino et al. (2016) applied basic preprocessing techniques with word tokenization to classify radiology reports. In addition to basic preprocessing techniques, researchers also applied few text normalization techniques using regular expressions to convert numbers or dates to common units such as *number* and *date*. The experimental findings showed that such text normalization techniques improve the classification accuracy and overcome the issue of dimensionality.

Table 2.3: Preprocessing tasks used in related literature

Studies	Preprocessing Techniques	Study Count
(Afzal et al., 2013; Danso et al., 2013; Garla et al., 2013; Greaves et al., 2013; Wagholikar et al., 2013; Danso et al., 2014; Nguyen & Patrick, 2014; Y. Zhou et al., 2014; Bates et al., 2015; X. Dai & M. Bikdash, 2015; Jindal & Taneja, 2015; Kasthurirathne et al., 2015; Kavuluru et al., 2015; Parlak & Uysal, 2015; Rani et al., 2015; H. Y. Zhou et al., 2015; Fragos & Skourlas, 2016; Hassanpour & Langlotz, 2016; Kalter et al., 2016; Lopprich et al., 2016; Napolitano et al., 2016; Saqlain et al., 2016; Sedghi et al., 2016; K. Yadav et al., 2016; Shin et al., 2017; Yoon et al., 2017)	These studies reported that the removal of stop-words, punctuation marks, white spaces, and converting text into lower case improves the classification performance.	26
(Jo, 2013; Adeva et al., 2014; B. Koopman, S. Karimi, et al., 2015; B. Koopman, G. Zuccon, et al., 2015; Sarker & Gonzalez, 2015; Kasthurirathne et al., 2016; Kocbek et al., 2016; Mujtaba et al., 2016; Parlak & Uysal, 2016b; Amrit et al., 2017; Kasthurirathne et al., 2017; Lucini et	These studies reported that the converting the text into lower case, converting the text into word tokens, applying stemming technique, and removing the stop-words, punctuation marks, and white spaces improve the classification performance.	15

Table 2.3: continued

Studies	Preprocessing Techniques	Study Count
al., 2017; Oleynik et al., 2017; Parlak & Uysal, 2018)		
(Buchan et al., 2017; Y. Wang et al., 2017)	These studies reported that the converting the text into lower case, converting the text into word tokens, applying stemming and lemmatization techniques, spell correction, and removing the stop-words, punctuation marks, and white spaces improve the classification performance.	4
(Luo et al., 2014; Imane & Mohamed, 2017; Lauren et al., 2017)	These studies reported that the converting the text into lower case, removing the stop-words, punctuation marks, and white spaces, and converting the text into sentence tokens improve the classification performance.	3
(Martinez et al., 2015; Masino et al., 2016)	These studies reported that the converting the text into lower case, converting the text into word tokens, and converting numeric measures and some units into common terms improve the classification performance.	2
(Clark et al., 2017)	These studies reported that the converting the text into lower case, converting the text into word tokens, applying spell checker, and removing the stop-words, punctuation marks, and white spaces improve the classification performance.	1
(Butt et al., 2013; Farshchi & Yaghoobi, 2013; Pineda et al., 2013; Wei et al., 2013; Zuccon et al., 2013; Alghoson, 2014; Danso et al., 2014; Gatta et al., 2014; Ye et al., 2014; Yeow et al., 2014; Comelli, Agnello, Vitabile, & Ieee, 2015; Deng et al., 2015; Miasnikof et al., 2015b; Pineda et al., 2015; L. Zhou et al., 2015; Zuccon et al., 2015; Mouriño-García et al., 2016; Barak-Corren et al., 2017; Hassanpour et al., 2017; Wu & Wang, 2017)	These studies have not reported any preprocessing techniques	21

2.8 Review of Feature Engineering Techniques

As discussed earlier in Section 2.3.1.3, the most important step in the SML-based ATC technique is feature engineering (Wolpert & Macready, 1995; Aggarwal & Zhai, 2012a; Domingos, 2012; Witten et al., 2016). The success or failure of any text classification model is heavily dependent upon the quality of features used in the classification task. Feature engineering is further subdivided into three substeps, namely, feature extraction, feature representation, and feature selection. The review of these three steps is presented in subsequent sections (from Section 2.8.1 to Section 2.8.3).

2.8.1 Review of Feature Extraction Techniques

Feature extraction is the process of extracting useful features from free-text clinical reports. Several features have been used to classify the clinical reports. The complete taxonomy of all these features has been defined in this thesis and is shown in Figure 2.9.

In the field of clinical text classification, computer-aided expert systems are required only if a certain condition, such as specific level of accuracy or quality, is met. In the related literature, researchers usually employed and empirically investigated two general approaches of feature extraction, namely, expert-driven (X. F. Dai & M. Bikdash, 2015; Sarker & Gonzalez, 2015; Sedghi et al., 2016; Barak-Corren et al., 2017; Clark et al., 2017) and fully automated feature extraction (Wei et al., 2013; Nguyen & Patrick, 2014; Bates et al., 2015; Comelli et al., 2015). Thus, this section aims to present details of both of these approaches with their subtypes (as shown in Figure 2.9). Moreover, it presents the related literature that compared both of these features to evaluate the classification performance.

2.8.1.1 Fully-automated feature extraction approaches

Here, the features are automatically extracted from given clinical reports by computer programs through various statistically approaches. In these approaches, no human or expert intervention is required. The literature shows that researchers used the automated feature extraction techniques to extract content-based features, concept-based features, structural features, and linguistic features. These features are discussed in subsequent paragraphs with the help of related studies.

(a) *Content-based features*

Features are usually extracted from the content of free-text clinical reports. These features include BoW, *n*-gram, and Word2Vec. The BoW and *n*-gram features are already defined in Section 2.3.1.3(a). Word2Vec exists in two models: skip-gram and continuous bag of words (CBoW) (Goldberg & Levy, 2014). The skip-gram model learns iteratively from existing words available in a sentence to predict the next word. By contrast, the CBoW model uses the neighboring words to predict the current word. In both skip-gram and CBoW, the parameter window size determines the limit on number of words used in the context.

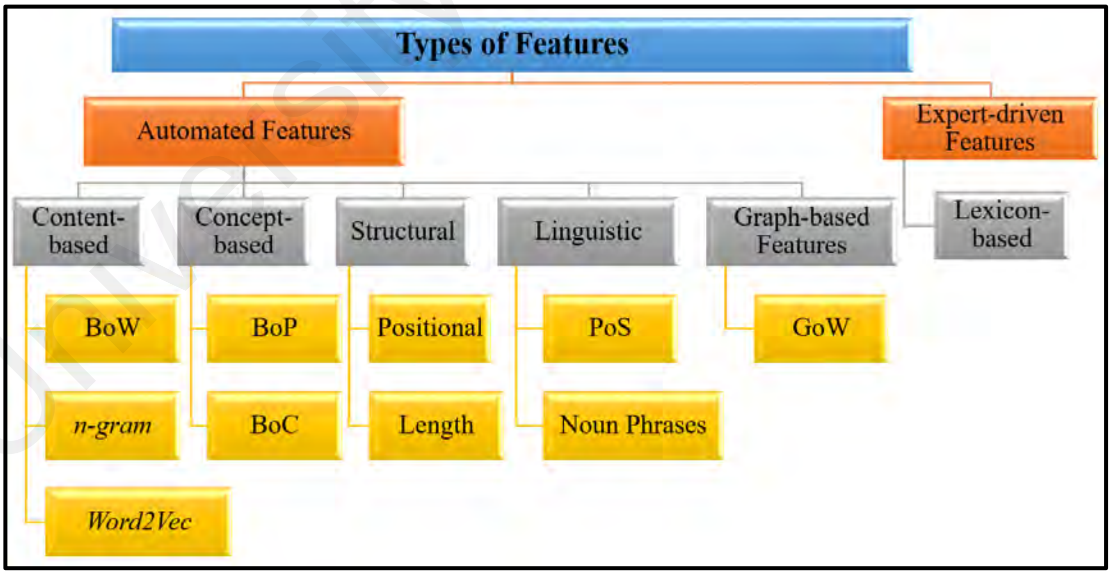


Figure 2.9: Features used in related literature

(b) *Concept-based features*

Medical experts (such as physician or surgeon) may use different terms to describe the same condition in free-text clinical reports. For instance, experts may use the term “heart attack" or "myocardial infarction" interchangeably. Although, both terms belong to the same medical concept, the content-based feature extraction techniques cannot identify the relationship between these two terms. To overcome this issue, specialized medical ontologies are developed such as SNOMED CT (Stearns et al., 2001; Donnelly, 2006) to extract medical concepts instead of terms from clinical reports. In the literature, two widely used concept-based features were identified, namely, BoP and BoC (discussed in Section 2.3.1.3(a)).

(c) ***Structural features***

These features exploit the structure or form of clinical documents for obtaining the discriminative features. These features include length of the clinical reports, number of sentences available in the clinical reports, number of sections available in the clinical reports, and position of the word in a given sentence.

(d) ***Linguistic features***

The linguistic features are used to determine the correct sense of given words used in the text. These features include parts of speech (POS) features. In POS, each word is represented with its POS tag.

(e) ***Graph-based features or graph of words (GoW)***

In the GoW model, content- or concept-based features are represented in graphs to capture word order. In GoW, a graph is the combination of V, E, and W, where V represents graph vertices with each vertex containing a distinct feature in the input text, E represents graph edges that connect co-occurring features, and W is the function that computes the edge weight by considering the co-occurrence frequency of adjacent vertices in the graph. The motivation in using the graph model is to consider word order in the input text.

Table 2.4 shows the type of automated features used in the related literature. In most of the studies, researchers used BoW features for classifying free-text clinical reports. For instance, in studies (Jo, 2013; Jindal & Taneja, 2015; Parlak & Uysal, 2015; Zhou, Q. R. Zhang, et al., 2015; Parlak & Uysal, 2016a, 2018), authors employed BoW features to classify Medline abstracts. Moreover, Garla et al. (2013) and Kasthurirathne et al. (2015) extracted and used BoW features for classifying cancer reports. Wu and Wang (2017) and Wang et al. (2017) obtained BoW features from death certificates to determine ICD-10 codes of reported CoDs. Oleynik et al. (2017) and Kasthurirathne et al. (2017) classified pathology reports using BoW features. Moreover, Yeow et al. (2014) used BoW features to predict CoDs from forensic autopsy reports and reported that BoW is useful for predicting CoDs. Although the BoW model is simpler and effective however, it ignores the word order. Therefore, to overcome the limitations of the BoW model, *n*-gram feature extraction was proposed (Cavnar & Trenkle, 1994).

Several studies have employed *n*-gram feature extraction to extract powerful features from clinical reports. For instance, Mujtaba et al. (2016) empirically investigated the performance of unigram, bigram, and trigram features to classify forensic autopsy reports. Their experimental results showed that unigram outperforms bigram and trigram features, but the performance of bigram was slightly lower than that of unigram. Moreover, Lucini et al. (2017) investigated the effectiveness of unigram, bigram, and trigram to predict hospital admission using free-text emergency department reports. The experimental results showed that trigram outperforms unigram and bigram. Zhou et al. (2014) employed *n*-gram features to classify radiology reports; they experimentally investigated *n*-gram from 1 to 8 and reported that *n*-gram (where $n = 4$) obtained the best classification accuracy. Masino et al. (2016) investigated the character *n*-gram and word *n*-gram features (where $n = 1$ to 3) to classify radiology reports. Their experimental results showed that word bi-gram and word-trigram demonstrate enhanced results. Moreover, Masino et al. (2016) compared the effectiveness of word-level unigram, bigram, and trigram and character-level unigram, bigram, and trigram to classify radiology reports. Their experimental results showed that word-level bigram and trigram outperform the other features.

Pineda et al. (2013) and Pineda et al. (2015) extracted *n*-gram features to classify influenza-like diseases using free-text clinical reports related to influenza. The authors reported that a combination of unigram and bigram features yields improved results. In studies (Zhou, A. W. Baughman, et al., 2015; Hassanpour & Langlotz, 2016; Hassanpour et al., 2017), authors extracted unigram, bigram, and trigram features from clinical reports, aggregated the extracted features, and used them in classification. The experimental findings showed that combining unigram, bigram, and trigram exhibited enhanced results. Danso et al. (2013) extracted *n*-gram features (unigrams and bigram) and linguistic features (PoS) from verbal autopsy reports to determine the time of death and CoD. Their experimental results showed that a combination of *n*-gram and PoS features is useful for classification tasks. Many studies have demonstrated the effectiveness of the *n*-gram approach, but this technique has three major limitations. First, the *n*-gram approach does not capture word inversion and subset matching. Second, this approach does not consider word-level synonymy and polysemy when applied on clinical text reports. Finally, the number of features increases enormously with increasing *n*, thereby resulting in dimensionality. To overcome these issues, researchers employed other kinds of features such as BoP and BoC.

Pineda et al. (2015) employed BoP features to extract useful medical phrases using MetaMap tool from influenza-related free-text clinical reports. Kocbek et al. (2016) used BoP and BoC features to predict the admission against disease via a combination of pathology, radiology, and admission-related patients' data. In studies (Wei et al., 2013;

Nguyen & Patrick, 2014; Bates et al., 2015; Comelli et al., 2015), authors employed BoW and BoC features to classify free-text clinical reports, and they reported that the combination of both BoW and BoC features enhances the classification accuracy. In recent studies (Martinez et al., 2015; Amrit et al., 2017; Buchan et al., 2017), researchers used the combination of BoW, *n*-gram, BoC, structural, and linguistic features to classify clinical reports. The experimental findings showed that structural and linguistic features obtain robust classification accuracy when combined with content-based features such as BoW, *n*-gram, and BoC. Recently, Yoon et al. (2017) used GoW to classify breast cancer and lung cancer cases. The authors transformed each cancer report into a graph, where graph vertices show the words available in the report and edge shows the two co-occurring words. To show the effectiveness of the GoW model, authors compared the proposed GoW with *n*-gram. The experimental results demonstrated that the proposed GoW model outperforms the *n*-gram model because graph representation provides flexibility and robustness on representing the natural language text compared with traditional *n*-gram. Moreover, the GoW approach can overcome the limitations of word co-occurrence and word order. Nonetheless, GoW is more effective than traditional BoW and *n*-gram but computationally expensive relative to BoW or *n*-gram.

2.8.1.2 Expert-driven feature extraction

The groups of experts are responsible for discovering the useful and discriminative features from the clinical reports. Moreover, the experts rank the extracted features on the basis of their discriminative power and store those features in lexicons for classification. This approach requires readily available expert knowledge in the form of decision rules, expert domain knowledge, and human expertise. Table 2.4 shows various studies that employed expert-driven features for classifying clinical reports.

X. F. Dai and M. Bikdash (2015) employed expert-driven approaches to manually extract features (related to medicine and alcohol) from tweets related to influenza. After extracting the features, the authors developed two lexicons: one for storing the features related to medicine and another for storing the features related to alcohol. Finally, the authors used the developed lexicons to classify influenza-related tweets. Sarker and Gonzalez (2015) extracted *n*-gram and BoC features from Twitter tweets, daily strength instances, and free-text clinical reports related to adverse drug reaction. Moreover, the authors developed expert-driven lexicons that contain some useful features to classify adverse drug reaction-related tweets or clinical reports. The experimental findings showed that a combination of n-gram features, BoC features, and manually created expert-driven lexicons leads to enhanced classification accuracy. Deng et al. (2015) employed expert-driven feature extraction to classify pathology reports and obtained good classification accuracy. Sedghi et al. (2016) developed expert-driven lexicons with the help of experts for migraine and stroke-related cases. Saqlain et al. (2016) prepared and used expert-driven lexicons for predicting the heart failure risk of heart patients. In the aforementioned studies, the authors reported that lexicon-based features result in high classification accuracy.

2.8.1.3 Expert-driven versus fully-automated features

To obtain a specific level of classification performance in the clinical text classification domain, several studies empirically investigated the effectiveness of expert-driven and fully automated approaches. Pineda et al. (2013) compared the classification performance of expert-driven features and fully automated features to classify influenza-related reports. Their experimental results showed no significant difference between the classification performance obtained through expert-driven and fully automated feature extraction approaches. Zuccon et al. (2013) employed the ATC technique to identify limb fracture radiology reports. The authors developed two different text classification models using expert-driven and fully automated features. Their experimental results showed that the text classification model developed using fully automated features obtained 3% more classification accuracy than that of the expert-driven model.

Ye et al. (2014) employed SML-based ATC techniques to classify influenza-related clinical reports. Features were extracted by domain experts from collected datasets. These extracted features were fed to a classifier to categorize the collected report. To compare the effectiveness of expert-driven features, authors also extracted the BoP features using automated tools, namely, TOPZ and MEDLEE. These tools extracted the medical phrases used in clinical reports. These extracted features were then fed to a classifier to organize the influenza-related reports. The experimental results showed that expert-driven features outperformed TOPZ and MEDLEE features. Koopman, S. Karimi, et al. (2015) developed an ATC model to classify death certificates through expert-driven and fully automated features. In expert-driven features, experts extracted the useful terms, features, or keywords from death certificates. Conversely, in the automated approach, the BoW and BoC features were extracted from death certificates. The experimental results showed a minute difference in performance between these two approaches; the performance obtained by the expert-driven approach was 1% higher than that of the fully automated approach.

Kalter et al. (2016) developed an ATC model to classify verbal autopsy reports using expert-driven and fully automated features. The experimental results showed that expert-drive features outperform fully automated features. Masino et al. (2016) developed a text classification model with and without the help of domain expert intervention to classify temporal bone-related radiology reports. Text classifiers with expert intervention obtained the highest classification accuracy than those without expert intervention. Kasthurirathne et al. (2016) employed expert-driven features and fully automated features to classify cancer-related pathology reports. No significant difference in classification performance was found between the expert-driven and fully automated approaches.

Table 2.4: Feature sets used in related literature

Study	Features	Feature Representation	Feature Selection
Afzal et al. (2013)	BoW	BR	Chi-Square
Garla et al. (2013)	Bow	BR	--
Jo (2013)	BoW	TF	--
Pineda et al. (2013)	*n*-gram	BR	--
Wei et al. (2013)	BoW and BoC	TF	IG
Butt et al. (2013)	*n*-gram and BoC	BR and TF	--
Zuccon et al. (2013)	*n*-gram and BoC	TF	--
Danso et al. (2013)	*n*-gram and LGF	N-TFiDF	--
Farshchi and Yaghoobi (2013)	BoW, STF, LGF	TF and TFiDF	--
Greaves et al. (2013)	EDF	TF	IG
Wagholikar et al. (2013)	EDF	BR	--
Danso et al. (2014)	*n*-gram	BR, TF, TFiDF and N-TFiDF	LSFS
Gatta et al. (2014)	BoW	TFiDF	--
Yeow et al. (2014)	BoW	TF	--
Y. Zhou et al. (2014)	*n*-gram	BR	--
Luo et al. (2014)	GoW	TF	--
Nguyen and Patrick (2014)	BoW and BoC	BR	--
Alghoson (2014)	EDF	BR	--
Ye et al. (2014)	*n*-gram and EDF	BR	--
Adeva et al. (2014)	*n*-gram	TFiDF	Chi-Square
Jindal and Taneja (2015)	BoW	TF	--
Kasthurirathne et al. (2015)	BoW	TF	--
Parlak and Uysal (2015)	BoW	TF	--
H. Y. Zhou et al. (2015)	BoW	TF	--
Rani et al. (2015)	*n*-gram	BR	--
Pineda et al. (2015)	*n*-gram	BR	--
Bates et al. (2015)	BoW and BoC	BR	MI
Comelli et al. (2015)	BoW and BoC	BR	--
Kavuluru et al. (2015)	*n*-gram and BoC	BR	BNSS
B. Koopman, S. Karimi, et al. (2015)	*n*-gram and BoC	BR	--
B. Koopman, G. Zuccon, et al. (2015)	*n*-gram and BoC	BR	IG
Martinez et al. (2015)	BoW, BoS, BoP, BoC, and STF	BR	PC
Rios and Kavuluru (2015)	W2V	TF	--
Zuccon et al. (2015)	*n*-gram and STF	BR	--
X. Dai and M. Bikdash (2015)	EDF	BR	ED
Deng et al. (2015)	EDF	BR	--
MacRae et al. (2015)	EDF	BR	--
Sarker and Gonzalez (2015)	EDF	TFiDF	--

Table 2.4: continued

Study	Features	Feature Representation	Feature Selection
Miasnikof et al. (2015b)	BoW and EDF	BR	--
L. Zhou et al. (2015)	*n*-gram	TF	--
Kasthurirathne et al. (2016)	BoW	TF	IG
Lopprich et al. (2016)	BoW	TF	--
Parlak and Uysal (2016b)	BoW	TF	GI and DFS
Hassanpour and Langlotz (2016)	*n*-gram	TFiDF	--
Masino et al. (2016)	*n*-gram	BR and TF	--
Napolitano et al. (2016)	BoS and BoP	BR and TF	--
Mouriño-García et al. (2016)	BoC	TF	--
K. Yadav et al. (2016)	*n*-gram and BoC	TF	--
Kocbek et al. (2016)	BoP and BoC	TF	IG
Fragos and Skourlas (2016)	BoW and BoP	TFiDF	--
Kalter et al. (2016)	EDF	BR	--
Saqlain et al. (2016)	EDF	TF	--
Sedghi et al. (2016)	EDF	BR	--
Imane and Mohamed (2017)	BoW	TFiDF	--
Kasthurirathne et al. (2017)	BoW	TF	IG
Oleynik et al. (2017)	BoW	TFiDF	--
Y. Wang et al. (2017)	BoW	BR, TF and TFiDF	--
Wu and Wang (2017)	BoW	TFiDF	--
Hassanpour et al. (2017)	*n*-gram	TFiDF	--
Yoon et al. (2017)	GoW	TF	--
Lauren et al. (2017)	W2V	TF	--
Shin et al. (2017)	W2V	BR, TF, TFiDF and N-TFiDF	--
Amrit et al. (2017)	BoW and STF	BR, TF and TFiDF	--
Buchan et al. (2017)	*n*-gram, BoC and LGF	N-TFiDF	--
Barak-Corren et al. (2017)	EDF	BR	--
Clark et al. (2017)	EDF	BR and TF	MI
Lucini et al. (2017)	*n*-gram	BR, TF and TFiDF	Chi-Square
Mujtaba et al. (2016)	*n*-gram	BR, TF, TFiDF and N-TFiDF	IG, Chi-Square and PC
Parlak and Uysal (2018)	BoW	TF and TFiDF	DFS

****BoW** (Bag of Words), **NGF** (*n*-gram Features), **BoS** (Bag of Sentences), **BoP** (Bag of Phrases), **BoC** (Bag of Concepts), **GoW** (Graph of words), **W2V** (Word2Vector), **STF** (Structural Features), **LGF** (Linguistics Features), **EDF** (Expert-driven Features)
** **BR** (Binary Representation), **TF** (Term Frequency), **TFiDF** (Term Frequency with inverse Document Frequency), **N-TFiDF** (Normalized TFiDF)
****IG** (Information Gain), **Chi** (Chi-Square), **PC** (Pearson Correlation), **GI** (Gini-Index), **LSFS** (Local Semi-Supervised Feature Selection), **ED** (Expert-driven), **MI** (Mutual Information), **MDA** (Multiple Discriminant Analysis), **BNSS** (Bi-Normal Separation Score), **PCA** (Principal Component Analysis), **DFS** (Distinguishing Feature Selector)

In the field of clinical text classification, researchers have comprehensively investigated the performance of classification models using expert-driven features and fully automated features. Moreover, varying results were obtained. Few studies showed that expert-driven features outperform fully automated features, and several studies reported that automated features outperform expert-driven features. Some studies reported no significant difference between the classification performance obtained through expert-driven and fully automated features. Therefore, one should empirically investigate the performance of both approaches on free-text clinical reports to evaluate the superior one.

Both approaches have their advantages and disadvantages. The expert-driven approach is flexible and can easily understand the importance of manually extracted features. Moreover, the misclassification error can be easily fixed when working with expert-driven features. Nonetheless, the major limitation of this approach is that it depends heavily on the deep skills and knowledge of domain experts for robustness and scalability. The expert-driven approach is not purely a scientific activity but more of a balancing act in black art, architecture, design, and development. This approach is time consuming and resource extensive. Finally, this approach is not easily extendable; for any new class or category, experts will be engaged to extend the functionality of the existing model. Nevertheless, this technique is effective in creating baseline results so that further automated methods can be designed and engineered to obtain accuracy similar or better than expert-driven approaches.

Fully automated feature extraction approaches are less time consuming and do not require any expert intervention to extract useful features from clinical reports. Nonetheless, the major limitation of these approaches is their requirement for a large number of labeled clinical reports for extracting useful features that correlate well with the class. Moreover, in medical domains, one cannot rely only on fully automated techniques, so a robust comparison of fully automated and expert-driven approaches is needed to evaluate their performance differences.

2.8.2 Review of Feature Representation Techniques

As discussed in Section 2.3.1.3(b), an important step after extracting features from clinical reports is transforming extracted features into numeric vectors for linear algebraic methods by employing feature value representation or feature weighing schemes (Debole & Sebastiani, 2004). In the related literature on clinical text classification, four types of feature value representation techniques were used: BR, TF, TFiDF, and N-TFiDF (discussed in Section 2.3.1.3(b)). Table 2.4 shows the study-wise frequency distribution of each feature representation technique.

As shown in Table 2.4, in most of the studies, researchers used either the BR or TF technique. Moreover, in studies (Butt et al., 2013; Masino et al., 2016; Napolitano et al., 2016), authors compared the performance of BR and TF to classify clinical reports and reported that BR outperforms TF. Clark et al. (2017) compared the performance of BR and TF to classify psychiatric evaluation reports. They concluded that TF outperforms BR. Kavuluru et al. (2015) compared the performance of BR and TFiDF to classify pathology and radiology reports. They found no significant difference between the findings of BR and TFiDF. Farshchi and Yaghoobi (2013) extracted the BoW features from a dataset of medical news articles and represented the extracted features using TF and TFiDF feature representation techniques. No significant difference was observed between the results obtained through TF and TFiDF. Moreover, Parlak and Uysal (2016a) and Parlak and Uysal (2018) compared the performance of TF and TFiDF to classify Medline abstracts; they revealed that TF outperforms TFiDF. Amrit et al. (2017) compared BR, TF, and TFiDF to classify child abuse consultation reports and reported that TF outperforms the two other techniques. Lucini et al. (2017) compared BR, TF, and TFiDF to classify emergency department reports and reported that TFiDF demonstrates the best performance. Wang et al. (2017) compared BR, TF, and TFiDF to classify incident reports and reported that BR outperforms the two other methods. Danso et al.

(2014) compared BR, TF, TFiDF, and N-TFiDF to classify verbal autopsy reports and found that N-TFiDF exhibits optimal performance. Moreover, Mujtaba et al. (2016) compared BR, TF, TFiDF, and N-TFiDF to classify forensic autopsy reports. They found that TF and TFiDF outperform BR and N-TFiDF.

The above discussion indicates that the choice of feature representation scheme affects the classification results, because all four feature representation schemes have a different design philosophy. Thus, one should always empirically investigate the use of all four types of feature representation schemes on clinical datasets to determine which one presents superior classification accuracy. The BR approach is easy to compute and constructs a basic binary numeric vector to differentiate between two documents. However, it is only suitable for datasets with controlled terminologies and slight conceptual differences. By contrast, the TF, TFiDF, and N-TFiDF approaches maybe suitable for datasets with uncontrolled vocabulary and can easily compute the similarity between two documents. However, these approaches cannot capture the position in a text and fail to extract the semantics and co-occurrences in different clinical reports.

2.8.3 Review of Feature Selection Techniques

As discussed in Section 2.3.1.3(c), feature selection techniques select the most relevant subset of features following certain selection criteria (Guyon & Elisseeff, 2003). Thus, feature selection is widely used for efficient clinical text classification. Nonetheless, in the related literature on clinic text classification, very few studies have employed feature selection to examine the effect of various subsets of features on classification accuracy. Most features for construction of text classification models have been used. In the related literature, the following feature selection techniques were employed.

- ***Information Gain (IG):*** It identifies the significance of a given feature *f* in a feature vector, and the expected reduction in entropy caused by segregating the data sample according to *f* (Yang & Pedersen, 1997).

- ***Chi-square (chi):*** The Chi-square test (χ^2) is the statistical test that measures the relevance of feature *f* with class *c* (Yang & Pedersen, 1997).

- ***Pearson Correlation (PC):*** It is a commonly used method for reducing feature dimensionality and evaluating the discrimination power of a feature in classification methods. It is also a straightforward method for choosing significant features. Pearson correlation measures the relevance of a feature by computing the Pearson correlation between it and a class. Pearson correlation coefficient measures the linear correlation between two attributes (Benesty et al., 2009).

- ***Local Semi-Supervised Feature Selection (LSFS):*** It utilizes class labels to define a margin for each data sample and selects the most discriminative features by maximizing the margins with regard to a feature weight vector(Xu, King, Lyu, & Jin, 2010).

- ***Expert-driven (ED):*** In ED, experts manually rank the discriminative features from the set of given features.

- ***Mutual Information (MI):*** MI is a measure of the amount of information that one random variable has about another variable (Cover & Thomas, 2012). It computes the amount of information of a feature *f* contributes in accurate classification decision. It gives a way to quantify the relevance of a feature subset with respect to the output vector *C*.

- ***Gini-Index (GI):*** It is a non-purity split method. It is widely used in decision trees. It calculates the heterogeneity from the sum of squared probabilities of each class from one (Loh, 2011).

- ***Distinguishing Feature Selector (DFS):*** It aims is to select distinctive features while eliminating uninformative ones considering some pre-determined criteria (Uysal & Gunal, 2012).
- ***Principal Component Analysis (PCA):*** It is a statistical method that utilizes orthogonal transformation to transform a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the smaller of the number of original variables or the number of observations (Wold, Esbensen, & Geladi, 1987).
- ***Multiple Discriminant Analysis (MDA):*** It is a statistical technique that is used to minimize the differences between variables in order to classify them into a set number of broad groups (Hair, Black, Babin, Anderson, & Tatham, 1998).
- ***Bi-Normal Separation Score (BNSS):*** It is defined as $F^{-1}(tp)-F^{-1}(fp)$ where F^{-1} is the inverse cumulative probability of standard normal distribution (Forman, 2003).

Table 2.4 shows the distribution of related studies based on feature selection techniques used. Few studies have employed feature selection techniques to discover discriminative feature subsets. Among those studies, IG, chi square, and Pearson correlation (PC) were mainly used. For instance, Mujtaba et al. (2016) compared three feature selection techniques, namely, information gain (IG), chi square, and PC for discovering discriminative feature subsets to classify forensic autopsy reports. Their experimental results showed that IG and chi-square outperform PC. Kasthurirathne et al. (2016) compared the performance of manually ranked features by experts with that of IG (fully automated feature selection scheme) to classify cancer reports. N significant difference was found between the expert-driven feature ranking and IG techniques. Amrit et al. (2017) used GI and chi-square feature selection schemes to classify child abuse consultation reports; their findings demonstrated that chi square outperforms GI. Parlak and Uysal (2016a) and Parlak and Uysal (2018) compared GI and DFS feature selection schemes to classify Medline abstracts. They revealed that DFS outperforms the GI technique. Buchan et al. (2017) applied PCA and MI feature selection schemes to classify diabetic patient history reports; PCA was found to outperform MI. Table 2.4 shows that IG and chi square are the most widely employed feature selection schemes for clinical text classification. Both techniques obtain robust classification results because they favor the most frequently used clinical terms in an available dataset. Moreover, both techniques use categorized information to discover useful feature subsets. These techniques also consider the information of clinical term absences to determine the category probability (Yang & Pedersen, 1997). The poor performance of MI or GI maybe due to their bias toward low frequent term features (Yang & Pedersen, 1997). In some studies, PC performed worse than chi square possibly because the former is more suitable for dichotomous data than the latter (Sebastiani, 2002; Nicolosi, 2008). Moreover, PC showed the lowest results with minimal numbers of feature subsets. PC selects the features that are most indicative of membership only, whereas chi square and IG consider the features most indicative of membership and non-membership, which maybe useful for classification performance (Yang & Pedersen, 1997; Forman, 2003).

2.9 Review of Machine Learning Algorithms

Table 2.5 shows the machine learning algorithms that were employed in related studies. In addition, it also shows the preeminent algorithm that obtained the highest classification results. Notably, in several studies, authors did not compare various machine learning algorithms but only employed one algorithm. In such studies, the third column value is empty.

Butt et al. (2013) employed SVM, NB, DT, and AdaBoost text classifiers to classify cancer-related death certificates. Their findings showed that SVM outperforms NB, DT, and AdaBoost by obtaining 98% F-measure. Farshchi and Yaghoobi (2013) classified medical news articles utilizing ANN via back propagation, *k*NN, NB, and SVM. ANN using back propagation obtained the highest overall accuracy of 86%. Pineda et al. (2013) employed efficient Bayesian multivariate classification (EBMC) to classify influenza-related clinical reports. Moreover, to demonstrate the effectiveness of EMBC, authors compared its performance with those of NB, BN, RF, SVM, LR, and ANN. EMBC was revealed to outperform all other classifiers by obtaining 99% AUC. Afzal et al. (2013) modified the C4.5 decision tree algorithm to develop a new classifier called My C to classify hepatobiliary disease and renal failure reports. This classifier builds a decision tree by recursively splitting samples using the chi-square test results. To show the effectiveness of My C, its performance was compared with those of C4.5 decision tree, SVM, and Ripper. My C obtained the highest sensitivity of 94%. Moreover, authors experimentally proved that the proposed My C classifier is computationally faster than existing decision tree algorithms. Kasthurirathne et al. (2015) investigated the performances of LR, NB, *k*NN, RF, and DT text classifiers to classify cancer reports. The LR, *k*NN, RF, and DT classifiers obtained the highest accuracy with no significant differences among them.

Pineda et al. (2015) compared the performance of seven different text classifiers, namely, NB, BN, EBMC, RF, SVM, LR, and ANN, to classify influenza-related clinical reports. They reported that the NB, LR, SVM, and ANN classifiers almost obtained similar results. In studies (Danso et al., 2013, 2014), authors compared three different text classifiers (SVM, NB, and RF) for the classification of verbal autopsy reports; SVM obtained the highest accuracy of 83%. Mujtaba et al. (2016) compared the performance of three different classifiers, namely, NB, SVM, and RF classifiers to classify forensic autopsy reports; SVM obtained the highest accuracy of 78% followed by RF and NB. Kasthurirathne et al. (2016) and Kasthurirathne et al. (2017) investigated the performance of NB, LR, DT, RF, and *k*NN text classifiers to classify cancer-related pathology reports. They reported that DT and RF outperform the other techniques by obtaining 90% F-measure. Kasthurirathne et al. (2017) developed and compared the classification models with and without domain-related ontologies to classify cancer-related pathology reports. No significant difference in classification performance was observed when models were developed with or without domain-related ontologies. As shown in Table 2.5, the majority of the studies used SML-based algorithms to classify free-text clinical reports. However, such algorithms may obtain good or poor classification results. Building an effective and efficient text classification model using SML-based algorithms depends on various factors. The most important factor is the extracted features (as discussed in Section 2.8.1) from free-text clinical reports.

In general, if the supervised machine learning algorithm is provided with several independent features that are positively correlated with the targeted class, then the classification performance will be good. Conversely, if the extracted features do not positively correlate with the targeted class, then the classification performance will be poor. Thus, in SML-based ATC models, researchers generally focus on feature engineering (Wolpert & Macready, 1995; Aggarwal & Zhai, 2012a; Domingos, 2012; Witten et al., 2016). SML-based ATC techniques are characterized by two major limitations. First is the knowledge bottleneck, in which a decent SML-based algorithm requires a large number of labeled clinical reports for constructing an accurate classification model (Hastie et al., 2009). Hence, many believe that the quality of SML-based algorithms heavily depends on data rather than algorithms. Another major limitation of SML-based ATC models is difficulty to fix reported quality bugs (Hastie et al., 2009). The developed model is usually a black box, and no direct expert intervention is available to fix the problem unless the constructed model is retrained with new features. However, in such models, there is no guarantee that the reported issue will be fixed well with retraining because the learning process needs to balance all the features in the newly constructed model.

Table 2.5: Machine learning algorithms used in related studies

Study	Preeminent Classifier	Compared with
(Afzal et al., 2013)	My C	DT, SVM, and Ripper
(Pineda et al., 2013)	EBMC	NB, BN, RF, SVM, LR, and ANN
(Butt et al., 2013)	SVM	NB, DT, and AdaBoost
(Danso et al., 2013)	SVM	ZeroR
(Farshchi & Yaghoobi, 2013)	ANN	*k*NN, NB, and SVM
(Zuccon et al., 2013)	RB	NB, and SVM
(Garla et al., 2013)	SVM	RB
(Greaves et al., 2013)	NB	DT, Bagging, and SVM
(Jo, 2013)	Proposed Table-based	*k*NN, NB, ANN, and SVM
(Wagholikar et al., 2013)	NB	RB
(Wei et al., 2013)	SVM	PLS-DA
(Adeva et al., 2014)	SVM	KNN, NB, and Rocchio
(Alghoson, 2014)	RB	--
(Danso et al., 2014)	SVM	NB, and RF
(Gatta et al., 2014)	ESA	Rocchio, and NB
(Luo et al., 2014)	SVM	--
(Nguyen & Patrick, 2014)	SVM	--
(Zhou et al., 2014)	DLM and NB	DLM and NB
(Ye et al., 2014)	BN tuned with expert	BN tuned with TOPZ, and with MEDLEE
(Yeow et al., 2014)	CBR	--
(Pineda et al., 2015)	NB, LR, SVM, and ANN	NB, BN, EBMC, RF, SVM, LR, and ANN
(Bates et al., 2015)	SVM	--
(Comelli et al., 2015)	*k*NN	--
(Deng et al., 2015)	RB	--
(Zuccon et al., 2015)	SVM	NB, LR, J48, RF, and LMT
(Zhou, Q. R. Zhang, et al., 2015)	*k*NN	--
(Jindal & Taneja, 2015)	L-*k*NN	*k*NN
(Kasthurirathne et al., 2015)	LR, *k*NN, RF, and DT	LR, NB, *k*NN, RF, and DT
(Kavuluru et al., 2015)	LR	NB, SVM, and LR
(Koopman, S. Karimi, et al., 2015)	RB	SVM
(Koopman, G. Zuccon, et al., 2015)	SVM	--
(Zhou, A. W. Baughman, et al., 2015)	DT	SVM, *k*NN, and RIPPER
(MacRae et al., 2015)	RB	Human Expert Classification
(Martinez et al., 2015)	SVM	BN, NB, and RF

Table 2.5: continued

Study	Preeminent Classifier	Compared with
(Miasnikof et al., 2015b)	NB	OTM, and Inter-VA4
(Parlak & Uysal, 2015)	BN	DT, and RF
(Rani et al., 2015)	RF	J48, NB, and LAD Tree
(Rios & Kavuluru, 2015)	CNN	NB, SVM, LR, AdaBoost, and Voted Classifier
(Sarker & Gonzalez, 2015)	SVM	NB, and MEM
(X. F. Dai & M. Bikdash, 2015)	NB	--
(Fragos & Skourlas, 2016)	Extended lf-igf-*k*NN	*k*NN, and lf-igf *k*NN
(Hassanpour & Langlotz, 2016)	SVM	--
(Yadav et al., 2016)	DT	--
(Kalter et al., 2016)	RB	Compare with fully automated
(Kocbek et al., 2016)	SVM	SVM, and NB
(Lopprich et al., 2016)	SVM	MEM
(Masino et al., 2016)	SVM and LR	DT, RF, and NB
(Mouriño-García et al., 2016)	BN	--
(Napolitano et al., 2016)	*k*NN	PAUM, and NB
(Parlak & Uysal, 2016a)	BN	DT
(Kasthurirathne et al., 2016)	RF, and DT	LR, NB, and KNN
(Saqlain et al., 2016)	NB	LR, SVM, ANN, RF, and DT
(Sedghi et al., 2016)	PART	NB, and SVM
(Amrit et al., 2017)	NB	RF, and SVM
(Barak-Corren et al., 2017)	NB	--
(Buchan et al., 2017)	NB	MaxEnt, and SVM
(Clark et al., 2017)	ANN	--
(Hassanpour et al., 2017)	SVM	--
(Imane & Mohamed, 2017)	DT	SVM, and AdaBoost
(Lauren et al., 2017)	ELM	--
(Lucini et al., 2017)	SVM	DT, RF, Random Trees, AdaBoost, LR, and NB
Mujtaba et al. (2016)	SVM	RF, and NB
(Oleynik et al., 2017)	SVM	--
(Parlak & Uysal, 2018)	BN	DT
(Shin et al., 2017)	CNN	LR, RF, and SVM
(Kasthurirathne et al., 2017)	RF, and DT	LR, NB, and *k*NN
(Wang et al., 2017)	SVM	LR
(Wu & Wang, 2017)	CNN	LR, NB, and SVM
(Yoon et al., 2017)	RF	NB, and LR

****NB** (Naive Bayes), **BN** (Bayesian Network), **SVM** (Support Vector Machine), **RF** (Random Forest), **DT** (Decision Tree), ***k*NN** (k-Nearest Neighbor), **ANN** (Artificial Neural Network), **CNN** (Convolutional Neural Network), **LR** (Linear Regression), **EBMC** (Efficient Bayesian Multivariate Classification), **ESA** (Entropy Scoring Algorithm), **CBR** (Case-based Reasoning), **RB** (Rule-based), and **AUC** (Area Under the Curve)

****Note:** Table showing the preeminent classifier, compared classifiers and the performance of preeminent classifier in related studies. This should be noted that in most of the related studies customized dataset was used. Thus, it is not viable to compare the performance values across different related studies.

