UNIVERSITY OF THE WITWATERSRAND

MSC RESEARCH REPORT

Application of text mining techniques to extract leukaemia information from reports of flow cytometry investigations conducted between 2015 and 2019 at a Johannesburg academic hospital

---



Student:
Sibonginkosi MAPOSA

Supervisor(s):
Mr. Michael MAPUNDU
Dr. Naseem CASSIM
Prof. Deborah GLENCROSS

A Research Report submitted to the Faculty of Health Sciences, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Master of Science in Epidemiology (Public Health Informatics)

October 27, 2025

Declaration of Authorship

I, Sibonginkosi MAPOSA, declare that this dissertation titled, "Application of text mining techniques to extract leukaemia information from reports of flow cytometry investigations conducted between 2015 and 2019 at a Johannesburg academic hospital" is my own, unaided work. It is being submitted for the Degree of Master of Science in Epidemiology (Public Health Informatics)at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

Signed:

_____

Date:

_____

University of The Witwatersrand

Abstract

Faculty of Health Sciences

School of Public Health
Division of Epidemiology and Biostatistics

Master of Science in Epidemiology (Public Health Informatics)

Application of text mining techniques to extract leukaemia information from reports of flow cytometry investigations conducted between 2015 and 2019 at a Johannesburg academic hospital

by Sibonginkosi MAPOSA

Leukaemia data for epidemiological and public health planning is limited in South Africa. The application of text mining techniques could improve leukaemia data representation. This study aimed to apply text mining techniques to extract clinical and epidemiological data from narrative laboratory reports of immunophenotyping flow cytometry completed at the Charlotte Maxeke Johannesburg Academic Hospital (CMJAH) laboratory from 2015 to 2019. An expert driven regular expressions (RegEx) based algorithm and topic modelling were used. The RegEx algorithm was used to place the reports into categories of Acute Lymphocytic Leukaemia, Acute Myeloid Leukaemia, Chronic Lymphoid Leukaemia or Chronic Myeloid Leukaemia. The algorithm achieved F1 scores comparable to those of similar tasks in the literature. It achieved 0.74 for leukaemia and 0.93 for non-leukaemia classes respectively. Topic modelling was evaluated using coherence scores, inspection of visualisation of clusters, and interpretability of topics. Topic modelling uncovered more information from the corpus, relative to the regex based algorithm. Age standardised incidence rates for leukaemia in the City of Johannesburg, calculated using extracted data ranged from 0.54 to 0.87 per 100 000 population, which was slightly lower than the figures reported by the National Cancer Registry. This study has shown that text mining techniques can be applied to improve representativity of leukaemia data. It is recommended that further work be carried out to refine both the RegEx algorithm and topic models.

Acknowledgements

Dear Lord Jesus - As I come to the end of this chapter, I cannot say thank you enough. Thank you for the Grace to do this degree, the opportunity to learn and grow.

I am grateful to Africa CDC for the financial support. I am in a new role where I am applying the knowledge from my studies. I feel more equipped and effective because of the learning and skills from the degree.

A special thank you to my Supervisors. I had the most patient, accommodating and conscientious team of Supervisors. I cannot mention all the wonderful things I am grateful for, but herewith a few:

- Mr Mapundu, thank you for encouraging me to keep going when I felt like giving up. Thank you for the meetings,your patience, and guidance helping me to navigate the world of text mining, and scripting, and encouraging me.
- Dr Naseem, you made me believe in myself, that I could do this. Thank you for your kindness, empathy, and encouragement. Thank you for always making time to answer my questions, and reading through my drafts.
- Professor Glencross, thank you for taking time to explain the technicalities of haematology, introducing me to the concepts, sitting with me to give me expert input in evaluating the models. Thank you for patiently working through and labelling the validation dataset, and patiently guiding me to prepare it. Your patience, enthusiasm, encouragement and compassion throughout the project have inspired me to aspire to grow.

I would also like to extend my heartfelt appreciation to staff in the School of Public Health. In particular I would like to thank Professors Levin and Musenge, and Dr Maposa. My love for Biostatistics grew in your classes. Dr Ndlovu, thank you for all the research support seminars, they were so informative and helpful. I enjoyed them and learnt a lot. Ms Inno Vukeya, thank you.

To my babies, Thandeka, Zandile, Manelisi and Andile: @Andile - Mummy has finally finished her project! @Thandeka and @Zandi, my girls, the wind beneath my wings, my biggest cheerleaders - thank you my babies, my friends and sisters, all in one. @Papa (Manelisi) - thank you the encouraging smiles. Last but not definitely not least, my husband, Melusi, thank you for always encouraging me, and sitting up late with me. Thank you for the cakes, and chocolate boxes...

You are all my village!

...

Contents

Declaration of Authorship i

Abstract ii

Acknowledgements iii

List of Figures vii

List of Tables ix

List of Abbreviations x

1 INTRODUCTION 1

1.1 Background . . . 1
1.2 Problem statement . . . 3
1.3 Justification . . . 4
1.4 Research Question . . . 4
1.5 Research Aim . . . 4
1.6 Specific Objectives . . . 4

2 LITERATURE REVIEW 6

2.1 Introduction . . . 6
2.2 Overview of text mining approaches and techniques . . . 6
2.2.1 Expert driven approaches . . . 6
2.2.2 Machine learning approaches . . . 8
2.2.3 Validation and evaluation of text mining algorithms . . . 9
2.3 Text mining in public health, biomedicine and clinical knowledge management . . . 10
2.3.1 Overview of applications of text mining . . . 10
2.3.2 Text mining cancer data . . . 11
2.3.3 Text mining and leukaemia . . . 12
2.4 Electronic medical records: A research and knowledge resource . . . 13
2.4.1 Text mining pathology reports . . . 14
2.5 Application of text mining to leukaemia knowledge in South Africa . . . 14
2.5.1 The epidemiology of leukaemia in South Africa . . . 14
2.5.2 Sources of leukaemia information in South Africa . . . 15

3 METHODOLOGY 16

3.1 Introduction . . . 16

3.2 Study design . . . 16
3.3 Computational environment . . . 16
3.3.1 Study population and sampling . . . 16
3.4 Data source and description . . . 17
3.5 Text mining work-flow, methods and techniques . . . 17
3.5.1 Data acquisition, inclusion and exclusion of reports . . . 17
3.5.2 Text pre-processing . . . 18
3.5.3 Information extraction . . . 19
Extraction of leukaemia information: Expert-driven regular expressions . . . 19
Topic Modelling: Latent Dirichlet Allocation . . . 20
Topic Modelling: BERTopic . . . 21
3.6 Epidemiology of leukaemia: Age Standardised Incidence Rates of cases worked up at CMJAH Laboratory . . . 21
3.7 Evaluation of the text mining . . . 22
3.7.1 Evaluation of the regex based algorithm . . . 22
3.7.2 Evaluation of the topic modelling . . . 22

4 RESULTS 24
4.1 Introduction . . . 24
4.2 Application of text mining techniques to extract leukaemia data . . . 24
4.2.1 Data exploration . . . 24
4.2.2 Effect of cleaning on the corpus . . . 26
Highest frequency words: Word clouds and bar charts . . . 26
Reduction in the size of the feature space . . . 28
N-grams in the cleaned corpus . . . 29
4.3 Extraction of leukaemia information . . . 30
4.3.1 Expert driven approach: Application of regular expressions in a rule based algorithm . . . 30
4.3.2 Application of an unsupervised learning algorithm: Latent Dirichlet Allocation implementation . . . 32
Modelling the corpus of reports with an incident positive haematological malignancy finding . . . 32
Modelling the corpus of reports of all new investigations . . . 34
4.3.3 Application of an unsupervised learning algorithm: BERTopic implementation 36
Modelling the corpus of reports with an incident positive haematological malignancy finding . . . 36
Modelling the corpus of reports of all new investigations . . . 37
4.3.4 Age specific incidence rates of the main leukaemia types in reports from CMJAH . . . 38
4.4 Evaluation of the performance of the text mining . . . 39
Performance of the regex based algorithm . . . 39
Performance of the topic models . . . 41