To overcome the aforementioned limitations of the SML-based ATC model, researchers developed ATC models using rule-based algorithms. Alghoson (2014) developed a rule-based classifier to classify Medline abstracts and obtained 60% precision. Moreover, Deng et al. (2015) developed a rule-based classifier to classify pathology reports and obtained 91% F-measure. Koopman, S. Karimi, et al. (2015) developed a rule-based classifier to classify death certificates and compared its performance with that of the SML-based classifier using SVM. Their experimental results showed a minor difference in performance between these two approaches. The rule-based obtained 95% F-measure, and SVM obtained the 94% F-measure. Kalter et al. (2016) developed a rule-based classifier to classify verbal autopsy reports. In the developed rule-based classifier, the authors used the rules defined by domain experts to determine the CoD. Moreover, the results of the developed rule-based classifier were compared with two fully automated systems, namely, Tariff and Inter-VA4. The developed rule-based classifier was revealed to outperform the automated systems by obtaining 80% overall accuracy.

The abovementioned studies showed good classification accuracy using rule-based classifiers, but this approach has its disadvantages and advantages. The rule-based approach is flexible, with rules that are easy to understand. Misclassification results are easier to fix in the rule-based approach than with other approaches. However, the major limitation of the rule-based approach is that it depends more on the deep skills and knowledge of domain experts and rule designers for robustness and scalability. The rule-based approach is not purely a scientific activity but more of a balancing act in architecture, design, and development.

The frequency count of preeminent machine learning algorithms in the related literature is shown in Figure 2.10. In most of the related studies, a customized dataset was used. Thus, comparing the performance values across different related studies is inadvisable. Nevertheless, when performance was analyzed, most of the studies found that the SVM algorithm outperforms many other algorithms, followed by NB, RF, DT, RB, BN, and *k*NN. The least used classifiers were LR and ANN. SVM with appropriate kernel function (such as poly kernel and RBF kernel) can learn good classification rules on linear and non-linear data. Moreover, SVM exhibits enhanced performance with high-dimensional data. The limitations of SVM include memory requirement, complexity, and interpretability (Cristianini & Shawe-Taylor, 2000). In many comparative studies, *k*NN showed the lowest classification performance. The *k*NN algorithm computes the similarity between a new clinical report and a training set of clinical reports. The *k*-most similar cases are retrieved in descending order. The new clinical report is assigned with a class label that belongs to majority of the retrieved *k* reports (Fukunaga, 2013). The modest classification performance of *k*NN maybe due to the linear scaling of features, which possibly inaccurately computed the *k*NN distance measures. Moreover, this assumption of linear scaling becomes misleading when the master feature vector contains non-discriminative features (Hastie et al., 2009; Fukunaga, 2013).

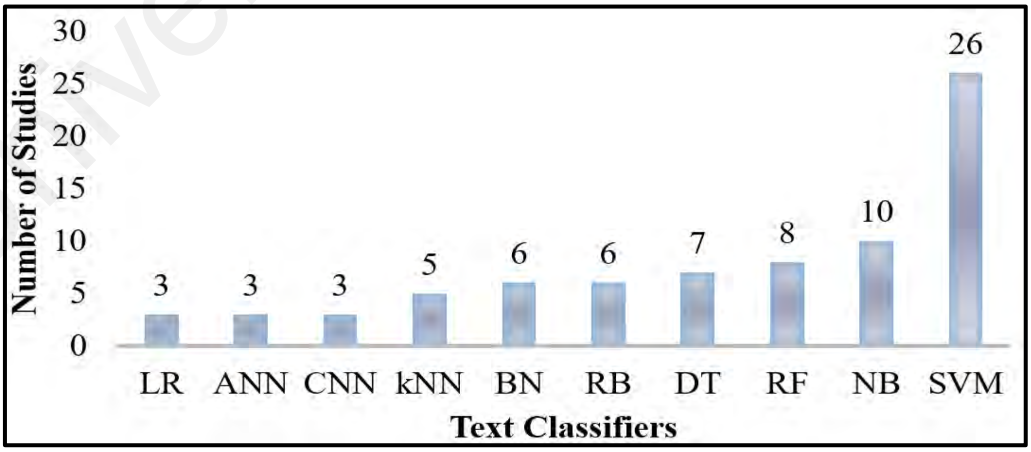


Figure 2.10: Frequency count of machine learning algorithms used in literature

2.10 Review of Performance Metrics

The related studies employed different types of performance metrics to evaluate the classification performance. The commonly used performance metrics for binary class problem were precision, recall, F-measure, accuracy, area under curve, sensitivity, specificity, true positive rate, and false positive rate. In multi-class problems, the commonly used performance metrics were micro or macro average precision, recall, and F-measure. A detailed discussion on these performance metrics can be found in (Sokolova & Lapalme, 2009). Table 2.6 shows the frequency count of performance measures used in each study. Majority of the studies either employed precision, recall, and F-measure or used F-measure and accuracy for binary class problems. For multi-class problems, the studies either used micro-averaging or macro-averaging. In general, macro-averaging is used to determine overall performance of a system across sets of data. Conversely, micro-averaging is effective when the datasets vary in size (Sokolova & Lapalme, 2009).

The analysis showed that the most commonly employed performance metrics ere precision, recall, and F-measure, but these metrics alone may not be sufficient to evaluate classifier performance correctly. For instance, the dataset was imbalanced in various studies. In such cases, the AUC should be the correct performance metric for evaluating classification performance correctly because AUC is suitable in computing the classification performance pertaining to individual class (Provost & Fawcett, 1997; Provost et al., 1998). For instance, Sarker and Gonzalez (2015) collected three different datasets (i.e., Twitter tweets, daily strength instances, and clinical reports) to develop a classification model for predicting adverse drug events. Moreover, the authors used accuracy and F-measure metrics to evaluate classification performance. The Twitter dataset comprised 11.4% tweets that mention ADR and 88.6% tweets that do not mention ADR. Moreover, daily strength dataset contained 23.7% instances that mention ADR and 76.3% instances that do not mention ADR. Finally, the clinical reports were composed of

29.0% ADR mentions and 71.0% that do not mention ADR. In the above example, all three datasets were imbalanced in nature. In such cases, accuracy or F-measure metrics maybe biased toward the majority class. Thus, AUC is a correct choice in such cases to determine the performance of classifiers accurately. Ye et al. (2014) collected the corpus of influenza-related clinical reports to develop a classifier for classifying influenza-related clinical reports. The collected corpus was imbalanced in nature and comprised 592 influenza-related reports and 29,092 non-influenza-related reports. Thus, the authors employed AUC to address the class imbalance problem and evaluate the performance of classifiers accurately.

Several studies employed simple precision, recall, and F-measure for multi-class classification (Farshchi & Yaghoobi, 2013; Jo, 2013; Gatta et al., 2014). However, the suitable performance metrics for multi-class classification problems are micro- and macro-averaging precision, recall, and F-measure (Sokolova & Lapalme, 2009). Mujtaba et al. (2016) developed a clinical text classification model to determine the CoD from a forensic autopsy dataset that comprised eight different CoDs. To evaluate the classification performance, authors employed macro-averaging precision, recall, and F-measure. Danso et al. (2014) developed a clinical text classification model for determining the CoD from a verbal autopsy dataset that comprised 16 different CoDs. To evaluate the classification performance, authors employed macro-averaging precision, recall, and F-measure. Yoon et al. (2017) developed a clinical text classification model to determine the cancer stage from pathology reports. The dataset comprised pathology reports related to cancer. These reports were related to four different stages of cancer: Grades I, II, III, and IV. Thus, to evaluate the performance of classifiers, authors used micro-averaging precision, recall, and F-measure.

Table 2.6: Frequency count of performance metrics used in related studies

Study	Metrics	Count
(Butt et al., 2013; Alghoson, 2014; Luo et al., 2014; Jindal & Taneja, 2015; B. Koopman, S. Karimi, et al., 2015; Martinez et al., 2015; L. Zhou et al., 2015; Kocbek et al., 2016; Mouriño-García et al., 2016; Napolitano et al., 2016; Barak-Corren et al., 2017; Imane & Mohamed, 2017; Lauren et al., 2017; Lucini et al., 2017; Y. Wang et al., 2017)	Recall, Precision, and F-Measure	**14**
(Farshchi & Yaghoobi, 2013; Greaves et al., 2013; Jo, 2013; Wagholikar et al., 2013; Wei et al., 2013; Zuccon et al., 2013; Bates et al., 2015; Sarker & Gonzalez, 2015; Zuccon et al., 2015; Fragos & Skourlas, 2016; Lopprich et al., 2016; Parlak & Uysal, 2016b, 2018)	Accuracy and F-Measure	**13**
(Gatta et al., 2014; Yeow et al., 2014; Y. Zhou et al., 2014; Comelli et al., 2015; X. Dai & M. Bikdash, 2015; Rani et al., 2015; Kalter et al., 2016; Shin et al., 2017; Wu & Wang, 2017)	Accuracy	**9**
(Danso et al., 2013, 2014; B. Koopman, G. Zuccon, et al., 2015; Mujtaba et al., 2016)	Macro Averaging of Accuracy, Recall, Precision, and F-Measure	**4**
(Ye et al., 2014; Kasthurirathne et al., 2015; Parlak & Uysal, 2015; Saqlain et al., 2016; Amrit et al., 2017; Hassanpour et al., 2017)	Accuracy, Recall, Precision, and F-Measure	**6**
(Garla et al., 2013; Hassanpour & Langlotz, 2016; Kasthurirathne et al., 2016; Masino et al., 2016; K. Yadav et al., 2016; Kasthurirathne et al., 2017)	F-Measure, Sensitivity, and Specificity	**6**
(Nguyen & Patrick, 2014; Deng et al., 2015; Rios & Kavuluru, 2015; H. Y. Zhou et al., 2015; Oleynik et al., 2017)	F-Measure	**5**
(Afzal et al., 2013; MacRae et al., 2015; Miasnikof et al., 2015b; Sedghi et al., 2016; Clark et al., 2017)	Sensitivity and Specificity	**5**
(Adeva et al., 2014; Kavuluru et al., 2015; Buchan et al., 2017; Yoon et al., 2017)	Micro Averaging of Accuracy, Recall, Precision, and F-Measure	**4**
(Pineda et al., 2013; Lucini et al., 2017)	Accuracy, F-Measure, and AUC	**2**
(Pineda et al., 2015)	AUC	**1**

2.11 Limitations Related to Existing Literature

This section presents the limitations identified from the reviewed literature. The major limitations were related to feature engineering. Other limitations were identified in dataset collection and performance metric selection, which are discussed in subsequent sections.

2.11.1 Limitations Related to Feature Engineering

Features are vital components in improving the effectiveness of machine learning algorithms (Domingos, 2012). Most of the discussed studies attempted to provide an effective SML-based ATC solution to classify free-text clinical reports by proposing significant features (see Table 2.4). Of these studies, some used fully automated features. However, for clinical report classification, one should not only depend on fully automated features, such as BoW and *n*-gram. These features may produce inaccurate classification results because of two major limitations. First, various experts may use different vocabulary terms to report any event or information in the clinical reports. Thus, these features do not consider word-level synonymy and polysemy when applied on clinical text reports (Yadav et al., 2014). Second, in these features, grammar and word order are disregarded but word frequency is maintained (Cavnar & Trenkle, 1994; Sebastiani, 2002; Yadav et al., 2014; Papadakis et al., 2016). In addition, these features do not capture word inversion and subset matching (Cavnar & Trenkle, 1994; Malliaros & Skianis, 2015; Papadakis et al., 2016; Witten et al., 2016).

To address the first limitation, researchers proposed two solutions. First, the researchers employed expert-driven features in which experts are responsible for extracting discriminative term features from the free-text clinical reports and storing them in lexicons. Experts also add synonyms or related term features in the lexicons to overcome the issue of word-level synonymy and polysemy. Several studies reported that the expert-driven features outperform BoW and *n*-gram features (Zuccon et al., 2013; Ye et al., 2014; Kalter et al., 2016). The second solution is the use of concept-based features (such as BoC and BoP) with content-based features. The studies reported the effectiveness of a hybrid of content-based and concept-based features in improving the classification accuracy of free-text clinical reports (Wei et al., 2013; Nguyen & Patrick, 2014; Bates et al., 2015; Comelli et al., 2015).

To address the second limitation, Yoon et al. (2017) employed GoW features and reported that these features outperform traditional *n*-gram and BoW features. Graph representation provides flexibility and robustness in representing the natural language text compared with traditional *n*-gram. Moreover, the GoW approach can overcome the limitation of word co-occurrence, word order, and word inversion. Nonetheless, GoW is more effective than traditional BoW and *n*-gram but computationally expensive compared with BoW or *n*-gram. Therefore, effective computation approaches for GoW-based features are required. The classification performance of the combination of GoW, BoW, BoP, and BoC features should be empirically investigated to overcome the issue of word order, word inversion, and word-level synonymy and polysemy to classify free-text clinical reports.

To classify free-text clinical reports, researchers should empirically investigate the use of both expert-driven and fully automated features to evaluate the performance of both features (Ye et al., 2014; Koopman, S. Karimi, et al., 2015; MacRae et al., 2015; Kalter et al., 2016). These studies argued that the ATC model constructed through expert-driven features can also serve as a benchmark for the ATC model constructed through fully automated features. Moreover, the ATC model constructed through fully automated features can be accepted for application in a real environment during one of two conditions. First, fully automated models should achieve a specific level of accuracy that is higher than that of expert-driven models. Second, no significant difference in classification performance should exist between both of these models (Ye et al., 2014; Koopman, S. Karimi, et al., 2015; MacRae et al., 2015; Kalter et al., 2016).

From the above discussion, one may question the need to develop an ATC model using fully automated features if the ATC model was already developed using expert-driven features. This argument can be answered in two possible ways. First, the creation of a benchmark is important so that the effectiveness of fully automated features can be compared and evaluated. Second, in various clinical text classification domains, the number of targeted classes increases daily. Therefore, engaging human experts to extract useful features is impractical, expensive, and time consuming. In such a scenario, fully automated features work effectively if the effectiveness of these features has already been evaluated with benchmark features (such as expert-driven features). For instance, in the autopsy domain, experts are responsible to collect autopsy findings from a dead body. On the basis of the autopsy findings, experts are responsible for determining the primary CoD and its ICD-10 code. In the past, experts used the International Classification of Disease Ninth Edition (ICD-9), which contains approximately 18,000 unique codes, to assign primary CoDs. However, ICD-9 was recently enhanced to ICD-10, which contains nine times more codes than ICD-9 (Organization, 1992; Sundararajan et al., 2004; Hazelwood & Venable, 2010; Control & Prevention, 2015). Therefore, the extraction of features from all of these categories is very tedious. For such domains, experts can be utilized to extract features from few categories, and the ATC model can be developed using those expert-driven features to establish a benchmark performance. Fully automated features can then be designed to obtain accuracy equal to or more than that of the benchmark. Once fully automated features obtain the specific level of classification performance, such features can be exploited for the remaining categories without generating expert-driven features.

2.11.2 Limitations Related to Datasets

Dataset size and dataset quality are positively correlated with text classification performance. As shown in Table 2.2, in most of the existing studies, researchers used homogenous–homogenous datasets to develop text classification models for classifying free-text clinical reports. However, several hospitals may have different medical documentation systems and patterns or styles, thereby producing hurdles in generalizing a constructed classifier to multiple hospitals. Hence, collecting clinical reports from more than one organization is recommended to develop more generic classification models (Martinez et al., 2015; Kasthurirathne et al., 2016; Sedghi et al., 2016; Barak-Corren et al., 2017; Kasthurirathne et al., 2017). Moreover, one disease can be reported into a variety of reports. For instance, cancer patients' findings can be reported in both pathology and radiology reports. Hence, multi-modal reports should be used in constructing an accurate text classification model (Kavuluru et al., 2015; Sarker & Gonzalez, 2015; Kocbek et al., 2016; Mujtaba et al., 2016; Hassanpour et al., 2017; Wang et al., 2017). In summary, for an effective classification model, multi-modal clinical reports collected from more than one organization are recommended.

2.11.3 Limitations Related to Performance Metrics

The major limitation in clinical text classification research is the collection of a balanced dataset with sufficient sample size for the training set to enable the text classifiers to learn effectively from the training set and determine the category on the test set. Several studies used imbalanced datasets where the samples of the predicting classes varied in size (Afzal et al., 2013; Kocbek et al., 2016; Amrit et al., 2017). In such cases, many studies did not place appropriate emphasis on employing a suitable validation approach to evaluate the classification performance (Sarker & Gonzalez, 2015). Therefore, in case of imbalanced class distribution datasets, appropriate sampling techniques, such as over-sampling, under-sampling, or SMOTING (Japkowicz, 2000;

Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Tang & Liu, 2005), should be utilized to balance the class distribution in the datasets. For imbalanced datasets, appropriate performance metrics (such as AUC) should be used to evaluate the classification performance accurately (Provost & Fawcett, 1997; Provost et al., 1998). Table 2.2 illustrated that many studies belonged multi-class single-label problems, and Table 2.6 shows very few studies that employed micro- or macro-averaging of precision, recall, and F-measure performance metrics for multi-class single-label problems. Thus, appropriate measures should be used for multi-class classification problems to accurately measure the classification performance (Sokolova & Lapalme, 2009).

2.12 Research Gap Analysis

The review of existing literature revealed only two studies on classifying forensic autopsy reports (Yeow et al., 2014; Mujtaba et al., 2016). Yeow et al. (2014) employed a homogenous–homogenous dataset to classify forensic autopsy reports using case-based reasoning coupled with NB classifier; 80% accuracy was obtained. However, in their experiments, the authors did not consider and use the features of all sections of forensic autopsy reports. Conversely, the authors only used one section (i.e., summary) in the classification learning process. Thus, one can argue that the developed classification model may not be reliable and accurate enough to apply in real-time environments. Moreover, the dataset used in their study was homogenous–homogenous in nature and may not be deployed and used on a wide scale.

Mujtaba et al. (2016) compared various SML-based ATC techniques to classify forensic autopsy reports. The authors used a heterogeneous–homogenous dataset comprising four MoDs and eight CoDs. The dataset included four different types of reports: autopsy reports, notes and remarks of pathologists, death scene information, and eyewitness information. Authors aggregated all these reports to form one report per case.

Subsequently, authors used existing SML-based ATC techniques to classify forensic autopsy reports. Their findings showed that unigram features outperform bigram and trigram features. Moreover, the TF and TFiDF feature representation techniques were found to outperform the BR and N-TFiDF techniques. Chi-square feature selection scheme outperforms IG and Pearson correlation schemes. In supervised machine learning algorithms, SVM outperforms RF, NB, *k*-NN, DT, and ensemble voting classifier by obtaining an overall accuracy of 78%. Thus, the findings of this study are of practical value and serve as references for future works. Moreover, the current findings will act as state-of-art techniques to compare future proposals with existing ATC techniques. To enhance the classification accuracy, other types of features such as expert-driven or fully automated features can be proposed to compare findings with current results.

2.13 Conclusion

This chapter presented a critical analysis of the clinical text classification domain by summarizing major research efforts to acquire a better awareness of the existing solutions in this domain. The related literature was reviewed from six rationale aspects: types of clinical reports used, dataset characteristics, pre-processing techniques, feature engineering techniques, machine learning algorithms, and performance metrics. In the clinical text classification domain, various types of free-text clinical reports were used. The most widely used clinical reports were pathology reports, radiology reports, and Medline biomedical documents. Moreover, in most of the selected primary studies, the authors used their own dataset, which mostly comprised only one type of report collected from only one organization. Nonetheless, in some studies, publicly available datasets such as the OHSUMED and i2b2 datasets were also used. Various pre-processing techniques were applied to remove noisy or irrelevant terms from the datasets. Some commonly used pre-processing techniques were stop word removal; removal of stop words, punctuation, and empty spaces; case conversion; tokenization; and normalization (such as stemming and lemmatization). Some studies reported that the removal of stop words reduces classification accuracy.

In fully-automated feature engineering techniques, a combination of BoW, BoP, and BoC yields robust results. Moreover, the representation of these features in graph structure further increases performance but increases computational time. In addition to these features, expert-driven features also perform well. Mixed results were found regarding the comparison of expert-driven and automated feature engineering approaches. In some studies, expert-driven features outperformed automated features, but several authors reported that automated features lead to better classification results. Therefore, one should always empirically investigate the performance of both automated and expert-driven approaches to evaluate which one will produce optimal results. In majority of the studies, the BR, TF, and TFiDF feature representation techniques were found useful. To remove redundant or non-discriminative features, various studies employed different kinds of feature selection schemes. Chi square and IG showed good results. For ATC, most of the studies either employed SML-based ATC techniques or rule-based ATC technique.

In SML-based ATC techniques, generative and discriminative models were used to classify free-text clinical reports. In most of the studies that employed generative models, NB showed good results. In discriminative models, SVM obtained superior results followed by RF, DT, and *k*NN. Rule-based classifiers also showed promising results. Some studies employed both rule-based and supervised machine learning-based classifiers. Few studies reported that supervised machine learning-based classifiers outperform rule-based classifiers, some reported that rule-based classifiers outperform supervised machine learning-based classifiers, and several studies reported no significant difference between the results obtained by both approaches. For performance metrics, most of the studies used precision, recall, F-measure, and accuracy to measure the classification performance in binary-class classification problems. In multi-class classification problems, authors used micro- and macro-averaging of precision, recall, and F-measure. Notably, several studies used simple precision, recall, F-measure, and accuracy to measure the classification performance of multi-class problems; however, these metrics are unsuitable for multi-class problems. Thus, for multi-class problems, accurate metrics such as micro- and macro-averaging of precision, recall, and F-measure are necessary. Several researchers used the accuracy metric with an imbalanced dataset, but the AUC metric would have been more appropriate in the given situation.

Finally, in the domain of free-text clinical text classification, work on classifying CoDs using complete forensic autopsy reports is limited. Only one study Mujtaba et al. (2016) employed complete forensic autopsy reports to classify CoDs. However, in that study, the authors only used existing feature engineering techniques and machine learning algorithms to classify CoD; a limited accuracy of approximately 78% was obtained. Moreover, the authors mentioned that the possible reason for the low performance classification is the use of irrelevant and non-discriminative features. Thus, robust feature engineering techniques are needed to enhance the classification performance.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

This chapter presents the research methodology for the proposed feature engineering techniques, namely, the semi-automated expert-driven feature engineering technique and fully automated conceptual graph-based feature engineering technique, for classifying the cause of death (CoD) from free-text forensic autopsy reports.

In the field of medical science, computer-aided expert systems are occasionally required only under certain conditions, such as a specific level of accuracy or quality. Thus, existing related literature (Sections 2.8.1 and 2.11.1) generally applied two techniques for developing such systems. These methods were semi-automated expert-driven technique and fully automated machine-learning-based technique. Previous studies comprehensively investigated and empirically evaluated the performances of both techniques (Ye et al., 2014; Koopman, S. Karimi, et al., 2015; MacRae et al., 2015; Kalter et al., 2016). The semi-automated expert-driven technique requires readily available expert knowledge in the form of decision rules, expert domain knowledge, or human expertise. Conversely, the fully automated machine learning-based technique learns the classification or prediction rules from a labeled dataset. Here, the machine learning algorithms explore the training dataset, uncover useful classification or prediction patterns within, and use these patterns to classify new or unlabeled cases. Hence, this study also proposes two effective feature engineering techniques, namely, semi-automated expert-driven and fully automated conceptual graph-based techniques, to comprehensively investigate their performances and suitability for classifying CoDs from forensic autopsy reports. In this thesis, the phrase “effective feature engineering techniques" means that the techniques are more accurate than the existing feature engineering techniques and consume minimal computational time and resources.

The methodology for this research comprised six different phases (Figure 3.1). In the succeeding sections, all these phases are presented precisely. Meanwhile, the specific research methodology and functionality of each proposed technique (semi-automated expert-driven feature engineering technique and fully automated conceptual graph-based feature engineering technique) are comprehensively explained in Chapters 4 and 5, respectively.

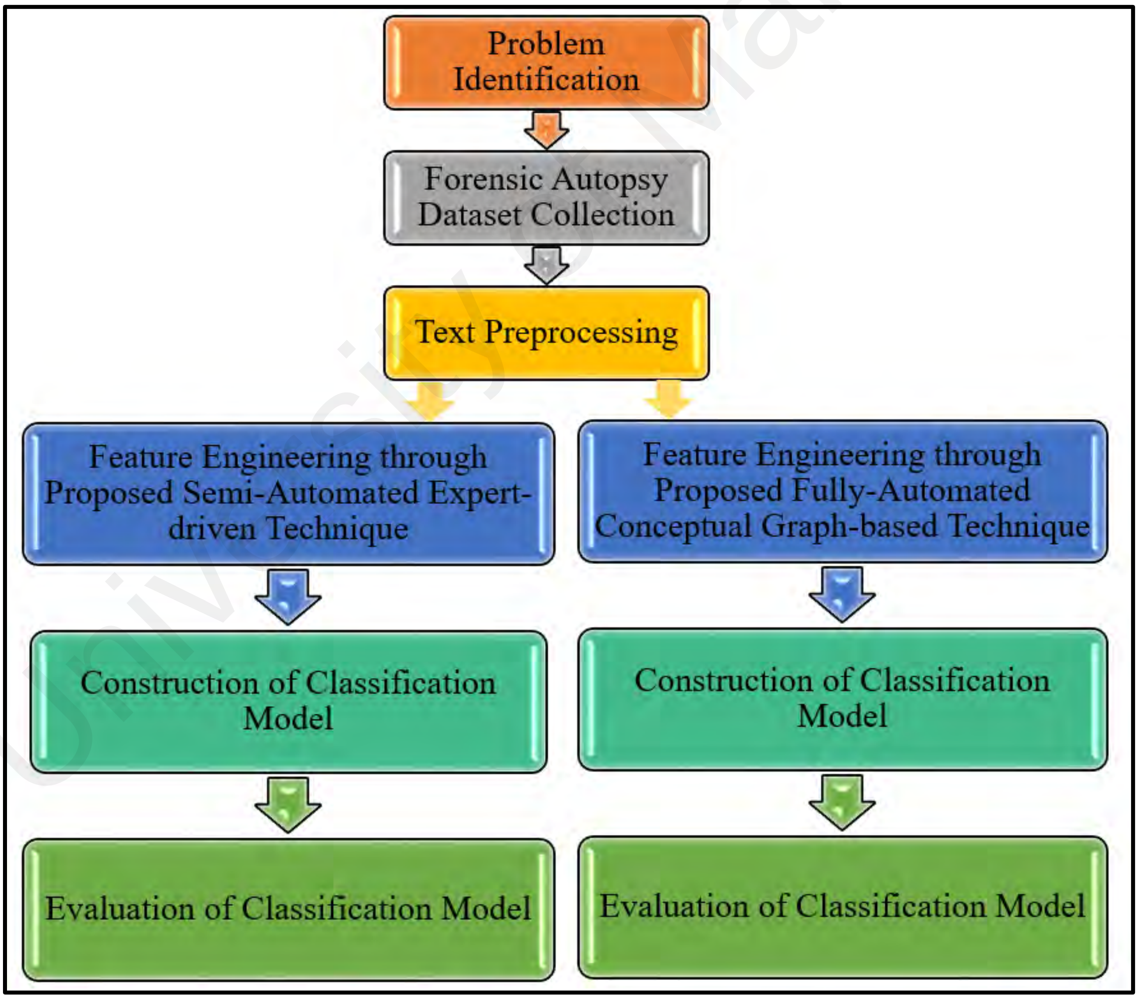

Problem Identification
Forensic Autopsy Dataset Collection
Text Preprocessing
Feature Engineering through Proposed Semi-Automated Expert-driven Technique
Feature Engineering through Proposed Fully-Automated Conceptual Graph-based Technique
Construction of Classification Model
Construction of Classification Model
Evaluation of Classification Model
Evaluation of Classification Model

Figure 3.1: Detailed research methodology and design

3.2 Problem Identification

As discussed in Sections 1.2 and 1.3, Danso et al. (2013), Yeow et al. (2014), and Mujtaba et al. (2016) applied existing feature engineering techniques, such as bag-of-words (BoW), and *n*-gram, to classify the CoDs from autopsy reports. Experimental results showed that these three aforementioned studies achieved low accuracy (between 57% and 80%). This finding proved that the traditional feature engineering techniques (such as BoW and *n*-gram) are inappropriate for extracting useful features from autopsy reports. This shortcoming was observed probably because such techniques ignore word context and order in free-text autopsy reports. Recently, several variants of BoW and *n*-gram feature engineering techniques have been proposed to overcome the limitations of traditional BoW and *n*-gram techniques. Examples of these variants include the skip-gram technique (Mikolov et al., 2013), continuous BoW (CBoW) technique (Wang, 2014), and entropy-optimized BoW (EO-BoW) technique (Passalis & Tefas, 2016). However, the aforementioned techniques involve word sequences and fail to capture word inversions and subset matching (Joachims, 1998b; Sebastiani, 2002; Papadakis et al., 2016) when applied to classify free-text documents (such as autopsy reports). Moreover, these aforementioned feature engineering techniques failed to handle the complex semantic information of texts, such as word-level synonymy and polysemy (Yadav et al., 2014; Malliaros & Skianis, 2015; Dasondi et al., 2016; Jiang et al., 2016; Papadakis et al., 2016). Therefore, effective (highly accurate and consuming minimal computational time and resources) feature engineering techniques remain to be developed to include word order, word context, and word-level synonymy and polysemy to classify free-text clinical reports (such as forensic autopsy reports). These methods are projected to enhance the accuracy of classifying CoDs from forensic autopsy reports.

3.3 Forensic Autopsy Dataset Collection

In this research, the forensic autopsy dataset was collected from Pusat Perubatan Universiti Malaya (PPUM), Kuala Lumpur, Malaysia. The ethics letter provided by PPUM is also shown in Appendix-A. The dataset included forensic autopsy reports involving all four kinds of manners of death (MoDs), namely, accident, suicide, homicide, and natural death. For each MoD, the forensic autopsy reports of the four most common CoDs were collected. A total of 1500 forensic autopsy reports involving 16 CoDs (S06, S38, T07, T75, X80, X74, T71, T14, X93, X99, Y00, Y09, I23, I24, I25, and Z11) and four MoDs were collected. The distribution of forensic autopsy reports based on MoD is shown in Figure 3.2. The detailed distribution of CoDs with several demographic details is provided in

Table 3.1. Notably, in this study, all ICD-10 codes were truncated at the three-character level. For instance, the code S06.9 (unspecified intracranial injury) was converted to simply S06 (intracranial injury). This three-character level truncation was applied because of two reasons. First, the present study aimed to classify the CoD from forensic autopsy reports by using automated text classification (ATC) techniques up to three-character levels. For instance, S06.1 (Traumatic cerebral edema), S06.2 (Diffuse traumatic brain injury), S06.3 (Focal traumatic brain injury), S06.4 (Epidural hemorrhage), and S06.9 (Unspecified intracranial injury) were all truncated to their upper levels, that is, S06 (Intracranial injury). A classifier was then constructed to classify the three-character level CoD from forensic autopsy findings. Second, for any specific CoD, autopsy reports are scarce in our dataset and insufficient to train and construct a robust and effective ATC model. Therefore, to achieve a reasonable training set, all the forensic autopsy reports were converted into three-character level codes.

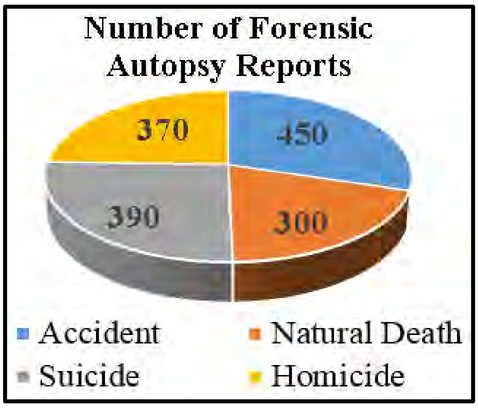


Figure 3.2: MoD-wise distribution of forensic autopsy reports

The MoD, CoD, and corresponding ICD-10 code for these reports were manually labeled unanimously by a team of pathologists. Each report consisted of the detailed examination of the dead body and included the deceased's personal information, external examination, injury-related information, and internal examination. This research also considered other kinds of related reports (when available), such as histopathology reports and toxicology reports. Furthermore, the previous medical history of the deceased was considered. Death scene-related eyewitness information was used in the classification model. Thus, the collected dataset was heterogeneous– homogeneous in nature, because the dataset included various reports, such as forensic autopsy findings, previous medical-related history of the deceased, death scene-related eyewitness information, histopathology report results, and toxicology report results. However, the data were collected from one hospital only (with many branches). In the subsequent paragraph, the details of the forensic autopsy report attributes are discussed. An autopsy report sample is also presented in Appendix-B.

Personal information: This section included the name of the deceased, unique identity number, gender, date of birth, date of death, age upon death, and nationality.

Injury-related information: This portion included the injury-related information, such as the size, location, and pattern of abrasion, laceration, and wound on the body.

Table 3.1: Details of forensic autopsy dataset collection

MoD	CoD	ICD-10 Code	No. of Reports	Gender	Age (Years)
Accident	Craniocerebral injury	S06	120	M: 84% F: 16%	Min: 6 Max: 86 Avg: 41
	Abdominal injury	S38	100	M: 92% F: 8%	Min: 20 Max: 50 Avg: 30
	Multiple injuries	T07	120	M: 87% F: 13%	Min: 14 Max: 87 Avg: 39
	Electrocution	T75	100	M: 88% F: 12%	Min: 5 Max: 44 Avg: 24
Suicide	Intentional self-harm by jumping from height	X80	120	M: 66% F: 34%	Min: 16 Max: 45 Avg: 28
	Intentional self-harm by stabbing	X74	75	M: 54% F: 46%	Min: 12 Max: 47 Avg: 19
	Intentional self-harm by hanging	T71	120	M: 78% F: 22%	Min: 17 Max: 48 Avg: 23
	Intentional self-harm by poisoning	T14	75	M: 68% F: 32%	Min: 18 Max: 43 Avg: 24
Homicide	Assault by handgun discharge	X93	75	M: 89% F: 11%	Min: 19 Max: 53 Avg: 32
	Assault by sharp object	X99	110	M: 71% F: 29%	Min: 17 Max: 59 Avg: 33
	Assault by blunt object	Y00	110	M: 79% F: 21%	Min: 18 Max: 57 Avg: 26
	Assault by unspecified means	Y09	75	M: 73% F: 27%	Min: 13 Max: 58 Avg: 23
Natural	Acute myocardial infar ction	I23	75	M: 63% F: 37%	Min: 23 Max: 64 Avg: 36
	Ischemic heart diseases	I24	75	M: 82% F: 18%	Min: 25 Max: 57 Avg: 39
	Chronic heart diseases	I25	75	M: 76% F: 24%	Min: 27 Max: 65 Avg: 37
	Pulmonary tuberculosis	Z11	75	M: 83% F: 17%	Min: 26 Max: 69 Avg: 34

External examination: This part includes the information about the deceased's external body parts, such as height, weight, eyes, ear, hands, feet, legs, nose, mouth, lips, teeth, and reproductive organs. Information about rigor mortis, hypostasis, and decomposition signs is also recorded here. In addition, any specific symbols or patterns on the body are noted.

Internal examination: This segment includes the anatomical examination of the brain, neck, thorax, cardiovascular system, respiratory system, gastrointestinal tract, liver, spleen, pancreas, endocrine system, kidneys, and urinary bladder.

Previous medical-related history of the deceased: This section records the previous medical history of the deceased in free-text format.

Death scene-related eyewitness information: This section includes the information given by eyewitnesses regarding the death scene.

MoD: This section is the output variable of the forensic autopsy report. Here, the experts determine the MoD, namely, accident, suicide, homicide, or natural death.

CoD: This section is also the output variable of the forensic autopsy report. Here, the experts process the autopsy findings, correlate the findings with previous cases, use their experience, and determine the primary CoD. Moreover, experts also assign the ICD-10 code to the determined CoD.

Each forensic autopsy report comprises three to seven pages, depending on the CoD. For example, a report on an accident-related CoD maybe longer than a report on natural death. An accident-related forensic autopsy report is longer because it contains additional information regarding external and internal injuries, whereas such information maybe unavailable when the CoD is natural. In the experiments, personal information was not used in the prediction of MoD and CoD because these features do not contribute to the prediction of the MoD and CoD. Furthermore, all other features, such as external examination, internal examination, history, and injury-related features, were concatenated in one string for the sake of simplicity.

The biggest obstacle during the forensic autopsy data collection was that the reports were available in hardcopy format, and these reports were typed and converted into softcopy format. Therefore, six postgraduate students were hired from the University of Malaya for data entry for 4 months from November 2015 to February 2016. Two were from the Faculty of Medicine, one was from the Faculty of Dentistry, and three were from the Faculty of Computer Science and Information Technology. The data entry job was supervised by two senior lecturers of the University of Malaya: one from the Faculty of Medicine and one from the Faculty of Computer Science and Information Technology.

3.4 Text Preprocessing

In this phase, the collected forensic autopsy reports were cleaned and prepared for the classification task. Here, several steps were taken to remove noisy text from the forensic autopsy reports by using Python and a natural language processing tool kit (Bird, Klein, & Loper, 2009). First, the sections of the forensic autopsy reports that do not contribute to the improvement in classification accuracy were removed. These sections included the personal information of the deceased, medicolegal case, and deceased identifying features. Furthermore, the remaining sections, such as external examination, internal examination, history, and injury-related features, were concatenated in one string for the sake of simplicity. The whole text was converted into lowercase after removing the special symbols, punctuations, stop words, and unnecessary empty lines and trimming the leading and trailing spaces.

An empirical investigation was performed to investigate the performance of the classification model in the presence or absence of stop words because some related studies showed that the removal of stop words decreases the classification performance (Danso et al., 2013, 2014; Sarker & Gonzalez, 2015; Lauren et al., 2017). However, many studies showed that the presence of stop words decreases the classification performance (Jo, 2013; Adeva et al., 2014; Koopman, S. Karimi, et al., 2015; Koopman, G. Zuccon, et al., 2015; Sarker & Gonzalez, 2015). Mujtaba et al. (2016) empirically investigated the presence and absence of stop-word removal for determining the CoD from forensic autopsy reports. Authors experimental results showed that the presence of stop words decreased the classification performance because of the noise factor. Thus, in this study, stop-words were removed from forensic autopsy reports to increase classification accuracy. The resulting text was then fed to PyEnchant spell checker library (Bird et al., 2009) to correct the misspelled words.

Finally, the Porter stemming algorithm (Porter, 1980; Willett, 2006) was applied on the resultant text to convert the variant forms of a word into its stem or root form. Several related studies reported that the stemming process improves the classification accuracy (Buchan et al., 2017; Wang et al., 2017). Nonetheless, several studies reported that the stemming process does not improve the classification accuracy (Clark et al., 2017; Lauren et al., 2017). Thus, an empirical investigation was performed by Mujtaba et al. (2016) to evaluate classification performance in the presence or absence of the stemming process for determining the CoD from forensic autopsy reports. Authors experimental results showed the effectiveness of the stemming process on the collected forensic autopsy dataset. After applying the stemming process, the resulting text was then fed as input to a feature engineering phase to create numeric MFV. The feature engineering phase is discussed in a subsequent section.

3.5 Feature Engineering

As discussed in Section 2.3.1.3, feature engineering is the key step in classifying the free-text clinical reports by using a supervised machine learning (SML)-based technique (Wolpert & Macready, 1995; Aggarwal & Zhai, 2012a; Domingos, 2012; Witten et al., 2016) because the success or failure of any text classification model heavily depends upon the quality of features used in the classification task. If the extracted features correlate well with the class, then the classification will be easy and accurate. By contrast, if the extracted features do not correlate well with the class, then the classification task will be difficult and inaccurate. The raw data are often not in a form that is amenable to learning, but features from it can be constructed for learning. Much of the effort in text classification goes to this task. It is often also one of the most interesting parts, where intuition and creativity are as important as the technical aspects. The construction of the classification model is often the quickest part compared to feature engineering. Feature engineering is difficult because it is domain specific, whereas learners are for a general purpose. Therefore, in this study, two effective feature engineering techniques, namely, the semi-automated expert-driven and fully automated conceptual graph-based feature engineering techniques, were proposed and developed for classifying forensic autopsy reports. The difference between both the proposed techniques is also shown in Table 5.8, Chapter 5, Section 5.5. Both techniques are precisely discussed in subsequent sections. However, the detailed discussion of these techniques with algorithm, experimental setup, experimental results, and discussion is available in Chapters 4 and 5. Moreover, the need to propose these two feature engineering techniques for classifying forensic autopsy reports is already justified in Sections 2.8.1.3 and 2.11.1.

3.5.1 Proposed Semi-Automated Expert-Driven Technique

In this feature engineering technique, expert pathologists prepared the discriminative features for all 16 types of forensic autopsy reports to determine the CoD. For these 16 kinds of forensic autopsy reports, 16 lexicons were prepared. Each lexicon stored the discriminative features (extracted by the experts) of each kind of forensic autopsy report. After the creation of lexicons, the next important step was to create a numeric MFV for use as input to construct a classification model. To create the MFV, each forensic autopsy report was compared with all 16 lexicons through the proposed expert-matching algorithm. This algorithm matched the occurrence of input report words with each lexicon to form a numeric MFV. Hence, each input report had 16 distinct values and one class value that was the CoD or type of forensic autopsy report. This numeric vector was then fed to the machine learning algorithm to develop the classification model. The detailed functionality of the semi-automated expert-driven feature engineering technique is discussed in Chapter 4.

3.5.2 Proposed Fully Automated Conceptual Graph-Based Technique

In this feature engineering technique, a graph theory was exploited to classify the forensic autopsy reports. Moreover, content-based and concept-based features were mined and represented through graphs. The proposed technique first converted all the autopsy reports (belonging to a particular CoD) into individual report graphs. Each vertex *V* represented a unique term, each edge *E* connected co-occurring terms in the input text, and the weight of an edge *W* was the frequency of the co-occurrence of terms. These report graphs (belonging to a particular CoD) were combined to form an aggregated CoD-level graph. The Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) ontology was used to extract the semantic concepts of the input nodes of the CoD-level aggregated graph. The multi-information obtained from the SNOMED CT ontology can be fully utilized by organizing the unique co-occurring terms, along with their SNOMED CT concepts and descriptors, in the CoD-level aggregated graph. Thus, the aggregated CoD-level graph was converted into a conceptual CoD-level aggregated graph. Finally, to create a numeric MFV, each report was compared with each conceptual

CoD-level aggregated graph to compute six metrics, namely, vertex similarity metric, vertex uniqueness metric, edge similarity metric, edge uniqueness metric, similar edge weight metric, and unique edge weight metric. Each input report in the numeric MFV contained $\left[(6 \times n)+1\right]$ values. Here, 6 indicates the number of six aforementioned metrics, *n* is the number of conceptual CoD-level aggregated graphs, and 1 indicates the class variable. The details of this technique are discussed in Chapter 5.

3.6 Construction of the Forensic Autopsy Report Classification Model

The outcome of the feature engineering phase is a numeric MFV. This MFV is then fed to machine learning algorithms to construct a classification model. However, deciding which classifier will outperform on a given dataset is quite challenging. In the "no free lunch" theorem (Wolpert & Macready, 1995), no single machine learning algorithm performs excellently in all application areas. Hence, a variety of machine learning algorithms should be tested because the philosophy of learning process varies from one machine learning algorithm to another. In this study, six different machine learning algorithms (namely, NB, SVM, *k*NN, J48, RF, and ensemble voting classifier) were applied to evaluate the classification performance of the forensic autopsy reports. Two points were used as guides to narrow down the selection of the machine learning algorithm to be used. First, a specific literature on free-text clinical reports was important to select a certain machine learning algorithm. The preeminence of a machine learning algorithm maybe circumscribed to a given domain (Macià, Bernadó-Mansilla, Orriols-Puig, & Ho, 2013). Therefore, the literature review in Chapter 2 was used as a guide to select the machine learning algorithm. Finally, the performance comparison of several machine learning algorithms on a comprehensive dataset (Fernández-Delgado et al., 2014) was used as basis to select the algorithm. Fernández-Delgado et al. (2014) empirically evaluated the performance of 179 machine learning algorithms on 121 different datasets and concluded that RF and SVM outperform the other algorithms, followed by decision trees, NB, and ensemble voting classifiers. Therefore, the six aforementioned machine learning algorithms were chosen to comprehensively investigate the classification performance of the forensic autopsy reports.

3.7 Evaluation of the Forensic Autopsy Report Classification Model

The performance of the forensic autopsy report classification model was evaluated using four performance measures, namely, macro precision, macro recall, macro F-measure, and overall accuracy. These performance measures were used because of the imbalanced class distribution, and these metrics permitted equal weights for each CoD category (Sokolova & Lapalme, 2009). These evaluation metrics are discussed briefly in subsequent sections. The performance metrics are discussed in detail in Section 2.3.1.5. The proposed feature engineering techniques were also compared with five state-of-the-art baseline feature engineering techniques, namely, the traditional BoW (Harris, 1954), EO-BoW (Passalis & Tefas, 2016), paragraph vector technique (Le & Mikolov, 2014), hybrid of BoW and Word2Vec technique (Enríquez, Troyano, & López-Solaz, 2016), and term graph technique (Papadakis et al., 2016), to show its significance.

3.8 Conclusion

This chapter presented the research methodology used in the design and implementation of the proposed feature engineering techniques for the classification of forensic autopsy reports. Here, the dataset was discussed in detail. Moreover, the preprocessing techniques were discussed to clean the collected dataset. This chapter also discussed briefly the proposed feature engineering techniques and the construction and evaluation of the classification models using the proposed feature engineering techniques. The specific details of the proposed feature engineering techniques and their contribution are provided in Chapters 4 and 5, respectively.

CHAPTER 4: PROPOSED SEMI-AUTOMATED EXPERT-DRIVEN FEATURE ENGINEERING TECHNIQUE

4.1 Introduction

As mentioned in Chapter 2, Mujtaba et al. (2016) comprehensively investigated the performance of traditional supervised machine learning-based (SML-based) automated text classification (ATC) techniques and feature engineering techniques to determine cause of death (CoD) from forensic autopsy reports. The experimental results showed the 78% classification accuracy for the classification of forensic autopsy reports. Hence, to enhance the classification accuracy and to overcome the issues of traditional SML-based ATC techniques and existing feature engineering techniques, this chapter presents an effective semi-automated expert-driven feature engineering technique for determining CoD from forensic autopsy reports. Section 4.2 discusses the technical details of proposed technique. In addition, Section 4.3 presents the underlying methodology for designing semi-automated expert-driven feature engineering technique. The experimental setup and implementation details behind the philosophy of proposed technique are discussed in Section 4.4. Section 4.5 reports the experimental results. The discussion on findings is presented in Section 4.6. Finally, this chapter is concluded in Section 4.7. The significant contribution is the extraction and representation of features from all 16 kinds of forensic autopsy reports with the help of expert intervention. Finally, the extracted expert-driven features are coupled with SML algorithms to construct a classification model for determining CoD from forensic autopsy reports.

4.2 Semi-Automated Expert-Driven Feature Engineering Technique

This section discusses in detail the functionality of proposed semi-automated expert-driven feature engineering technique. In this proposed technique, experts were given the set of forensic autopsy reports along with some summary statistics information to design the feature sets for all kinds of forensic autopsy reports. Once, the feature sets are prepared by the experts, the MFV was developed that was then fed to machine learning algorithms to construct the forensic autopsy reports classification model to determine CoD. Figure 4.1 depicts the detailed functionality of expert-driven approach. In addition, the detailed algorithm of expert-driven feature engineering approach is also shown in Figure 4.2. The subsequent paragraphs comprehensively present the functionality of proposed expert-driven technique.

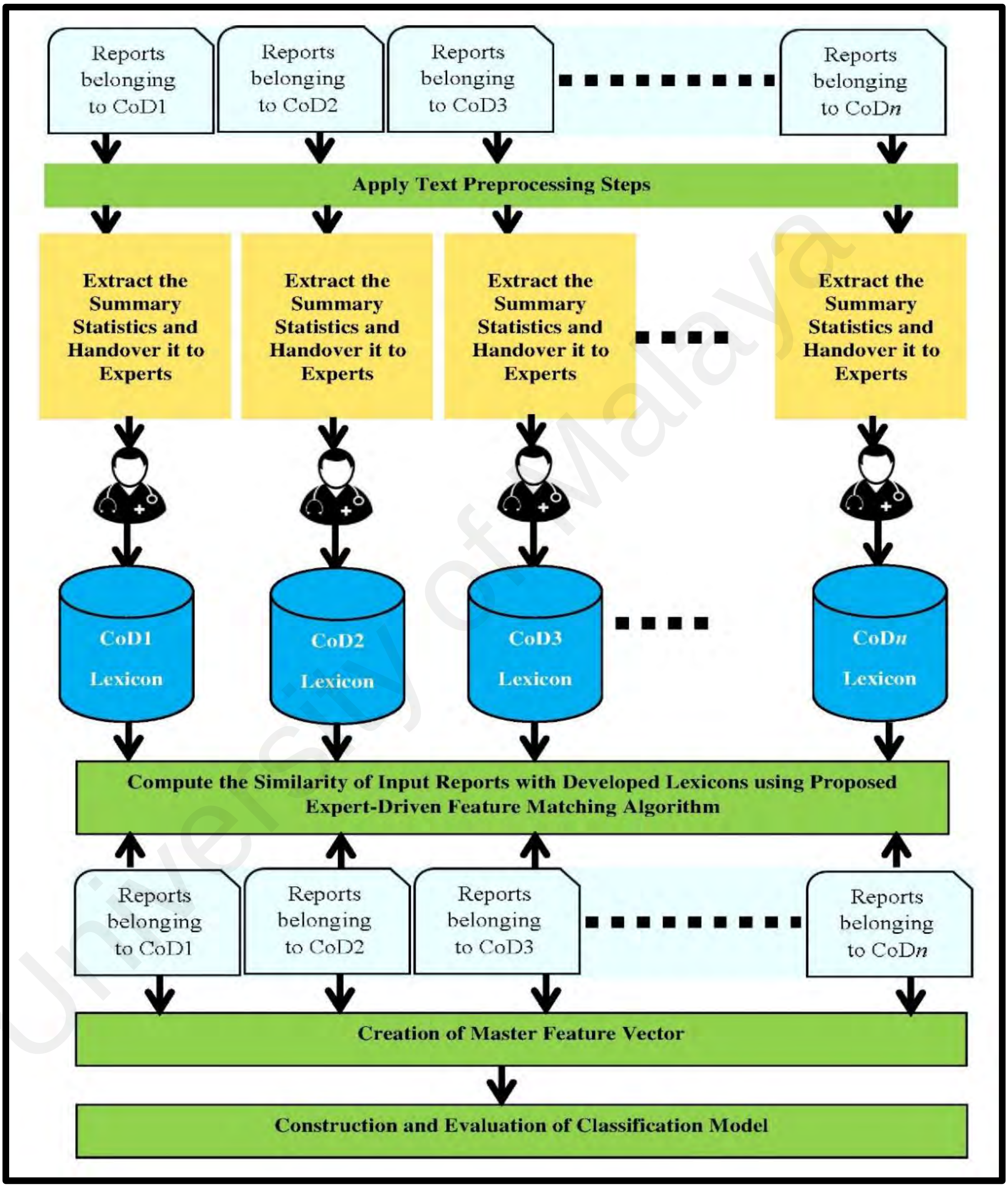
In semi-automated expert-driven feature engineering technique, three experts were involved from Faculty of Medicine, University of Malaya for feature engineering process. These experts were responsible for discovering the discriminative features from all sixteen kinds of forensic autopsy reports. Initially, the experts were given all 1500 autopsy reports classified by CoD. Moreover, the experts were also given the summary statistics for each token to consider this information with their own domain expertise to select and prioritize features for classifying forensic autopsy reports. This summary statistics information was retrieved from pre-processed forensic autopsy reports. A Java program was written for extracting this summary information from preprocessed forensic autopsy reports. The summary statistics supplied to domain experts to aid in discovering the discriminative features includes the following information:

(a) Number of forensic autopsy reports across each Manner of MoD (MoD): This information includes number of forensic autopsy reports belonging to each MoD.

(b) Number of forensic autopsy reports across each CoD: This information includes number of forensic autopsy reports across each CoD.

(c) All distinct tokens with their frequencies in specific MoD: This information includes the list of all distinct tokens with their frequencies across each MoD.

(d) All distinct tokens with their frequencies in specific CoD: This information includes the list of all distinct tokens with their frequencies across each CoD.



Reports belonging to CoD1
Reports belonging to CoD2
Reports belonging to CoD3
Reports belonging to CoD*n*
Apply Text Preprocessing Steps
Extract the Summary Statistics and Handover it to Experts
Extract the Summary Statistics and Handover it to Experts
Extract the Summary Statistics and Handover it to Experts
Extract the Summary Statistics and Handover it to Experts
CoD1 Lexicon
CoD2 Lexicon
CoD3 Lexicon
CoD*n* Lexicon
Compute the Similarity of Input Reports with Developed Lexicons using Proposed Expert-Driven Feature Matching Algorithm
Reports belonging to CoD1
Reports belonging to CoD2
Reports belonging to CoD3
Reports belonging to CoD*n*
Creation of Master Feature Vector
Construction and Evaluation of Classification Model

Figure 4.1: Functionality of expert-driven feature engineering technique

(e) Number of forensic autopsy reports that contain token 'T': This information includes the number of forensic autopsy reports that contains token 'T'.

(f) Number of forensic autopsy reports belonging to specific MoD that contain token 'T': This information includes the number of forensic autopsy reports that belong to same MoD and contains token 'T'.

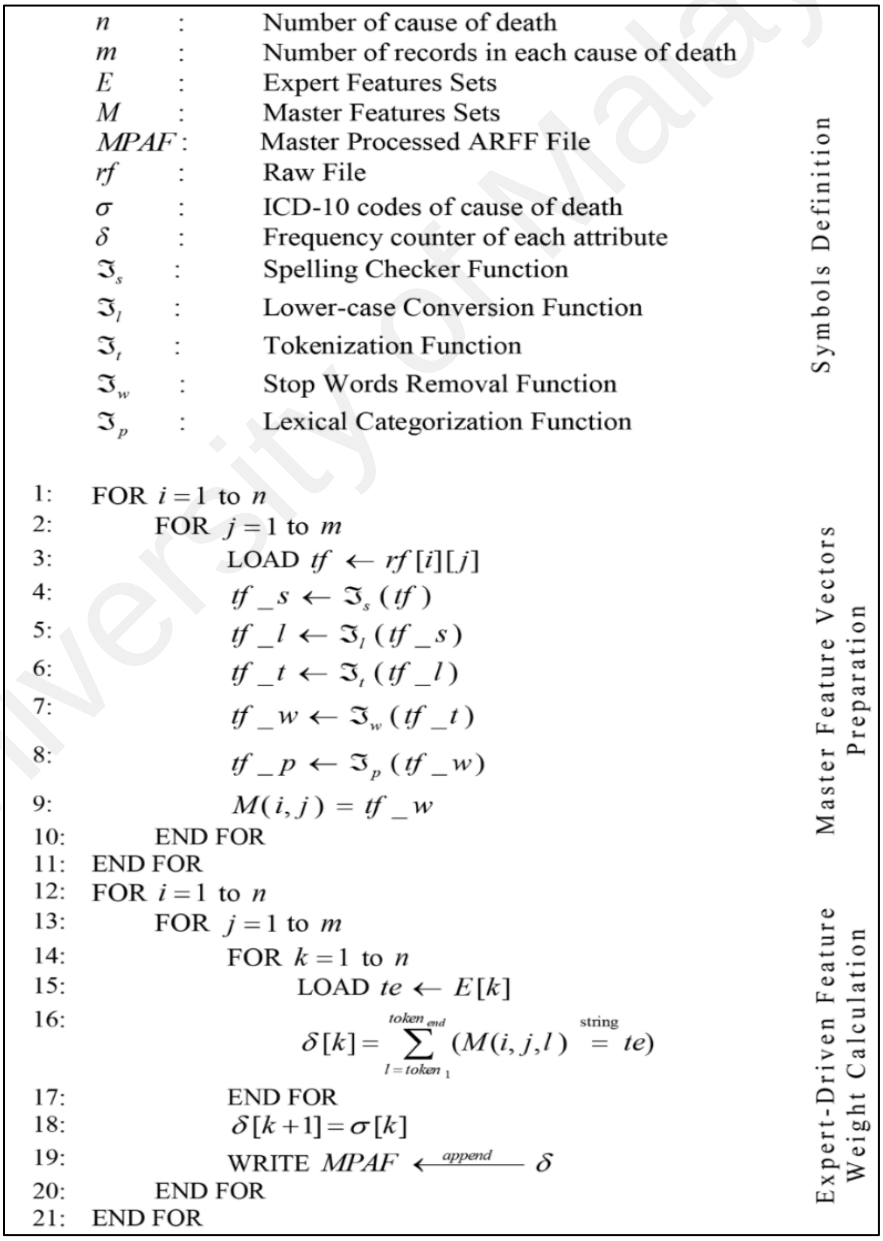
(g) Number of forensic autopsy reports belonging to specific CoD that contain token 'T': This information includes the number of forensic autopsy reports that belong to same CoD and contains token 'T'.

(h) All unigram, bigram, and trigram features with their frequencies related to specific CoD: This information includes the list of all unique unigrams, bigrams, and trigrams with their frequencies across all CoD.

The experts were given all 1500 forensic autopsy reports with aforementioned supplementary information of summary statistics to discover useful features across all different kinds of forensic autopsy reports. Moreover, they were also required to rank the discovered features based on their relevancy. Thus, experts were responsible for designing sixteen different kinds of lexicons (one for each kind of forensic autopsy reports), whereby, each lexicon contains the unique and distinct feature set of each kind of CoD or forensic autopsy reports. A sample of four accident related CoD lexicons are shown in Table 4.1. Here, top 30 expert-driven features are shown for four accident related CoDs. This whole process of creation of feature sets took approximately four months to complete. All three experts individually prepared and ranked the 16 feature sets (one for each CoD). For any disagreements, voting approach was used either to include or exclude the features, or to rank the features in the feature sets. Moreover, in case of ties in voting approach, a fourth expert served as a tiebreaker. The detailed expert-driven algorithm is also shown in Figure 4.2 and explained below.

Suppose, n different number of CoDs having unique σ ICD-10 CoD code are needed to classify. Each cause of death comprises of m number of autopsy reports which are available in the rf raw files. For each cause of death, one expert feature set, E, exists.

E contains the most discriminative features with ranked order list across all n. This E was prepared independently by three experienced domain experts. Moreover, in E, the possible synonyms and alternative words for the selected features were also added. Furthermore, all three domain experts created the prioritized list of features that would predict the accurate CoD from medical autopsy reports. Afterward, the domain experts matched their feature ranking and resolved their conflicts. A fourth pathologist was consulted to resolve the conflicts in case of disagreements. In this manner, n number of E was created. The sample list of expert –driven features is also shown in Table 4.1.



Symbols Definition

n	:	Number of cause of death
m	:	Number of records in each cause of death
E	:	Expert Features Sets
M	:	Master Features Sets
$MPAF$	:	Master Processed ARFF File
rf	:	Raw File
σ	:	ICD-10 codes of cause of death
δ	:	Frequency counter of each attribute
$\Im_s$	:	Spelling Checker Function
$\Im_l$	:	Lower-case Conversion Function
$\Im_t$	:	Tokenization Function
$\Im_w$	:	Stop Words Removal Function
$\Im_p$	:	Lexical Categorization Function

Master Feature Vectors Preparation

```
FOR $i = 1$ to $n$
    FOR $j = 1$ to $m$
        LOAD $tf \leftarrow rf[i][j]$
        $tf\_s \leftarrow \Im_s(tf)$
        $tf\_l \leftarrow \Im_l(tf\_s)$
        $tf\_t \leftarrow \Im_t(tf\_l)$
        $tf\_w \leftarrow \Im_w(tf\_t)$
        $tf\_p \leftarrow \Im_p(tf\_w)$
        $M(i, j) = tf\_w$
    END FOR
END FOR
```

Expert-Driven Feature Weight Calculation

```
FOR $i = 1$ to $n$
    FOR $j = 1$ to $m$
        FOR $k = 1$ to $n$
            LOAD $te \leftarrow E[k]$
                $\delta[k] = \sum_{l = token_1}^{token_{end}} (M(i, j, l) \overset{string}{=} te)$
        END FOR
        $\delta[k+1] = \sigma[k]$
        WRITE $MPAF \xleftarrow{append} \delta$
    END FOR
END FOR
```

Figure 4.2: Expert-driven feature engineering algorithm

Table 4.1: A sample of four lexicons showing top 30 expert-driven features

S06	S38	T07	T75
Scalp	Subconjuctival	Acute	Face
Temporal	Symphysis	Subscalpal	Congestion
Frontal	Pubic	Bruise	Collapsed
Subarachnoid	Renal	Cervical	Blister
Arachnoids	Vessel	Thoracic	Parietal
Hemorrhage	Adernal	Vertebrae	Lpeural
Leptomeninges	Kidney	Sacroiliac	Thickening
Extradual	Parenchyma	Joint	Aspirated
Intracerebral	Pale	Shock	Gastric
Tentorial	Wound	Grazed	Contents
Herniation	Paraumbilical	Ramous	Petechial
Aneurysm	Area	Calcification	Electricity
Ventricles	Nail	Thrombosis	Haemorrhage
Venous sinuses	Beds	Embolism	Oedamatous
Cranial	Bluish	Stenosis	Hand
Odema	Discolouration	Subendocardial	Charring
Nerves	Abdominal	Sternum	Burn
Cerebro	Wall	Perennial	Marks
Spinal	Blunt	Bladder	Shock
Fluid	Penetrating	Hemothoran	Red
Cerebrospinal	Trauma	Infarction	Voltage
Cerebral	Obstruction	Fibrosis	Breath
Vessels	Instestine	Pulmonary	Electric
Thrombosis	Rupture	Embolism	Skin
Preorbital	Distention	Thrombo	Heels
Mandible	Gastrointestinal	Oedema	High
Sphenoid	Contusions	Mural	Bones
Circle	Thoracic	Thrombi	Heart
Willis	Pain	Limbs	Numbness
Ruptured	Bike	Ulna	Tingling

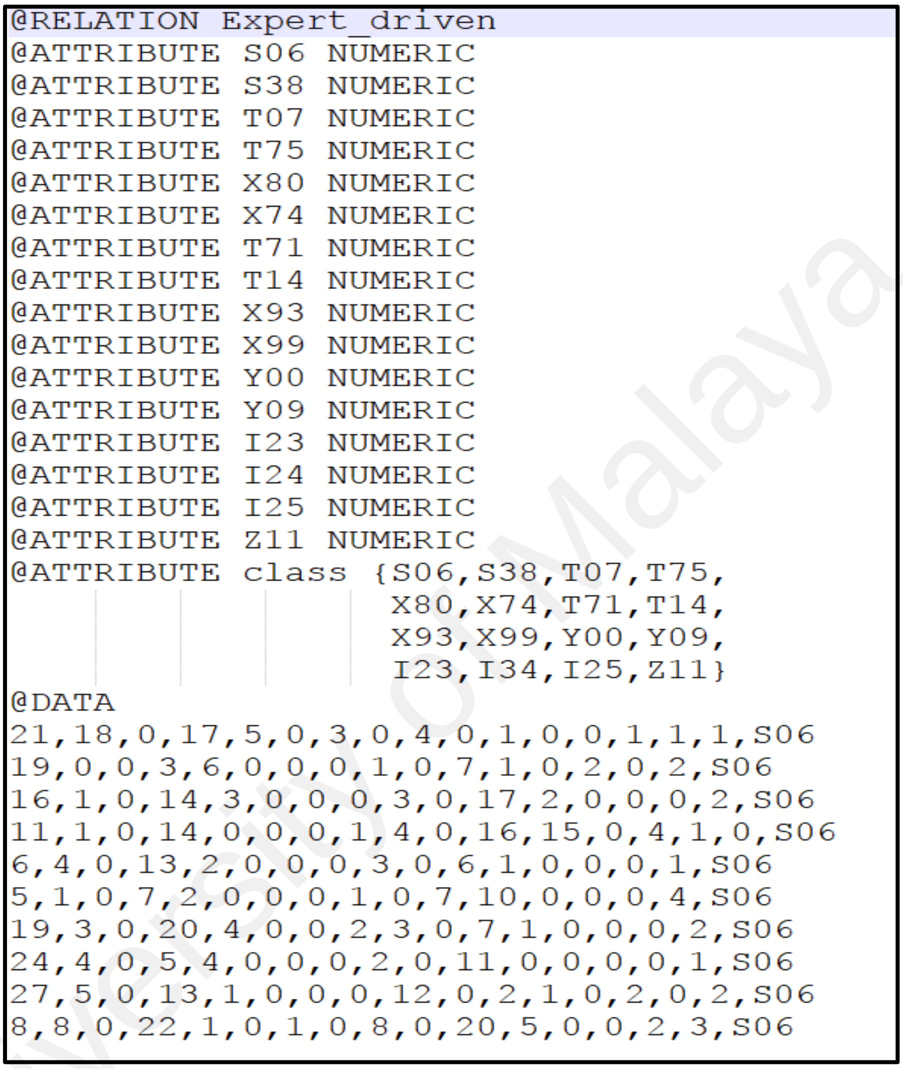
Once the n number of rf are obtained, m number of autopsy reports in each rf , and n number of E , then the n number of M master feature vectors are created. To create the M , first one rf was loaded into memory and performed five different pre-processing tasks on each m in the rf to extract useful features from it. First, $\Im_s$ function was applied on m in rf to correct the misspelled words. Second, $\Im_l$ function was applied on m in rf to convert all words into lower case. Third, $\Im_t$ function was applied on each m in the

rf to tokenize the autopsy reports into unique tokens. Fourth, $\Im_w$ function was applied on m in the rf to remove the most common words which do not contribute in the classification task. Finally, $\Im_p$ function was applied on each m in the rf to assign a lexical category or parts of speech tagging to each token. Finally, the processed m in the rf was stored in M . As such, all n number of rf were converted into n number of M that contained processed m autopsy reports.

After the creation of M and the preparation of E , M and E were loaded into memory. Equation 4.1 was applied on M and E to further process the M and form an ARFF file for classification. As shown, Equation 4.1 matches the tokens of m of M with each E and maintains the frequency count of the features of each E matched with the feature of m of M . Afterward, a unique ICD-10 cause of death (σ) was assigned to m of M , and this m of M was added to the ARFF file to create the training set.

$$ExpertDrivenFeatureWeight = \sum_{l=token_1}^{token_{end}} \left(M(i,j,l) \overset{string}{==} te \right) \tag{4.1}$$

To summarize, once all feature sets were prepared, numeric MFV was then constructed. For creation of MFV, each forensic autopsy report R was taken as an input from forensic autopsy dataset D and it was then preprocessed (as discussed in Section 3.4). Afterwards, unigram features were extracted from pre-processed R . The reason behind the extraction of unigram features is that Mujtaba et al. (2016) comprehensively investigated the performance of unigram, bigram, and trigram features for classifying forensic autopsy reports and reported that unigram features outperformed bigram, and trigram features for classifying forensic autopsy reports. Afterwards, the extracted unigram features were compared will all 16 feature sets (each belonging to one CoD) to count the occurrences of features of R matched with features of feature set1 to feature set 16. Finally, after the conversion of forensic autopsy report into numeric vector, at the end of numeric vector of R , CoD is appended to show the class of that R . Likewise, all remaining reports $(R_1,...R_n)$ of forensic autopsy dataset D were converted into numeric vectors to form numeric MFV. Figure 4.3 shows the sample of MFV. This MFV was then fed to classifiers to construct and evaluate classification model (discuss in Section 4.3).



```
@RELATION Expert_driven
@ATTRIBUTE S06 NUMERIC
@ATTRIBUTE S38 NUMERIC
@ATTRIBUTE T07 NUMERIC
@ATTRIBUTE T75 NUMERIC
@ATTRIBUTE X80 NUMERIC
@ATTRIBUTE X74 NUMERIC
@ATTRIBUTE T71 NUMERIC
@ATTRIBUTE T14 NUMERIC
@ATTRIBUTE X93 NUMERIC
@ATTRIBUTE X99 NUMERIC
@ATTRIBUTE Y00 NUMERIC
@ATTRIBUTE Y09 NUMERIC
@ATTRIBUTE I23 NUMERIC
@ATTRIBUTE I24 NUMERIC
@ATTRIBUTE I25 NUMERIC
@ATTRIBUTE Z11 NUMERIC
@ATTRIBUTE class {S06,S38,T07,T75,
                   X80,X74,T71,T14,
                   X93,X99,Y00,Y09,
                   I23,I34,I25,Z11}
@DATA
21,18,0,17,5,0,3,0,4,0,1,0,0,1,1,1,S06
19,0,0,3,6,0,0,0,1,0,7,1,0,2,0,2,S06
16,1,0,14,3,0,0,0,3,0,17,2,0,0,0,2,S06
11,1,0,14,0,0,0,1,4,0,16,15,0,4,1,0,S06
6,4,0,13,2,0,0,0,3,0,6,1,0,0,0,1,S06
5,1,0,7,2,0,0,0,1,0,7,10,0,0,0,4,S06
19,3,0,20,4,0,0,2,3,0,7,1,0,0,0,2,S06
24,4,0,5,4,0,0,0,2,0,11,0,0,0,0,1,S06
27,5,0,13,1,0,0,0,12,0,2,1,0,2,0,2,S06
8,8,0,22,1,0,1,0,8,0,20,5,0,0,2,3,S06
```

Figure 4.3: Sample of MFV created after running expert-driven technique

4.3 Experimental Design

This section presents the experimental design of construction of classification model for forensic autopsy reports though proposed semi-automated expert-driven feature engineering technique. An extensive set of experiments were run to measure the performance of proposed expert-driven feature engineering technique with state-of-the-arts baseline feature engineering techniques. The complete flow of experimental design is shown in Figure 4.4.

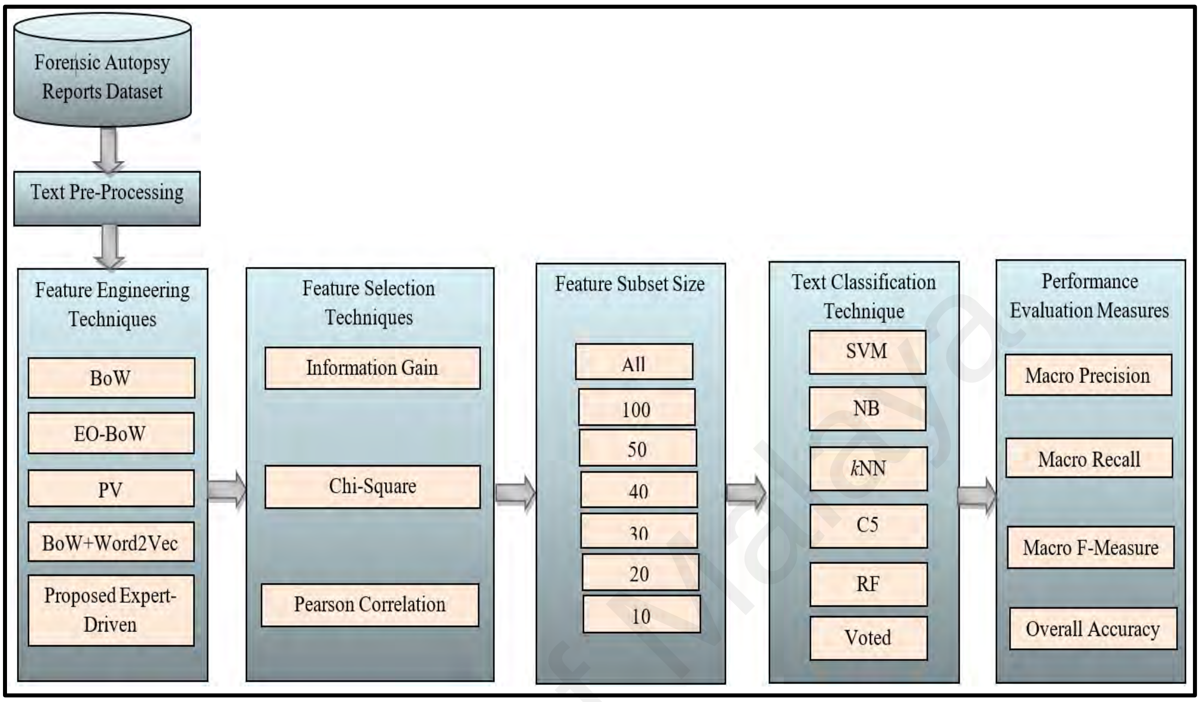


Figure 4.4: Experimental design for evaluation of expert-driven technique

As shown here, the performance of proposed expert-driven technique was evaluated comprehensively. For experiments forensic autopsy dataset (discussed in Section 3.3) was used. Moreover, all reports were preprocessed to remove irrelevant and uninformative features (discussed in Section 3.4). Afterwards, discriminative features were extracted from preprocessed forensic autopsy reports using proposed expert-driven feature engineering technique to create a numeric MFV. This MFV was then fed to six different classifiers (namely, SVM, NB, *k*NN, C5, RF, and ensemble-voted) to evaluate the most suitable classifier for classifying autopsy reports. The justification for the selection of these classifiers is also given in Chapter 3, and Section 3.6. The effect of feature selection on the overall performance of the classification model was also investigated empirically. Thus, to evaluate the effect of feature selection on the overall performance of classification model, three different feature selection schemes were employed and compared namely, information gain (Guyon & Elisseeff, 2003), Chi-square(Guyon &

Elisseeff, 2003), and Pearson correlation (Guyon & Elisseeff, 2003). Finally, the proposed expert-driven feature engineering technique is also compared with four baseline feature engineering techniques to show its significance.

To evaluate the performance of proposed semi-automated expert-driven feature engineering technique, all aforementioned experiments were performed systematically in four different settings. These are:

I. Proposed expert-driven technique and basic classifiers: Here, the discriminative features were extracted from forensic autopsy reports through expert-driven feature engineering technique. The extracted features are then fed to machine learning algorithms for construction of classification models (see Figure 4.5). In this setting, in total six analyses (1 feature engineering technique × six text classifiers) were run to evaluate the performance of classification models.

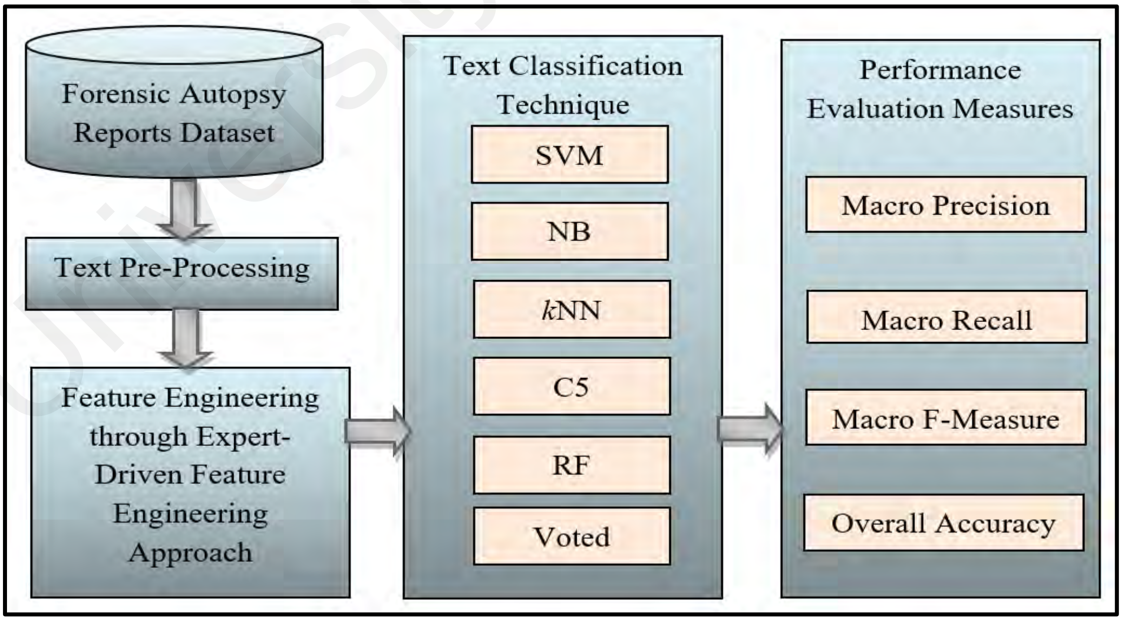

Forensic Autopsy Reports Dataset
Text Pre-Processing
Feature Engineering through Expert-Driven Feature Engineering Approach
Text Classification Technique
SVM
NB
*k*NN
C5
RF
Voted
Performance Evaluation Measures
Macro Precision
Macro Recall
Macro F-Measure
Overall Accuracy

Figure 4.5: Experiments using setting I

II. Proposed expert-driven technique, feature selection schemes, and basic classifiers: It was hypothesized that various subsets of features would produce different performance results in terms of Precision$_M$, Recall$_M$, F-measure$_M$, and overall accuracy. Thus, to evaluate this proposition, three different feature selection schemes namely, information gain (Guyon & Elisseeff, 2003), chi-square (Guyon & Elisseeff, 2003), and Pearson correlation were employed on expert-driven feature sets (Guyon & Elisseeff, 2003). The feature subset sizes of 10, 20, 30, 40, 50, 100, and "all" were selected after performing the sensitivity analysis (discussed in Section 4.4). In addition, these subsets were extracted because of their implementation feasibility, thereby allowing the evaluation of classifier performance within a suitable operating range. Finally, these extracted feature subsets were fed as an input to six different text classifiers to construct effective classification model (Figure 4.6). In this setting, in total 126 analyses (1 feature engineering technique × 3 feature selection schemes × 7 feature subsets × 6 text classifiers) were run to evaluate the performance of classification models.