5 Discussion 44

5.1 Introduction 44
5.2 Applicability of text mining techniques to flow cytometry narrative reports 44
5.2.1 Outcomes and performance of text mining techniques 44
5.2.2 Data pre-processing and feature engineering 46
5.2.3 Development of accurate regular expressions algorithm and topic models . . 47
5.3 Age Standardised Incidence Rates of Leukaemia in Johannesburg from 2015 to 2019 47
5.4 Limitations of the study 48

6 Conclusions and Recommendations **49**
6.1 Introduction 49
6.2 Text mining techniques can improve data representivity for leukaemia 49
6.3 Recommendations for further work towards routine implementation of smart tools . . 50

Bibliography **51**

A Appendix 1 **61**

B Appendix 2 **63**

C Appendix 3 **65**
C.1 Top words in topics from the corpus of reports with haematological malignancy findings 65

D Appendix 4 **71**
D.1 Top words in topics from the corpus of all reports 71

E Appendix 5 **77**

List of Figures

1.1 Celllines . . . 2

3.1 Mining Workflow . . . 18

4.1 Age distribution of records with missing patient gender . . . 25
4.2 Distribution of patients' ages . . . 25
4.3 Age distribution of patients' ages by sex . . . 26
4.4 Word Cloud of the full corpus of reports as received . . . 26
4.5 Word Clouds of the bone marrow corpus before and after cleaning . . . 27
4.6 Top twenty words in the bone marrow corpus before and after cleaning . . . 27
4.7 Representation of the corpus of incident haematological malignancy reports . . . 28
4.8 Representation of text species in corpus of incident haematological malignancy reports before and after cleaning . . . 28
4.9 A comparison of a report before and after cleaning showing the effect of pre-processing 29
4.10 Top n-grams in the pre-processed corpus . . . 30
4.11 An illustration of the steps in the regular expressions rule based algorithm . . . 31
4.12 Features with the top twenty tf-idf scores in the corpus of reports with an incident positive haematological malignancy finding . . . 32
4.13 Plot of coherence scores against number of topics from the corpus of reports with an incident positive haematological malignancy finding . . . 33
4.14 Visualisation of the ten topics in the chosen model of the corpus of reports with an incident positive haematological malignancy finding . . . 34
4.15 Features with the top twenty tf-idf scores in the corpus of reports of all new investigations . . . 34
4.16 Plot of coherence scores against number of topics for the corpus of reports of all new investigations . . . 35
4.17 Visualisation of the 12 topics from chosen LDA model of the corpus of reports of all new investigations . . . 35
4.18 Barcharts of top words in topics from chosen model of the corpus of reports with an incident positive haematological malignancy finding . . . 36
4.19 Intertopic distance map of topics from the chosen model of the corpus of reports with an incident positive haematological malignancy finding . . . 36
4.20 Barcharts of top words in topics from BERTopic model of the corpus of reports of all new investigations . . . 37
4.21 Intertopic distance map of topics from the corpus of reports of all new investigations 38
4.22 The leukaemia ASIR from flow cytometry data from CMJAH from 2015 to 2019 . . . 39
4.23 The confusion matrix for the binary task at evaluation with a labelled dataset . . . 40

4.24 The confusion matrix for the different leukaemia classes at evaluation with a labelled dataset 41

C.1 Top words in Topic 1 65
C.2 Top words in Topic 2 66
C.3 Top words in Topic 3 66
C.4 Top words in Topic 4 67
C.5 Top words in Topic 5 67
C.6 Top words in Topic 6 68
C.7 Top words in Topic 7 68
C.8 Top words in Topic 8 69
C.9 Top words in Topic 9 69
C.10 Top words in Topic 10 70

D.1 Top words in Topic 1 71
D.2 Top words in Topic 2 72
D.3 Top words in Topic 3 72
D.4 Top words in Topic 4 72
D.5 Top words in Topic 5 73
D.6 Top words in Topic 6 73
D.7 Top words in Topic 7 74
D.8 Top words in Topic 8 74
D.9 Top words in Topic 9 75
D.10 Top words in Topic 10 75
D.11 Top words in Topic 11 76
D.12 Top words in Topic 12 76

List of Tables

3.1 Expert terms and regular expressions for leukaemia categories 20
3.2 Other terms and their regular expressions 20
3.3 Schematic representation of the confusion matrix 22

4.1 Number of reports by province 24
4.2 Descriptions of the corpus and numbers of reports at different stages of the algorithm 32
4.3 Observed and Modelled Age Standardised Incidence Rates for 2015 to 2019 38
4.4 Performance matrix of the expert driven algorithm on the binary task 39
4.5 Performance matrix of the expert driven algorithm on the multi class task 40
4.6 Comparison of coherence scores of the different topic models 41
4.7 Topics identified by the domain expert from models of the corpus of reports with an incident positive haematological finding 42
4.8 Topics identified by the domain expert from models of the corpus of reports of all new investigations 43

5.1 ASIR reported by the NCR for histologically diagnosed leukaemia in 2015 to 2019 . 48

List of Abbreviations

AARQA	**A**cademic **A**ffairs **R**esearch and **Q**uality **A**ssurance
ACRN	**A**frican **C**ancer **R**egistry **N**etwork
AI	**A**rtificial **I**ntelligence
ALL	**A**cute **L**ymphoid **L**eukaemia
AML	**A**cute **M**yeloid **L**eukaemia
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
CANSA	**C**ancer **A**ssociation of **S**outh **A**frica
CDW	**C**orporate **D**ata **W**arehouse
CLL	**C**hronic **L**ymphoid **L**eukaemia
CMJAH	**C**harlotte **M**axeke **J**ohannesburg **A**cademic **H**ospital
CML	**C**hronic **M**yeloid **L**eukaemia
COVID-19	**C**oronavirus **D**isease 20**19**
CPU	**C**entral **P**rocessing **U**nit
cTAKES	**c**linical **T**ext **A**nalysis and **K**nowledge **E**xtraction **S**ystem
EMR	**E**lectronic **M**edical **R**ecords
HIV	**H**uman **I**mmunodeficiency **V**irus
HSC	**H**aemopoietic **S**tem **C**ell
IARC	**I**nternational **A**gency for **R**esearch on **C**ancer
IBM	**I**nternational **B**usiness **M**achines
ICD-O-3	**1**nternational **C**lassification of **D**iseases for **O**ncology **3**rd Edition
IDE	**I**ntegrated **D**evelopment **E**nvironment
IE	**I**nformation **E**xtraction
LDA	**L**atent **D**irichlet **A**llocation
LIS	**L**aboratory **I**nformation **S**ystem
LSA	**L**atent **S**emantic **A**nalysis
ML	**M**achine **L**earning
MVP	**M**inimally **V**iable **P**roduct
NCR	**N**ational **C**ancer **R**egistry
NER	**N**amed **E**ntity **R**ecognition
NHLS	**N**ational **H**ealth **L**aboratory **S**ervice
NICD	**N**ational **I**nstitute for **C**ommunicable **D**iseases
NLP	**N**atural **L**anguage **P**rocessing
NLTK	**N**atural **L**anguage **T**ool **K**it
NMF	**N**on **N**egative **M**atrix **F**actorisation
NPV	**N**egative **P**redictive **V**alue
PPV	**P**ositive **P**redictive **V**alue
RED	**R**egular **E**xpressions **D**iscovery
RegEx	**R**egular **E**xpressions
SAS	**S**tatistical **A**nalysis **S**oftware
SVM	**S**upport **V**ector **M**achine
TF-IDF	**T**erm **F**requency **I**nverse **D**ocument **F**requency
WHO	**W**orld **H**ealth **O**rganisation
UK	**U**nited **K**ingdom