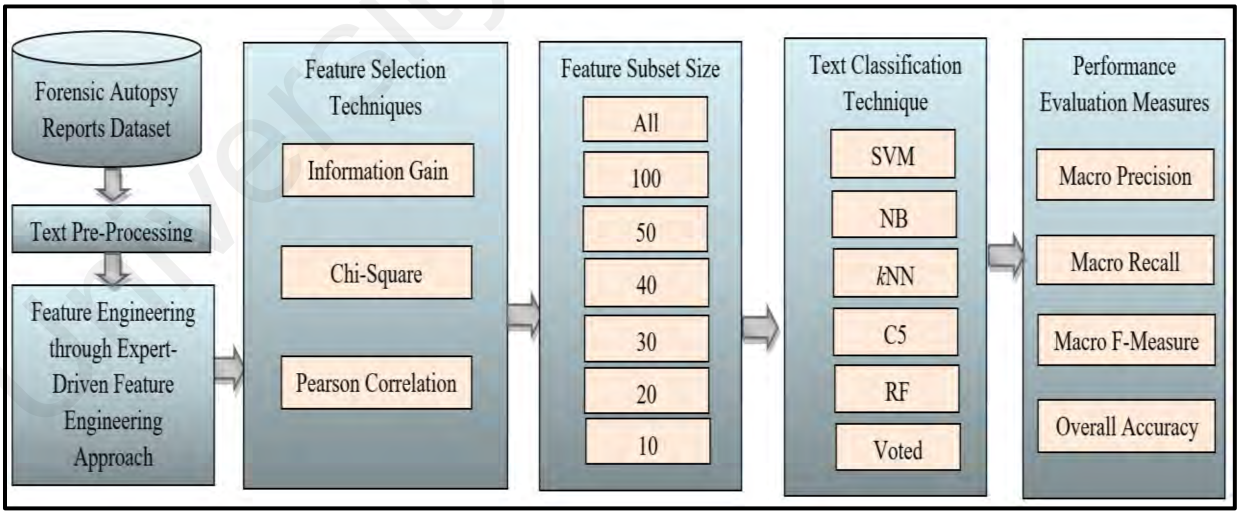

Forensic Autopsy Reports Dataset
Text Pre-Processing
Feature Engineering through Expert-Driven Feature Engineering Approach
Feature Selection Techniques
Information Gain
Chi-Square
Pearson Correlation
Feature Subset Size
All
100
50
40
30
20
10
Text Classification Technique
SVM
NB
kNN
C5
RF
Voted
Performance Evaluation Measures
Macro Precision
Macro Recall
Macro F-Measure
Overall Accuracy

Figure 4.6: Experiments using setting II

III. Comparison of proposed expert-driven technique with state-of-the-art baseline feature engineering techniques: Given the restrictions brought about by privacy or ethical considerations, no public dataset was available for testing the significance of the proposed approach. To examine such significance, four baselines were created from the collected dataset for this research, namely, the traditional BoW technique (Harris, 1954), the entropy-optimized feature-based BoW (EO-BoW) technique (Passalis & Tefas, 2016), the paragraph vector (PV) technique (Le & Mikolov, 2014), the hybrid of BoW and Word2Vec (BoW + Word2Vec) technique (Enríquez et al., 2016). In these experiments, the features from preprocessed forensic autopsy reports were extracted and represented using abovementioned four baseline techniques and one proposed expert-driven technique. Thus, five numeric MFVs were prepared. Afterwards, these five MFVs were then fed as input to six machine learning algorithms to construct classification models (see Figure 4.7). The experiments were conducted to measure the overall accuracy of all six classifiers using these four baseline feature engineering techniques. The baseline accuracy was compared with the accuracy of the proposed expert-driven feature engineering technique using the "all" feature subset size. Hence, in this setting, in total 30 analyses (5 feature engineering techniques × 6 machine learning algorithms) were run to evaluate the performance of proposed expert-driven feature engineering technique with existing baseline feature engineering techniques.

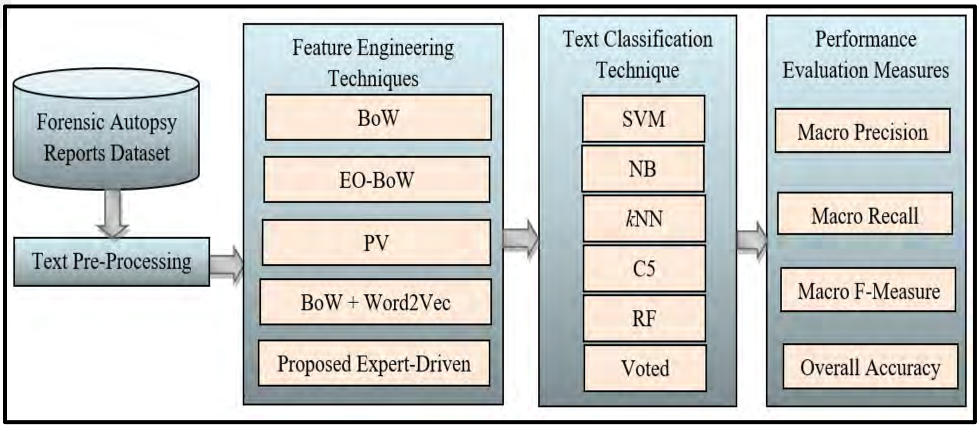


Figure 4.7: Experiments using setting III

IV. In medical autopsy, suitably annotated and statistically independent samples of autopsy reports for the construction and evaluation of classifier are inadequate and expensive. In addition, ethical considerations often restrict the number of autopsy reports collection. Thus, sample size planning is an important aspect in the design of experiments. Hence, to find the optimum sample size for each class, various experiments were performed to examine a range of sample size from 10 to a number of instances where no further improvement in accuracy was observed. Here, all the experiments were performed using expert-driven feature engineering technique and best performed machine learning algorithm (as per experiments in setting I).

All classification experiments were performed in Java using Weka API (Hall et al., 2009; Witten et al., 2016) except the C5. This is because Weka does not provide the implementation of C5. Thus, for this purpose C5 was applied using R programming language. In addition, the selected six machine learning algorithms were run using the parameters shown in Table 4.2. These parameters were used because Mujtaba et al. (2016) rigorously performed the comparative study on classification of forensic autopsy reports using various machine learning algorithms and reported that the machine learning algorithms with these reported parameters outperformed.

Furthermore, in experiments, the 10-fold cross validation was used because that is a standard approach in the field. In this approach, the data were first randomized (shuffled) and then stratified into 10 folds. This randomization and stratification has been performed in advance and remained fixed for all algorithms to make sure all the tests run under the same conditions (the same ordering). The use of cross validation allowed us to obtain average evaluation of the experiments. The more detailed discussion on cross validation can be found in (Kohavi, 1995; Refaeilzadeh et al., 2009). For performance evaluation, four evaluation metrics were used namely, Precision$_M$, Recall$_M$, F-measure$_M$, and overall accuracy. These performance measures were used because of imbalanced class distribution, and these metrics permit equal weights for each cause of death category (Sokolova & Lapalme, 2009). These evaluation metrics are discussed in detail in Chapter 2, Section 2.3.1.5. Finally, statistical significant tests were applied to see the significant different between the analyses results of aforementioned four experimental settings. A pair-wise McNemar statistical test (McCrum-Gardner, 2008; Adedokun & Burgess, 2011; Ott & Longnecker, 2015) was performed (using significance level of alpha = 0.05) to compare the classification performance obtained by six classifiers using proposed expert-driven technique and baseline techniques.

Table 4.2: Parameters selected for machine learning algorithms in experiments

Classifier	**Parameters**
NB	batchSize = 100;debug=false;displayModelInOldFormat=false;doNotCheckCapabilities=false; numDecimalPlaces=2;useKernelEstimator=false;useSupervisedDiscretization=false
SVM	batchSize=100;buildCalibrationModels=False;c=1.0;calibrator=Logistic;checksTurnedOff=False; debug=False;doNotCheckCapabilities=False;epsilon=1.0E-12;filterType=Normalize;kernel=PolyKernel; numDecimalPlaces=2;numFolds=-1;randomSeed=1;toleranceParameter=0.001
*k*NN	KNN=1;batchSize=100;crossValidate=False;debug=False;distanceWeighing=No distance weighing;doNotCheckCapabilities=False;meanSquared=false; nearestNeighbourSearchAlgorithm=LinearNNSearch;numDecimalPlaces=2;windowSize=0
C5	Standard parameters provided in R programming language
RF	bagSizePercent=100;batchSize=100;breakTiesRandomly=False;calcOutOfBag=False;debug=False;doNotCheckCapabilities=False;maxDepth=0;numDecimalPlaces=2;numExecutionSlots=1;numFeatures=0;numIterations=100;outputOutOfBagComplexityStatistics=False; printClassifiers=False;seed=1;storeOutOfBagPrediction=False
Voted	Combination of all aforementioned five classifiers with aforementioned parameters

4.4 Experimental Results

This section presents the results of all the experiments discussed in Section 4.3. The results are presented as per four experimental settings. First, the results of proposed expert-driven feature engineering technique coupled with machine learning algorithms were obtained. Second, the results of these classification models with five different feature selections were obtained. Third, the results of expert-driven technique and baseline techniques were obtained. Fourth, the results for effectiveness of proposed expert-driven techniques and machine learning algorithms on different data samples were obtained. All these results are reported in subsequent subsections.

4.4.1 Experimental Setting I Results

This section presents the results of experimental Setting-I, whereby, the features extracted by expert-driven approach were fed to six different machine learning algorithms namely, NB, SVM, *k*NN, C5, RF, and ensemble voted classifier. The Precision$_{M}$, Recall$_{M}$, F-measure$_{M}$ and Overall Accuracy of six analyses (1 feature engineering technique × 6 machine learning algorithms) are shown in Table 4.3. Here, it can be seen that the values of overall accuracy, Precision$_{M}$, Recall$_{M}$, and F-measure$_{M}$ ranges from 81% to 90%. In addition, it shows that SVM, and RF machine learning algorithms outperformed NB, *k*NN, C5, and voted classifiers by obtaining the overall accuracy between 89% to 90%. Moreover, the lowest performance was observed in *k*NN, and NB machine learning algorithms which produced overall accuracy between 81% to 83%, followed by C5, and voted machine learning algorithms (overall accuracy between 86% to 88%). It can be noticed that there is very minor difference in performance measures obtained by SVM, and RF. Moreover, it can be concluded that NB, and *k*NN classifiers are not suitable for classifying forensic autopsy reports. A pair-wise McNemar statistical test (McCrum-Gardner, 2008; Adedokun & Burgess, 2011; Ott & Longnecker, 2015) was performed (using significance level of alpha = 0.05) to compare the overall accuracy of SVM classifier with all other five classifiers. The statistical difference was observed between SVM and all other classifiers ($p < 0.01$) except RF ($p > 0.05$).

Table 4.3: Results of experimental setting-I

MLA	Overall Accuracy	Precision$_M$	Recall$_M$	F-measure$_M$
NB	82.63%	0.828	0.839	0.831
SVM	89.81%	0.904	0.902	0.903
***k*NN**	81.18%	0.812	0.812	0.812
C5	86.13%	0.865	0.869	0.867
RF	89.13%	0.896	0.896	0.896
Voted	87.85%	0.882	0.867	0.874

** **MLA** = Machine Learning Algorithms

4.4.2 Experimental Setting II Results

This setting presents the results of 126 analyses (1 feature engineering technique × 3 feature selection schemes × 7 feature subsets × 6 machine learning algorithms) that were run to evaluate the performance of classification models. The Precision$_M$, Recall$_M$, F-measure$_M$ and overall accuracy of 126 analyses are shown in Figure 4.8 to Figure 4.11 respectively. In all 126 analyses, the highest Precision$_M$ (0.928), Recall$_M$ (0.926), F-measure$_M$ (0.927), and overall accuracy (92.72%) was produced by Chi-square, followed by information gain and Pearson correlation using SVM machine learning algorithm. The performance of information gain is slightly lower than the performance of chi-square. An increasing trend in Precision$_M$, Recall$_M$, F-measure$_M$ and overall accuracy was observed for the feature subset size 10 to 40 in all three feature selection schemes and six machine learning algorithms. This trend decreased with feature subset size of 50, 100 and “all”. However, in some experiments “all” feature subset size showed better performance in the SVM machine learning algorithm. In machine learning algorithms, SVM produced the highest performance with feature subset sizes of 30 and 40 with the three types of feature selection schemes. Furthermore, the performances of all six machine learning algorithms decreased dramatically when they used 10 features for the classification task, specifically in NB and *k*NN using Pearson correlation. To conclude, the highest performance of classification of CoD was obtained by SVM with the feature subset sizes of 30 and 40, and with chi-square as a feature selection scheme.

Figure 4.8 shows the overall accuracy of all 126 analyses. In the figure, the chi-square significantly outperformed, followed by information gain and Pearson correlation feature selection scheme. In addition, a slight difference in the results of information gain and Chi-square was observed. The lowest results were shown by Pearson correlation scheme. A fluctuating trend was found in the feature subset size. However, the lowest accuracy was observed in the “all” and 10 feature subset sizes. The reasonable accuracy was found in the feature sub set sizes of 30, and 40, respectively. SVM and RF classifiers outperformed in all three feature selection schemes by producing the highest accuracy of 92.23%-92.65% (with a feature subset size of 30 and 40). Moreover, the lowest performance was observed in *k*NN, and NB machine learning algorithms across all three feature selection schemes.

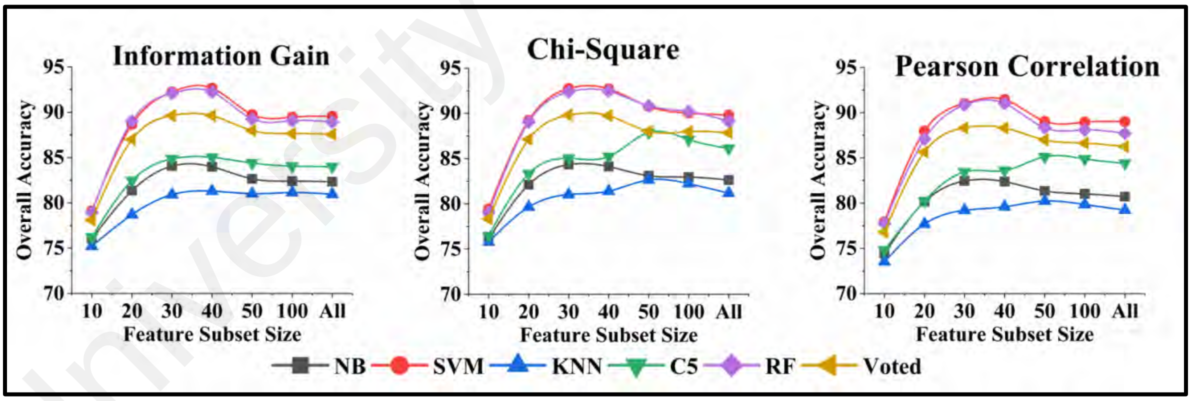


Figure 4.8: Overall accuracy of 126 analyses in setting-II

Figure 4.9 shows the Precision$_M$ of all 126 analyses. As shown here, in all three feature selection schemes, Chi-Square produced the highest Precision$_M$, followed by Information gain and Pearson correlation. The Precision$_M$ of Chi-square yielded a hair greater than that of information gain. Furthermore, in chi-square, SVM, and RF produced the highest Precision$_M$ of 92.82%, and 92.47%, respectively with the feature subset size of “30”. In addition, *k*NN, and NB produced the lowest Precision$_M$ of 75.28% and 76.34%, respectively, with the feature subset of "10" using Pearson Correlation.

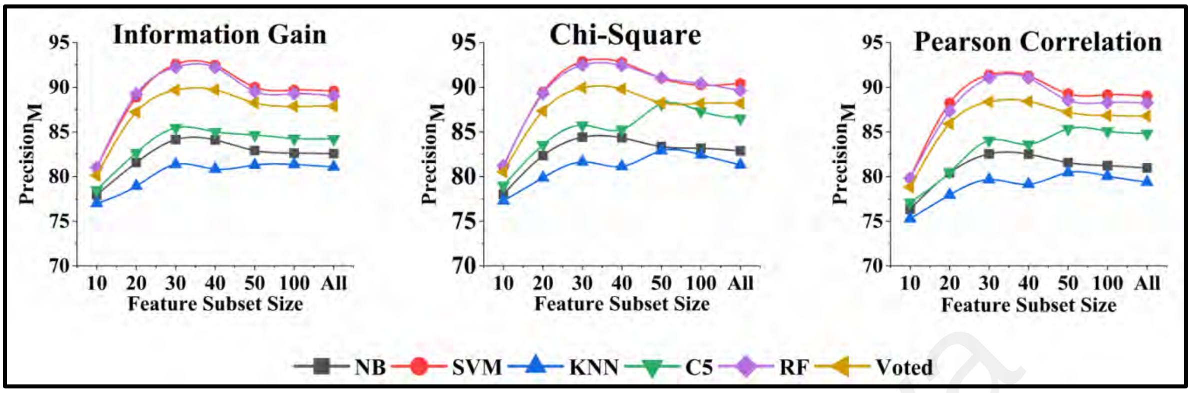


Figure 4.9: Precision$_M$ of 126 analyses in setting-II

Figure 4.10 shows the Recall$_M$ of all 126 analyses. The figure shows that the chi-square outperformed the information gain, and Pearson correlation. Moreover, the lowest Recall$_M$ was observed in Pearson correlation using *k*NN and NB classifiers. Majority of the developed models yielded the lowest Recall$_M$ with feature subset sizes of "all" and 10 and the highest Recall$_M$ with feature subset sizes of 30 and 40. The Recall$_M$ of chi-square yielded a hair greater than that of information gain. Furthermore, in chi-square, SVM, and RF produced the highest Recall$_M$ of 92.68%, and 92.39%, respectively with the feature subset size of "30". In addition, *k*NN, and NB produced the lowest Recall$_M$ of 75.31% and 76.14%, respectively, with the feature subset of "10" using Pearson correlation.

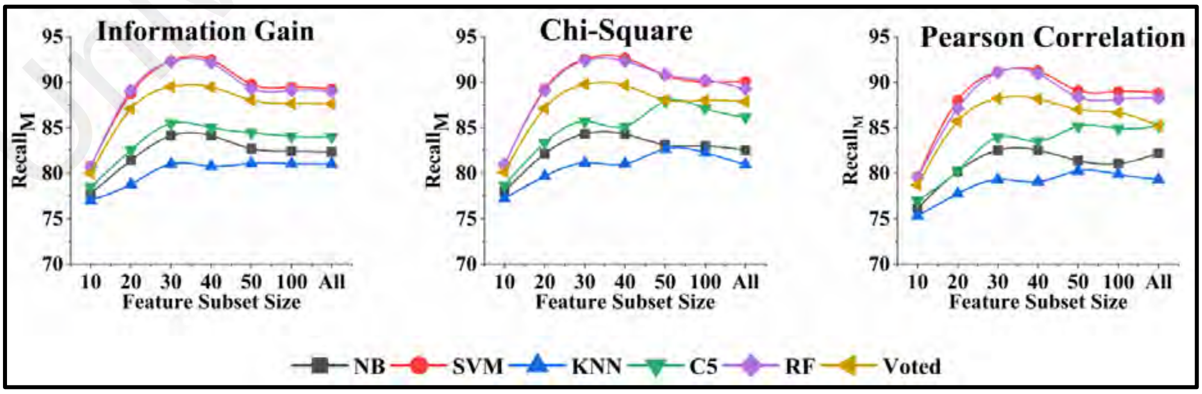


Figure 4.10: Recall$_M$ of 126 analyses in setting-II

Figure 4.11 shows the F-measure$_M$ of all 126 analyses. Here, the highest F-measure$_M$ was produced by the chi-square, followed by information gain and Pearson correlation.

The lowest F-Measure$_M$ was observed in Pearson correlation. In many classification models, both information gain and Chi-square feature selection schemes produced similar results with extremely minute fluctuations. Majority of the classification models yielded the lowest F-measure$_M$ with feature subset sizes of "all" and "10" and highest F-measure$_M$ with feature subset sizes of "30", and "40". In chi-square, SVM, and RF produced the highest F-measure$_M$ of 92.73% and 92.43% respectively, with a feature subset size of "30" and "40". Moreover, the lowest F-measure$_M$ of 75.30% and 76.24 was observed in *k*NN and NB machine learning algorithms with the feature subset size of "10" using Pearson correlation.

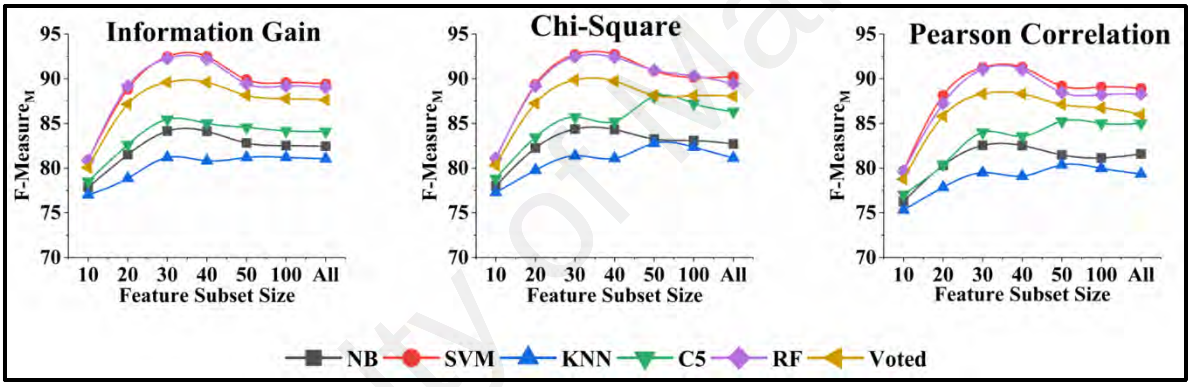


Figure 4.11: F-Measure$_M$ of 126 analyses in setting-II

4.4.3 Experimental Setting III Results

This section reports the findings of 30 analyses that were performed to evaluate the performance of the proposed expert-driven technique compared with four baseline feature engineering techniques. Figure 4.12 presents the accuracy of all 30 analyses. As shown here, in all four baselines, SVM and RF consistently showed a promising accuracy and the lowest accuracy was observed in *k*NN, and NB. In addition, compared to all these baselines, the proposed expert-driven feature engineering technique showed the promising results, followed by PV, and EO-BoW feature engineering techniques. A pair-wise McNemar statistical test (McCrum-Gardner, 2008; Adedokun & Burgess, 2011; Ott & Longnecker, 2015) was performed (using significance level of alpha = 0.05) to compare the overall accuracy of proposed expert-driven technique with all other five baseline techniques using SVM classifier. The statistical difference was observed between expert-driven technique and all other baseline techniques ($p < 0.01$).

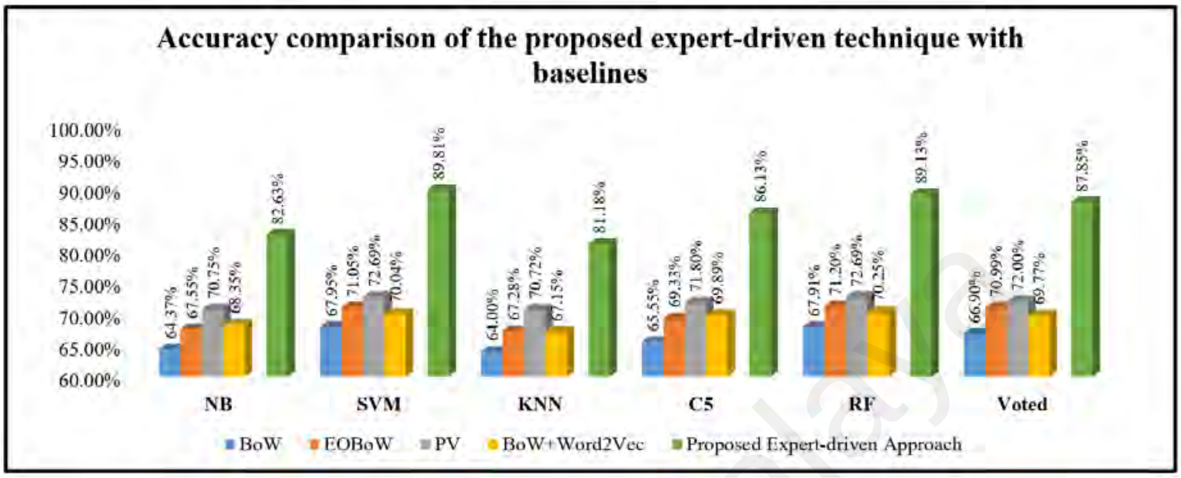


Figure 4.12: Overall accuracy comparison of expert-driven technique with baselines

4.4.4 Experimental Setting IV Results

In forensic autopsy, properly annotated reports for constructing and evaluating classifiers are insufficient and expensive. Moreover, ethical concerns often restrict the amount of report collections. In addition, planning of sample size is main aspect in designing experiments. Therefore, progressive sampling method was used to determine the optimum learning curve. This learning curve describes the performance of classification model with respect to different training set size. In general, there are three phases to determine the optimum learning curve (Figueroa, Zeng-Treitler, Kandula, & Ngo, 2012; Beleites, Neugebauer, Bocklitz, Krafft, & Popp, 2013). In the first phase, classification performance increased swiftly as the size of training set increases. In the second phase, this increment in classification performance is not so rapid even though the training set size increases. In the third phase, there is no or very marginal improvement in classification is observed with increasing training set size. In general, on third phase, the classification model has reached its performance threshold (Figueroa et al., 2012; Beleites et al., 2013). Hence, to determine the ideal sample size for every CoD, several experiments were conducted to determine a range of sample sizes from 10 to a number of instances until accuracy no longer increased. Here, all the experiments were performed using expert-driven feature engineering technique and SVM machine learning algorithm because SVM showed the highest performance.

Figure 4.13 shows these experimental results. As shown in here, the lowest accuracy of 43.63%–67.15% was observed when reports were 10 to 30 in number. Accuracy of 78.67%–89.23% was observed with 40 to 60 autopsy reports. A slight variation of 0.39% in accuracy was examined when dataset size increased from 60 to 70. A very marginal improvement of 0.19% in accuracy was observed when forensic autopsy reports were increased in size from 70 to 100. The consistent accuracy of 89.81% was observed when autopsy reports were 80, 90, and 100 Therefore, it can be concluded that a minimum of 70 to 100 autopsy reports are good enough to construct a forensic autopsy classification model.

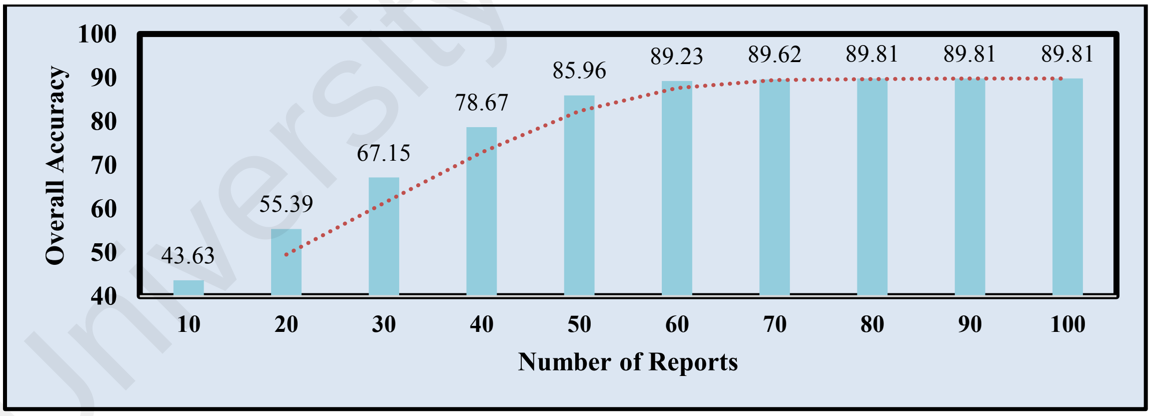


Figure 4.13: Classification accuracy versus number of forensic autopsy reports

4.5 Discussion

The experimental results of this research study show that SML-based ATC techniques coupled with proposed semi-automated expert-driven feature engineering technique can determine the CoD from free text forensic autopsy reports with performance measures between 70%–90%. Furthermore, a considerable difference was observed in most of the analyses. From the experimental findings, different combinations were determined to optimize the performance of each measurement.

To optimize the overall accuracy, Precision$_M$, Recall$_M$, and F-measure$_M$, SVM machine learning algorithm built with expert-driven feature engineering technique, and chi-square feature selection schemes using a subset of 40 features is recommended. In most of the experiments, information gain produced results that are slightly lower with those returned by Chi-square. Pearson Correlation showed the lowest performance results in all of the experiments.

This section provides the theoretical analysis of the SML-based ATC techniques used in this study, namely, proposed and baseline feature engineering techniques, feature selection schemes, and machine learning algorithms.

4.5.1 Theoretical Analysis of Proposed and Baseline Techniques

The performance of a classification task primarily depends on the quality of features. Irrelevant and inadequate features typically produce unsatisfactory performance results. Therefore, the key task in clinical text classification is to identify the most relevant, discriminative, and powerful features using state-of-the-art feature engineering techniques (Domingos, 2012). Accordingly, the proposed expert-driven feature engineering technique was compared with four existing state-of-the-art baseline feature engineering techniques. The experimental results demonstrated the effectiveness of proposed expert-driven technique compared with existing baselines namely, BoW, EO-BoW, PV, and BoW+Word2Vec.

The proposed expert-driven technique performed better than existing baselines because it overcomes the limitation of existing feature engineering techniques, such as word-level synonymy and polysemy. The pathologists might have used different synonyms and vocabulary while preparing the autopsy reports. For instance, many pathologists used the tokens “abrasion,” “graze,” and “trauma” interchangeably. Hence, in the proposed expert-driven approach, the intervention of experts (for feature engineering) overcomes the limitation of word-level synonymy and polysemy by considering word synonyms and concepts. In expert-driven technique, it was suggested to experts during the creation of expert-driven features to select the features that were the most discriminative to a particular CoD. In addition, experts were also suggested to come up with a possible set of synonyms of selected features. Hence, the resultant expert-based feature space comprised of rich set of discriminative features for each CoD under consideration. Therefore, potential researchers should not only rely on results produced by automated feature engineering techniques but should also explore more features with the help of domain experts.

The possible reason behind the unsatisfactory performance of the BoW technique is that it disregards word semantics (Lewis, 1992; Nigam, McCallum, Thrun, & Mitchell, 2000; Sebastiani, 2002). The hybrid of BoW + Word2Vec technique produced the lowest results possibly because this technique uses the voting approach between the BoW and Word2vec techniques. Therefore, BoW results may affect the classification result of the Word2Vec technique. The EO-BoW technique performed better than the BoW and the hybrid Bow + Word2Vec techniques because it calculates the uncertainty of any given features with clinical reports categories. The EO-BoW technique associates features with the clinical report category that obtains the lowest entropy value for that feature. The PV technique performed better than the BoW, hybrid BoW + Word2Vec, and EO-BoW techniques because it learns vector representation for variable length paragraphs of text. Vector representation is learnt to predict the surrounding words in context samples from a paragraph. Hence, this technique captures more semantics from a paragraph than Word2Vec, Bow, or EO-BoW.

The proposed expert-driven feature engineering technique was much faster than all existing baseline approaches. Such result was caused by techniques, such as EOBOW, PV, Word2Vec, in automated feature engineering which consider the whole dataset in determining the most discriminative features by applying various computational methods. However, in expert-driven feature engineering technique, the features were already engineered by experts, hence, this approach only calculated the expert-driven feature weighted from forensic autopsy reports and prepared the classification data using frequency count. Furthermore, the classification file or numeric MFV prepared by expert-driven feature engineering technique was much smaller in size compared with that of baseline techniques. The classification file prepared by expert-driven technique only contained the number of attributes equivalent to the number of classes. Conversely, the other four baseline techniques counted each token as one feature after tokenization, and the number of attributes was equal to the number of unique tokens. Therefore, the baseline techniques required longer classification time. Finally, the proposed expert-driven feature engineering technique can be used in classifying any kind of clinical reports. The only thing required by this technique is the features from an expert.

4.5.2 Theoretical Analysis of Feature Selection Schemes

The accuracy of classification task usually depends upon the quality of features set. The inadequate, extraneous, and irrelevant features may generate less accurate and incomprehensible results. Therefore, it is an important task to remove irrelevant and non-discriminative feature subset from master feature set by using feature subset selectors algorithms prior to classification (Hall & Smith, 1998). The purpose of feature subset selection is to decide which number of features to include in classification and which to remove. For this research, it was hypothesized that various subsets of features would produce different performance results in terms of Precision_M, Recall_M, F-measure_M, and overall accuracy. To evaluate this proposition, it was aimed to determine the best feature subset size for the classification of forensic autopsy reports from all 16 expert-driven lexicons to improve the classification performance using aforementioned three feature selection schemes. To discover the best feature subset size, initially, the subset of 10 features were selected using all aforementioned three feature selection schemes to evaluate the performance of all six classifiers. The number of features were increased up to the point where no further improvement in performance was found to determine the optimal learning curve. In addition, the performance of six machine learning algorithms using 'all' features were also evaluated. In most of the experiments, it was noticed that increasing the size of feature subset from 10 to 40 led to considerable improvements in experimental results. Conversely, increasing the size of feature subset from 40 to 100 and "all" did not cause considerable improvements in the results. As a result, it can be inferred that a feature subset of larger size may not positively affect the results and may cause over-fitting during the classification. Therefore, selecting appropriate and relevant features can reduce over-fitting by the machine learning algorithms in the training dataset (Sebastiani, 2002).Thus, to determine an optimum size of features in feature vector, researchers are suggested to perform sensitivity analysis to examine a range of feature sizes from point 10 to a point where no improvement in accuracy is observed to obtain the optimal learning curve.

In feature selection schemes, chi-square performed better than information gain and Pearson correlation, most likely because it measures divergence from the expected distribution, assuming that feature occurrence is independent of class value (Forman, 2003). Moreover, in this study, the classification task was a multi-class classification task, and chi-square generalizes the multi-class classification task well (Nicolosi, 2008). In addition, chi-square is easier to compute than information gain and Pearson correlation and works better with categorical values (Forman, 2003; Dasgupta, Drineas, Harb, Josifovski, & Mahoney, 2007). The Pearson correlation performed worse than chi-square because it is suitable for dichotomous or contend data (Sebastiani, 2002; Nicolosi, 2008). Moreover, Pearson correlation showed the lowest results with minimal numbers of feature subsets, such as 5, 10, and 20. This poor performance maybe because Pearson correlation selects the features that are most indicative of membership only, whereas chi-square and information gain feature selection consider the features most indicative of membership and non-membership that maybe valuable for classification outcomes (Yang & Pedersen, 1997; Forman, 2003). Nonetheless, chi-square and information gain performed well because of independent feature scoring. However, independent feature scoring renders chi-square and information gain susceptible to distraction by strongly discriminating features for easier classes; therefore, the two techniques cannot select valuable features for difficult classes (Dasgupta et al., 2007; Nicolosi, 2008).

4.5.3 Theoretical Analysis of Machine Learning Algorithms

According to the "no free lunch" theorem (Wolpert & Macready, 1995), there is no single machine learning algorithm that performs best in all application areas. Hence, a variety of machine learning algorithms should be tested. Therefore, the performance of six different machine learning algorithms (NB, SVM, *k*NN, C5, RF, and voted) was evaluated to classify free-text forensic autopsy reports.

Among these six machine learning algorithms, the SVM classifier showed the best performance because the task of predicting CoD is not linearly divisible and SVM utilizes threshold functions to split classes with margins. Furthermore, SVM is not vulnerable to over-fitting because of its independence among features; hence, SVM classifier does not suffer from the higher number of features (Joachims, 1998a). The results obtained by RF were hair less than the results of SVM. This is because; RF performance can be compromised when the dataset contains enormous features and very few numbers of informative features. Hence, the resultant trees in a forest populated by less powerful features that eventually produce the incorrect predictions (Xu et al., 2012). The results obtained with ensemble-voted algorithm are slightly lower than those of SVM and RF, but are slightly higher than those of NB, C5, and *k*NN, because ensemble-voted classifier combines several machine learning algorithms and selects the classification results based on majority voting (Kodovsky, Fridrich, & Holub, 2012). Therefore, in our experiments, this classifier includes the strengths and weakness of the other five classifiers i.e., NB, SVM, *k*NN, C5, and RF.

The lowest performance was observed among the NB, *k*NN, and C5 algorithms. This maybe because, NB classifier supposes conditional independence among features probably invalid for the current dataset (Lewis, 1998b). This conditional dependence in features becomes more complicated as the number of features increases, thus negatively affecting the NB performance. *k*NN had the worst performance because it is a lazy learner algorithm that does not learn from a training set and instead utilizes the training set itself for classification. Therefore, the *k*NN does not generalize the classification problem effectively and is not robust for noisy data (Liu, Moore, Yang, & Gray, 2004). Moreover, to predict CoD for new forensic autopsy cases, *k*NN will locate the *k*-nearest neighbors to the new instance from the training set, and the predicted class label will be assigned as the most common label in the *k*-nearest neighbors (Liu et al., 2004). C5 decision tree exhibited the lowest performance in predicting CoD from forensic autopsy reports because all attributes in the MFV represented continuous data that hinders finding the optimal thresholds needed to construct the C5 decision tree (Dreiseitl et al., 2001). Hence, C5 maybe unsuitable for classifying forensic autopsy reports.

An error analysis was performed for misclassified cases. About 39% of the misclassified cases were due to the general problem with report ambiguity. About 36%

of the misclassified cases were due to the Spelling variations. For instance, in many forensic autopsy reports the word ‘Leptomeninges’ was also written as ‘Leptomeniges’, and ‘Laptomeneges’. About 7% of the misclassified cases were due to the the present of negation besides the available features. In addition, certain features of injury findings (such as degree of displacement of a skull fracture) were not explicitly reported in few of the reports related to Craniocerebral injury and multiple injury classes, and therefore difficult to detect by developed classification model. This issue caused around 11% of misclassification error. About 7% of the misclassified cases were due to the report conversion error. As discussed in Section 3.3, the forensic autopsy reports were available in hardcopy format. Thus, these reports were typed by students to convert hardcopy reports into softcopy format. During the conversion, few issues (such as missing of sentences and discriminative words that were originally available in the report) were observed in some sections of the autopsy reports.

4.6 Strengths and Limitations of Expert-driven Technique

There are several strengths and weaknesses of proposed expert-driven feature engineering technique. This section briefly describes its strengths and weaknesses.

The main strength of proposed expert-driven technique is that it obtained the highest accuracy when compared to existing baseline techniques. This is because, human experts (pathologists) were responsible for extracting and ranking useful features that belonging to specific CoD forensic autopsy reports. Moreover, in SML-based ATC techniques, one of the crucial performance measures is the computational time taken by machine learning algorithm in building the classification model.

Figure 4.14 shows the average computational time for all six machine learning algorithms in all seven feature subset sizes (i.e. 10, 20, 30, 40, 50, 100, and ‘all’ selected using Chi-Square) by using proposed expert-driven technique and four baseline techniques. All 126 analyses were run on Corei7 system having 2.80 GHZ clock speed and a 16-gigabyte memory. As shown here, the classification time of proposed expert-driven technique is much faster than the baseline techniques. Moreover, in baseline techniques, BoW and EO-BoW proved to be faster than PV and BoW+Word2Vec technique. In classifiers, *k*NN and NB required the least time to construct the decision model. Nevertheless, in the majority of the experiments, SVM and RF showed the highest accuracy, Precision$_{\mathrm{M}}$, Recall$_{\mathrm{M}}$, and F-measure$_{\mathrm{M}}$, however, they both took the longest computational time to build a classification model.

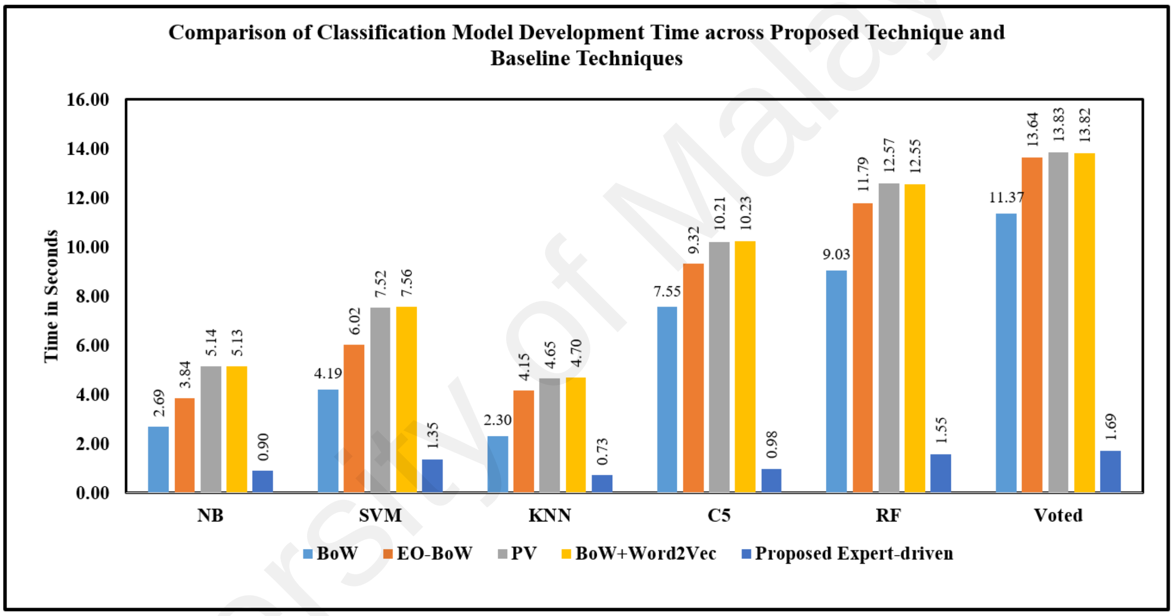


Figure 4.14: Comparison of classification time

Though the semi-automated expert-driven feature engineering technique outperformed the existing baseline techniques, however, some of the limitations were also identified in the proposed expert-driven technique. First, results of the proposed expert-driven technique depend heavily on the domain knowledge of the experts and their familiarity with forensic autopsy findings. It is believed that in the current study, the engagement of pathologists yielded experimental results that can be reflected across other medical systems. Second, the presented findings are exclusive to the free-text forensic autopsy reports obtained from PPUM, one of the largest hospital in Kuala Lumpur, Malaysia. It is also believed that the quality of the extracted reports is sufficiently heterogeneous, diverse, and comprehensive compared with the data gathered by other medical systems and therefore should produce acceptable results across other healthcare systems. Third, in the expert-driven technique, human experts (pathologists) are responsible for extracting and ranking useful features that belong to specific CoD autopsy reports. Hence, such feature engineering requires ample amount of time for extracting discriminative features for each CoD and rank them. Finally, the developed model can only detect sixteen CoDs. There are thousands of CoDs in each MoD (DiMaio & DiMaio, 2001). Hence, it is very difficult for pathologists to dig out useful features from each type of CoD. Moreover, these CoDs are increasing in number from time to time. For instance, in ICD-9, there were 14025 codes were available, however, in ICD-10, total codes available are approximately 68823 (CDC, 2015). Hence, it is infeasible for experts to manually dig out the features for each and every CoD and rank accordingly. Therefore, effective fully-automated feature engineering techniques can be useful and yet to be developed to achieve the strengths of proposed expert-driven technique and overcome its aforementioned limitations. Thus, in next chapter (Chapter 5), a fully-automated conceptual graph-based feature engineering technique is proposed, developed and evaluated. This proposed technique has the strengths of semi-automated expert-driven technique and also overcomes the limitations of expert-driven technique. The detail of conceptual graph-based feature engineering technique is presented in chapter 5.

4.7 Conclusion

In this chapter, semi-automated expert-driven feature engineering technique was presented to predict the CoD from free text forensic autopsy reports. Moreover, the state-of-the-art SML-based ATC techniques with feature selection schemes were used to classify the CoD from free-text forensic autopsy reports. It was discovered that the proposed semi-automated expert-driven feature engineering technique outperformed in terms of performance measures exceeding 90% when compared with baseline feature engineering techniques. Moreover, SVM and RF machine learning algorithms were found to be suitable for the classification of forensic autopsy reports with a feature subset size of 30 and 40. Based on the results, the proposed system proved to be more robust and more accurate when it was compared with four baselines. Furthermore, the promising results indicate that the pathologists can use the proposed system as a source of second opinion, assisting them in more accurately and rapidly determining the CoD. Nonetheless, the proposed semi-automated expert-driven technique showed better performance, several weaknesses of proposed technique are also identified (discussed in Section 4.6). Hence, to overcome such weaknesses, a fully-automated conceptual graph-based feature engineering technique is proposed and presented in Chapter 5.

CHAPTER 5: PROPOSED FULLY-AUTOMATED CONCEPTUAL GRAPH-BASED FEATURE ENGINEERING TECHNIQUE

5.1 Introduction

To overcome the limitations of semi-automated expert-driven feature engineering technique (discussed in Chapter 4, Section 4.6), this chapter presents an effective fully-automated conceptual graph-based feature engineering technique (CGFE) for determining MoD and CoD from forensic autopsy reports. This technique exploits the graphs to overcome the issues (such as word-order, word-context, and word-level synonymy and polysemy) of existing feature engineering technique such as BoW, *n*-gram, etc. (discussed in Section 2.8.1). Moreover, this proposed technique does not require any type of expert intervention to extract and rank the features of forensic autopsy reports. Conversely, it employs the use of medical ontologies to overcome the issue of word-level synonymy and polysemy.

The graph-based technique has been recently adopted in several machine learning sub-domains, including information retrieval (Giannakopoulos, Karkaletsis, Vouros, & Stamatopoulos, 2008), text summarization (Bronselaer & Pasi, 2013), and text classification (Bleik, Mishra, Huan, & Song, 2013; Papadakis et al., 2016). The motivation in using the graph-based technique is to consider word order in the input text for classifying autopsy reports. In graph-based text classification technique, graph is the combination of V, E, and W, where V represents graph vertices with each vertex containing a distinct term in the input text, E represents graph edges that connect co-occurring terms, and W is the function that computes the edge weight by considering the co-occurrence frequency of adjacent vertices in the graph.

Bleik et al. (2013) employed the graph-based text classification technique to classify biomedical text documents using controlled vocabulary. Though, the authors experimental results showed the promising results. However, their proposed technique has two major weaknesses. First, it suffers from high computational cost to classify the new biomedical document. This is because, their proposed approach compares the input graph with all documents graphs to discover the most similar graph for classification. Finally, it computes only the graph similarity based on relationship between the nodes. However, other important parameters for comparison maybe useful for classification such as, exact and fuzzy match, etc. The weighted frequent subgraphs were also employed in (Jiang, Coenen, Sanderson, & Zito, 2010) to extract useful features for classification and to reduce computational cost. Aery and Chakravarthy (2005)employed exact and approximate graph matching to improve classification results. Gee and Cook (2005) used the linguistic features of phrases in the text to encode text as graph and discover substructure and patterns for classification. Authors experimental results showed better results. However, their work was applicable for very small texts containing 8 to 13 terms. Papadakis et al. (2016) proposed the graph-based text classification technique to classify the web documents. Though, the technique used by Papadakis et al. (2016) was computationally effective however, their proposed technique has two major weaknesses. First, their proposed technique only used edge matching metric for document matching and classification. However, this single metric may prove insufficient for accurately classifying the forensic autopsy reports. Second, it fails to capture word-level synonymy and polysemy. For example, pathologists use the terms *heart attack* and *myocardial infarction* interchangeably in the collected forensic autopsy reports. Thus, in such cases, their proposed technique may not identify these two phrases as similar. Therefore, influenced by the work done by (Papadakis et al., 2016) and to overcome the limitations of the aforementioned graph-based techniques, this study develops a CGFE technique for classifying forensic autopsy reports.

There are three main strengths of proposed CGFE technique compare to other graph-based techniques for classifying text documents. First, the proposed CGFE technique can address the issues of order and context of words in the text, as well as of word-level synonymy and polysemy, when classifying forensic autopsy reports. Second, it considers the similarity and uniqueness metrics to compute the similarity between nodes, edges, and edge weights. This is because, in forensic autopsy reports, there is a very little difference among various CoDs. For instance, S06, S38, T07, and X80 are very close to each other and these CoDs usually share many relevant terms and concepts. Therefore, in such cases, uniqueness features may prove useful to differentiate between related CoDs. For instance, '*laceration wound*', '*abrasion wound*', '*Grazed abrasion*' are common in S06, S38, T07, and X80. However, '*abrasion forehead*', '*wound scalp*', '*right scalp*' are less frequent in S38, X80 and more frequent in S06, and T07. Therefore, the proposed similarity and uniqueness features may provide useful results. Finally, the proposed CGFE technique is computationally effective compared to other graph-based text classification techniques in the literature. This is because, for classification of new forensic autopsy reports, the proposed CGFE technique does not compare it with all *n* report level graphs where *n* represents the number of forensic autopsy reports in the dataset. Conversely, it only compares with *m* conceptual aggregated CoD level graphs where *m* represents the number of classes in the dataset (the detailed functionality of proposed CGFE technique is discussed in Section 5.2).

The major contributions of this portion of study are listed below.

1. An effective fully-automated conceptual graph-based feature engineering (CGFE) technique is developed to classify forensic autopsy reports from four manners of death

(MoD) and sixteen causes of death (CoD). Content-based and Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT) based conceptual features were mined and represented through graphs. These features were then used to train a two-level text classification model. The first level was responsible for predicting MoD and the second level was responsible for predicting CoD using the proposed CGFE technique. The proposed CGFE technique converts all forensic autopsy reports (belonging to a particular CoD) into individual report graphs, where each vertex *V* represents a unique term, each edge *E* connects co-occurring terms in the input text, and the weight of an edge *W* is the frequency of the co-occurrence of terms. Afterwards, these report graphs (belonging to a particular CoD) are combined to form an aggregated CoD-level graph. Afterwards, SNOMED-CT ontology is used to extract the semantic concepts of input nodes of CoD-level aggregated graph. The multi-information obtained from SNOMED-CT ontology can be fully utilized by organizing unique co-occurring terms, along with their SNOMED-CT concepts and descriptors, in the CoD-level aggregated graph. Thus, an aggregated CoD-level graph is converted into conceptual CoD-level aggregated graph.

2. To create a numeric master feature vector (MFV), and to construct a classification model, each report is compared with each conceptual CoD-level aggregated graph to compute six metrics namely, vertex similarity metric (VSM), vertex uniqueness metric (VUM), edge similarity metric (ESM), edge uniqueness metric (EUM), similar edge weight metric (SEWM), and unique edge weight metric (UEWM).
3. Extensive experiments of the proposed CGFE technique along with supervised machine learning algorithms are performed to achieve high-performance classification of 16 CoDs from free-text forensic autopsy reports. Six supervised machine learning algorithms, namely, naïve Bayes (NB), support vector machine (SVM), *k*-nearest neighbor (*k*NN), decision tree (C5.0), random forest (RF), and

ensemble voted classifier (ensemble of NB, SVM, *k*NN, C5.0, and RF), are compared using the proposed CGFE technique to evaluate their performances using four performance measures. These performance measures are macro recall (Recall$_M$), macro precision (Precision$_M$), macro F-measure (F-measure$_M$), and overall accuracy.

4. To demonstrate the significance of proposed CGFE technique, its performance is compared with six existing state-of-the-art baseline feature engineering techniques, namely, the traditional BoW technique (Harris, 1954), the entropy-optimized feature-based BoW (EO-BoW) technique (Passalis & Tefas, 2016), the paragraph vector (PV) technique (Le & Mikolov, 2014), the hybrid BoW and Word2Vec (BoW + Word2Vec) technique (Enríquez et al., 2016), the term-based graph (TG) technique (Papadakis et al., 2016), and the semi-automated expert-driven (ED) technique (discussed in Chapter 4). The experimental results show that the proposed CGFE technique outperforms other fully automated baseline feature engineering techniques.

The rest of this chapter is structured as follows. Section 5.2 discusses the technical details of proposed CGFE technique. In addition, Section 5.3 presents the underlying methodology for designing fully-automated CGFE technique including, experimental setup and implementation details behind the philosophy of proposed CGFE technique. Section 5.4 reports the experimental results. The discussion on findings is presented in Section 5.5. Finally, this chapter is concluded in Section 5.6.

5.2 Fully-Automated Conceptual Graph-based Feature Engineering Technique

This section discusses the proposed CGFE technique in detail. In this technique, graph-based approach is exploited to represent the natural text of autopsy reports. Graph representation is exploited because it provides flexibility and robustness on representing the natural language text compared to traditional *n*-gram, and BoW. Moreover, it can also prove suitable to overcome the limitations (such as, word co-occurrence, word ordering, and word inversion) of traditional feature engineering techniques such as, BoW, *n*-gram. Furthermore, SNOMED-CT is also used with graph representation to extract concept-based features from extracted graph of word (GoW) features to address the issue of word-level synonymy and polysemy.

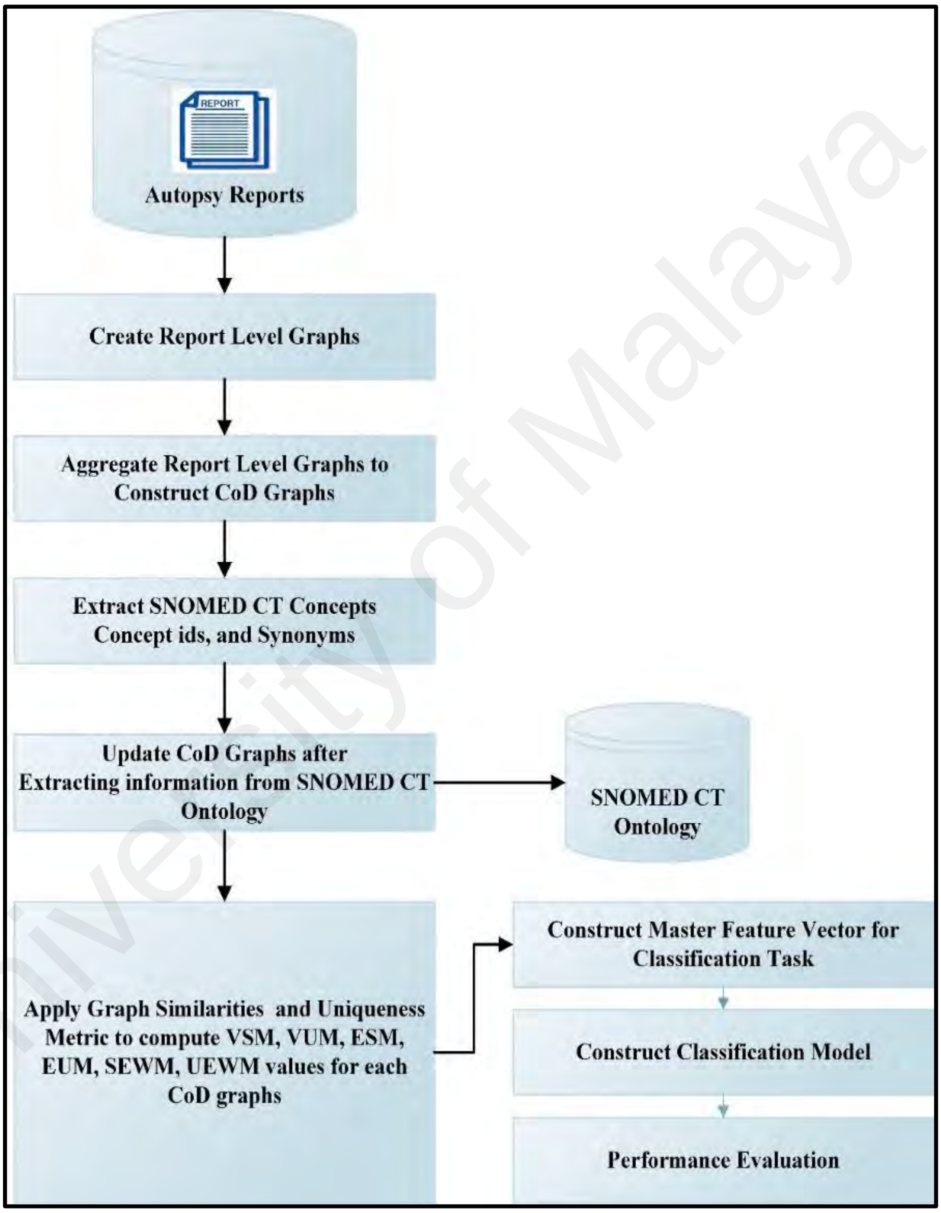

REPORT
Autopsy Reports
Create Report Level Graphs
Aggregate Report Level Graphs to Construct CoD Graphs
Extract SNOMED CT Concepts Concept ids, and Synonyms
Update CoD Graphs after Extracting information from SNOMED CT Ontology
SNOMED CT Ontology
Apply Graph Similarities and Uniqueness Metric to compute VSM, VUM, ESM, EUM, SEWM, UEWM values for each CoD graphs
Construct Master Feature Vector for Classification Task
Construct Classification Model
Performance Evaluation

Figure 5.1: Flowchart of the functionality of the proposed CGFE technique

The proposed technique can be constructed by applying five steps: creation of report graphs, creation of aggregated CoD graphs, addition of SNOMED-CT concepts and descriptors to the aggregated CoD graphs, computation of graph similarity and graph uniqueness metrics, and creation of an MFV for classification. These steps are discussed in detail in subsequent sections. In addition, Figure 5.1 shows the functionality flowchart of the proposed CGFE technique and Table 5.1 shows the detailed CGFE technique algorithm. As can be seen here, initially, the preprocessing steps such as $\Im_{lc}$, $\Im_{sw}$, and $\Im_{sc}$ will be applied to forensic autopsy reports belong to a particular CoD. Afterwards, each pre-processed report will be converted into ug, bg, and tg . After that, all the report level ug, bg, and tg will aggregated together to form G_c. After producing G_c, the $\Im_{SNOMED\_CT}$ will be applied to form G^S. Once, all G^S for all CoDs are produced then master feature vector is prepared. For that, each report n will be taken and that n will be converted to G_r. Afterwards, $\Im_{SNOMED\_CT}$ will be applied to convert G_r into conceptual report level graph. Finally, for this conceptual report level graph six metrics from each G^S will be computed using $\Im_{VEW}$ to generate the master feature vector and this vector is fed to machine learning algorithm to generate classification model. This algorithm is further explained in detail in subsequent sub sections.

Table 5.1: Proposed CGFE algorithm

m	:	Number of distinct causes of death (CoD)
n	:	Number of forensic reports in each CoD
tc	:	Temporary variable holding cause of death record
tr	:	Temporary variable holding forensic autopsy report data
ug	:	Unigram graph
bg	:	Bigram graph
tg	:	Trigram graph
G_r	:	Report level graph
G_c	:	Aggregated class level graph
G^S	:	SNOMED-CT Aggregated class level graph
$\Im_{lc}$	:	Lower Case Function
$\Im_{sc}$	:	Spell Checker Function
$\Im_{sw}$	:	Stop Words Function
$\Im_{ug}$	:	Unigram Graph Function
$\Im_{bg}$	:	Bigram Graph Function
$\Im_{tg}$	:	Trigram Graph Function

Symbols Definition

Table 5.1: continued

$\Im_{VEW}$: Vertex (V), Edge (E) and Weight of an edge (W) Generator Function

$\Im_{SNOMED\_CT}$: SNOMED-CT Function

```
FOR $i = 1$ to $m$
    LOAD $tc \leftarrow COD[i]$
    FOR $j = 1$ to $n$
        $tr \leftarrow tc[j]$
        $tr\_lc \leftarrow \Im_{lc}(tr)$
        $tr\_sc \leftarrow \Im_{sc}(tr\_lc)$
        $tr\_sw \leftarrow \Im_{sw}(tr\_sc)$
        $T_{ug} \leftarrow \Im_{ug}(tr\_sw)$
        $T_{bg} \leftarrow \Im_{bg}(tr\_sw)$
        $T_{tg} \leftarrow \Im_{tg}(tr\_sw)$
        $G_{ug}\_r_j(V, E, W) \leftarrow \Im_{VEW}(T_{ug})$
        $G_{bg}\_r_j(V, E, W) \leftarrow \Im_{VEW}(T_{bg})$
        $G_{tg}\_r_j(V, E, W) \leftarrow \Im_{VEW}(T_{tg})$
        $G_{ug}\_c_i(V, E, W) \leftarrow G_{ug}\_c_i(V, E, W) \bigcup G_{ug}\_r_j(V, E, W)$
        $G_{bg}\_c_i(V, E, W) \leftarrow G_{bg}\_c_i(V, E, W) \bigcup G_{bg}\_r_j(V, E, W)$
        $G_{tg}\_c_i(V, E, W) \leftarrow G_{tg}\_c_i(V, E, W) \bigcup G_{tg}\_r_j(V, E, W)$
    END FOR
END FOR
FOR $i = 1$ to $m$
    $G^S_{ug}\_c_i(V,E,W) \leftarrow \Im_{SNOMED\_CT}(G_{ug}\_c_i(V,E,W))$
    $G^S_{bg}\_c_i(V,E,W) \leftarrow \Im_{SNOMED\_CT}(G_{bg}\_c_i(V,E,W))$
    $G^S_{tg}\_c_i(V,E,W) \leftarrow \Im_{SNOMED\_CT}(G_{tg}\_c_i(V,E,W))$
END FOR
LOAD $MFV$
FOR $i = 1$ to $n$
    $V_{ug}\_r \xleftarrow{Separate} G_{ug}\_r_i(V, E, W)$
    $E_{ug}\_r \xleftarrow{Separate} G_{ug}\_r_i(V, E, W)$
```

Table 5.1: continued

```
    FOR $j = 1$ to 𝕞
        $V_{ug}\_c \xleftarrow{Separate} G^{S}_{ug\_}c_j\ (V, E, W)$
        $E_{ug}\_c \xleftarrow{Separate} G^{S}_{ug\_}c_j\ (V, E, W)$
        $W_{ug}\_c \xleftarrow{Separate} G^{S}_{ug\_}c_j\ (V, E, W)$
        $VSM_{(G_{ug}\_r, G_{ug}\_c)} = \frac{\left|V_{ug}\_r \cap V_{ug}\_c\right|}{\left|V_{ug}\_r\right|}$
        $VUM_{(G_{ug}\_r, G_{ug}\_c)} = \frac{\left[V_{ug}\_r - \left(\left|V_{ug}\_r \cap V_{ug}\_c\right|\right)\right]}{\left|V_{ug}\_r\right|}$
        $ESM_{(G_{ug}\_r, G_{ug}\_c)} = \frac{\left|E_{ug}\_r \cap E_{ug}\_c\right|}{\left|E_{ug}\_r\right|}$
        $EUM_{(G_{ug}\_r, G_{ug}\_c)} = \frac{\left[E_{ug}\_r - \left(\left|E_{ug}\_r \cap E_{ug}\_c\right|\right)\right]}{\left|E_{ug}\_r\right|}$
        $SEWM_{(G_{ug}\_r, G_{ug}\_c)} = \frac{\sum_{e \in E_{ug}\_r} \left( \frac{\min\left(W\left(E_{ug}\_r\right), W\left(E_{ug}\_c\right)\right)}{\max\left(W\left(E_{ug}\_r\right), W\left(E_{ug}\_c\right)\right)} \right)}{\left|E_{ug}\_r\right|}$
        $UEWM_{(G_{ug}\_r, G_{ug}\_c)} = 1 - \left( \frac{\sum_{e \in E_{ug}\_r} \left( \frac{\min\left(W\left(E_{ug}\_r\right), W\left(E_{ug}\_c\right)\right)}{\max\left(W\left(E_{ug}\_r\right), W\left(E_{ug}\_c\right)\right)} \right)}{\left|E_{ug}\_r\right|} \right)$
        $tf_{Data} \xleftarrow{append} strcat\ (VSM, VUM, ESM, EUM, SEWM, UEWFM)$
        END FOR
    $tf_{Data} \xleftarrow{append} strcat\ (",\mathrm{CoD}", atoi\ (\delta[i]), "\backslash n")$
    WRITE $MFV \xleftarrow{append} tf_{Data}$
    DELETE $tf_{Data}$
END FOR
CLOSE $MFV$
```

5.2.1 Creation of Simple Graphs

To explain this section, assume that we have a dataset for forensic autopsy reports (R) that belong to m distinct CoDs, namely, (CoD_1, CoD_2, …, CoD_m). Each CoD contains n number of forensic autopsy reports (r_1, r_2, r_3, r_4, …, r_n). In the proposed CGFE technique, several pre-processing steps (discussed in Chapter 3, Section 3.4) were first applied to each individual report. After the report pre-processing step, each autopsy report r_i that belonged to a particular CoD, e.g., CoD_i, was converted into unigram, bigram, and trigram unidirectional report graphs. In the unigram, bigram, and trigram report graphs, each vertex comprised one, two, and three terms, respectively. These graphs were purposely designed to evaluate their performances and determine which one could provide the best results. For instance, Figure 5.2 (a) and Figure 5.2 (b) show two report level graphs of two forensic autopsy reports belonging to same CoD. In these report graphs, each vertex V represents a unique term features (such as, $w_1, w_2, w_3, ...w_n$), each edge E connects the co-occurring terms in the report, and the weight of an edge W shows the co-occurrence frequency of terms in the report (such as, $f_{r1}, f_{r2}, f_{r2}, ...f_{rn}$). Figure 5.3 (a) and Figure 5.3 (b) show the samples of constructed unigram, and bigram report graphs from real dataset. As shown in here, the two vertices are connected by hyphen ('- ') sign and underscore sign ('_') separates the second vertex and weight of two vertices. For instance, as shown in Figure 5.3 (a) the vertex 'grazed' is connected with the vertex 'abrasion' approximately 19 times.

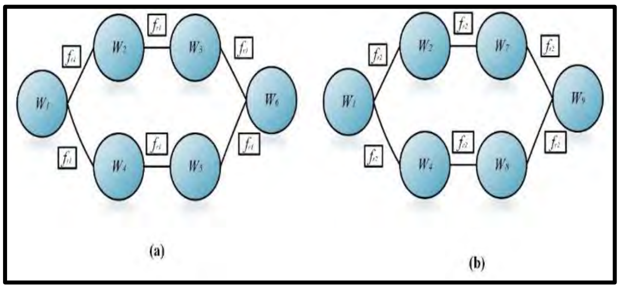
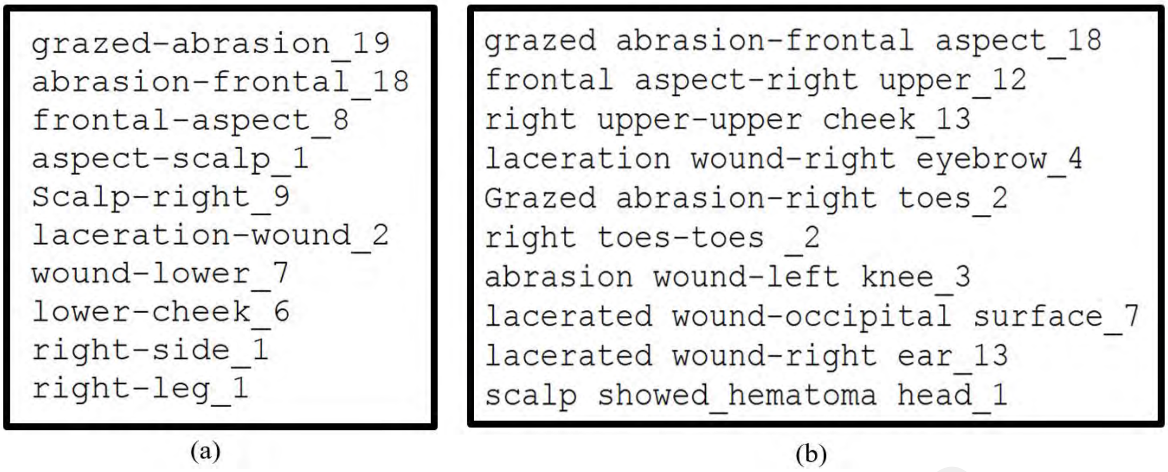


Figure 5.2: Report level graphs



```
grazed-abrasion_19
abrasion-frontal_18
frontal-aspect_8
aspect-scalp_1
Scalp-right_9
laceration-wound_2
wound-lower_7
lower-cheek_6
right-side_1
right-leg_1
```

(a)

```
grazed abrasion-frontal aspect_18
frontal aspect-right upper_12
right upper-upper cheek_13
laceration wound-right eyebrow_4
Grazed abrasion-right toes_2
right toes-toes _2
abrasion wound-left knee_3
lacerated wound-occipital surface_7
lacerated wound-right ear_13
scalp showed_hematoma head_1
```

(b)

Figure 5.3: Sample of partial unigram and bigram report graphs from real dataset

5.2.2 Creation of Class-Level Aggregated Graphs

After all the reports r_n of a particular CoD_i were converted into unigram, bigram, and trigram report graphs, all the unigram graphs were combined to form an aggregated CoD_i unigram graph. Similarly, all the bigram and trigram report graphs were combined to form aggregated CoD_i bigram and trigram graphs, respectively. The graph aggregation process is described as follows. Given a set of autopsy reports r_n that belong to CoD_i, the i^{th} report $r_i \in CoD_i$ is transformed into a report graph $Gr_i = \left(V_{r_i}, E_{r_i}, W_{r_i}\right)$. Therefore, for all r_n reports, we have n report graphs $Gr_1 = \left(V_{r_1}, E_{r_1}, W_{r_1}\right), Gr_2 = \left(V_{r_2}, E_{r_2}, W_{r_2}\right), ...Gr_n = \left(V_{r_n}, E_{r_n}, W_{r_n}\right)$. An initially empty aggregated CoD-level graph G_{CoD_i} was built, and then all previously constructed report graphs G_{r_n} were merged into a single aggregated G_{CoD_i} graph, i.e., $G_{CoD_i} = \left(V_{CoD_i}, E_{CoD_i}, W_{CoD_i}\right)$, where $V_{CoD_i} = V_{CoD_i} \bigcup V_{r_i}$, $E_{CoD_i} = E_{CoD_i} \bigcup E_{r_i}$, and $W_{CoD_i}(e) = W_{CoD_i}(e) + (W_{r_i}(e) - W_{CoD_i}(e)) \times 1/i$. However, if the co-occuring nodes of report level graph do not match with CoD level aggregated graph, then new vertices are added in CoD level aggregated graph and their edge weight will be computed using aforementioned formula. In addition, the weight of edges of CoD level aggregated graph incrementally converge to their overall average due to the division with *i* in the formula. In this fashion, CoD level graphs encapsulate features common in the content of specific CoD. The Figure 5.4 shows the example of aggregated CoD level graph that is constructed by combining report graph 1, and report graph 2 (as shown in Figure 5.2 (a), and Figure 5.2 (b)). Hence, the aggregated CoD graph contains features that will be common in an entire category of autopsy reports that belong to that particular CoD. Moreover, three aggregated CoD graphs were prepared for each CoD, namely, unigram, bigram, and trigram aggregated CoD graphs. Hence, if there are n CoDs, then the aggregated CoD graphs will be $n \times 3$.

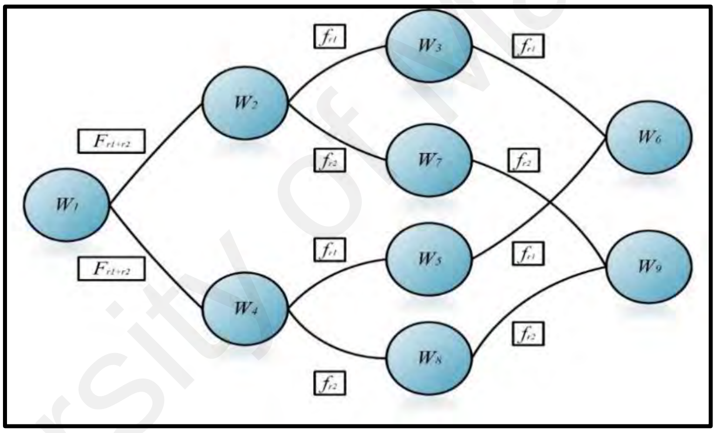


Figure 5.4: Aggregated CoD level graph

5.2.3 Addition of SNOMED-CT Concepts

Once the aggregated CoD graphs were constructed, SNOMED-CT (Donnelly, 2006) concepts and descriptors of each vertex were mapped from SNOMED-CT ontology. SNOMED-CT is a standardized and multilingual vocabulary of clinical Ontology used by physicians and other healthcare providers for the electronic exchange of clinical health information (Donnelly, 2006). The high-level structure of SNOMED-CT ontology is illustrated in Figure 5.5. Each medical term has various similar concepts, and each concept has a unique concept id and concept description. In addition, each concept has three types of descriptors, namely, fully specified name (FSN), preferred name (PN), and synonyms. Each concept also has parent concepts, child concepts, and relationship to other related concepts. For example, the term *heart attack* has six related medical concepts in SNOMED CT ontology. Among these concepts, the top related concept is *myocardial infarction*, which has a unique concept id of *22298006*. The FSN of this concept is *myocardial infarction (disorder)*; the PN is *myocardial infarction*; the synonyms are *cardiac infarction*, *infarction of the heart*, and *MI-myocardial infarction*; and the parent concept is *stroke risk (135877001)*.

There were several reasons for employing SNOMED CT ontology in the proposed work. First, the SNOMED CT covers the long range of clinical terms (Cornet & de Keizer, 2008; Lee, Cornet, Lau, & De Keizer, 2013). Second, it is an international standard, and for this reason it is good for semantic interoperability (Saripalle, 2010). Third, SNOMED CT has the ability to rationally associate terms from several concepts to describe the clinical findings (Stearns et al., 2001). Fourth, it is organized into series of hierarchies of medical terms including clinical findings, events, and procedures to obtain related atomic-level terms (Lee et al., 2013). Finally, in several studies, authors have shown the effectiveness of using SNOMED CT ontology to dig out related medical terms and concepts (Chin & Kim, 2003; Halland & Britz, 2011; Zuccon et al., 2013; Kasthurirathne et al., 2017).

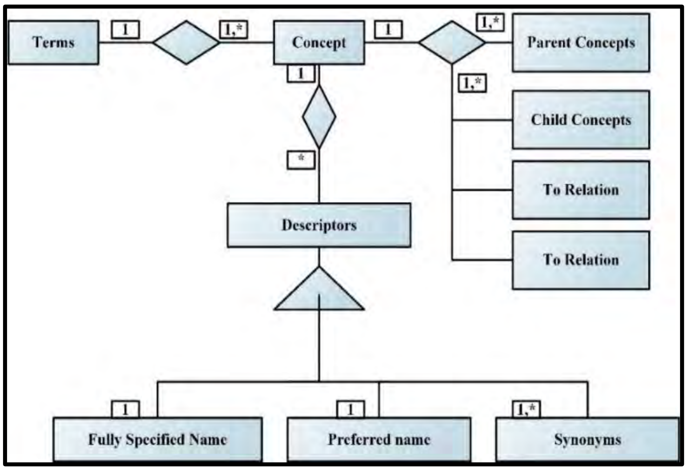


Figure 5.5: SNOMED CT ontology

The complete MySQL scripts of SNOMED-CT ontology was downloaded from (NLM, 2017) and configured on a local computer. Afterward, a Java program was written to extract the SNOMED-CT concepts from SNOMED CT ontology by matching the nodes of aggregated CoD-level graph. All the extracted concept ids of the input vertices were added to the aggregated conceptual CoD-level graph as new vertices, where their edges, weight, and co-occurring vertices will be the same as those of the original vertex. Hence, the updated aggregated graph contains semantically rich information after adding related concepts. Accordingly, the aggregated graph can now resolve the issue of word-level synonymy and polysemy. However, if no concept id is matched then the node of aggregated conceptual level CoD graph contains the same term which was in the aggregated CoD level graph. Figure 5.6 shows the aggregated conceptual CoD level graph that is constructed from aggregated CoD level graph (as shown in Figure 5.4). Here, $s_{1_{w1}}$ represents the conceptual node of w_1 containing the concept identifier of w_1. Figure 5.7 shows the sample of aggregated conceptual CoD-level graph. Here, it can be seen that the concept ids of vertices are added in the aggregated CoD-level graph to transform it into aggregated conceptual CoD-level graph. For instance, '81654009' concept id was added for the vertex 'frontal'.

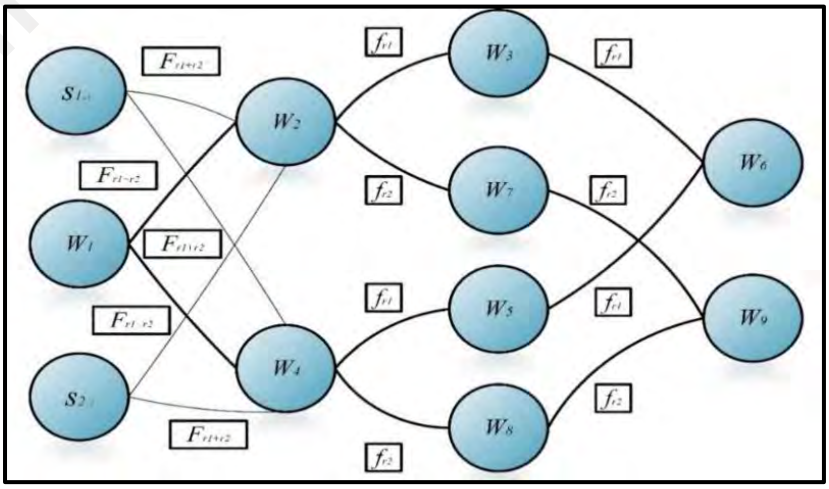
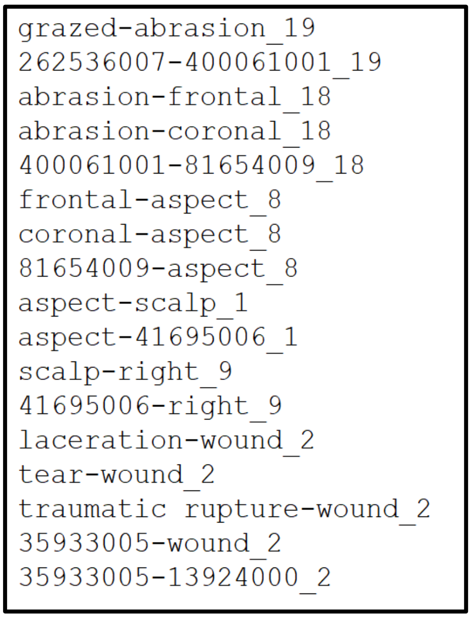


Figure 5.6: Aggregated conceptual CoD level graph



```
grazed-abrasion_19
262536007-400061001_19
abrasion-frontal_18
abrasion-coronal_18
400061001-81654009_18
frontal-aspect_8
coronal-aspect_8
81654009-aspect_8
aspect-scalp_1
aspect-41695006_1
scalp-right_9
41695006-right_9
laceration-wound_2
tear-wound_2
traumatic rupture-wound_2
35933005-wound_2
35933005-13924000_2
```

Figure 5.7: Sample of conceptual aggregated graph

5.2.4 Graph Similarity and Uniqueness Metrics

The following similarity metrics were used to determine the similarity between the report-level graph and the class-level aggregated graph.