INTRODUCTION

1.1 Background

Leukaemia is a blood cancer in which there is abnormal accumulation of lymphoid or myeloid cells in the bone marrow and or peripheral blood. Haemopoiesis is the process of production and maintenance of blood components. The haemopoietic system is a network of organs, tissues and cells that include the bone marrow, spleen and lymphoid tissues, regulated by hormones such as erythropoietin and interleukin-3. Haematopoiesis in the bone marrow begins at five months gestation of a human embryo and continues throughout adulthood. Prior to five months gestation, haemopoiesis takes place in the developing liver and spleen, and their precursor cells in the yolk sac of the embryo (Hoffbrand, A.V. and Moss, 2011).

In the bone marrow, haemopoietic stem cells (HSCs) divide and differentiate into the different blood cells. Haemopoiesis begins with a pluripotential stem cell that can self-renew, and also give rise to new cell lines. The ability to self renew allows marrow cellularity to be maintained in a healthy steady state (Hoffbrand, A.V. and Moss, 2011). The pluripotential stem cell can differentiate into a common myeloid or lymphoid progenitor cell. Further cell lines differentiate from these progenitor cells. The common myeloid cells can differentiate into red cells, platelets, monocytes, neutrophils, eosinophils and basophils (Hoffbrand, A.V. and Moss, 2011). Common lymphoid cells on the other hand can become B or T lymphocytes or natural killer cells (Hoffbrand, A.V. and Moss, 2011). As the cells differentiate, they lose the ability to self renew (Hoffbrand, A.V. and Moss, 2011). Figure 1.1 provides an illustration of the pluripotent stem cell and lines that arise from it.

Leukaemia can be classified as acute or chronic, reflecting the rate at which disease develops. In acute leukaemias, symptoms develop suddenly with rapid disease progression. There is accumulation of blasts, which are immature, undifferentiated or minimally differentiated cells (Zuckerman and Rowe, 2015). This is associated with secondary suppression of normal development of cells from haematopoietic stem cells in the bone marrow (Zuckerman and Rowe, 2015). In chronic leukaemia, mature cells accumulate in the blood or bone marrow. At least initially, there is no suppression of haemopoiesis and disease progresses slowly over time. The acute and chronic diseases can be further classified by the lineage of cells affected, resulting in four main groups: Acute Myeloid Leukaemia (AML), Chronic Myeloid Luekaemia (CML), Acute Lymphoid Leukaemia (ALL) and Chronic Lymphoid Leukaemia (CLL) (Zuckerman and Rowe, 2015). The World Health Organisation (WHO) provides a comprehensive classification of the types and sub-types of leukaemia (Arber et al., 2016). The WHO classifications are continually reviewed and updated to reflect increasing knowledge and understanding about the diseases by researchers and clinicians (Li, 2022b). These classification efforts have been reported to put the leukaemias into different prognostic groups, and for chronic forms

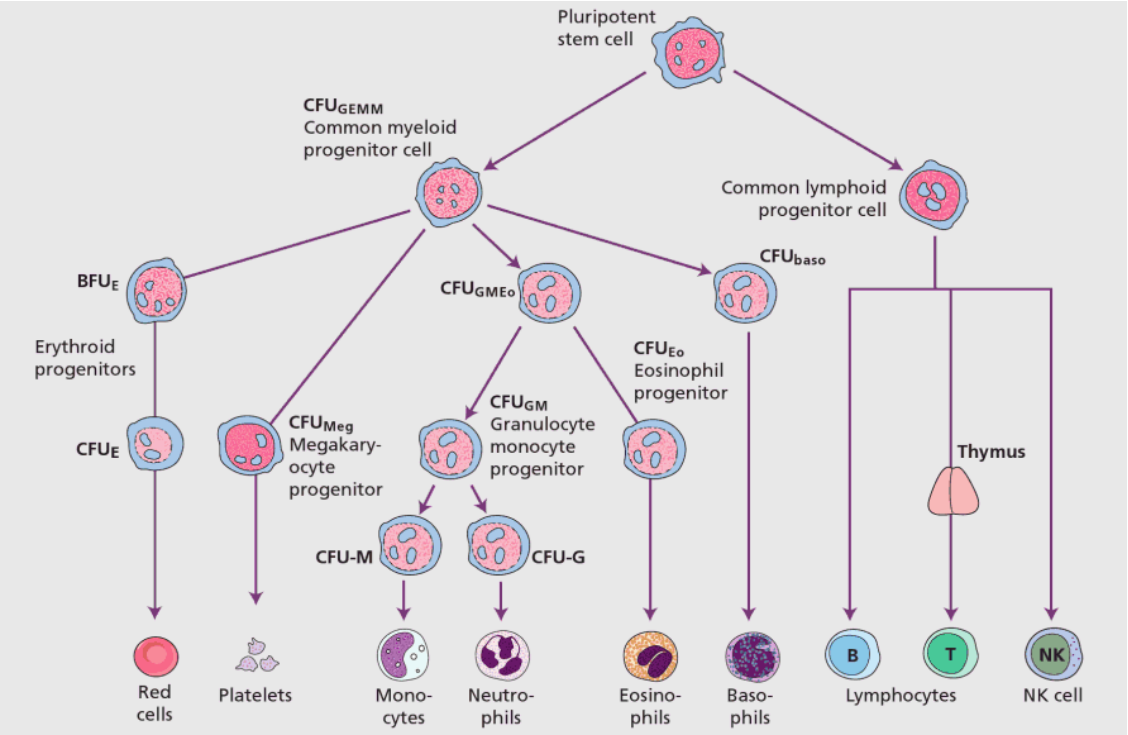


FIGURE 1.1: Diagrammatic representation of the bone marrow pluripotent stem cell and lines arising from it
Source: Hoffbrand, A.V. and Moss, 2011

of leukaemia, led to improvements in treatment and outcomes (Zuckerman and Rowe, 2015).

The International Agency for Research on Cancer (IARC) ranked leukaemia as the 15th most frequent cancer amongst South Africans with 1 922 projected cases for 2020 (IARC, 2020). The IARC estimates incident cases using data from the South African National Cancer Registry (NCR). Incidence is modelled using mortality to incidence ratios from cancer registry data of neighbouring countries (IARC, 2020). A local study analysed data from 2000 to 2006, for individuals fifteen years and older, and reported that blood cancers accounted for six percent of cancers. The blood cancers appeared to be under-ascertained mainly due to the reporting system (Schonfeld et al., 2016). Overall, leukaemia was the second most commonly reported haematological malignancy (Schonfeld et al., 2016). Leukaemia occurs more frequently in children, accounting for thirty percent of childhood cancers (Madhusoodhan, Carroll, and Bhatla, 2016).

The diagnosis and treatment of leukaemia relies on laboratory investigations. Diagnosis makes use of disease induced morphological and immunophenotypic changes in white blood cells (Musaigwa et al., 2021; Hallek and Al-Sawaf, 2021). The changes enable differentiation between types and sub-types of leukaemia. The laboratory investigations include morphologic, immunologic, cytogenetic, and molecular tests (CANSA, 2021). Morphologic analysis of bone marrow smears, though tedious and labour intensive requiring highly qualified pathologists, is still an important initial step for leukaemia diagnosis because it is economical and relatively convenient (Zuckerman and Rowe, 2015). Findings from the investigations are interpreted by pathologists who compile a text patient report.

This study used reports of flow cytometry immunophenotyping investigations from Charlotte Maxeke

Johannesburg Academic Hospital (CMJAH) laboratory for the period 2015 to 2019. CMJAH laboratory is one of eight national central facilities in a network of 265 National Health Laboratory Service (NHLS) laboratories (Stevens et al., 2016). Laboratories are networked into a Laboratory Information System (LIS). Text reports, housed in the centralised LIS, are transferred to, and stored in, the Corporate Data Warehouse (CDW) (NHLS, 2019a). This project investigated the application of text mining techniques for the timely extraction of leukaemia information from the text reports. Information was extracted as structured data and used to study the epidemiology of the disease.

1.2 Problem statement

There is limited clinical and epidemiological data on leukaemia in South Africa. The scarcity of data imposes limitations on understanding of the disease burden and associated epidemiology of leukaemia. The impact ranges all the way from detection and mitigation of risk factors to the planning and distribution of health services including skills required to ensure equitable access to care (NHLS, 2019a);(D.K Glencross 2021, personal communication, 14th January).

The NCR is the main source of cancer surveillance data in South Africa. There are a number of challenges with the current process of tracking the leukaemia disease burden, and reporting by the NCR as outlined below.

1. The NCR only reports diagnoses recorded on anatomical pathology (histology) reports, and due to this, a significant amount of information is lost during the transfer of patient data from the clinical diagnostic reports. Historically, data has been taken from narrative reports on the LIS. This is based on the outdated premise that only anatomical pathology patient reports contain diagnostic information (ACRN, 2021), D.K Glencross 2021, personal communication, 14th January). Consequently, data from other diagnostic platforms like flow cytometry and cytogenetics do not get captured. The modern methods also yield substantively more detail on the diagnosis. The more sophisticated molecular diagnostics, in some instances facilitate diagnosis without the patient undergoing biopsy for microscopic review. The diagnoses from such instances are effectively excluded in NCR documentation if no trephine biopsy sample is performed and submitted for review (D.K Glencross 2021, personal communication, 14th January).

2. Clinical information captured in narrative histology reports is lost as only ICD-0-3 codes are reported by the NCR. A large amount of information giving the detailed picture of disease associated changes in cells is left in the text reports. The text data remains unavailable to researchers and health planners. This is a problem given the limited data on the burden and epidemiology of leukaemias (Schonfeld et al., 2016).

3. All leukaemia subtypes, with different clinical and prognostic outcomes are lumped into a single category of ‘leukaemia’ as seen in the published reports (National Health Laboratory Service: National Cancer Registry, 2023). Thus, more diagnostic information is lost. The lack of detailed reporting of leukaemia types and subtypes means that epidemiological trends for different leukaemias over time cannot be examined (NICD, 2020). In the analysis of the 2000 to 2006 NCR data, the unavailability of data on subtypes of haematologic malignancies limited understanding of epidemiology (Schonfeld et al., 2016).

4. The tedious and time consuming nature of uploading the information leads to delays in the update of incidence reports on the registry. Trained coders read the entire report, and upload the ICD-O-3 code to register the cancer. This is a challenge that the NCR has faced over the years. At one point, the backlog was ten years for incidence reports (Singh et al., 2016). As of 2020, the backlog was three years with the 2019 data being the latest that has been published (NICD, 2020).

1.3 Justification

The narrative phenotyping text reports have large amounts of leukaemia information. This project focused on narrative reports of immunophenotyping flow cytometry performed on bone marrow aspirate samples. Flow cytometry is a modern technique that can provide rapid multi-parameter analysis of single cells or particles (Li, 2022a). The technique enables rapid and accurate phenotypic characterisation of cell populations facilitating diagnosis or evaluation of disease progression. In the clinical setting, the quick turnaround time of flow cytometry lends itself to timely diagnosis or selection of relevant tests (Li, 2022a).

It is time-consuming to manually extract and analyse data from diagnostic text reports to obtain leukaemia diagnoses and other information. There is need to apply smart tools and techniques to extract data from text for meaningful outcomes such as epidemiological studies. The availability of information has been shown to reduce the morbidity and mortality from leukaemia (Ning et al., 2020). Structured data, at a fine granular level, for epidemiological analysis and other applications can be of immense value to the management of leukaemia. The information on the disease burden can be used to formulate disease control strategies, and provides a basis for further study of underlying determinants (Schonfeld et al., 2016). In further applications, an accurate text mining algorithm could facilitate timely update of the cancer registry as the process can be automated.

1.4 Research Question

How can the application of text mining techniques improve knowledge representation of leukaemia data and incidence reporting through the accurate extraction of data from immunophenotyping flow cytometry narrative reports?

1.5 Research Aim

The aim of this project was to apply text mining techniques to extract diagnostic data from narrative laboratory reports of immunophenotyping flow cytometry completed at the Charlotte Maxeke Johannesburg Academic Hospital (CMJAH) laboratory from 2010 to 2019.

1.6 Specific Objectives

1. To apply text mining techniques to extract leukaemia data from immunophenotyping flow cytometry narrative reports.
2. To ascertain age-standardised incidence rates of leukaemia.

3. To evaluate and validate the performance of the text mining algorithm.

LITERATURE REVIEW

2.1 Introduction

The rapid accumulation of knowledge in textual format has driven the application of text mining (TM)to keep track of, extract, and discover knowledge from semi-structured and unstructured documents. TM is the extraction of 'previously unknown, understandable, potential and practical patterns or knowledge' from unstructured documents (Zhang, Chen, and Liu, 2015). TM is a form of Natural Language Processing (NLP), a branch of artificial intelligence (AI) in which computers are programmed to understand, generate and process both written and spoken human languages (Moreno and Redondo, 2016). Advances in text mining make the semi-structured and unstructured data computationally accessible (Simmons, Singhal, and Lu, 2016) and can provide timely access to explicit and implicit associations between information in text (Renganathan, 2017).

This chapter is an overview of the knowledge landscape on mining of biomedical text, with a focus on cancer and particularly leukaemia. The chapter begins with an outline of TM approaches and techniques. This is followed by a brief summary of the various applications of TM. The opportunities and challenges presented by electronic health records as a resource for health research, and implications for TM are explored. Lastly, the case for application of TM to semi- and unstructured reports containing leukaemia information, to improve data representation of the disease is made. This is based on an analysis of limited data on leukaemia disease burden and epidemiology, existing pathology work-up algorithm, data representation and information management in South Africa.

2.2 Overview of text mining approaches and techniques

TM has been developing over the last few decades. A 1958 study (Luhn, 1958) reporting on a business intelligence system that automatically extracted, coded and classified documents is viewed as the prototype of TM (Zhang, Chen, and Liu, 2015). As far back as 2015, TM was a maturing discipline, moving from an experimental to practical or application stage as evidenced by the number of successful systems. These included IBM's Intelligent Miner for Text, SAS Text Miner and VisualText (Zhang, Chen, and Liu, 2015). The field has since grown with a wide range of processes and software becoming available in various disciplines, including biomedicine (Renganathan, 2017).