Vertex similarity metric (VSM): This metric indicates the proportion of nodes shared between the report graph $\left(G_{r_i}\right)$ and the CoD graph $\left(G_{CoD_i}\right)$. The mathematical definition of VSM is given in Equation 1:

$$VSM(G_{r_i}, G_{C_i}) = \frac{\left|V_{r_i} \cap V_{C_i}\right|}{\left|V_{r_i}\right|}. \qquad \textbf{(1)}$$

Vertex uniqueness metric (VUM): This metric indicates the proportion of the vertices of the report graph $\left(G_{r_i}\right)$ that do not match the vertices of the CoD graph $\left(G_{CoD_i}\right)$. The mathematical definition of VUM is given in Equation 2:

$$VUM(G_{r_i}, G_{C_i}) = \frac{\left[V_{r_i} - \left(\left|V_{r_i} \cap V_{C_i}\right|\right)\right]}{\left|V_{r_i}\right|}. \qquad \textbf{(2)}$$

Edge similarity metric (ESM): This metric shows the proportion of edges shared between the report graph $\left(G_{r_i}\right)$ and the CoD graph $\left(G_{CoD_i}\right)$. The mathematical definition of ESM is given in Equation 3:

$$ESM(G_{r_i}, G_{C_i}) = \frac{\left|E_{r_i} \cap E_{C_i}\right|}{\left|E_{r_i}\right|}. \qquad \textbf{(3)}$$

Edge uniqueness metric (EUM): This metric indicates the proportion of edges of the report graph $\left(G_{r_i}\right)$ that do not match the edges of the CoD graph $\left(G_{CoD_i}\right)$. The mathematical definition of EUM is given in Equation 4:

$$EUM(G_{r_i}, G_{C_i}) = \frac{\left[E_{r_i} - \left(\left|E_{r_i} \cap E_{C_i}\right|\right)\right]}{\left|E_{r_i}\right|}. \qquad \textbf{(4)}$$

Similar edge weight metric (SEWM): This metric computes the proportion of the overall weight of the report graph $\left(G_{r_i}\right)$ edges that match the edges of the CoD graph $\left(G_{CoD_i}\right)$. The mathematical definition of SEWM is given in Equation 5:

$$SEWM(G_{r_i}, G_{CoD_j}) = \frac{\sum_{e \in E_{r_i}} \left(\frac{\min\left(W\left(e, G_{r_i}\right), W\left(e, G_{CoD_j}\right)\right)}{\max\left(W\left(e, G_{r_i}\right), W\left(e, G_{CoD_j}\right)\right)} \right)}{\left|E_{r_i}\right|} \qquad \textbf{(5)}$$

Unique edge weight metric (UEWM): This metric computes the proportion of the overall weight of the report graph $\left(G_{r_i}\right)$ edges that do not match the edges of the CoD graph $\left(G_{CoD_i}\right)$. The mathematical definition of UEWM is given in Equation 6:

$$UEWM(G_{r_i}, G_{CoD_j}) = 1 - \left(\frac{\sum_{e \in E_{r_i}} \left(\frac{\min\left(W\left(e, G_{r_i}\right), W\left(e, G_{CoD_j}\right)\right)}{\max\left(W\left(e, G_{r_i}\right), W\left(e, G_{CoD_j}\right)\right)} \right)}{\left|E_{r_i}\right|} \right) \qquad \textbf{(6)}$$

The similarity and uniqueness metrics are used because in the forensic dataset there is a very little difference among various CoDs. For instance, S06, S38, T07, and X80 are very close to each other and these CoDs usually share many relevant features in collected dataset. Thus, in such cases, uniqueness features may prove useful to differentiate between related CoDs. For instance, '*laceration wound*', '*abrasion wound*', '*Grazed abrasion*' are common in S06, S38, T07, and X80. However, '*abrasion forehead*', '*wound scalp*', '*right scalp*' are less frequent in S38, X80 and more frequent in S06, and T07. Therefore, the proposed similarity and uniqueness features may provide useful results. In another example, I23, I24, and I25 CoDs are very close to each other and these CoDs usually share many relevant features in collected dataset. Thus, in such cases, uniqueness features may prove useful to differentiate between related CoDs. For instance, '*pericardium*', '*epicardium*', '*myocardium*', '*endocardium*', are very are common in I23, I24, and I25 CoDs. However, co-occurrence of '*myocardium*' and '*fibrosis*', is more common in I24, and I25 and less common in I23. Similarly, the co-occurrence of '*myocardium*' and '*infarction*' and '*myocardium*' and '*hemorrhage*' is more frequent in I23, less frequent in I24, and very less frequent in I25. Finally, the co-occurrence of '*epicardium*' and '*dimpling*' can be found in I25 but maybe less frequently found in I23, and I25. Therefore, in such cases, the proposed similarity and uniqueness features may provide useful results.

To summarize these metrics, the values of the VSM and VUM metrics represent the number of common features that match and do not match, respectively, between the input report-level graph and the aggregated class-level graph. Similarly, the values of the ESM and EUM metrics denote the co-occurrence of matched and unmatched features in the input report-level graph and the aggregated class-level graph. Lastly, the values of SEWM and UEWM metrics indicate the sum of frequency of co-occurring *n*-grams matched and unmatched between the input report-level graph and the aggregated CoD-level graph. In addition, all the aforementioned metrics consider the co-occurring *n*-grams instead of individual *n*-grams. Finally, section 5.4.5 also shows the significance of similarity and uniqueness metrics with the help of experiments.

5.2.5 Creation of Master Feature Vector

For any classification task, the primary input for the supervised machine learning algorithm is the master feature vector (MFV). The MFV is a numeric vector where each row represents one instance or one record (in the case of current study, one forensic autopsy report) and each column represents a feature. In this study, the following procedure was followed to create an MFV from a dataset of labelled forensic autopsy reports (D_l).

1- Aggregated class-level graphs ($G_{C_1}, G_{C_2}, G_{C3}, ...G_{C_N}$) were constructed from D_l (as discussed in Section 5.2.2). N represents the number of distinct classes in D_l.

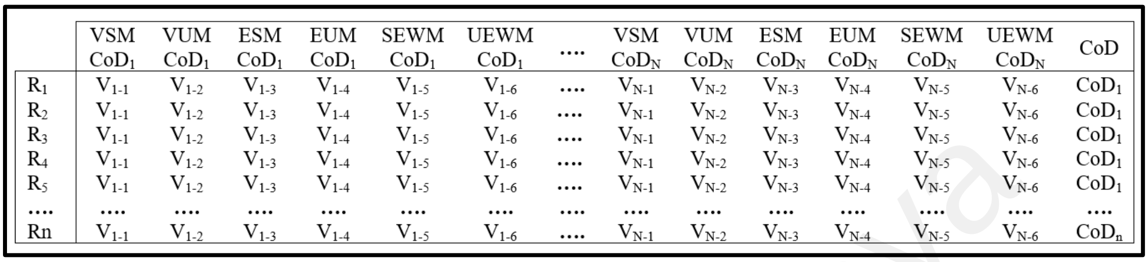
2- The corresponding report-level graphs ($G_{r_1}, G_{r_2}, G_{r3}, ...G_{r_N}$) were built for each autopsy report (as discussed in Section 5.2.1). N indicates the number of autopsy reports in D_l.

3- For each report-level graph ($G_{r_1}, G_{r_2}, G_{r3}, ...G_{r_N}$), SNOMED-CT concept descriptions, concepts ids, and synonyms were extracted from each vertex of the graph using SNOMED CT ontology. The obtained concept descriptions, concepts ids, and synonyms were added to the corresponding graphs (as discussed in Section 5.2.3).

4- Each report-level graph was compared with the *N* aggregated class-level graphs to compute the values of the aforementioned six graph similarity and uniqueness metrics (i.e., VSM, VUM, ESM, EUM, SEWM, and UEWM). Hence, for each report-level graph, $6 \times N$ values were computed to form a feature vector. Each value was separated by a comma (,).

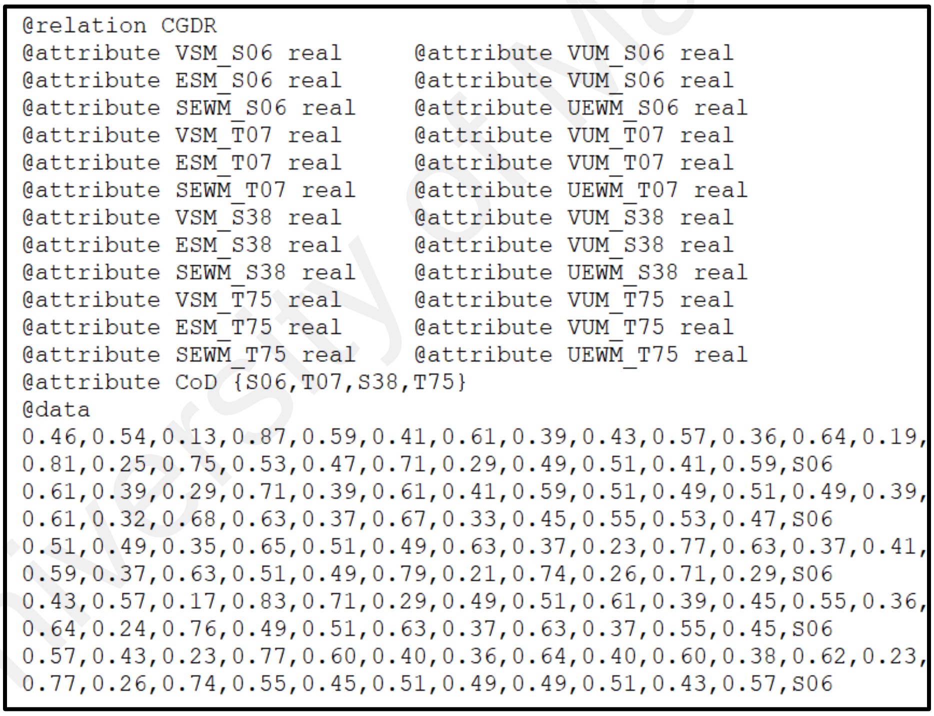
5- The class of each report was appended after $6 \times N$ values.

The sketch of the resultant MFV is shown in Figure 5.8. In addition, Figure 5.9 shows the sample of numeric MFV of accident related CoDs from real dataset. This numeric MFV is then fed to the classifier as input to construct the classification model.



	VSM CoD$_1$	VUM CoD$_1$	ESM CoD$_1$	EUM CoD$_1$	SEWM CoD$_1$	UEWM CoD$_1$	....	VSM CoD$_N$	VUM CoD$_N$	ESM CoD$_N$	EUM CoD$_N$	SEWM CoD$_N$	UEWM CoD$_N$	CoD
R$_1$	V_{1-1}	V_{1-2}	V_{1-3}	V_{1-4}	V_{1-5}	V_{1-6}	....	V_{N-1}	V_{N-2}	V_{N-3}	V_{N-4}	V_{N-5}	V_{N-6}	CoD$_1$
R$_2$	V_{1-1}	V_{1-2}	V_{1-3}	V_{1-4}	V_{1-5}	V_{1-6}	....	V_{N-1}	V_{N-2}	V_{N-3}	V_{N-4}	V_{N-5}	V_{N-6}	CoD$_1$
R$_3$	V_{1-1}	V_{1-2}	V_{1-3}	V_{1-4}	V_{1-5}	V_{1-6}	....	V_{N-1}	V_{N-2}	V_{N-3}	V_{N-4}	V_{N-5}	V_{N-6}	CoD$_1$
R$_4$	V_{1-1}	V_{1-2}	V_{1-3}	V_{1-4}	V_{1-5}	V_{1-6}	....	V_{N-1}	V_{N-2}	V_{N-3}	V_{N-4}	V_{N-5}	V_{N-6}	CoD$_1$
R$_5$	V_{1-1}	V_{1-2}	V_{1-3}	V_{1-4}	V_{1-5}	V_{1-6}	....	V_{N-1}	V_{N-2}	V_{N-3}	V_{N-4}	V_{N-5}	V_{N-6}	CoD$_1$
....	....	....	....	....	....	....	....	....	....	....	....	....	....	....
Rn	V_{1-1}	V_{1-2}	V_{1-3}	V_{1-4}	V_{1-5}	V_{1-6}	....	V_{N-1}	V_{N-2}	V_{N-3}	V_{N-4}	V_{N-5}	V_{N-6}	CoD$_n$

Figure 5.8: Sketch of the resultant MFV



```
@relation CGDR
@attribute VSM_S06 real     @attribute VUM_S06 real
@attribute ESM_S06 real     @attribute VUM_S06 real
@attribute SEWM_S06 real    @attribute UEWM_S06 real
@attribute VSM_T07 real     @attribute VUM_T07 real
@attribute ESM_T07 real     @attribute VUM_T07 real
@attribute SEWM_T07 real    @attribute UEWM_T07 real
@attribute VSM_S38 real     @attribute VUM_S38 real
@attribute ESM_S38 real     @attribute VUM_S38 real
@attribute SEWM_S38 real    @attribute UEWM_S38 real
@attribute VSM_T75 real     @attribute VUM_T75 real
@attribute ESM_T75 real     @attribute VUM_T75 real
@attribute SEWM_T75 real    @attribute UEWM_T75 real
@attribute CoD {S06,T07,S38,T75}
@data
0.46,0.54,0.13,0.87,0.59,0.41,0.61,0.39,0.43,0.57,0.36,0.64,0.19,
0.81,0.25,0.75,0.53,0.47,0.71,0.29,0.49,0.51,0.41,0.59,S06
0.61,0.39,0.29,0.71,0.39,0.61,0.41,0.59,0.51,0.49,0.51,0.49,0.39,
0.61,0.32,0.68,0.63,0.37,0.67,0.33,0.45,0.55,0.53,0.47,S06
0.51,0.49,0.35,0.65,0.51,0.49,0.63,0.37,0.23,0.77,0.63,0.37,0.41,
0.59,0.37,0.63,0.51,0.49,0.79,0.21,0.74,0.26,0.71,0.29,S06
0.43,0.57,0.17,0.83,0.71,0.29,0.49,0.51,0.61,0.39,0.45,0.55,0.36,
0.64,0.24,0.76,0.49,0.51,0.63,0.37,0.63,0.37,0.55,0.45,S06
0.57,0.43,0.23,0.77,0.60,0.40,0.36,0.64,0.40,0.60,0.38,0.62,0.23,
0.77,0.26,0.74,0.55,0.45,0.51,0.49,0.49,0.51,0.43,0.57,S06
```

Figure 5.9: Sample of numeric MFV of accident related CoDs from real dataset

5.3 Experimental Design

This section presents the experimental design of construction of classification model for forensic autopsy reports though proposed fully-automated CGFE technique. An extensive set of experiments were run to measure the performance of proposed CGFE technique with state-of-the-arts baseline feature engineering techniques. The complete flow of experimental design is shown in Figure 5.10. As shown here, the performance of proposed CGFE technique was evaluated comprehensively. For experiments forensic autopsy dataset (discussed in Section 3.3) was used. Moreover, all reports were preprocessed to remove irrelevant and uninformative features (discussed in Section 3.4). Afterwards, discriminative features were extracted and represented from preprocessed forensic autopsy reports through proposed CGFE technique to form a numeric MFV. This numeric MFV was then fed to six different supervised machine learning algorithms (namely, SVM, NB, *k*NN, C5, RF, and ensemble-voted) to evaluate the most suitable algorithm for classifying forensic autopsy reports using CGFE technique. The justification for the selection of these classifiers are also given in Chapter 3, Section 3.6.

In addition, the proposed CGFE technique was also compared with six baseline feature engineering techniques namely, the traditional BoW technique (Harris, 1954), the entropy-optimized feature-based BoW (EO-BoW) technique (Passalis & Tefas, 2016), the paragraph vector (PV) technique (Le & Mikolov, 2014), the hybrid of BoW and Word2Vec (BoW + Word2Vec) technique (Enríquez et al., 2016), Term graph (TG) technique (Papadakis et al., 2016), and semi-automated expert-driven technique (discussed in Chapter 4) to show its significance.

Finally, statistical significant tests were applied to see the significant different between the analyses results of aforementioned experiments. A Friedman statistical test (Demšar, 2006; McCrum-Gardner, 2008; Ott & Longnecker, 2015) along with Nemenyi post hoc statistical tests (Demšar, 2006; Ott & Longnecker, 2015) was performed (using significance level of alpha = 0.05) to compare the overall accuracies for six classifiers across all four datasets.

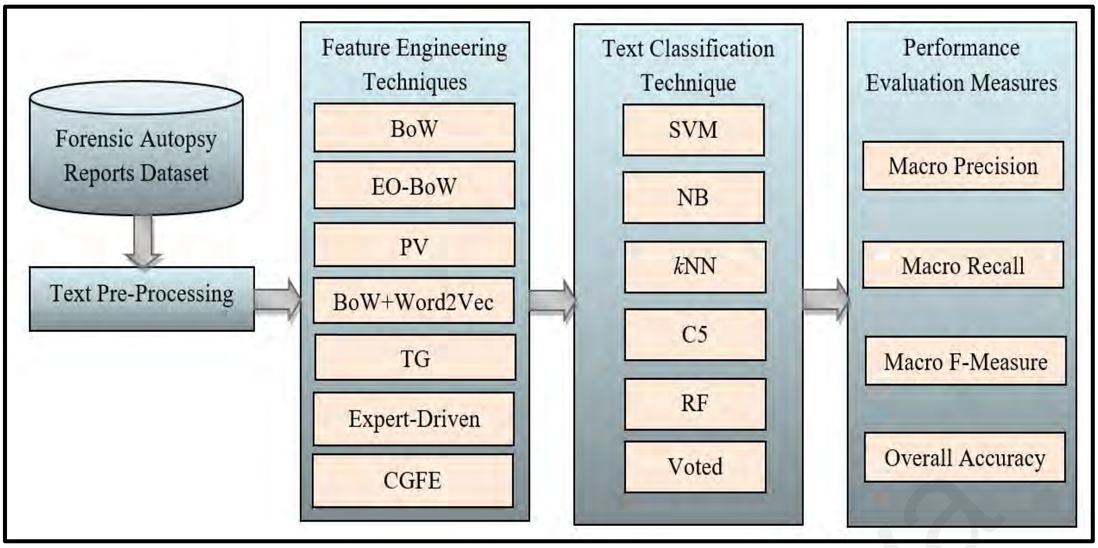


Figure 5.10: Experimental design for evaluation of CGFE technique

Moreover, for determining MoD, and CoD, hierarchical classification method was employed (as shown in Figure 5.11). In the hierarchical classification, the first-level classifier was trained to determine MoD, whereas the second-level classifier was trained to determine CoD. Therefore, five classification models were constructed, namely, one MoD-level classifier, and four CoD-level classifiers (i.e., accident-related CoD, homicide-related CoD, natural death-related CoD, and suicide-related CoD). The aforementioned classification models were deployed in a two-level cascade architecture where the autopsy report was first processed by the MoD-level classification model to determine MoD and then by the CoD-level classification model to determine the CoD. For example, if the MoD-level classification model determines the "accident" MoD for any given report, then the "accident"-related CoD classification model will be activated and further process the report to determine accident-related CoD (such as S06, S38, T07, and T75). The flowchart of the functionality of the hierarchical classification method is shown in Figure 5.11.

For MoD-level classification, all 1500 autopsy reports are grouped into four classes, namely, *accident*, *suicide*, *homicide*, and *natural death*, to create a training set. Afterward, various preprocessing steps (discussed in Section 3.4) were applied to the training set to remove irrelevant and noise features from this set. Unigram features were then extracted from the training set, and these features were represented using the TFiDF feature representation scheme. The unigram features represented by the TFiDF scheme were used because Mujtaba et al. (2016) compared term-based unigram, bigram, and trigram features to determine CoD from autopsy reports. In their study, these authors reported that unigram features produced better results than bigram and trigram features. In addition, they compared four feature representation techniques, namely, binary representation (BR), term frequency (TF), TFiDF, and NTFiDF, and found that TF and TFiDF outperformed BR and NTFiDF. Therefore, only term-based unigram features were extracted in MoD-level classification, and these features were represented by the TFiDF feature representation scheme to create an MFV. This MFV was used as the input for the supervised machine learning algorithms. Six supervised machine learning algorithms (namely, NB, SVM, *k*NN, C5.0, RF, and ensemble voted algorithms) were used to predict MoD at the first classification level. Hence, six analyses (1 MFV × 6 supervised machine learning algorithms) were performed to determine the results of MoD-level classification.

For the second-level classification task, all 1500 reports were grouped into their respective CoDs. Accordingly, 4 datasets were prepared, with each dataset belonging to one MoD (i.e., accident, suicide, homicide, and natural death). Each dataset contains forensic autopsy reports that belong to four distinct CoDs. Each dataset was converted into MFV by applying the proposed CGFE technique. Moreover, three different MFVs were produced from each dataset by applying unigram, bigram, and trigram CGFE models. Hence, 12 MFVs were produced. Afterward, 6 supervised machine learning algorithms were applied to the 12 constructed MFVs to determine CoD. Hence, a total of 72 analyses (12 MFVs × 6 supervised machine learning algorithms) were performed to determine the results of CoD-level classification.

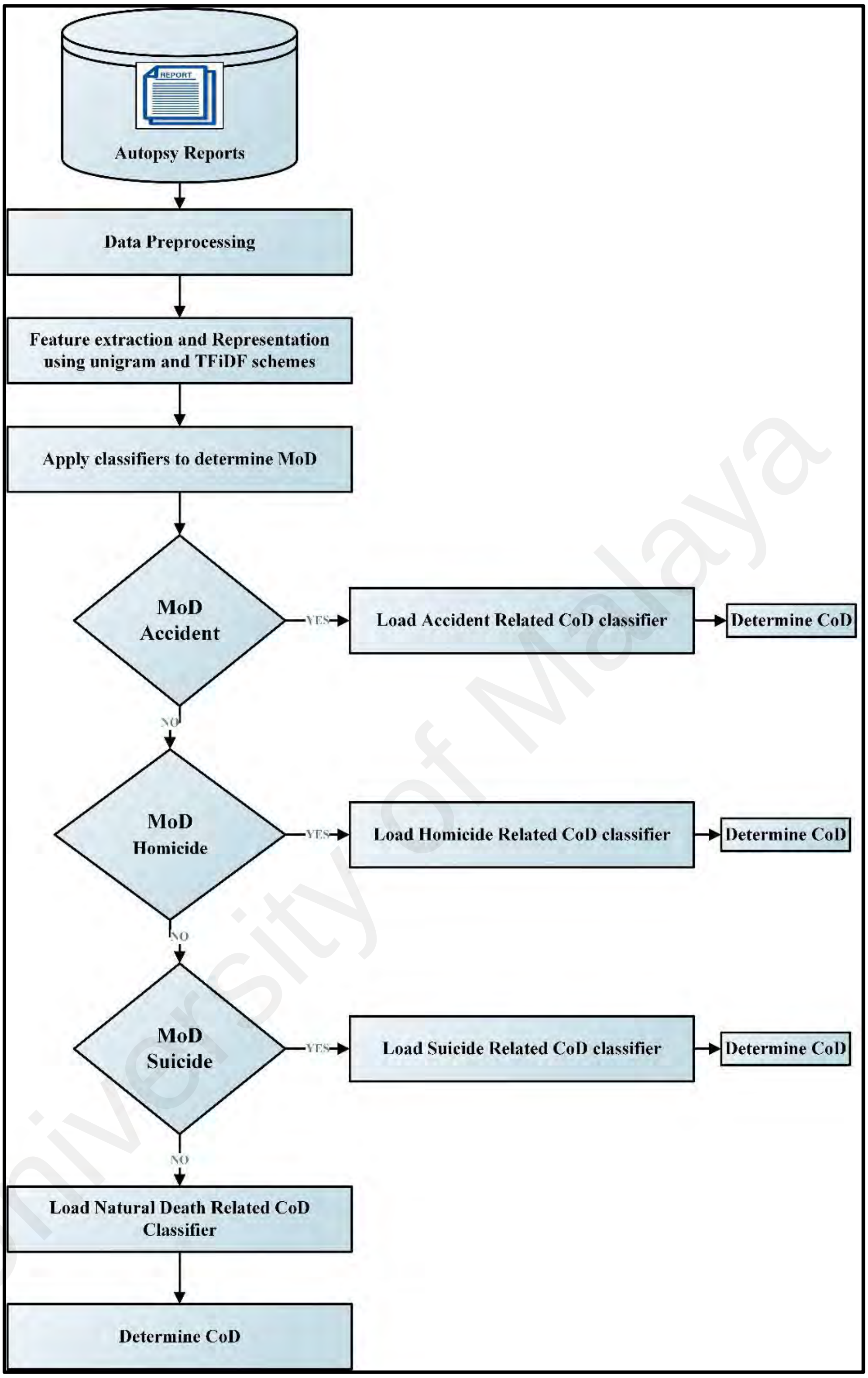

REPORT
Autopsy Reports
Data Preprocessing
Feature extraction and Representation using unigram and TFiDF schemes
Apply classifiers to determine MoD
MoD Accident
YES
Load Accident Related CoD classifier
Determine CoD
NO
MoD Homicide
YES
Load Homicide Related CoD classifier
Determine CoD
NO
MoD Suicide
YES
Load Suicide Related CoD classifier
Determine CoD
NO
Load Natural Death Related CoD Classifier
Determine CoD

Figure 5.11: Flowchart of the functionality of the hierarchical classification method

The performance of the constructed classification model through proposed CGFE technique was then compared with six state-of-the-art feature engineering techniques to increase understanding of its behavior. These state-of-the-art feature engineering techniques are the BoW technique (Harris, 1954), the EO-BoW technique (Passalis & Tefas, 2016), the PV technique (Le & Mikolov, 2014), the hybrid BoW + Word2Vec technique (Enríquez et al., 2016), the TG technique (Papadakis et al., 2016), and the expert-driven feature engineering technique (discussed in Chapter 4). All the techniques were applied to four MoD-related datasets to construct 24 MFVs. Then, the best performing supervised machine learning algorithm was applied to the 24 constructed MFVs to determine CoD. Therefore, 24 analyses (4 datasets × 6 MFVs × 1 supervised machine learning algorithm) were performed to compare the results of the proposed CGFE model with those of six aforementioned baselines.

Basic preprocessing steps, such as converting reports into lowercase, removing stop words, and spell checking, were written in Python using NLTK (Bird, 2006). The proposed CGFE technique was then applied to the preprocessed files. The proposed CGFE technique was implemented using Java programming language and MySQL database (for SNOMED-CT ontology). All the classification experiments (discussed in Section 5.3) were performed using the Weka workbench (Hall et al., 2009) except the C5. This is because Weka does not provide the implementation of C5. Thus, for this purpose C5 was applied using R programming language. In addition, the selected six machine learning algorithms were run using the parameters shown in Chapter 4, Table 4.2. These parameters were used because Mujtaba et al. (2016) rigorously performed the comparative study on classification of forensic autopsy reports using various machine learning algorithms and reported that the machine learning algorithms with these reported parameters outperformed. All the experiments were conducted using 10-fold cross-validation (Kohavi, 1995). To evaluate the performances of all the analyses, Precision$_M$, Recall$_M$, F-measure$_M$, and overall accuracy metrics were used (Sokolova & Lapalme, 2009). These metrics were selected due to an imbalanced class distribution in the dataset (Sokolova & Lapalme, 2009).

5.4 Experimental Results

This section presents the first-level classification, second-level classification, and baseline comparison results.

5.4.1 First-level Classification Results

The first-level classification results are presented in Table 5.2. This table shows the Precision_M, Recall_M, F-measure_M, and overall accuracy of the six supervised machine learning algorithms. It indicates that the SVM outperformed the other five algorithms by achieving the highest accuracy of 95.41%. However, a marginal difference in performance results was observed among the SVM, voted, and RF classifiers. The lowest accuracy of 79.35% was demonstrated by the *k*NN algorithm, followed by the C5.0 (88.99%) and NB (89.44%). A pair-wise McNemar statistical test (McCrum-Gardner, 2008; Adedokun & Burgess, 2011; Ott & Longnecker, 2015) was performed (using significance level of alpha = 0.05) to compare the overall accuracy of SVM classifier with all other five classifiers. The statistical difference was observed between SVM and all other classifiers ($p < 0.01$).

Table 5.2: First-level (MoD-level) classification results

Classifier	F-measure_M	Overall Accuracy
NB	0.888	89.44%
SVM	0.95	95.41%
C5.0	0.87	88.99%
***k*NN**	0.775	79.35%
RF	0.933	93.57%
Voted	0.932	94.03%

5.4.2 Second-level Classification Results

This section presents the second-level classification results using the proposed CGFE technique. As discussed in Section 5.3, 72 analyses were performed to evaluate the performance of the second-level classification. This section presents the results of all 72 analyses in Table 5.3 to Table 5.6. Each table shows a particular MoD (i.e., accident, homicide, natural death, and suicide). Moreover, each table indicates overall accuracy and F-measure$_M$ across three types of CGFE models (unigram, bigram, and trigram) and six supervised machine learning algorithms. As shown in these tables, the unigram CGFE model outperformed the bigram and trigram CGFE models in all 4 datasets. The lowest performance results were observed in the trigram CGFE model. In accident-related CoDs, the highest accuracy of 89.97% was achieved through the SVM, followed by the RF (89.84%) and voted (89.03%) algorithms using the unigram CGFE model. Moreover, in accident-related CoDs, the lowest accuracy of 72.23% was obtained by the *k*NN classifier using the trigram CGFE model. Table 5.3 presents all the results acquired using the accident-related CoD dataset.

In homicide-related CoDs, the highest accuracies of 89.45%, 89.04%, and 88.69% were obtained through the SVM, RF, and voted classifiers, respectively, using the unigram CGFE model. Moreover, in homicide-related CoDs, the lowest accuracy of 72.23% was obtained by the *k*NN classifier using the trigram CGFE model. Table 5.4 presents all the results acquired using the homicide-related CoD dataset.

Table 5.3: Results obtained from the accident-related CoD dataset

Classifier	**Unigram CGFE**		**Bigram CGFE**		**Trigram CGFE**	
	ACC	**FM$_M$**	**ACC**	**FM$_M$**	**ACC**	**FM$_M$**
NB	85.04%	0.849	82.52%	0.826	74.92%	0.747
SVM	89.97%	0.900	84.39%	0.845	78.74%	0.787
C5.0	85.13%	0.849	82.41%	0.824	75.07%	0.751
***k*NN**	82.81%	0.827	80.49%	0.803	72.23%	0.725
RF	89.84%	0.899	83.93%	0.839	78.39%	0.785
Voted	89.03%	0.891	83.47%	0.834	77.63%	0.775

** ACC (Overall Accuracy), FM$_M$ (F-Measure$_M$)

In natural death-related CoDs, the highest accuracy of 88.59% was achieved by the SVM algorithm using the unigram CGFE model. However, the accuracies of the voted and RF algorithms very closely follow that of SVM. Furthermore, in natural death-related CoDs, the lowest accuracy of 71.41% was obtained by the *k*NN using the trigram CGFE model. Table 5.5 presents all the results acquired using the natural death-related CoD dataset.

Table 5.4: Results obtained from the homicide-related CoD dataset

Classifier	Unigram CGFE		Bigram CGFE		Trigram CGFE	
	ACC	FM$_M$	ACC	FM$_M$	ACC	FM$_M$
NB	84.15%	0.840	79.39%	0.791	75.79%	0.759
SVM	89.45%	0.894	82.54%	0.825	77.63%	0.776
C5.0	84.29%	0.842	79.38%	0.794	76.45%	0.763
kNN	79.88%	0.800	78.13%	0.781	74.90%	0.749
RF	89.04%	0.889	82.33%	0.823	77.01%	0.769
Voted	88.69%	0.885	82.25%	0.821	76.87%	0.767

** ACC (Overall Accuracy), FM$_M$ (F-Measure$_M$)

In suicide-related CoDs, the highest accuracy of 95.33% was achieved by SVM using the unigram CGFE model. Moreover, the accuracies produced by the RF and voted classifiers very closely follow that of SVM. In addition, the lowest accuracy of 76.65% was demonstrated by the *k*NN classifier using the trigram CGFE model. Table 5.6 presents all the results acquired using the suicide-related CoD dataset.

Table 5.5: Results obtained from the natural death-related CoD dataset

Classifier	Unigram CGFE		Bigram CGFE		Trigram CGFE	
	ACC	FM$_M$	ACC	FM$_M$	ACC	FM$_M$
NB	83.25%	0.833	81.05%	0.804	74.29%	0.740
SVM	88.55%	0.885	85.33%	0.847	77.85%	0.775
C5.0	83.54%	0.835	81.16%	0.807	74.89%	0.745
kNN	79.06%	0.792	76.59%	0.761	71.44%	0.711
RF	88.05%	0.880	84.89%	0.844	77.21%	0.769
Voted	88.09%	0.881	84.76%	0.842	77.09%	0.765

** ACC (Overall Accuracy), FM$_M$ (F-Measure$_M$)

Table 5.6: Results obtained from the suicide-related CoD dataset

Classifier	Unigram CGFE		Bigram CGFE		Trigram CGFE	
	ACC	FM$_M$	ACC	FM$_M$	ACC	FM$_M$
NB	89.01%	0.889	84.79%	0.846	78.33%	0.784
SVM	95.33%	0.951	89.48%	0.894	81.00%	0.810
C5.0	89.97%	0.900	86.73%	0.867	78.31%	0.783
kNN	86.20%	0.861	80.04%	0.805	76.65%	0.765
RF	94.99%	0.950	89.15%	0.890	80.80%	0.808
Voted	94.69%	0.946	88.98%	0.890	80.49%	0.804

** ACC (Overall Accuracy), FM$_M$ (F-Measure$_M$)

In summary, the highest performance results were observed using the unigram CGFE model, followed by the bigram and trigram CGFE models in the second-level classification. Furthermore, among the supervised machine learning algorithms, SVM outperformed the other five algorithms. However, an insignificant difference was observed among the SVM, RF, and voted classifiers. The lowest performance was demonstrated by the *k*NN classifier. In addition, the highest performance results were observed in the suicide-related CoD dataset compared with the other three MoD datasets.

A Friedman statistical test (Demšar, 2006; McCrum-Gardner, 2008; Ott & Longnecker, 2015) was performed (using significance level of alpha = 0.05) to compare the overall accuracies for six classifiers across all four datasets. The statistical difference was found between the classifiers and datasets. Furthermore, Nemenyi post hoc statistical tests (Demšar, 2006; Ott & Longnecker, 2015) were performed and statistical difference was observed between SVM and all the classifiers across all the datasets ($p < 0.01$) except SVM and RF in accident-related CoDs dataset ($p > 0.05$).

5.4.3 Comparison of Proposed CGFE Technique with Baselines

This section reports the findings of 24 analyses that were performed to evaluate the performance of the proposed CGFE technique compared with six baseline techniques for feature engineering (as discussed in Section 5.3). Table 5.7 presents the accuracies of all 24 analyses using the SVM algorithm. As shown in the Table 5.7, the proposed CGFE technique performed better than the BoW, EO-BoW, BoW+Word2Vec, PV, and TG feature engineering techniques. However, the accuracy of the proposed CGFE technique is slightly lower than that of the semi-automated expert-driven technique. This result is attributed to the expert-driven technique being a semi-automated model, in which features are manually engineered and ranked by expert pathologists for each CoD. However, the expert-driven technique has two major limitations: the dependency on human experts and the time required for feature engineering and ranking. Moreover, manually engineering and ranking the features of all available ICD-10-related CoDs are nearly impossible for experts. Although the accuracy of the proposed CGFE technique is marginally lower than that of the semi-automated expert-driven technique, the proposed CGFE technique is fully-automated, and thus, it does not require any expert intervention for feature engineering or ranking. Furthermore, features can be engineered more rapidly in the proposed CGFE model than in the semi-automated expert-driven model.

Table 5.7: Overall accuracy comparison of CGFE technique with baselines

Feature Engineering Techniques	**Accident Dataset**	**Homicide Dataset**	**Natural Death Dataset**	**Suicide Dataset**
Fully automated BoW technique (Harris, 1954)	68.77%	69.68%	69.45%	73.27%
Fully automated EO-BoW technique (Passalis & Tefas, 2016)	73.27%	70.72%	70.45%	75.89%
Fully automated PV technique (Le & Mikolov, 2014)	75.55%	72.81%	72.18%	76.15%
Fully automated BoW+Word2Vec technique (Enríquez et al., 2016)	71.08%	69.71%	70.45%	73.49%
Fully automated TG technique (Papadakis et al., 2016)	77.63%	76.65%	76.23%	81.67%
Semi-automated expert-driven technique (proposed in Chapter 4)	**92.09%**	**91.45%**	**93.12%**	95.06%
Fully-automated unigram CGFE model	89.97%	89.45%	88.59%	**95.33%**

A Friedman statistical test was performed using significance level of alpha = 0.05) to compare the overall accuracies for seven methods (six baseline methods and one proposed CGFE technique) across all four datasets. The statistical difference was found across all seven techniques and four datasets. Furthermore, Nemenyi post hoc statistical tests were performed and statistical difference was observed between the proposed CGFE technique and BoW, EO-BoW, PV, BoW+Word2Vec, and TG techniques ($p < 0.01$) across all four datasets. Moreover, the statistical difference between the ED technique and proposed CGFE technique was observed across three datasets ($p < 0.01$) except suicide related dataset ($p > 0.05$).

Figure 5.12 shows the average computational time for classification model construction constructed through the SVM machine learning algorithm across the proposed and baseline techniques. The experiments were run on Corei7 system having 2.80 GHZ clock speed and a 16-gigabyte memory. As shown here, the classification time of proposed expert-driven and proposed CGFE technique is much faster than the baseline techniques. Moreover, in baseline techniques, BoW and EO-BoW proved to be faster than PV and BoW+Word2Vec technique.

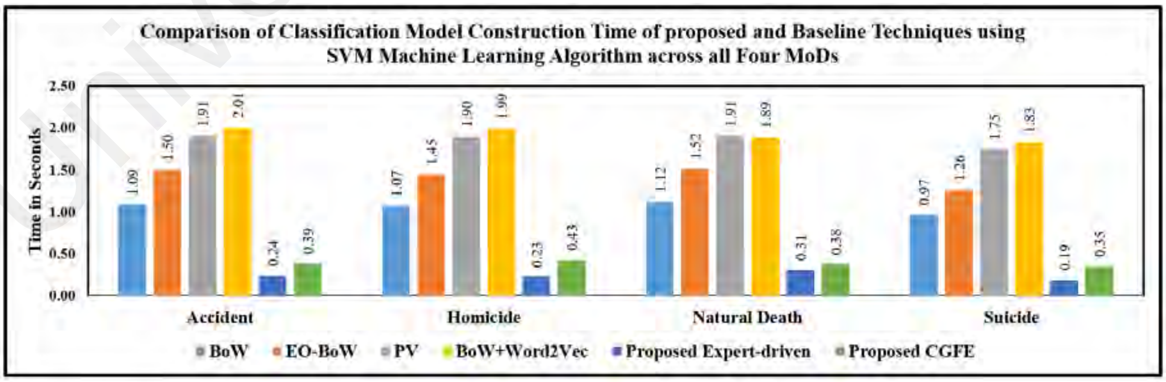


Figure 5.12. Classification time comparison between proposed CGFE and baselines

5.4.4 The Significance of using Graph-based Representation

To evaluate the significance of proposed CGFE technique, its performance was compared to with and without using graphs. In this set of experiments, concept-based features were mapped from collected reports using SNOMED CT ontology. Moreover, the extracted conceptual-based features were used to evaluate the classification results. Furthermore, in this set of experiments, the extracted concept-based features were not represented into graph structure. In experiments, SVM classifier was used to evaluate the classification performance. Furthermore, overall accuracy performance metric was used for performance evaluation. The experimental results are shown in Figure 5.13. As can be seen here, proposed CGFE technique outperformed when its results were compared with this new set of experiments. The possible reason for low performance of concept-based features when used without graph-based representation is that the SNOMED CT-based conceptual features overcomes the issue of polysemy and synonymy. However, it has no standardize functionality to construct statements with complex interrelationship between the terms or concepts. Thus, to achieve this, the graph-based approach was used and proven useful in classifying forensic autopsy reports.