2.2.1 Expert driven approaches

In AI, computer systems are developed to mimic the ability of human intelligence. The development or training of these systems requires human expert knowledge and input. AI has found application in expert systems, NLP, speech recognition and others. The role of experts is critical in establishing a

basis, and using existing knowledge in the domain. The AI system does not replace but complements human involvement or facilitates time saving by speeding processes up. In the NLP domain, expert knowledge can be applied to amongst other things, search for entities of interest or extract information from unstructured text resources. The importance of this domain knowledge for the success of TM has been demonstrated (Zweigenbaum et al., 2007; Song-Quan Ong, Maisarah Binti Mohamed Pauzi, 2022). In biomedical applications, information extraction tasks include Named Entity Recognition (NER), event identification, and relation extraction.

NER is a process that assigns a name tag to a recognised entity using rules and heuristics (Song-Quan Ong, Maisarah Binti Mohamed Pauzi, 2022). Assignment of the label follows entity boundary identification and class determination. Pre-trained libraries such as Stanford NER, spaCy and Natural Language Tool Kit (NLTK) NE_Chunk, are available for general NER of entities such as names, locations and organisations. Recognition of domain specific entities requires extensive labelling to enable NER processes to recognise genes and diseases, for example. Application of NER in biomedicine faces challenges arising from typing and grammatical errors, multiple forms of medical terms such as epilepsy and atrophy, which refer to the same disease, and ambiguity in abbreviations amongst other issues (Song-Quan Ong, Maisarah Binti Mohamed Pauzi, 2022).

Entity recognition contrasts with, and can be applied in Information Extraction (IE). In NER, any entity can appear as text and a report can contain several or no entities of a specific class, but the entities are always explicitly mentioned. In contrast, information extraction extracts information of interest, in a structured form, from unstructured data (Song-Quan Ong, Maisarah Binti Mohamed Pauzi, 2022). The goal of IE is to complete a list of information needs with correct values. It is known beforehand the information sought in the text (Weegar, Nygård, and Dalianis, 2017).

NLP methods can be divided into rule, dictionary and machine learning approaches (Sun et al., 2018). In rule based approaches, specifications for the features or entities of interest need to be defined. The rule based tools are written by experts describing how the machine should classify the text reports (Pironet et al., 2021). Rule based methods can be time consuming as there are usually many ways to express the information of interest (Pironet et al., 2021). On the other hand, rules make the approaches transparent as it is clear how values were arrived at. The rule based approaches work well when there is some structure in the text. Rule, and dictionary-based approaches require domain expertise to manually construct the rule templates and dictionaries. The dictionary approaches work well for large scale annotation and indexing, with fuzzy matching lending itself well to handling large numbers of variants in medical terminology. The ProMiner system (Hanisch et al., 2005), follows a rule based approach to identify potential name occurrences and link back to protein and gene databases. The clinical Text Analysis and Knowledge Extraction System (cTAKES) combines rule-based and machine learning techniques (Sun et al., 2018). cTAKES is an open-source NLP system specifically built for the clinical domain to create rich linguistic and semantic annotations (Savova et al., 2010). It was initially developed by a team of physicians, computer scientists and software engineers at Mayo Clinic, and development is continuing (https://ctakes.apache.org/history.htmlpublications-and-acknowledgments, 2023).

One of the ways that variations in writing have been addressed in clinical literature is the use of regular expressions (RegEx) (Bui and Zeng-Treitler, 2014). RegEx are typically developed manually by human experts (Bui and Zeng-Treitler, 2014). The expression is a sequence of special characters that

act as a shorthand notation. The generation or testing of RegEx is not standardised, and this makes their maintenance and extension challenging (Bui and Zeng-Treitler, 2014). Algorithms have been developed to automate the discovery of expressions. Two contrasting approaches have been taken. One is top-down, relying on human experts to provide seed expressions. The algorithm is fine tuned to achieve the desired evaluation metrics. In contrast, the bottom up approach generalises similarities or patterns in textual data to build RegEx (Bui and Zeng-Treitler, 2014). Whilst the top down approach has the disadvantage of relying on manual generation of seed expressions, the up down algorithms tend to be task constrained (Bui and Zeng-Treitler, 2014). It has been shown that combining automatically generated RegEx, discovered using a general purpose algorithm Regular Expressions Discovery (RED), with a Support Vector Machine (SVM) classifier improved the performance of the SVM. The RED algorithm was designed to be generalisable compared to other bottom up algorithms.

RegEx are useful because of their expressiveness and flexibility, with a compact, tolerant syntax. Research has shown that the expressions are difficult to compose, hard to understand (Michael et al., 2019) and error-prone (Spishak, Dietl, and Ernst, 2012). The lack of full compatibility between RegEx matching engines adds another layer of complexity (Zheng et al., 2021). It is important that the expressions are correct as errors in the sequences can cause failure of applications built on them. Expressions need to be syntactically, and semantically correct. Semantic correctness means the expression matches exactly the strings it is intended to match (Zheng et al., 2021).

2.2.2 Machine learning approaches

In machine learning (ML) approaches, a learning algorithm is used to perform entity recognition using statistical characteristics and parameters of sample data (Sun et al., 2018). The ML approaches of NER can be subdivided into conventional and deep learning (Pironet et al., 2021). The ML approach is data-driven and application-oriented. ML based NER methods have shown good performance on clinical text with algorithms and features being key considerations (Sun et al., 2018).

Topic modelling is an unsupervised ML technique that has gained popularity, lending itself well to the handling of large amounts of unstructured text data. Topic modelling is typically applied to manage large collections of text or documents to understand underlying themes, or get an overview of information in the collection, summarised as themes or topics. Algorithms look for patterns in the corpus to identify themes in the data. The input is unlabelled data, from which categories are generated by the algorithm. The underlying premise is that if a document is on a given topic, certain words will occur frequently (Cho, 2019). Topics from the model are thus groups of words, associated with the theme. A document can have multiple topics, in varying proportions (Cho, 2019).

Topic modelling has been applied for text summarisation, recommendation systems and spam filters amongst others. The applications of AI in medicine and health care have been uncovered and summarised by using topic modelling to analyse 27,451 eligible articles from the Web of Science database, published between 1977 and 2018 (Tran et al., 2019). The study investigated productivity, workflow, topics, and research landscape, identifying three leading platforms where AI was being applied (Tran et al., 2019). Specifically TM has been used to map the evolution of topics on women's health in news articles (Cho, Kim, and Woo, 2019), uncover research themes and trends in emergency medicine (Porturas and Taylor, 2021), gynaecological oncology (Grubbs et al., 2023), and paediatrics (Levy_Mendelovich et al., 2021). Topic modelling applied to initial consultation clinical notes and

genetic data revealed a number of genotype-phenotype relationships that have been independently identified in other work (Chan et al., 2013). With the prolific use of social media, topic models have been a useful tool for the analysis of social media data for health research, and efforts are ongoing to improve these methodologies (Narasimhulu et al., 2022). Topic models have been used in prediction tasks, and examples include predicting in-patient violence (Rijcken et al., 2022), early readmission in psychiatry (Rumshisky et al., 2016), and mortality in patients with dementia (Wang et al., 2019).

There are a number of algorithms available for topic modelling including Latent Dirichlet Allocation (LDA), Latent Sentiment Analysis (LSA), Non-Negative Matrix Factorisation (NMF) and others. LDA is the more popular and widely used of the topic modelling algorithms. It is a generative probabilistic three level Bayesian model that assumes each document is a combination of topics and every word is associated with one of the topics (Blei, Ng, and Jordan, 2003). LDA will model each document as a topic distribution, topic as a distribution over words, and words over topics. Bidirectional Encoder Representations from Transformers (BERT) approaches aim to overcome the loss of semantic meaning that occurs in approaches such as LDA (Grootendorst, 2022). The topic model algorithm BERTopic, thus begins by capturing document level information using pre-trained language models. This is followed by dimensionality reduction, clustering, and extraction of topic representations using a custom class-based variation of TF-IDF (Grootendorst, 2022).