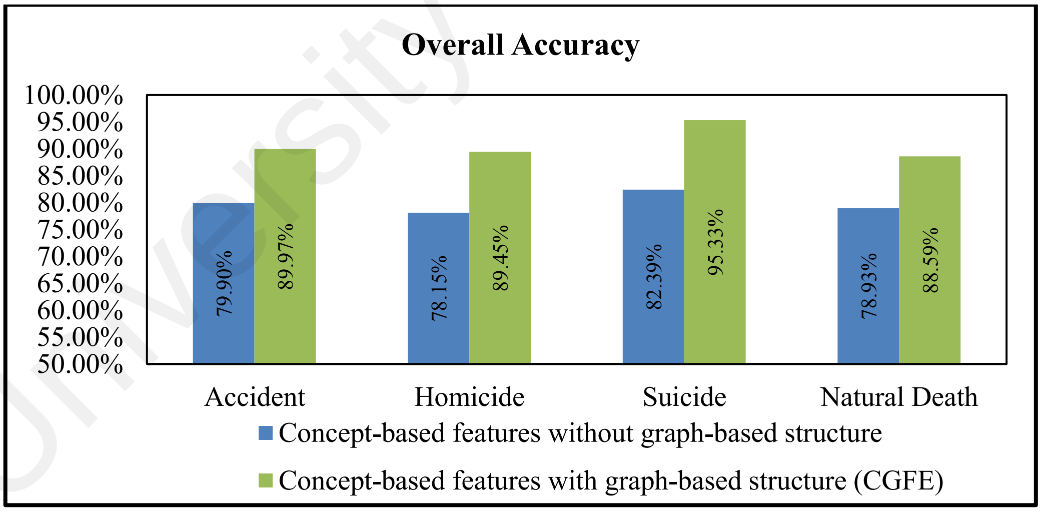


Figure 5.13. Significance of graph-based structure

For instance, consider the autopsy finding taken from forensic autopsy reports belonging to I23 CoD "The myocardium showed presence of infarction and absence of fibrosis". Moreover, consider the autopsy finding taken from forensic autopsy reports belonging to I25 CoD "The myocardium showed presence of fibrosis and absence of infarction.". It can be observed from aforementioned autopsy findings that in I23 CoD, in most of the cases, fibrosis and infarction is found on myocardium layer of the heart. Moreover, in some reports belonging to I23 CoD, the infarction was found however, fibrosis may or may not be found on myocardium layer of the heart. Conversely, in I25 CoD, in most of the cases, fibrosis was found in myocardium layer of the heart, however, infarction may or may not be found on myocardium layer of the heart. Thus, in such cases, the traditional approaches (such as BoW) will come up with list of features containing (absence, fibrosis, infarction, myocardium, presence, showed) features and transform the above-mentioned findings into these numeric vectors [(statement1 = 1,1,1,1,1, I23), (statement2 = 1,1,1,1,1, I25)] and may fail to accurately classify autopsy reports. Moreover, in such cases, SNOMED CT ontology can capture all the functional and topographic terms and concept ids from the aforementioned findings. However, it cannot capture the complex interrelationship between the medical terms or medical concepts. Hence, to achieve this and to achieve word context and word ordering, graph-based approach was exploited along with SNOMED-CT ontology by connecting co-occurring concepts with each other.

5.4.5 The Significance of Similarity and Uniqueness Metrics

It is important to ensure that the proposed features are effective features and not the redundant features. This is because the redundant features may increase the computationally complexity as well as cause the classification model overfitting. Conversely, providing sufficient features to train the machine learning algorithms is also one of the most significant step towards constructing effective classification model (Domingos, 2012). Therefore, to show the effectiveness of similarity, and uniqueness metrics, four experiments were run on four different datasets (namely accident, suicide, homicide, and natural death) without adding the uniqueness metrics. Hence, only 12 features were used (3 similarity features × 4 CoDs). For experiments, SVM classifier was used because it produced better accuracy in first-level and second-level classification models. The experimental results are shown in Figure 5.14.

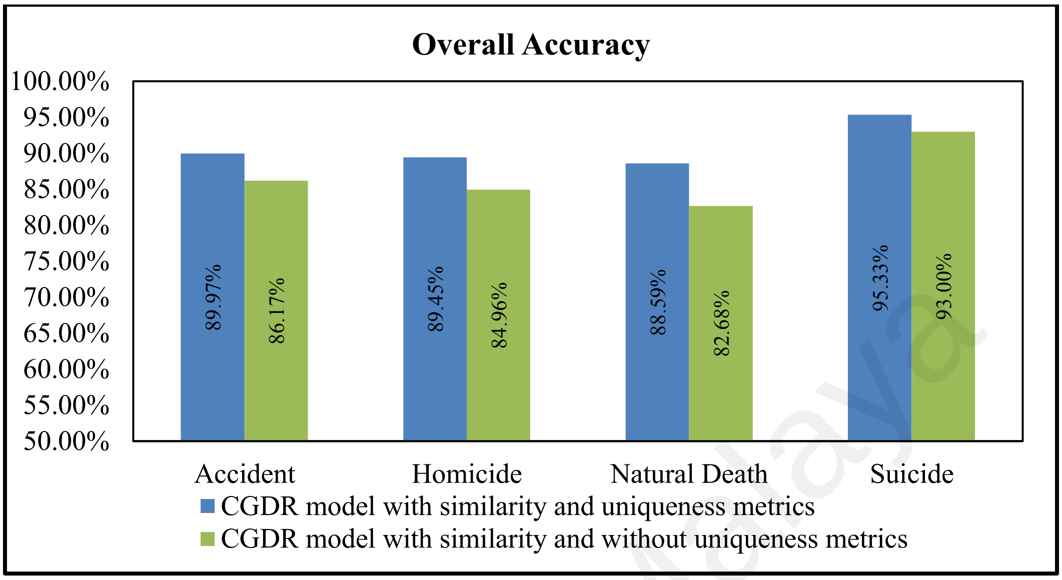
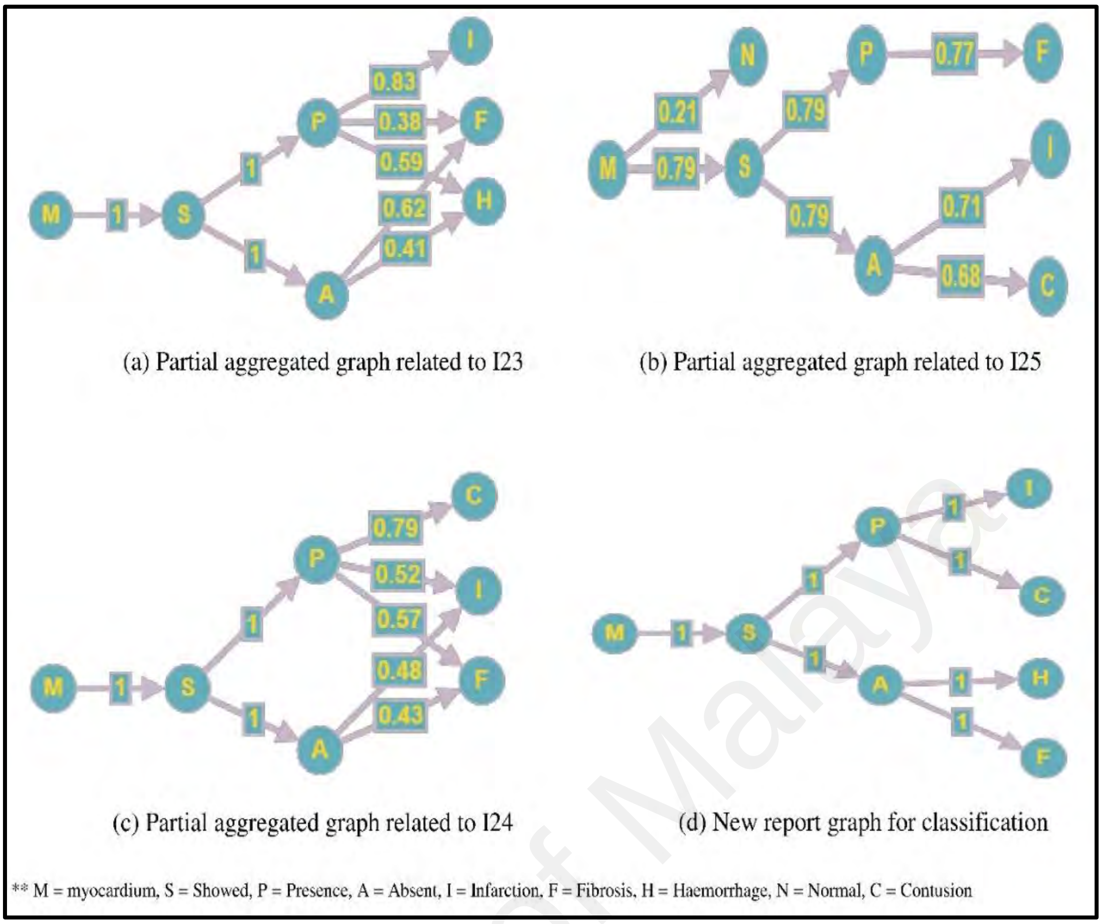


Figure 5.14. Significance of similarity and uniqueness metrics

As can be seen from Figure 5.14, the combination of uniqueness and similarity metrics produced the better results. This maybe because the combination of both metrics provide more context that may help classifiers in learning better classification rules. Moreover, in the forensic dataset, there is a very little difference among various CoDs. For instance, S06, S38, T07, and X80 are very close to each other and these CoDs usually share many relevant features in collected dataset. Thus, in such cases, the combination of uniqueness and similarity metrics may prove useful to differentiate between related CoDs. For instance, '*laceration wound*', '*abrasion wound*', '*Grazed abrasion*' are common in S06, S38, T07, and X80. However, '*abrasion forehead*', '*wound scalp*', '*right scalp*' are less frequent in S38, X80 and more frequent in S06, and T07. Therefore, the proposed similarity and uniqueness features may provide useful results. Consider another example where the terms '*pericardium*', '*epicardium*', '*myocardium*', '*endocardium*', are very common in I23, I24, and I25 CoDs. However, co-occurrence of '*myocardium*' and '*fibrosis*', is more common in I24, and I25 and less common in I23. Similarly, the co-occurrence of '*myocardium*' and '*infarction*' and '*myocardium*' and '*hemorrhage*' is more frequent in I23, less frequent in I24, and very less frequent in I25. Finally, the co-occurrence of '*epicardium*' and '*dimpling*' can be found in I25 but maybe less frequently found in I23, and I25. Therefore, in such cases, the proposed similarity and uniqueness features may provide better learning rules to the classifier.

Consider the Figure 5.15 where graphs (a), (b), and (c) represents the partial aggregated CoD level graphs related to I23, I25, and I24 CoDs respectively. These graphs encode the information related to myocardium layer of the heart. For instance, graph (a) encodes that most of the time, I23 is caused by the presence of infarction on myocardium. However, in the case of I23, the fibrosis or haemorrhage may or may not be found on myocardial layer. Conversely, graph (b) encodes that most of the time, I25 is caused by the presence of fibrosis on myocardium. However, in the case of I25, the infarction or contusion may not be found on myocardial layer. Finally, graph (c) encodes that most of the time, I24 is caused by the presence of contusion on myocardium. However, in the case of I24, the infarction or fibrosis may or may not be found on myocardial layer. In addition to graph (a), (b), and (c), the graph (d) in Figure 5.15 shows the new report level graph that needs to be classified. It can be seen from graph (d) that many vertices and edges are common in all three aggregated graphs. However, these can be discriminated on the basis of edge weights. For instance, the co-occurrence of (P → I) and (A → F) is more common in I23, and less common in I24. Moreover, the edge (P → C) is only available in I24 and the (A → H) is only available in I23 CoD. In addition, in such cases, the combination of similarity and uniqueness metrics may provide better learning rules to the machine learning classifiers and may produce better classification performance.



(a) Partial aggregated graph related to I23

(b) Partial aggregated graph related to I25

(c) Partial aggregated graph related to I24

(d) New report graph for classification

** M = myocardium, S = Showed, P = Presence, A = Absent, I = Infarction, F = Fibrosis, H = Haemorrhage, N = Normal, C = Contusion

Figure 5.15. Sample aggregated graphs belonging to I23, I24, and I25 CoDs

5.5 Discussion

The findings of this study indicated that supervised machine learning techniques can effectively classify forensic autopsy reports using the proposed CGFE technique with performance measures reaching 88% to94%. In the experiments, a few combinations were found to optimize classification performance. Among these combinations, the SVM classifier built with the proposed unigram CGFE model is the most effective for classifying forensic autopsy reports.

In the experimental results, the highest performance was observed in the suicide-related CoD dataset compared with the homicide-, accident-, and natural death-related CoD datasets. This result is attributed to the CoDs in the suicide-related dataset being different from one another, and thus, can be easily classified using supervised machine learning algorithms by determining the discriminative features of each class. For example, CoD X80 (intentional self-harm by jumping from height) and CoD T71 (intentional self-harm by hanging) are quite different from each other and can easily be discriminated using a supervised machine learning classifiers. CoD X80 may comprise injury-related information, such as injury on the head, abdomen, chest, arms, and legs. However, such information maybe unavailable in CoD T71. Thus, these CoDs can be easily differentiated using a machine learning algorithms. In the other three datasets (accident, homicide, and natural death), CoDs are similar in nature and share various features. Therefore, the classification task is more challenging for machine learning classifiers in these three datasets. For example, CoD S06 (Craniocerebral injury) and CoD T07 (multiple injuries) in the accident-related dataset are similar to each other. Furthermore, these CoDs contain numerous common features. Hence, discriminating between them is more challenging for machine learning classifiers.

The performance of a classification task primarily depends on the quality of features. Irrelevant and inadequate features typically produce unsatisfactory performance results. Therefore, the key task in free-text clinical reports classification is to identify the most relevant, discriminative, and powerful features using state-of-the-art feature engineering techniques (Domingos, 2012). Accordingly, the proposed CGFE technique was compared with six existing state-of-the-art feature engineering techniques in the experiments. The experimental results demonstrated the effectiveness of our proposed CGFE technique compared with existing baselines. The proposed CGFE technique performed better than existing baselines because it overcomes the limitations of existing feature engineering techniques, such as BoW, *n*-gram, EO-BOW, and Word2Vec. These techniques have three major limitations: word ordering, word inversion, and word synonymy and polysemy. In the proposed CGFE technique, the use of graph solves the issues of word ordering and word inversion. Moreover, the proposed CGFE technique overcomes the limitation of word synonymy and polysemy by considering word synonyms and concepts using SNOMED-CT ontology.

Two possible reasons are behind the unsatisfactory performance of the BoW technique. First, this technique does not consider the ordering of words. Second, it disregards word semantics (Lewis, 1992; Nigam et al., 2000; Sebastiani, 2002). The hybrid of BoW + Word2Vec technique produced the lowest results possibly because this model uses the voting approach between the BoW and Word2vec techniques. Therefore, BoW results may affect the classification result of the Word2Vec model. The EO-BoW technique performed better than the BoW and the hybrid Bow + Word2Vec techniques because it calculates the uncertainty of any given feature with document categories. The EO-BoW technique associates features with the document category that obtains the lowest entropy value for that feature.

The PV technique performed better than the BoW, hybrid BoW + Word2Vec, and EO-BoW techniques because it learns vector representation for variable length paragraphs of text. Vector representation is learnt to predict the surrounding words in context samples from a paragraph. Hence, this technique captures more semantics from a paragraph than Word2Vec, Bow, or EO-BoW. The TG technique performed better than the BoW, hybrid BoW + Word2Vec, EO-BoW, and PV models because it overcomes the issue of word ordering and word inversion by representing a textual document in a graph. However, two major limitations of the TG technique are evident. First, it does not consider the issue of word-level synonymy and polysemy and only considers the input content of a textual document. Second, the TG technique considers only edge matching and edge value to compute graph similarity. However, edge matching and edge value maybe insufficient to classify a text document. By contrast, our proposed CGFE technique overcomes the issue of word-level synonymy and polysemy by using the SNOMED CT database. Moreover, the proposed CGFE technique considers six metrics (i.e., VSM, VUM, ESM, EUM, SEWM, and UEWM) to compute graph similarity and represent input forensic autopsy report discriminatively.

The proposed CGFE technique could not beat the accuracy of the semi-automated expert driven technique because the expert-driven technique is semi-automated and expert-dependent, whereas the proposed CGFE technique is fully-automated. In the expert-driven technique, human experts (pathologists) are responsible for extracting and ranking useful features that belong to specific CoD autopsy reports. Therefore, beating the performance of the expert driven technique is challenging. However, as shown in Table 5.7, our proposed CGFE technique outperformed the other five fully automated baseline techniques. Moreover, the accuracy obtained using the proposed CGFE technique is approximately only 4% less than that of the semi-automated expert driven technique. In the suicide-related CoD dataset, the difference between the accuracies of the expert driven technique and the proposed CGFE technique is very marginal. The major reason for this is that the SNOMED-CT program did not identify some important concepts form the content-based features. This is because, the SNOMED-CT ontology covers the whole medical terms not specifically the autopsy terms. By contrast, *n*-gram features cover the entire text, but often do not map to medical concepts using SNOMEDT-CT ontology. The difference between both the proposed expert-driven and CGFE techniques is shown in Table 5.8.

Among the supervised machine learning algorithms, SVM outperformed the other five algorithms used in this study. The possible reason for the superior performance of SVM is its capability to produce improved results with all the available features in the MFV because SVM is not prone to overfitting (Joachims, 1998a). The accuracy of RF was slightly lower than that of SVM possibly because the prediction results of RF can be affected by a huge number of non-discriminative features but few discriminative features. Such combination of features may generate trees with less powerful and redundant features in a forest. These trees may produce incorrect classification results (Xu et al., 2012).

Table 5.8: Expert-driven versus CGFE

Proposed Expert-driven Technique	Proposed CGFE Technique
It is a semi-automated technique	It is a fully-automated technique
It is expert-dependent	It does not involve any expert intervention
The features are ranked and prioritize by domain experts (pathologists)	The features are automatically extracted through the use of graph-based approach and medical ontologies (SNOMED-CT)
The issue of word-level synonymy and polysemy is overcome by adding the similar words in the lexicons by the domain experts	The issue of word-level synonymy and polysemy is overcome through the help of medical ontology (SNOMED-CT)
Domain Experts did not consider the word-order	The word-order was obtained through the use of graph-based approach
The master feature vector comprised of number of features equal to number of CoDs	The master feature vector comprised of 6 × number of CoDs features, where 6 shows the number of similarity and uniqueness metrics (please refer to section 5.2.4)
It is time consuming and labour intensive	It is not time-consuming and labour intensive
It takes enormous time for classification of more CoDs because the experts will dig out and rank the features for all CoDs. Thus, the generalizability is difficult using this technique.	It will not take time for classification of more CoDs because the experts are not responsible for digging out and ranking the features for all CoDs. Thus, the generalizability is easy for any number of CoDs using this technique.

The accuracy of the ensemble voted classifier was slightly lower than that of SVM probably because the former considers the classification decision of all five classifiers and chooses the final decision based on majority voting. In most cases, the C5.0 classifier performed worse than the SVM, RF, and voted classifiers. The possible reason for the poorer performance of C5.0 is because all attributes in the MFV represent continuous data, which makes finding the required optimal thresholds for constructing the C5.0 decision tree difficult (Dreiseitl et al., 2001). The lowest performance results were obtained by the *k*NN and NB classifiers. The NB classifier assumes conditional independence among various features and is probably unsuitable for the collected dataset (Lewis, 1998b). In addition, this dependency becomes more complex with an increasing number of features and can negatively affect the performance of NB. The *k*NN classifier obtained the lowest classification results because of its default assumption of linearly scaling features, which may lead to imprecise distance computation measures. Moreover, this assumption proves deceptive with low discriminative features (Gutierrez-Osuna, 2002; Bhatia, 2010).

5.5.1 Role of MoD Classifier

The hierarchical classification method was designed to improve the accuracy of CoD prediction. To ascertain the efficacy of the hierarchical classification method, experiments were performed to compare the results of one-level classification with two-level classification. In one-level classification, all 1500 reports were labeled with their respective CoDs. Moreover, the proposed CGFE technique was used for feature engineering and MFV creation. The SVM classifier was also applied to compute the classification results. Table 5.9 shows the classification results obtained from the one–level and two–level classification models.

As shown in Table 5.9, in three accident-related CoDs (S06, S38, and T07), two-level classification exhibited approximately 5% to 6% improvement in accuracy compared with one-level classification. Moreover, in homicide-related CoDs, two-level classification demonstrated 6% to 8% improvement in accuracy compared with one-level classification. In two suicide-related CoDs (X80 and X74), two-level classification presented 8% to 12% improvement in accuracy compared with one-level classification. The evident reason for this improvement in accuracy is the sharing of several features among accident-, homicide-, and suicide-related CoDs. For example, CoDs T07 and X80 may share several features, and thus, one-level classification may produce a false positive rate. Furthermore, in several CoDs (e.g., T75, I23, I24, I25, Z11, T71, and T14) a slight improvement in performance results between one-level and two-level classification was observed. The reason for such improvement is that the aforementioned CoDs differ from one another, and thus, can be easily differentiated using machine learning classifiers. For example, CoDs T71 and T14 are different from all the other CoDs in the dataset.

Table 5.9: Accuracy of one-level vs. two-level classification

CoD	One Level Classification	Two Level Classification
S06	82.61%	87.86%
S38	83.89%	88.66%
T07	82.56%	87.40%
T75	93.48%	95.95%
X93	86.85%	90.76%
X99	84.00%	90.32%
Y00	80.25%	88.66%
Y09	79.60%	88.07%
I23	87.11%	88.41%
I24	87.13%	88.29%
I25	87.19%	88.59%
Z11	88.22%	89.05%
X80	81.39%	94.91%
X74	88.61%	94.70%
T71	92.76%	94.82%
T14	93.56%	96.88%

5.6 Conclusion

This chapter presented an effective fully-automated conceptual graph-based feature engineering (CGFE) technique to classify 16 types of forensic autopsy reports using the hierarchical text classification method. The experimental results showed that the proposed CGFE technique outperformed five (5) existing state-of-the-art fully-automated feature engineering techniques. The promising results of the proposed CGFE technique suggest that pathologists can adopt the proposed system as their basis for second opinion, thereby supporting them in effectively determining CoD. Furthermore, the proposed CGFE technique can be applied to classify other types of free-text clinical reports. The proposed technique can reduce the time and effort required for preparing public healthcare reports (more specifically forensic autopsy reports).

The limitation of the proposed CGFE technique is that its accuracy is slightly lower than that of the semi-automated expert-driven feature engineering technique (proposed in Chapter 4) because the features in the latter are manually extracted and ranked by human experts. Although the expert-driven feature engineering technique exhibits better accuracy, this technique is not the optimum solution for classifying all types of autopsy report because human experts are necessary to extract and rank useful features for all types of available forensic autopsy reports. To overcome the issues of the expert-driven feature engineering technique, a fully automated CGFE technique was developed in this Chapter.

CHAPTER 6: CONCLUSION

6.1 Introduction

In this thesis, existing feature engineering techniques were explored and investigated to classify free-text clinical reports. Moreover, conventional and state-of-the-art feature engineering techniques coupled with various supervised machine learning (SML) algorithms were employed and empirically investigated to classify free-text forensic autopsy reports. The need for classifying forensic autopsy reports is justified in Chapter 1 (Section 1.2). For the experiments, the forensic autopsy dataset was obtained from one of the largest emergency hospitals in Kuala Lumpur, Malaysia, and comprised four manners of death (MoDs) and 16 causes of death (CoDs). After an extensive set of experiments on the collected dataset, the existing feature engineering techniques coupled with various SML algorithms obtained a classification accuracy of 70%–80%. Thus, to improve the classification performance, this thesis proposed and developed two effective feature engineering techniques, namely, the semi-automated expert-driven feature engineering technique and fully automated conceptual graph-based feature engineering (CGFE) technique.

To obtain a specific level of classification performance in the clinical text classification domain, several studies (from the reviewed literature) have empirically investigated the effectiveness of the semi-automated expert-driven and fully automated techniques. Some reported that the expert-driven technique outperforms the fully automated technique; some reported the opposite result. Some studies reported no significant difference in the classification performance of the two techniques. Therefore, one should empirically investigate the performance of both techniques on free-text clinical report corpus to evaluate which one is better. The detailed justification for proposing the two techniques is also presented in Chapter 2 (Sections 2.8.1.3 and 2.11.1).

The experimental results showed that the proposed feature engineering techniques outperformed the existing feature engineering techniques in terms of classification performance. Moreover, the classification performance obtained through the fully automated CGFE technique was slightly lower than that of the semi-automated expert-driven technique.

Although the classification performance obtained through the fully automated CGFE technique was marginally lower than that of the semi-automated expert-driven technique, the fully automated CGFE technique was recommended for real-time deployment. In the autopsy domain, the number of targeted classes is increasing with the advancement of the forensic and legal medicine field. For instance, in recent years, experts were using the International Classification of Disease Ninth Edition (ICD-9), which contains 18,000 unique codes, in assigning primary CoD. However, ICD-9 has been enhanced to ICD-10, and it contains nine times more codes than ICD-9 (Organization, 1992; Sundararajan et al., 2004; Hazelwood & Venable, 2010; Control & Prevention, 2015). Therefore, extracting features from all of these categories is tedious for experts. In the autopsy or related and similar clinical domain, experts can be utilized to extract features from few categories, and a classification model can be developed using these expert-driven features to create a benchmark performance. Fully automated features can then be designed to obtain an accuracy equal to or more than that of the benchmark. Once the fully automated features obtain the specific level of classification performance, such features can be exploited for the remaining categories without generating the expert-driven features. A more detailed discussion on this issue is presented in Chapter 2 (Section 2.11.1).

Within the context of this study, each individual research question (RQ) is answered and discussed in Chapters 2, 3, 4, and 5. This thesis concludes by revisiting the research objectives and RQs presented in Chapter 1 and describing how they are achieved. The core contributions of this thesis and the limitations and future research directions are also discussed.

6.2 Reappraisal of the Research Objectives and Research Questions

This section revisits the research objectives and RQs for this study. Moreover, it discusses the findings of each RQ of each objective briefly.

Objective 1: To investigate the existing feature engineering techniques for the classifying free-text clinical reports, including forensic autopsy reports.

To achieve this objective, the academic literature in the field of "automated text classification (ATC) in free-text clinical reports" was reviewed by exploiting the procedural decision analysis in six aspects, namely, types of clinical reports, datasets and their characteristics, preprocessing techniques, feature engineering techniques, machine learning algorithm, and performance metrics. To achieve the first objective, a total of 69 primary studies from eight different bibliographic databases (namely, Web of Science, Scopus, IEEE Xplore, PubMed, Medline, ScienceDirect, ACM Digital Library, and SpringerLink) were systematically selected and rigorously reviewed from the perspective of the six aforementioned aspects. The findings of each RQ of objective 1 are given below.

RQ1: What are the existing feature engineering techniques for classifying free-text clinical reports?

In the literature review, several existing feature engineering techniques were identified, including BoW, n-gram, and Word2Vec. The detailed answer is given in Chapter 2 (Section 2.8).

RQ2: How feasible are the existing feature engineering techniques in terms of their performance in classifying forensic autopsy reports and determining the MoDs and CoDs from free-text forensic autopsy reports?

The literature review revealed that the existing feature engineering techniques coupled with SML approaches can classify forensic autopsy reports and determine the MoD and CoD from the forensic autopsy reports. However, the maximum accuracy of 70%–80% can be obtained through the existing feature engineering techniques (see Chapter 2).

RQ3: What are the limitations of the existing feature engineering techniques in determining the MoDs and CoDs from free-text forensic autopsy reports?

Three major limitations of the existing feature engineering techniques were identified, namely, losing word order, losing word context, and ignoring word-level synonymy and polysemy. Given these limitations, these approaches obtained a low classification performance for classifying forensic autopsy reports. More detailed limitations of the existing feature engineering techniques are discussed in Chapter 2 (Section 2.8).

Objective 2: To develop an effective semi-automated expert-driven feature engineering technique for addressing the issue of word-level synonymy and polysemy to classify CoDs from forensic autopsy reports.

To achieve this objective, three experts extracted and ranked the discriminative features from each kind of forensic autopsy report (16 kinds of reports in this study). These experts were provided with a summary of the statistical information related to the forensic autopsy dataset to assist them in extracting and ranking the features. Finally, these expert-driven features were used as input for different machine learning algorithms to construct the classification models. The details of this technique are discussed in Chapter 4. The findings of each RQ of objective 2 are given below.

RQ4: *How much of the classification performance of forensic autopsy reports can be enhanced through the proposed semi-automated expert-driven feature engineering technique?*

An overall classification accuracy of approximately 92% was obtained through the semi-automated expert-driven features, which was 12%–22% more than that of the existing feature engineering techniques. More detailed experiments and results are presented in Chapter 4 (Sections 4.3 and 4.4, respectively). The justification of achieving a high accuracy through the proposed semi-automated expert-driven technique is also presented in Chapter 4 (Section 4.5).

RQ5: *How important is the proposed semi-automated feature engineering technique in classifying forensic autopsy reports?*

In the literature review (Chapter 2), several studies argued that researchers should empirically investigate the use of both expert-driven and fully automated features to evaluate the performance of both of these features in classifying free-text clinical reports (Ye et al., 2014; Koopman, S. Karimi, et al., 2015; MacRae et al., 2015; Kalter et al., 2016). These studies argued that the classification model constructed through expert-driven features can also serve as a benchmark for the classification models constructed through fully automated features. Moreover, classification models constructed through fully automated features can be deployed in a real environment in either of two conditions: first, when the fully automated models achieve a specific level of accuracy that is higher than that of the expert-driven models; and second, when no significant difference is observed in the classification performance of both models (Ye et al., 2014; Koopman, S. Karimi, et al., 2015; MacRae et al., 2015; Kalter et al., 2016). Thus, in this thesis, the semi-automated expert-driven model served as a benchmark for the fully automated feature engineering techniques. The details are discussed in Sections 2.8.1.3, 2.11.1, and 4.6.

***RQ6:** What are the limitations of the proposed semi-automated expert-driven feature engineering technique?*

Several limitations of the proposed semi-automated expert-driven feature engineering technique were identified and listed below. The detailed limitations are discussed in Chapter 4 (Section 4.6).

1. The results of the proposed expert-driven technique depended heavily on the domain knowledge of the experts and their familiarity with forensic autopsy findings.
2. This technique heavily depended upon the experts; thus, an ample amount of time was required to extract the discriminative features for each CoD and rank them.
3. The developed model can only detect 16 CoDs, but each MoD contained thousands of CoDs (DiMaio & DiMaio, 2001). Hence, identifying useful features from each type of CoD was extremely difficult for the pathologists.

Objective 3: To develop an effective fully automated CGFE technique to address word order, word context, and word-level synonymy and polysemy in the text for classifying CoDs from forensic autopsy reports.

This objective aimed to overcome the limitations of the proposed semi-automated expert-driven techniques (Chapter 4, Section 4.6). To achieve this objective, a fully automated CGFE technique was proposed. In CGFE, graph theory and Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) ontology were employed to extract the discriminative features from the forensic autopsy reports. Graph theory was employed because graph representation provides flexibility and robustness on representing the natural language text compared with the traditional *n*-gram. Moreover, it overcomes the limitation of word co-occurrence, word ordering, and word inversion. Furthermore, SNOMED CT was used to extract the concept-based features from the extracted graph of word features to address word-level synonymy and polysemy. In addition, six different metrics were used to form a master numeric vector. Finally, this master numeric vector was fed as input to different machine learning algorithms to construct classification models through the CGFE feature engineering technique. The details of this technique with its functionality, experimental setup, results, and discussion are discussed in Chapter 5. The findings of each RQ of objective 2 are given below. The findings of each RQ of objective 3 are given below.

RQ7: *How much of the classification performance of the forensic autopsy reports can be enhanced through the fully automated feature engineering technique without human expert intervention?*

The experimental results in Chapter 5 (Section 5.4) showed that the proposed CGFE technique outperformed the existing fully automated techniques by obtaining approximately 89% overall accuracy. The justification of achieving a high accuracy through the fully automated CGFE technique is also presented in Chapter 5 (Sections 5.4.4, 5.4.5, and 5.5).

RQ8: *How can graph theory concepts be exploited in obtaining word order and word context from free-text forensic autopsy reports?*

Graph representation provides flexibility and robustness on representing the natural language text compared with the traditional *n*-gram. Moreover, it overcomes the limitation of word co-occurrence, word ordering, and word inversion. Here, the words were represented in vertices, and the co-occurring words were represented in adjacent vertices. An edge can be used to connect to co-occurring words. Moreover, edge value can be used as weight to determine the frequency of the co-occurrence of words. The significance of using graph-based structure and similarity and uniqueness metrics to classify forensic autopsy reports is given in Chapter 5 (Section 5.4.4 and 5.4.5).

RQ9: *How can existing medical or clinical ontologies be utilized to extract word-level synonymy and polysemy from free-text forensic autopsy reports?*

SNOMED CT is a standardized and multilingual vocabulary of clinical ontology used by physicians and other healthcare providers for the electronic exchange of clinical health information (Donnelly, 2006). In the medical field, the doctors may use a variety of words interchangeably while reporting the patient records in clinical reports. Thus, SNOMED CT ontology can be utilized to avoid word-level synonymy and polysemy and capture all terms that belong to the same medical concept. The detailed structure and discussion on SNOMED CT ontology are presented in Chapter 5 (Section 5.2.3).

Objective 4: To evaluate the performance of the proposed feature engineering techniques by using real-world forensic autopsy reports and by comparing the proposed techniques' performance with those of baseline feature engineering techniques.

To achieve this objective, the performance of the proposed feature engineering techniques was compared with five existing state-of-the-art feature engineering techniques. The experimental results showed that the proposed feature engineering techniques outperformed the baseline techniques. The details are discussed in Chapters 4 and 5 (Sections 4.4.3 and 5.4.3, respectively).

RQ10: *How can the performance of the proposed feature engineering techniques be evaluated?*

The performance of the proposed feature engineering techniques was evaluated using four performance metrics, namely, overall accuracy, macro precision, macro recall, and macro F-measure. The justification of using these metrics is discussed in Chapter 2 and 3 (Sections 2.10 and 3.7, respectively).

RQ11: *How much of the performance of the proposed feature engineering techniques is improved relative to those of the conventional and state-of-the-art feature engineering techniques?*

The performance of the proposed feature engineering techniques was compared with five existing state-of-the-art feature engineering techniques. The experimental results showed that the proposed feature engineering techniques obtained 9%–13% more overall accuracy compared with the baseline techniques. The details are discussed in Chapters 4 and 5 (Sections 4.4.3 and 5.4.3, respectively).

RQ12: *Which of the proposed semi-automated expert-driven and fully automated CGFE techniques is effective?*

The experimental results showed that the classification performance obtained through the fully automated CGFE technique was approximately 3% lower than that of the semi-automated expert-driven technique. Although the classification performance obtained through the fully automated CGFE technique was marginally lower than that of the semi-automated expert-driven technique, the fully automated CGFE technique was recommended for real-time deployment. The justification of this recommendation is also presented in Chapters 2 and 5 (Sections 2.11.1 and 5.5, respectively).

6.3 Limitations

Certain limitations were identified in the current work:

1. The current classification models can only classify 16 kinds of forensic autopsy domains (Section 3.3 and Table 3.1). Thus, they can determine either of the 16 CODs from the forensic autopsy reports. For more CoDs, more forensic autopsy reports can be gathered, and the classification models can be retrained using the proposed feature engineering techniques coupled with SML algorithms. In addition, it is believed that the proposed techniques have potential to show the similar classification performance on more CoDs.
2. In this study, all ICD-10 codes were truncated at the three-character level. For instance, the code S06.9 (unspecified intracranial injury) was converted to simply S06 (intracranial injury). This three-character level truncation was employed for two reasons. First, the aim of this study was to classify the forensic autopsy reports by using SML-based ATC techniques up to three character levels. For instance, S06.1 (traumatic cerebral edema), S06.2 (diffuse traumatic brain injury), S06.3 (focal traumatic brain injury), S06.4 (epidural hemorrhage), and S06.9 (unspecified intracranial injury) were all truncated to their upper level, that is, S06 (intracranial injury). A classification model was then constructed using the proposed feature engineering techniques to determine the three-character level CoD from the autopsy findings. Second, very few forensic autopsy reports were

available in the collected dataset for any specific CoD, so these reports were not sufficient to train and construct a robust and effective classification model. For a reasonable train set, all the forensic autopsy reports were converted into three-character level codes. In future work, the scope of determining the CoD can be enhanced to a deeper level.

3. The weaknesses of the proposed expert-driven feature engineering technique are already discussed in Chapter 4 (Section 4.6). To overcome the limitations, a fully automated CGFE technique was proposed in Chapter 5. The classification performance of CGFE was very close to those of the expert-driven techniques. However, in the future, more fully automated feature engineering techniques can be proposed and developed to improve classification performance compared with the semi-automated expert-driven techniques.
4. The developed classification models can only classify forensic autopsy reports written in English. Nonetheless, the proposed techniques can also be adapted to classify CoD from forensic autopsy reports written in languages other than English, provided these techniques will be trained on the features of the respective languages.

6.4 Future Research Directions

This section identifies the future research direction for the classification of forensic autopsy reports.

6.4.1 Quality of Dataset and the Use of Big Data

Hospitals may have different medical documentation systems and patterns or styles of preparing forensic autopsy reports. This difference may produce hurdles in generalizing a constructed classification model to multiple hospitals across the country or world. Thus, in the future, a heterogeneous–heterogeneous type of dataset (Chapter 2, Section 2.6) will be considered to construct a classification model that is generic and applicable at a wide scale. Such multimodal data require big data tools and techniques to overcome the heterogeneity issue. However, grouping different forensic autopsy datasets collected from multiple institutions is a major challenge, primarily because of patient privacy and security concerns (Coates, Souhami, & El Naqa, 2016). Thus, appropriate data-sharing protocols are required to apply big data analysis techniques (Coates et al., 2016). The resultant classification models developed using multimodal forensic autopsy datasets will be highly accurate and generic and can be trained and tested on a large volume of data.

6.4.2 Classifying Forensic Autopsy Reports using Unsupervised Learning

Human learning is essentially unsupervised. The structure of the world is discovered by observing it and not by being told the name of every object. Nevertheless, unsupervised machine learning has been overshadowed by the successes of supervised learning (LeCun, Bengio, & Hinton, 2015). This gap in the literature maybe because nearly all current studies rely on manually labeled data as input to the supervised algorithm for classifying the classes; thus, finding patterns between two or more classes using unsupervised grouping remains difficult. Intensive research is required to develop fully automated unsupervised algorithms that can classify forensic autopsy reports and obtain better classification performance than the techniques proposed in this thesis.

6.4.3 Deep Learning for Classifying Forensic Autopsy Reports

Deep learning has recently attracted the attention of many researchers in different fields. Natural language understanding is a new area in which deep learning is poised to make a large effect over the next few years (LeCun et al., 2015). Deep learning allows computational methods with a number of processing layers to learn data representation with different levels of abstraction (LeCun et al., 2015; Zhang, Pueyo, Wendt, Najork, & Broder, 2017). The main benefit of deep learning is that the features are not engineered by human experts. Conversely, these features are learned automatically from training data through general-purpose learning processes. The deep learning algorithms may prove beneficial in classifying free-text forensic autopsy reports with high-dimensional data where human-engineered features may not essentially imitate learning vectors from training data.

6.4.4 Ontology-based Forensic Autopsy Report Classification

In the future, researchers can develop ontologies specifically related to the autopsy domain and classify free-text forensic autopsy reports by using these ontologies. Moreover, an adaptive ontology can be planned and created from the classification result that can be developed and customized on the basis of the end user's report.

6.4.5 Multilinguistic Classification Model

In the future, researchers can consider the forensic autopsy reports written in a variety of languages, such as English, Malay, French, and Chinese. Moreover, the features that assist in classifying forensic autopsy reports in more than one language must be identified and proposed.

6.4.6 Dynamic Updating of the Feature Set

In the future, researchers may contribute in designing feature engineering techniques that enable the incremental addition or removal of features without rebuilding the entire model to keep up with new trends in forensic autopsy report classification.

6.5 Conclusion

This chapter concluded the entire thesis by revisiting the research objectives and RQs. This chapter also discussed the various limitations of existing studies and presented the future research directions in the field of automated classification of forensic autopsy reports.

REFERENCES

Abacha, A., Chowdhury, M. F. M., Karanasiou, A., Mrabet, Y., Lavelli, A., & Zweigenbaum, P. (2015). Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug-drug interaction extraction and classification. *Journal of Biomedical Informatics, 58*, 122-132.

Adedokun, O. A., & Burgess, W. D. (2011). Analysis of paired dichotomous data: A gentle introduction to the McNemar test in SPSS. *Journal of MultiDisciplinary Evaluation, 8*(17), 125-131.

Adeva, J. G., Atxa, J. P., Carrillo, M. U., & Zengotitabengoa, E. A. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications, 41*(4), 1498-1508.

Aery, M., & Chakravarthy, S. (2005). *Infosift: adapting graph mining techniques for text classification.* Paper presented at the FLAIRS Conference.

Afzal, Z., Schuemie, M. J., van Blijderveen, J. C., Sen, E. F., Sturkenboom, M., & Kors, J. A. (2013). Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. *Bmc Medical Informatics and Decision Making, 13*, 11.

Aggarwal, C. C., & Zhai, C. (2012a). *Mining text data*: Springer Science & Business Media.

Aggarwal, C. C., & Zhai, C. (2012b). A survey of text classification algorithms. *Mining text data*, 163-222.

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information processing & management, 39*(1), 45-65.

Al-garadi, M. A., Khan, M. S., Varathan, K. D., Mujtaba, G., & Al-Kabsi, A. M. (2016). Using online social networks to track a pandemic: A systematic review. *Journal of Biomedical Informatics, 62*, 1-11.

Alabbas, W., Al-Khateeb, H. M., & Mansour, A. (2016). *Arabic text classification methods: Systematic literature review of primary studies.* Paper presented at the Information Science and Technology (CiSt), 2016 4th IEEE International Colloquium on.

Alghoson, A. M. (2014). Medical Document Classification Based on MeSH. In R. H. Sprague (Ed.), *2014 47th Hawaii International Conference on System Sciences* (pp. 2571-2575). New York: Ieee.

Amrit, C., Paauw, T., Aly, R., & Lavric, M. (2017). Identifying child abuse through text mining and machine learning. *Expert Systems with Applications, 88*, 402-418.

Aronson, A. R. (2001). *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.* Paper presented at the Proceedings of the AMIA Symposium.

Bao, Y., Ishii, N., & Du, X. (2004). Combining multiple k-nearest neighbor classifiers using different distance functions *Intelligent Data Engineering and Automated Learning–IDEAL 2004* (pp. 634-641): Springer.

Barak-Corren, Y., Castro, V. M., Javitt, S., Hoffnagle, A. G., Dai, Y., Perlis, R. H., et al. (2017). Predicting Suicidal Behavior From Longitudinal Electronic Health Records. *American Journal of Psychiatry, 174*(2), 154-162.

Barbantan, I., Porumb, M., Lemnaru, C., & Potolea, R. (2016). Feature Engineered Relation Extraction - Medical Documents Setting. *International Journal of Web Information Systems, 12*(3), 336-358.

Bates, J., Fodeh, S. J., Brandt, C. A., & Womack, J. A. (2015). Classification of radiology reports for falls in an HIV study cohort. *Journal of the American Medical Informatics Association, 23*(e1), e113-e117.

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. *Analytica chimica acta, 760*, 25-33.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient *Noise reduction in speech processing* (pp. 1-4): Springer.

Bhatia, N. (2010). Survey of nearest neighbor techniques. *arXiv preprint arXiv:1007.0085*.

Bird, S. (2006). *NLTK: the natural language toolkit.* Paper presented at the Proceedings of the COLING/ACL on Interactive presentation sessions.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*: " O'Reilly Media, Inc.".

Bleik, S., Mishra, M., Huan, J., & Song, M. (2013). Text categorization of biomedical data sets using graph kernels and a controlled vocabulary. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10*(5), 1211-1217.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Bronselaer, A., & Pasi, G. (2013). *An approach to graph-based analysis of textual documents.* Paper presented at the 8th European Society for Fuzzy Logic and Technology (EUSFLAT-2013).

Buchan, K., Filannino, M., & Uzuner, O. (2017). Automatic prediction of coronary artery disease from clinical narratives. *Journal of Biomedical Informatics, 72*, 23-32.

Burger, G., Abu-Hanna, A., de Keizer, N., & Cornet, R. (2016). Natural language processing in pathology: a scoping review. *Journal of clinical pathology, 69*(11), 949-955.

Butt, L., Zuccon, G., Nguyen, A., Bergheim, A., & Grayson, N. (2013). Classification of cancer-related death certificates using machine learning. *Australasian Medical Journal, 6*(5), 292-300.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. *Ann Arbor MI, 48113*(2), 161-175.

CDC, N. (2015). International Classification of Diseases, (ICD-10-CM/PCS) Transition - Background. Retrieved 29 September, 2017, from https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm

Chakravarthy, S., Venkatachalam, A., & Telang, A. (2010). *A graph-based approach for multi-folder email classification.* Paper presented at the Data Mining (ICDM), 2010 IEEE 10th International Conference on.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

Chin, H. J., & Kim, S. (2003). Standardization of main concept in chief complaint based on SNOMED CT for utilization in electronic medical record. *Journal of Korean Society of Medical Informatics, 9*(3), 235-247.

Chomutare, T. (2014). Text Classification to Automatically Identify Online Patients Vulnerable to Depression. In P. Cipresso, A. Matic & G. Lopez (Eds.), *Pervasive Computing Paradigms for Mental Health* (Vol. 100, pp. 125-130). Berlin: Springer-Verlag Berlin.

Clark, C., Wellner, B., Davis, R., Aberdeen, J., & Hirschman, L. (2017). Automatic classification of RDoC positive valence severity with a neural network. *Journal of Biomedical Informatics*.

Coates, J., Souhami, L., & El Naqa, I. (2016). Big data analytics for prostate radiotherapy. *Frontiers in oncology, 6*.

Comelli, A., Agnello, L., Vitabile, S., & Ieee. (2015). *An Ontology-Based Retrieval System for Mammographic Reports*. New York: Ieee.

Control, C. f. D., & Prevention. (2015). International Classification of Diseases (ICD-10-CM/PCS), transition: background. *Available at: cdc. gov/nchs/icd/icd10cm_pcs_ background. htm. Accessed November, 6*.

Cornet, R., & de Keizer, N. (2008). Forty years of SNOMED: a literature review. *Bmc Medical Informatics and Decision Making, 8*(1), S2.

Costache, M., Lazaroiu, A. M., Contolenco, A., Costache, D., George, S., Sajin, M., et al. (2014). Clinical or postmortem? The importance of the autopsy; a retrospective study. *Maedica (Buchar), 9*(3), 261-265.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*: John Wiley & Sons.

Cristianini, N., & Shawe-Taylor, J. (1999). An introduction to SVM: Cambridge University Press, Cambridge.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*: Cambridge university press.

Dai, X., & Bikdash, M. (2015). *Hybrid classification for tweets related to infection with influenza.* Paper presented at the Conference Proceedings - IEEE SOUTHEASTCON.

Dai, X. F., & Bikdash, M. (2015). Hybrid Classification for Tweets Related to Infection with Influenza *Ieee Southeastcon 2015*. New York: Ieee.

Danso, S., Atwell, E., & Johnson, O. (2013). Linguistic and statistically derived features for cause of death prediction from verbal autopsy text *Language processing and knowledge in the web* (pp. 47-60): Springer.

Danso, S., Atwell, E., & Johnson, O. (2014). A comparative study of machine learning methods for verbal autopsy text classification. *arXiv preprint arXiv:1402.4380*.

Dasgupta, A., Drineas, P., Harb, B., Josifovski, V., & Mahoney, M. W. (2007). *Feature selection methods for text classification.* Paper presented at the Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining.

Dasondi, V., Pathak, M., & Singh, N. P. (2016). *An implementation of graph based text classification technique for social media.* Paper presented at the Colossal Data Analysis and Networking (CDAN), Symposium on.

de la Iglesia, D., Garcia-Remesal, M., Anguita, A., Munoz-Marmol, M., Kulikowski, C., & Maojo, V. (2014). A Machine Learning Approach to Identify Clinical Trials Involving Nanodrugs and Nanodevices from ClinicalTrials.gov. *Plos One, 9*(10), 15.

Debole, F., & Sebastiani, F. (2004). Supervised term weighting for automated text categorization *Text mining and its applications* (pp. 81-97): Springer.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of machine learning research, 7*(Jan), 1-30.

Deng, Y., Groll, M. J., & Denecke, K. (2015) Rule-based Cervical Spine Defect Classification Using Medical Narratives. *Vol. 216. Studies in health technology and informatics* (pp. 1038).

DiMaio, D., & DiMaio, V. J. (2001). *Forensic pathology*: CRC press.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87.

Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform, 121*, 279.

Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., & Binder, M. (2001). A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *Journal of Biomedical Informatics, 34*(1), 28-36.

Enríquez, F., Troyano, J. A., & López-Solaz, T. (2016). An approach to the use of word embeddings in an opinion classification task. *Expert Systems with Applications, 66*, 1-6.

Farshchi, S. M. R., & Yaghoobi, M. (2013). Categorization of Medical Documents Using Hybrid Competitive Neural Network with String Vector, a Novel Approach. In Z. Y. Du (Ed.), *Intelligence Computation and Evolutionary Computation* (Vol. 180, pp. 1045-1054). Berlin: Springer-Verlag Berlin.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861-874.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems. *J. Mach. Learn. Res, 15*(1), 3133-3181.

Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M. A., & Meira Jr, W. (2011). Word co-occurrence features for text classification. *Information Systems, 36*(5), 843-858.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *Bmc Medical Informatics and Decision Making, 12*(1), 1.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research, 3*(Mar), 1289-1305.

Fragos, K., & Skourlas, C. (2016). *Smoothing Class Frequencies for KNN Medical Article Classification.* Paper presented at the Proceedings of the 20th Pan-Hellenic Conference on Informatics.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*: Academic press.

Garla, V., Taylor, C., & Brandt, C. (2013). Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. *Journal of Biomedical Informatics, 46*(5), 869-875.

Gatta, R., Vallati, M., De Bari, B., & Ozsahin, M. (2014). *The impact of different training sets on medical documents classification*.

Gee, K. R., & Cook, D. J. (2005). *Text Classification Using Graph-Encoded Linguistic Elements.* Paper presented at the FLAIRS Conference.

Giannakopoulos, G., Karkaletsis, V., Vouros, G., & Stamatopoulos, P. (2008). Summarization system evaluation revisited: N-gram graphs. *ACM Transactions on Speech and Language Processing (TSLP), 5*(3), 5.

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. *Journal of Medical Internet Research, 15*(11), 9.

Gutierrez-Osuna, R. (2002). Pattern analysis for machine olfaction: a review. *IEEE Sensors journal, 2*(3), 189-202.

Gutierrez-Sacristan, A., Bravo, A., Portero-Tresserra, M., Valverde, O., Armario, A., Blanco-Gandia, M. C., et al. (2017). Text mining and expert curation to develop a database on psychiatric diseases and their genes. *Database-the Journal of Biological Databases and Curation*, 9.

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

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate data analysis* (Vol. 5): Prentice hall Upper Saddle River, NJ.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning.* The University of Waikato.

Hall, M. A., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*(1), 10-18.

Hall, M. A., & Smith, L. A. (1998). Practical feature subset selection for machine learning.

Halland, K., & Britz, K. (2011). Investigations into the use of SNOMED CT to enhance an OpenMRS health information system. *South African Computer Journal, 47*(1), 33-45.

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning, 45*(2), 171-186.

Harris, Z. S. (1954). Distributional structure. *Word, 10*(2-3), 146-162.

Hassanpour, S., & Langlotz, C. P. (2016). Predicting High Imaging Utilization Based on Initial Radiology Reports: A Feasibility Study of Machine Learning. *Academic Radiology, 23*(1), 84-89.

Hassanpour, S., Langlotz, C. P., Amrhein, T. J., Befera, N. T., & Lungren, M. P. (2017). Performance of a Machine Learning Classifier of Knee MRI Reports in Two Large Academic Radiology Practices: A Tool to Estimate Diagnostic Yield. *AJR Am J Roentgenol, 208*(4), 750-753.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Overview of supervised learning *The elements of statistical learning* (pp. 9-41): Springer.

Hayes, P. J., & Weinstein, S. P. (1990). *CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories.* Paper presented at the IAAI.

Hazelwood, A. C., & Venable, C. A. (2010). ICD-10-CM and ICD-10-PCS Preview. *ICD-10-CM and ICD-10-PCS Preview, second edition/AHIMA, American Health Information Management Association*.

Hoelz, B. W., Ralha, C. G., & Geeverghese, R. (2009). *Artificial intelligence applied to computer forensics.* Paper presented at the Proceedings of the 2009 ACM symposium on Applied Computing.

Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., & Verspoor, K. (2014). Biomedical text mining: state-of-the-art, open problems and future challenges *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics* (pp. 271-300): Springer.

Hotho, A., Maedche, A., & Staab, S. (2002). Ontology-based text document clustering. *KI, 16*(4), 48-54.

Imane, A., & Mohamed, B. A. (2017). *Multi-label Categorization of French Death Certificates using NLP and Machine Learning.* Paper presented at the Proceedings of the 2nd international Conference on Big Data, Cloud and Applications.

Iqbal, E., Mallah, R., Jackson, R. G., Ball, M., Ibrahim, Z. M., Broadbent, M., et al. (2015). Identification of Adverse Drug Events from Free Text Electronic Patient Records and Information in a Large Mental Health Case Register. *Plos One, 10*(8), 14.

James, S. H., Nordby, J. J., & Bell, S. (2002). *Forensic science: an introduction to scientific and investigative techniques*: CRC press.

Japkowicz, N. (2000). *The class imbalance problem: Significance and strategies.* Paper presented at the Proc. of the Int'l Conf. on Artificial Intelligence.

Jiang, C., Coenen, F., Sanderson, R., & Zito, M. (2010). Text classification using graph mining-based feature extraction. *Knowledge-Based Systems, 23*(4), 302-308.

Jiang, Z., Li, L., & Huang, D. (2016). An Unsupervised Graph Based Continuous Word Representation Method for Biomedical Text Mining. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 13*(4), 634-642.

Jindal, R., & Taneja, S. (2015). A Lexical Approach for Text Categorization of Medical Documents. In P. Samuel (Ed.), *Proceedings of the International Conference on Information and Communication Technologies, Icict 2014* (Vol. 46, pp. 314-320). Amsterdam: Elsevier Science Bv.

Jivani, A. G. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl, 2*(6), 1930-1938.

Jo, T. (2013). Application of Table based Similarity to Classification of Bio-Medical Documents. *2013 Ieee International Conference on Granular Computing (Grc)*, 162-166.

Joachims, T. (1998a). *Text categorization with support vector machines: Learning with many relevant features.* Paper presented at the European conference on machine learning.

Joachims, T. (1998b). Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, 137-142.

Jouhet, V., Defossez, G., Burgun, A., le Beux, P., Levillain, P., Ingrand, P., et al. (2012). Automated Classification of Free-text Pathology Reports for Registration of Incident Cases of Cancer. *Methods of Information in Medicine, 51*(3), 242-251.

Kalter, H. D., Perin, J., & Black, R. E. (2016). Validating hierarchical verbal autopsy expert algorithms in a large data set with known causes of death. *J Glob Health, 6*(1), 010601.

Kasthurirathne, S. N., Dixon, B. E., Gichoya, J., Xu, H. P., Xia, Y. N., Mamlin, B., et al. (2016). Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection. *Journal of Biomedical Informatics, 60*, 145-152.

Kasthurirathne, S. N., Dixon, B. E., Gichoya, J., Xu, H. P., Xia, Y. N., Mamlin, B., et al. (2017). Toward better public health reporting using existing off the shelf approaches: The value of medical dictionaries in automated cancer detection using plaintext medical data. *Journal of Biomedical Informatics, 69*, 160-176.

Kasthurirathne, S. N., Dixon, B. E., & Grannis, S. J. (2015) Evaluating Methods for Identifying Cancer in Free-Text Pathology Reports Using Various Machine Learning and Data Preprocessing Approaches. *Vol. 216* (pp. 1070).

Kaurova, O., Alexandrov, M., & Blanco, X. (2011). Classification of free text clinical narratives (short review). *Scient. Book "Information Science and Computing", Publ. House ITHEA*.

Kavuluru, R., Rios, A., & Lu, Y. (2015). An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artif Intell Med, 65*(2), 155-166.

Kim, Y.-M., & Delen, D. (2016). Medical informatics research trend analysis: A text mining approach. *Health Informatics Journal*, 1460458216678443.

Ko, Y., & Seo, J. (2000). *Automatic text categorization by unsupervised learning.* Paper presented at the Proceedings of the 18th conference on Computational linguistics-Volume 1.

Kocbek, S., Cavedon, L., Martinez, D., Bain, C., Mac Manus, C., Haffari, G., et al. (2016). Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources. *Journal of Biomedical Informatics, 64*, 158-167.

Kodovsky, J., Fridrich, J., & Holub, V. (2012). Ensemble classifiers for steganalysis of digital media. *IEEE Transactions on Information Forensics and Security, 7*(2), 432-444.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Paper presented at the Ijcai.

Koopman, Karimi, S., Nguyen, A., McGuire, R., Muscatello, D., Kemp, M., et al. (2015). Automatic classification of diseases from free-text death certificates for real-time surveillance. *Bmc Medical Informatics and Decision Making, 15*, 10.

Koopman, Zuccon, G., Nguyen, A., Bergheim, A., & Grayson, N. (2015). Automatic ICD-10 classification of cancers from free-text death certificates. *International Journal of Medical Informatics, 84*(11), 956-965.

Koopman, B., Karimi, S., Nguyen, A., McGuire, R., Muscatello, D., Kemp, M., et al. (2015). Automatic classification of diseases from free-text death certificates for real-time surveillance. *Bmc Medical Informatics and Decision Making, 15*, 10.

Koopman, B., Zuccon, G., Nguyen, A., Bergheim, A., & Grayson, N. (2015). Automatic ICD-10 classification of cancers from free-text death certificates. *International journal of medical informatics, 84*(11), 956-965.

Lauren, P., Qu, G., Zhang, F., & Lendasse, A. (2017). Discriminant document embeddings with an extreme learning machine for classifying clinical narratives. *Neurocomputing*.

Le, Q. V., & Mikolov, T. (2014). *Distributed Representations of Sentences and Documents.* Paper presented at the ICML.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444.

Lee, D., Cornet, R., Lau, F., & De Keizer, N. (2013). A survey of SNOMED CT implementations. *Journal of Biomedical Informatics, 46*(1), 87-96.

Lewis, D. D. (1992). *Feature selection and feature extraction for text categorization.* Paper presented at the Proceedings of the workshop on Speech and Natural Language.

Lewis, D. D. (1998a). *Naive (Bayes) at forty: The independence assumption in information retrieval.* Paper presented at the European conference on machine learning.

Lewis, D. D. (1998b). Naive (Bayes) at forty: The independence assumption in information retrieval *Machine learning: ECML-98* (pp. 4-15): Springer.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18-22.

Lin, C., Karlson, E. W., Canhao, H., Miller, T. A., Dligach, D., Chen, P. J., et al. (2013). Automatic Prediction of Rheumatoid Arthritis Disease Activity from the Electronic Medical Records. *Plos One, 8*(8), 10.

Liu, T., Moore, A. W., Yang, K., & Gray, A. G. (2004). *An investigation of practical approximate nearest neighbor algorithms.* Paper presented at the Advances in neural information processing systems.

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 14-23.

Lopprich, M., Krauss, F., Ganzinger, M., Senghas, K., Riezler, S., & Knaup, P. (2016). Automated Classification of Selected Data Elements from Free-text Diagnostic Reports for Clinical Research. *Methods of Information in Medicine, 55*(4), 373-380.

Lucini, F. R., Fogliatto, F. S., da Silveira, G. J., Neyeloff, J. L., Anzanello, M. J., Kuchenbecker, R. d. S., et al. (2017). Text mining approach to predict hospital admissions using early medical records from the emergency department. *International Journal of Medical Informatics, 100*, 1-8.

Luo, Y., Sohani, A. R., Hochberg, E. P., & Szolovits, P. (2014). Automatic lymphoma classification with sentence subgraph mining from pathology reports. *Journal of the American Medical Informatics Association, 21*(5), 824-832.

Mabotuwana, T., Lee, M. C., & Cohen-Solal, E. V. (2013). An ontology-based similarity measure for biomedical data - Application to radiology reports. *Journal of Biomedical Informatics, 46*(5), 857-868.

Macià, N., Bernadó-Mansilla, E., Orriols-Puig, A., & Ho, T. K. (2013). Learner excellence biased by data set selection: A case for data characterisation and artificial data sets. *Pattern Recognition, 46*(3), 1054-1066.

MacRae, J., Love, T., Baker, M. G., Dowell, A., Carnachan, M., Stubbe, M., et al. (2015). Identifying influenza-like illness presentation from unstructured general practice clinical narrative using a text classifier rule-based expert system versus a clinical expert. *Bmc Medical Informatics and Decision Making, 15*(1), 78.

Malliaros, F. D., & Skianis, K. (2015). *Graph-based term weighting for text categorization.* Paper presented at the Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1): Cambridge university press Cambridge.

Marafino, B., Davies, J. M., Bardach, N. S., Dean, M. L., & Dudley, R. A. (2014). N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *Journal of the American Medical Informatics Association, 21*(5), 871-875.

Martinez, D., Ananda-Rajah, M. R., Suominen, H., Slavin, M. A., Thursky, K. A., & Cavedon, L. (2015). Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans. *Journal of Biomedical Informatics, 53*, 251-260.

Masino, A. J., Grundmeier, R. W., Pennington, J. W., Germiller, J. A., & Crenshaw, E. B. (2016). Temporal bone radiology report classification using open source machine learning and natural langue processing libraries. *Bmc Medical Informatics and Decision Making, 16*(1), 65.

McCrum-Gardner, E. (2008). Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery, 46*(1), 38-41.

Meadow, C. T. (1992). *Text information retrieval systems*: Academic Press, Inc.

Miasnikof, P., Giannakeas, V., Gomes, M., Aleksandrowicz, L., Shestopaloff, A. Y., Alam, D., et al. (2015a). Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. *BMC Med, 13*, 286.

Miasnikof, P., Giannakeas, V., Gomes, M., Aleksandrowicz, L., Shestopaloff, A. Y., Alam, D., et al. (2015b). Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. *BMC Medicine, 13*, 9.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.* Paper presented at the Advances in neural information processing systems.

Mitchell, M. (1998). *An introduction to genetic algorithms*: MIT press.

Mouriño-García, M., Pérez-Rodríguez, R., Anido-Rifón, L., & Gómez-Carballa, M. (2016). *Bag-of-Concepts Document Representation for Bayesian Text Classification.* Paper presented at the Computer and Information Technology (CIT), 2016 IEEE International Conference on.

Mujtaba, G., Shuib, L., Raj, R. G., Rajandram, R., & Shaikh, K. (2016). *Automatic text classification of ICD-10 related CoD from complex and free text forensic autopsy reports.* Paper presented at the Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on.

Nam, S., Kim, S.-K., Kim, H.-G., Ngo, V., & Zong, N. (2016). Structuralizing biomedical abstracts with discriminative linguistic features. *Computers in biology and medicine, 79*, 276-285.

Napolitano, G., Marshall, A., Hamilton, P., & Gavin, A. T. (2016). Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. *Artificial Intelligence in Medicine, 70*, 77-83.

Nguyen, D. H., & Patrick, J. D. (2014). Supervised machine learning and active learning in classification of radiology reports *J Am Med Inform Assoc, 21*(5), 893-901.

Nicolosi, N. (2008). Feature selection methods for text classification: November.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning, 39*(2-3), 103-134.

NLM. (2017). SNOMED CT International Edition. 2017, from https://www.nlm.nih.gov/healthit/snomedct/international.html

Oleynik, M., Patrão, D. F. C., & Finger, M. (2017) Automated Classification of Semi-Structured Pathology Reports into ICD-O Using SVM in Portuguese. *Vol. 235. Studies in health technology and informatics* (pp. 256-260).

Organization, W. H. (1979). Medical certification of cause of death: instructions for physicians on use of international form of medical certificate of cause of death.

Organization, W. H. (1992). *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines* (Vol. 1): World Health Organization.

Ott, R. L., & Longnecker, M. T. (2015). *An introduction to statistical methods and data analysis*: Nelson Education.

Paice, C. D. (1994). *An evaluation method for stemming algorithms.* Paper presented at the Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval.

Papadakis, G., Giannakopoulos, G., & Paliouras, G. (2016). Graph vs. bag representation models for the topic classification of web documents. *World Wide Web-Internet and Web Information Systems, 19*(5), 887-920.

Parlak, B., & Uysal, A. K. (2015). Classification of Medical Documents According to Diseases *2015 23rd Signal Processing and Communications Applications Conference* (pp. 1635-1638). New York: Ieee.

Parlak, B., & Uysal, A. K. (2016a). The impact of feature selection on medical document classification. In A. Rocha, L. P. Reis, M. P. Cota, O. S. Suarez & R. Goncalves (Eds.), *2016 11th Iberian Conference on Information Systems and Technologies*. New York: Ieee.

Parlak, B., & Uysal, A. K. (2016b). *The impact of feature selection on medical document classification.* Paper presented at the Iberian Conference on Information Systems and Technologies, CISTI.

Parlak, B., & Uysal, A. K. (2018) On feature weighting and selection for medical document classification. *Vol. 718. Studies in Computational Intelligence* (pp. 269-282).

Passalis, N., & Tefas, A. (2016). Entropy Optimized Feature-Based Bag-of-Words Representation for Information Retrieval. *Ieee Transactions on Knowledge and Data Engineering, 28*(7), 1664-1677.

Pineda, A. L., Tsui, F.-C., Visweswaran, S., & Cooper, G. F. (2013). Detection of patients with influenza syndrome using machine-learning models learned from emergency department reports. *Online journal of public health informatics, 5*(1).

Pineda, A. L., Ye, Y., Visweswaran, S., Cooper, G. F., Wagner, M. M., & Tsui, F. (2015). Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. *Journal of Biomedical Informatics, 58*, 60-69.

Plisson, J., Lavrac, N., & Mladenić, D. (2004). A rule based approach to word lemmatization.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130-137.

Porumb, M., Barbantan, I., Lemnaru, C., & Potolea, R. (2015). *REMed - Automatic relation extraction from medical documents*.

Provost, F. J., & Fawcett, T. (1997). *Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions.* Paper presented at the KDD.

Provost, F. J., Fawcett, T., & Kohavi, R. (1998). *The case against accuracy estimation for comparing induction algorithms.* Paper presented at the ICML.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*(1), 81-106.

Ramos, J. (2003). *Using tf-idf to determine word relevance in document queries.* Paper presented at the Proceedings of the first instructional conference on machine learning.

Rani, G. J. J., Gladis, D., & Mammen, J. (2015). Classification and Prediction of Breast Cancer Data derived Using Natural Language Processing. *Proceeding of the Third International Symposium on Women in Computing and Informatics (Wci-2015)*, 250-255.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation *Encyclopedia of database systems* (pp. 532-538): Springer.

Renganathan, V. (2017). Text mining in biomedical domain with emphasis on document clustering. *Healthc Inform Res, 23*(3), 141-146.

Rios, A., & Kavuluru, R. (2015). *Convolutional neural networks for biomedical text classification: Application in indexing biomedical articles.* Paper presented at the BCB 2015 - 6th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics.

Safavian, S. R., & Landgrebe, D. (1990). A survey of decision tree classifier methodology.

Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). *A Bayesian approach to filtering junk e-mail.* Paper presented at the Learning for Text Categorization: Papers from the 1998 workshop.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management, 24*(5), 513-523.

Saqlain, M., Hussain, W., Saqib, N. A., & Khan, M. A. (2016). *Identification of Heart Failure by Using Unstructured Data of Cardiac Patients.* Paper presented at the Proceedings of the International Conference on Parallel Processing Workshops.

Saripalle, R. (2010). Current status of ontologies in Biomedical and Clinical informatics. *International Journal of Science and Information*.

Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics, 53*, 196-207.

Schuemie, M. J., Sen, E., t Jong, G. W., van Soest, E. M., Sturkenboom, M. C., & Kors, J. A. (2012). Automating classification of free-text electronic health records for epidemiological studies. *Pharmacoepidemiology and Drug Safety, 21*(6), 651-658.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR), 34*(1), 1-47.

Sedghi, E., Weber, J. H., Thomo, A., Bibok, M., & Penn, A. M. (2016). A new approach to distinguish migraine from stroke by mining structured and unstructured clinical data-sources. *Network Modeling Analysis in Health Informatics and Bioinformatics, 5*(1), 30.

Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison, 52*(55-66), 11.

Shin, B., Chokshi, F. H., Lee, T., & Choi, J. D. (2017). *Classification of radiology reports using neural attention models.* Paper presented at the Neural Networks (IJCNN), 2017 International Joint Conference on.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437.

Spasić, I., Livsey, J., Keane, J. A., & Nenadić, G. (2014). Text mining of cancer-related information: review of current status and future directions. *International Journal of Medical Informatics, 83*(9), 605-623.

Stearns, M. Q., Price, C., Spackman, K. A., & Wang, A. Y. (2001). *SNOMED clinical terms: overview of the development process and project status.* Paper presented at the Proceedings of the AMIA Symposium.

Sundararajan, V., Henderson, T., Perry, C., Muggivan, A., Quan, H., & Ghali, W. A. (2004). New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *Journal of clinical epidemiology, 57*(12), 1288-1294.

Tang, L., & Liu, H. (2005). *Bias analysis in text classification for highly skewed data.* Paper presented at the Data Mining, Fifth IEEE International Conference on.

Toman, M., Tesar, R., & Jezek, K. (2006). Influence of word normalization on text classification. *Proceedings of InSciT, 4*, 354-358.

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research, 2*(Nov), 45-66.

Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems, 36*, 226-235.

Vangay, P., Steingrimsson, J., Wiedmann, M., & Stasiewicz, M. J. (2014). Classification of Listeria monocytogenes Persistence in Retail Delicatessen Environments Using Expert Elicitation and Machine Learning. *Risk Analysis, 34*(10), 1830-1845.

Vij, K. (2014). *Textbook of Forensic Medicine & Toxicology: Principles & Practice-e-book*: Elsevier Health Sciences.

Wagholikar, A., Zuccon, G., Nguyen, A., Chu, K., Martin, S., Lai, K., et al. (2013). Automated classification of limb fractures from free-text radiology reports using a clinician-informed gazetteer methodology. *Australas Med J, 6*(5), 301-307.

Wang. (2014). Introduction to Word2vec and its application to find predominant word senses. *URL: http://compling. hss. ntu. edu. sg/courses/hg7017/pdf/word2vec% 20and% 20its% 20appli cation% 20to% 20wsd. pdf.*

Wang, Coiera, E., Runciman, W., & Magrabi, F. (2017). Using multiclass classification to automate the identification of patient safety incident reports by type and severity. *Bmc Medical Informatics and Decision Making, 17*(1), 84.

Wang, Y., Coiera, E., Runciman, W., & Magrabi, F. (2017). Using multiclass classification to automate the identification of patient safety incident reports by type and severity. *Bmc Medical Informatics and Decision Making, 17*(1), 84.

Wei, Z., Ju, Z. X., Chun, X., Hua, J., & Jin, P. (2013). *An automatic electronic nursing records analysis system based on the text classification and machine learning.* Paper presented at the Proceedings - 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2013.

Willett, P. (2006). The Porter stemming algorithm: then and now. *Program, 40*(3), 219-223.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*(1-3), 37-52.

Wolpert, D. H., & Macready, W. G. (1995). No free lunch theorems for search: Technical Report SFI-TR-95-02-010, Santa Fe Institute.

Wu, H., & Wang, M. D. (2017). *Infer Cause of Death for Population Health Using Convolutional Neural Network.* Paper presented at the Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics.

Xu, Guo, X., Ye, Y., & Cheng, J. (2012). An improved random forest classifier for text categorization. *Journal of Computers, 7*(12), 2913-2920.

Xu, King, I., Lyu, M. R.-T., & Jin, R. (2010). Discriminative semi-supervised feature selection via manifold regularization. *IEEE Transactions on Neural networks, 21*(7), 1033-1047.

Xu, Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R., & Denny, J. C. (2010). MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association, 17*(1), 19-24.

Yadav, Sarioglu, E., Choi, H. A., Cartwright, W. B. t., Hinds, P. S., & Chamberlain, J. M. (2016). Automated Outcome Classification of Computed Tomography Imaging Reports for Pediatric Traumatic Brain Injury. *Acad Emerg Med, 23*(2), 171-178.

Yadav, Sharan, A., & Joshi, M. L. (2014). *Semantic graph based approach for text mining.* Paper presented at the Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on.

Yadav, K., Sarioglu, E., Choi, H. A., Cartwright, W. B. t., Hinds, P. S., & Chamberlain, J. M. (2016). Automated Outcome Classification of Computed Tomography Imaging Reports for Pediatric Traumatic Brain Injury. *Acad Emerg Med, 23*(2), 171-178.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information retrieval, 1*(1-2), 69-90.

Yang, Y., & Pedersen, J. O. (1997). *A comparative study on feature selection in text categorization.* Paper presented at the Icml.

Ye, Y., Tsui, F., Wagner, M., Espino, J. U., & Li, Q. (2014). Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. *Journal of the American Medical Informatics Association, 21*(5), 815-823.

Yeow, W. L., Mahmud, R., & Raj, R. G. (2014). An application of case-based reasoning with machine learning for forensic autopsy. *Expert Systems with Applications, 41*(7), 3497-3505.

Yoon, H. J., Roberts, L., & Tourassi, G. (2017). Automated Histologic Grading from Free-Text Pathology Reports using Graph-of-Words Features and Machine Learning. *2017 Ieee Embs International Conference on Biomedical & Health Informatics (Bhi)*, 369-372.

Zhang, A., Pueyo, L. G., Wendt, J. B., Najork, M., & Broder, A. (2017). Email Category Prediction.

Zhou, Amundson, P. K., Yu, F., Kessler, M. M., Benzinger, T. L., & Wippold, F. J. (2014). Automated classification of radiology reports to facilitate retrospective study in radiology. *J Digit Imaging, 27*(6), 730-736.

Zhou, Baughman, A. W., Lei, V. J., Lai, K. H., Navathe, A. S., Chang, F., et al. (2015). Identifying Patients with Depression Using Free-text Clinical Documents. *Stud Health Technol Inform, 216*, 629-633.

Zhou, Zhang, Q. R., Wang, H. X., & Zhang, D. (2015). *Feature selection in medical text classification based on differential evolution algorithm.* Paper presented at the Electronics, Information Technology and Intellectualization - International Conference on Electronics, InformationTechnology and Intellectualization, EITI 2014.

Zhou, H. Y., Zhang, Q. R., Wang, H. X., & Zhang, D. (2015). *Feature selection in medical text classification based on differential evolution algorithm.* Paper presented at the Electronics, Information Technology and Intellectualization - International Conference on Electronics, InformationTechnology and Intellectualization, EITI 2014.

Zhou, L., Baughman, A. W., Lei, V. J., Lai, K. H., Navathe, A. S., Chang, F., et al. (2015). Identifying Patients with Depression Using Free-text Clinical Documents. *Stud Health Technol Inform, 216*, 629-633.

Zhou, Y., Amundson, P. K., Yu, F., Kessler, M. M., Benzinger, T. L., & Wippold, F. J. (2014). Automated classification of radiology reports to facilitate retrospective study in radiology. *J Digit Imaging, 27*(6), 730-736.

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning, 3*(1), 1-130.

Zuccon, G., Khanna, S., Nguyen, A., Boyle, J., Hamlet, M., & Cameron, M. (2015). Automatic detection of tweets reporting cases of influenza like illnesses in Australia. *Health information science and systems, 3*(S1), S4.

Zuccon, G., Wagholikar, A. S., Nguyen, A. N., Butt, L., Chu, K., Martin, S., et al. (2013). Automatic Classification of Free-Text Radiology Reports to Identify Limb Fractures using Machine Learning and the SNOMED CT Ontology. *AMIA Jt Summits Transl Sci Proc, 2013*, 300-304.

Zurada, J. M. (1992). *Introduction to artificial neural systems* (Vol. 8): West St. Paul.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Full Article Publication (ISI-Indexed) Articles

1. **Mujtaba, G**., Shuib, L., Raj, R. G., Rajandram, R., & Shaikh, K. (2017). Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study. ***Journal of Forensic and Legal Medicine***. **(Published)**
2. **Mujtaba, G.**, Shuib, L., Raj, R. G., Rajandram, R., & Shaikh, K. (2017). Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. ***PloS one***, *12*(2), e0170242. **(Published)**
3. **Mujtaba, G.**, Shuib, L., Raj, R. G., Rajandram, R., & Shaikh, K. (2017, July). Classification of Forensic Autopsy Reports through Conceptual- Graph-based Feature Engineering Technique. ***Journal of Biomedical Informatics*** **(Recently Submitted the Revisions)**
4. **Mujtaba, G.**, Shuib, L., Raj, R. G., Rajandram, R., & Shaikh, K. (2018). Free-Text Clinical Reports Classification: A Systematic Review. ***Expert Systems with Applications***. **(Under Review)**
5. **Mujtaba, G.**, Shuib, L., Raj, R. G., Majeed, N., & Al-Garadi, M. A. (2017). Email Classification Research Trends: Review and Open Issues. ***IEEE Access***. **(Published)**

Conference paper

1. **Mujtaba, G.**, Shuib, L., Raj, R. G., Rajandram, R., & Shaikh, K. (2016, December). Automatic text classification of ICD-10 related CoD from complex and free text forensic autopsy reports. In ***Machine Learning and Applications***

***(ICMLA)**, 2016 15th IEEE International Conference on* (pp. 1055-1058). IEEE. **(Published and Presented at LoS Angeles, United States of America)**

2. **Mujtaba, G.**, Shuib, L., Raj, R. G., Rajandram, R., & Shaikh, K. (2017, July). Hierarchical Text Classification of Autopsy Reports to Determine MoD and CoD through Term-Based and Concepts-Based Features. In ***Industrial Conference on Data Mining*** (pp. 209-222). Springer, Cham. **(Published and Presented at New York, United States of America)**