2.2.3 Validation and evaluation of text mining algorithms

The value of a TM algorithm lies in its ability to perform the required task to a satisfactory level of accuracy. The evaluation metrics Sensitivity, Specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are traditionally used to evaluate screening and diagnostic tests. They have also been used to compare automated TM with manual searches to validate and evaluate the performance of the automated approaches (Jones-diette et al., 2019).

The metrics indicate the performance of the test in identifying cases and non-cases. Sensitivity is a measure of the ability of a test to correctly identify true cases from a population. It is thus the proportion of cases that are correctly identified by the test, the true positives. Specificity on the other hand gives the proportion of negative or non-diseased that are correctly identified. There is often a trade-off between sensitivity and specificity that needs to be taken into consideration based on the purpose of the evaluation (Trevethan, 2017). Predictive values are confidence measures in the outcome of a test or model. These predictive values are affected by the prevalence of a condition (Wong and Lim, 2011). A study showed that low prevalence resulted in good negative predictive values possibly adversely impacting the positive predictive values (Jones-diette et al., 2019).

In automated processes, the metrics are translated to evaluate the performance of a system, algorithm or application. Recall is thus the proportion of cases that the model classifies correctly, that is true positives as a proportion of true positives and false negatives (Huilgol, 2019). Precision is the proportion of true cases (true positives) amongst those that the model has classified as positive (True positives and false positives) (Huilgol, 2019). Precision and recall are combined in the F1-score to give an overall measure of model performance. The score is the harmonic mean of precision and recall and a more useful measure when datasets are imbalanced compared to accuracy (Huilgol, 2019).

The evaluation of unsupervised techniques is challenging due to the absence of a ground truth, subjectivity and lack of consistent alignment between evaluation metrics and human interpretability of

the outcomes (Meaney et al., 2023). As an illustration of the discordance between evaluation metrics and human interpretation, no correlation between interpretability and predictive performance has been found with psychiatry clinical notes (Rijcken et al., 2022). Human interpretability is a useful feature of topics models. In healthcare, interpretability of topic models is important for implementation and utilisation (Rijcken et al., 2022). Interpretability is critical when topics are used as features in predictive tasks for text classification (Rijcken et al., 2022).

Evaluation is therefore often one, or a combination of, a metric denoting distinctiveness of a cluster and human judgement. In topic modelling, the coherence score of a model, a measure of semantic similarity between the words in a theme, is a commonly used metric for evaluation. The coherence score helps to identify interpretable topics and distinguish them from those that are statistical inference artefacts. The score is regarded as an indicator of human interpretability of the underlying themes identified by the algorithm, although these do not always align (Meaney et al., 2023). Considerations in the evaluation and fine tuning of models thus include handling over fitting, non-linearity, and discovery of too many words that are not useful for identifying the topics because the number of clusters or topics is not fixed.

2.3 Text mining in public health, biomedicine and clinical knowledge management

2.3.1 Overview of applications of text mining

TM has been applied in several ways to a variety of fields and diseases including medical decision support, predicting disease risk and evolution, mobile health, personalized healthcare and detecting drug reactions (Sun et al., 2018). TM was seen early on as a tool for identifying studies for systematic reviews, offering time and labour savings (O'Mara-Eves et al., 2015a). In the face of rapidly growing bodies of knowledge, TM techniques have been used to understand research trends and gaps in biomedicine (Fleuren and Alkema, 2015), pharmacology, gynaecology (Wang and Lo, 2021), infectious diseases, respiratory disease (Piedra, Ferrer, and Gea, 2014), suicide research (Boggs and Kafka, 2022) and psychiatry (Adeline Abbe, Cyril Grouin, Pierre Zweigenbaum, 2016) amongst others. In pharmacololoy, TM has found application in drug discovery (Zheng et al., 2019), drug interactions (Heng-Yi Wu, Chien-Wei Chiang and Li, 2014), and detection of adverse reactions from electronic patient records (Warrer et al., 2012). In neuroscience, many of the computational tools developed to manage the growing body of knowledge are for textual data. These tools were summarised in 2012 (Kyle H. Ambert, 2012). In yet another application, TM has been used to discover time trends, topic clusters, streams and trends for mHealth research (Przybyla et al., 2016). More recently, as a knowledge management tool for the rapidly accumulating publications on Coronavirus Disease (COVID-19), TM facilitated documentation of a wealth of resources at the disposal of practitioners (Wang and Lo, 2021). Other notable applications are listed below.

- Recently, identification of molecular level associations and common markers led to discovery of links between periodontal and cardiovascular disease (Trindade et al., 2021).
- Review of more that 400 patented probiotic preparations revealed 19 clusters, including new innovative products such as probiotic layered condoms (Patel et al., 2022).

- LDA has been used to identify patterns of associated conditions. Co-occurrence of medical conditions affects prognosis. Topics from the modelling were aligned with distinct medical phenomena of co-occuring conditions, with the high probability conditions known to occur together (Bhattacharya et al., 2016).

2.3.2 Text mining cancer data

TM has been applied to improve representation of cancer data. The momentum seen in application of TM to unstructured clinical data has also been seen in cancer-specific NLP technology (Savova et al., 2019), including knowledge management for cancer. Some examples are listed below.

- Literature on cancer tumour types and anti-cancer drugs has been summarised on an interactive web based platform (https://knowledgebase.microdiscovery.de/heatmap). The platform was created for the 30 most common cancer types and 270 drugs, and it can be used by researchers to glean new insights (Bauer et al., 2021).
- Research trends on child and adolescent cancer survivors' were recently analysed from 1 677 eligible abstracts, identified from a search of articles published in five major web-based databases from 2016 to 2020 (Kim et al., 2022).
- Researchers have shown that, in the case of pathology reports that require manual annotation, a resource intensive and human error prone exercise, cNLP can facilitate faster, large-scale, and standardized (accurate) extraction of data for various purposes (Savova et al., 2019).
- LDA has shown promise as an approach for classification of tissue types based on gene expression data in cancer studies. Studies of differential gene expression in cancer and cellular processes have been a challenge because of the large number of genes and inter-related biological processes (Kho et al., 2017). The application of LDA enabled mitigation of over-fitting of high-dimensionality gene expression data, and facilitated understanding of the associated pathways. LDA was used for clustering, as well as to classify cancer and healthy tissues using lung cancer and breast cancer messenger RNA (mRNA) sequencing data. The study achieved high accuracy compared to alternative approaches, identified a number of genes and pathways that could potentially be involved in differentiating tumour samples from normal (Kho et al., 2017).
- Free text interview data has been coded from transcripts. This was done in a study investigating quality of life priorities for patients undergoing radical cystectomy and urinary diversion for bladder cancer (Li et al., 2019). The bladder cancer study used LDA to ensure statistical rigour and consistency in automating the interpretation of text data. The topic modelling showed the change in priorities before and after the procedures. Patients were concerned about surgery and resuming employment prior to the procedures (Li et al., 2019). Six months after the intervention, the concerns were over cancer recurrence, attaining a sense of normalcy, resuming work, and living more fully (Li et al., 2019).

TM of cancer data has seen advances contributing to the precision medicine quest to find clinically actionable granularity in diagnosis and risk classification (Simmons, Singhal, and Lu, 2016). Examples of this include:

- Building biomedical networks using breast, prostate and urothelial cancer in case study. This highlighted how smart techniques assisted researchers to deal with the high throughput assays that human curators would not cope with. TM allows efficient automatic identification of interactions as well as relationships between entities (Conceição and Couto, 2021).
- Prioritisation of cancer genes and pathways. The various TM based tools for this application are summarised in a review of TM in the field (Yuan Luo and Szolovits, 2014).
- A topic model developed to handle clinical reports, allowing individual patient timelines, has been able to identify patterns of clinical events in a cohort of brain cancer patients (Arnold and Speier, 2012).

Researchers in South Africa have investigated the application of TM to biomedical records. HIV results were successfully extracted from the clinical history section of cancer pathology reports by means of key words to infer the patients' conditions (Dhokotera et al., 2019). In prostate cancer work, the Gleason Score (GS) is used to assign patients to risk categories. The score has been successfully extracted from pathology reports using a custom TM algorithm (Cassim et al., 2021).

2.3.3 Text mining and leukaemia

There has been a rapid accumulation of literature on TM and leukaemia over the last two decades. A bibliometric analysis of the database Scopus showed that there was a consistent increase in the number of articles from 1999 to 2020 (Bhosale et al., 2021). The first article on leukaemia and TM in Scopus reportedly appeared in 1987. The number of articles began increasing slowly, remaining below 20 until 2006, followed by a rapid increase to nearly 200 articles by 2020 (Bhosale et al., 2021).

The rapid increase in the rate of accumulation of knowledge on TM and leukaemia coincided with the increased use of smart tools for the understanding of unstructured cancer data in the mid 2010s (Savova et al., 2019). The trend is also seen by El Alaoui et al., 2022 who reported a spike in publications focussing on AI and leukaemia in 2019. The spike is followed by AI, deep learning and classification trends in 2020 indicating progression in the field (El Alaoui et al., 2022). Analysis of the top ten most frequently used keywords by scholars in the field of AI and haematology reflects the focus of work the field (El Alaoui et al., 2022). The most frequent words were ML (29%), leukaemia (15%), classification (12%), deep learning(11%), artificial intelligence (8%), and 5% each for segmentation, flow cytometry, image processing, AML and ALL (El Alaoui et al., 2022). This provides insight into the prominent tools and techniques, data sources or laboratory tests (flow cytometry), and malignancy types.

The application of smart tools and techniques to facilitate the personalised treatment of leukaemia for improved patient outcomes is a growing field (Banjar et al, 2018). This is being fuelled by the rapid accumulation of omics data and the efficient manipulation and analysis offered by smart approaches. There has been a realisation of the importance of the genotype in addition to, if not over, the phenotype in leukaemia disease risk and treatment management (Walter et al., 2021). Researchers argue that efforts correlating Electronic Medical Records (EMRs) with omics data are not incorporating clinical text. This is significant as clinical records are an important source of phenotype information for oncology patients (Dhokotera et al., 2019; Cassim et al., 2021). AI has facilitated the consolidation of the unstructured with structured data from various disciplines in applications to extract leukaemia

data, diagnose disease and manage patients in the clinical setting for personalised interventions that lead to more favourable outcomes (El Alaoui et al., 2022). As an example, AML has seen, and continues to see ongoing, rapid, multi-disciplinary accumulation of knowledge on disease markers and mechanisms, prognostic factors and therapeutic interventions (Sun et al., 2023; Machado-Neto et al., 2023).

Other examples of the application of smart tools to leukaemia include those listed below.

- ML models trained on blood cell counts have been used to predict CML which accounts for an estimated 15% of adult leukaemia cases, up to five years earlier than the disease was detected by conventional tests (Hauser et al., 2021). The models performance improved with data collected closer to the diagnosis.
- A NLP tool was developed to extract information on the drugs administered to patients diagnosed with AML 30 days before death (Hanahlyn M Park, Vicky Sandhu, Paul Fearn, Kathleen Shannon Dorcy, Elihu H. Estey, 2014). Subsequent work aimed to extract data from records to augment studies based on other records.
- The aMILE study in Portugal was set to apply TM to pdf copies of clinical reports of patients with AML to create and validate algorithms that can be used to extract relevant clinical data. The algorithms would enable these records to be transformed to structured data in a reliable, safe and confidential manner (Rb-Silva and Karimova, 2021).
- A customised rule based algorithm, I2E TM platform has accurately extracted data from bone marrow biopsy reports of patients with myeloid neoplasms (Kunz et al., 2018).

There is still need for more work in the various domains and data sources as use cases have shown that domain and text record nature impact the performance of algorithms (Savova et al., 2019).

2.4 Electronic medical records: A research and knowledge resource

EMRs comprising clinical notes, surgical, radiology, pathology and discharge reports amongst others are a rich and valuable data resource for large scale health analysis. The records typically contain the patient's condition, diagnostic information, procedures performed, and treatment results (Sun et al., 2018).

The free text and semi structured components of the EMRs present some challenges relative to structured data. The free text is generally unstructured, may contain grammatical and spelling errors, as well as semantic ambiguities. In some languages, aspects of local language or dialects that are not standard may also be present (Sun et al., 2018). There are also elements of redundancy, incompleteness and privacy requirements with regard to medical records. These qualities of EMRs make the pre-processing of data for mining and analysis more challenging relative to structured formats (Sun et al., 2018).

In South Africa, the National Health Laboratory Service (NHLS) is the custodian of all laboratory testing data. The data are collated from the network of NHLS laboratories through the Laboratory Information System (LIS) and stored in the Corporate Data Warehouse (CDW), a server with massive storage capability. The NHLS serves over 87% of testing needs in the public sector (Nyasulu, Paszko,

and Mbelle, 2014). Thus, the bulk of national laboratory results are stored in one system. These EMRs are a valuable resource that can be utilised for surveillance, research and epidemiological applications. This is in addition to the primary purpose of facilitating the management of testing and reporting that the system was set up for. The management system of LIS is used by the national network of laboratories to manage data related to samples received, instruments used for testing, and result distribution, storage, and quality control among other functions (Stevens et al., 2016). The LIS is used to generate data that is accessible to the requesting clinician. Data is also used for appropriate and accurate billing, and planning purposes by the NHLS (Nyasulu, Paszko, and Mbelle, 2014)

2.4.1 Text mining pathology reports

The primary purpose of pathology reports is for the pathologist to communicate their findings to the doctor treating the patient (Weegar, Nygård, and Dalianis, 2017). However, pathology reports also contain other useful information. Information extraction has been applied to pathology reports of various cancers including breast (Currie et al., 2006; Weegar and Dalianis, 2015; Achilonu et al., 2021), colon and colorectal (Wagland R, Recio-Saucedo A, Simon M, Bracher M, Hunt K, Foster C, Downing A, Glaser A, 2016; Shelly Soffer , Eyal Klang, Noam Tau, Roni Zemet, Shomron Ben-Horin and Kopylov, 2020; Achilonu et al., 2021), lung (Nguyen et al., 2011) and prostate (Currie et al., 2006; Cassim et al., 2021; Achilonu et al., 2021) and skin (Pan, Zhang, and Liu, 2018) among others. These applications of TM have mostly been done in English using rule based, ML and combinations of methods, for a variety of reasons.

TM has also been done in other languages. Rule based approaches have been applied to Norwegian pathology reports (Weegar, Nygård, and Dalianis, 2017; Weegar and Dalianis, 2015; Dahl, Özkan, and Dalianis, 2016). It has been shown that automated extraction of information from unstructured pathology text can be done with a high degree of accuracy and relatively small effort (Weegar, Nygård, and Dalianis, 2017). This would ease the workload of trained coders that manually process the reports at the Cancer Registry of Norway (Weegar, Nygård, and Dalianis, 2017). The tool employed used a combination of RegEx and a set of heuristic rules, and flagged inconsistencies for human expert review. Exclusion of reports with inconsistencies improved the performance of the system.

When text reports come from different sources, there will be heterogeneity (Pironet et al., 2021). At the Belgian Cancer Registry, the reports could be either in French or Dutch or both in the same report, with different naming conventions, writing styles, and use of punctuation. In cases where there is heterogeneity, ML methods are expected to perform better than rule based ones (Pironet et al., 2021). This has been demonstrated by the application of a range of ML classifiers to automatically extract breast cancer biomarkers from pathology reports (Pironet et al., 2021).

2.5 Application of text mining to leukaemia knowledge in South Africa

2.5.1 The epidemiology of leukaemia in South Africa

In 2019, there were 275 and 299 new leukaemia cases amongst females and males respectively in South Africa (National Health Laboratory Service: National Cancer Registry, 2023). Males had a higher lifetime risk (1 in 869 individuals) of being diagnosed with leukaemia relative to females where the risk was 1 in 1127 (National Health Laboratory Service: National Cancer Registry, 2023).

In 2018, leukaemias were the most common diagnosed cancers in children aged 0-14 years in South Africa. The NCR is working to include data from all possible sources in future reports on childhood cancers (NCR-SA, 2022). The incidence of cancer is expected to increase in South Africa with 112 921 new cases predicted to be diagnosed in 2030 compared to 77 440 in 2012 (International Agency for Research on Cancer GLOBOCAN 2012 cited in Singh et al., 2016. There are limitations to understanding of the epidemiology of leukaemia with scarcity of data attributed to systemic issues (Schonfeld et al., 2016).

2.5.2 Sources of leukaemia information in South Africa

The NCR, a division of the NHLS and the only source of national cancer incidence and prevalence data in South Africa, collates electronic data from public and private pathology laboratories. It documents the data and releases reports by sex, age and ethnic group (Singh et al., 2016). The evolution and growth of the NCR is potentially a platform to address the under-reporting of cancer in South Africa (International Agency for Research on Cancer. GLOBOCAN 2012 cited in Singh et al., 2016). Cancer registries have been shown to be useful in understanding the public health issues around cancer (Höglund, Sandin, and Simonsson, 2015), albeit with limitations imposed by the accuracy and coverage of registries (Höglund, Sandin, and Simonsson, 2015).

It is vital that, as the country embarks on the implementation of the National Health Insurance (NHI), the burden of cancer is quantified for national resource planning (Singh et al., 2014). South Africa's public sector only has a limited number of pathologists (NHLS, 2019b) and these scarce skills and workforce are not equitably distributed between rural and urban areas (NHLS, 2019a). The limited availability of pathologists has been reported to affect the effectiveness of public health systems (Fleming et al., 2018). Resources need to be availed at the right place and time, which needs up to date, comprehensive epidemiological understanding of the disease.

There are other potential sources of leukaemia data in South Africa. These include clinical notes, hospital records, death records and research databases. TM and other NLP techniques can be employed to improve the data representation of leukaemia. In addition, the techniques could be used to automate the update of the cancer registry, making data available timeously for epidemiological studies and surveillance.

METHODOLOGY

3.1 Introduction

This chapter presents the methods used in the study. The chapter begins by describing the study design, computational environment, study population, and source of data. The TM work-flow, tools and techniques implemented are subsequently outlined. These were an expert-driven RegEx based algorithm and unsupervised topic models used to extract the leukaemia information. The methodology used to determine Age Standardised Incidence Rates (ASIR) using the data obtained from the RegEx based algorithm is also described. The chapter concludes with an account of the methods used to evaluate the TM.

3.2 Study design

The project was a retrospective cohort study using secondary data. The data was from text reports of flow cytometry investigations performed at CMJAH laboratory between 2015 and 2019 on bone marrow aspirate samples. Reports from 2010 to 2014 could not be retrieved from the system, hence the variation in dates. The study employed an agile development methodology, characterised by continuous interaction, iteration and delivery of solutions that add value as quickly as possible. In this approach, the solution is deployed rapidly, in the form of minimally viable products (MVPs) (Appugliese, Nathan, and Roberts, 2020). These MVPs and iterative cycles that lead to them are important components of the fast-paced problem solving that drives the agile approach (Appugliese, Nathan, and Roberts, 2020). In this study, the approach fostered collaboration between the researcher, data science and domain experts.

3.3 Computational environment

Computation for the project was done on a Microsoft Windows computer with an Intel Core i7 processor, and 8GB memory. Data management and computing from extraction, pre-processing, text mining and modelling was done using Anaconda Navigator Python Notebook scripting (version 3.9.7) running on 64-bit Microsoft Windows 11. The environment gives access to data science libraries for the TM on the Jupyter Notebook Integrated Development Environment (IDE).

3.3.1 Study population and sampling

The study population consisted of patients with clinically suspected leukaemia for whom haematology flow cytometry work-up on bone marrow aspirate samples was done at the CMJAH Laboratory

between 2015 and 2019. CMJAH is part of the NHLS. The laboratory processes more than 4 000 samples a day, providing support to surrounding public hospitals and primary health care facilities. In 2018, the laboratory performed 3 054 bone marrow morphologic evaluations and 38 071 automated differential counts. It is also a national reference laboratory for both the private and public sectors. The following variables were extracted from the CDW: episode number, unique patient identifier, age, gender, race group, health facility name, province, health district, municipality, testing laboratory name, date of testing, and narrative flow cytometry report. The subset of reports from bone marrow aspirate samples was taken from the full dataset received. A randomly selected 10% sample of the dataset was used to evaluate the expert-driven algorithm, against coding by experts. This allowed a large proportion of the data to be used for the development of the RegEx algorithm, balancing this with an acceptable for testing generalisation to an unseen subset. Allocating a sufficient proportion of data to training or in this case development of the algorithm whilst ensuring acceptable testing is important in AI (Sivakumar, Parthasarathy, and Padmapriya, 2024)

3.4 Data source and description

Data were obtained from the CDW of the NHLS. The dataset was received as a comma separated Microsoft Excel file, and read into a Pandas dataframe for cleaning and analysis. The variables included the unique patient identifier, gender, birth date, age at time of testing, dates of sample collection, receipt by the laboratory, and review by the pathologist, province and a number of others containing information on the health facility that referred the patient. The flow cytometry text reports were contained in the variable 'comments'. A column for age, in completed years, was created from the age, in years and months, using RegEx. Exploratory analysis of the data gave indication of the cleaning steps that would be required to pre-process the corpus.

3.5 Text mining work-flow, methods and techniques

The TM was implemented following the general framework of tasks, starting with data acquisition, pre-processing, modelling, validation and evaluation of the algorithm and topic modelling. The tasks are described in detail below, and the workflow is illustrated in Figure 3.1. The script used to is available,on request at https://github.com/SibonginkosiM/LeukamiaTM (Maposa, 2025). The code was written by the researcher using resources of two main introductory python texts, Python for Everybody (Severance, 2018) and Learning the Pandas library (Harrison and Prentiss, 2016), and online resources such as stack overflow and Medium to help understand the text modelling and coding concepts.

3.5.1 Data acquisition, inclusion and exclusion of reports

The Pandas dataframe was created in the Jupyter Notebook IDE using the Pandas read_ csv method. Reports were included in the study if they were flow cytometry investigations on bone marrow aspirate samples, initial diagnosis work-ups, and written in English. Reports of investigations on other samples and follow ups were excluded. A report typically contained information on the sample analysed, cell types of interest and amounts, a diagnosis, and differential diagnoses. The cell types and their features characterised the suspected malignancy type.

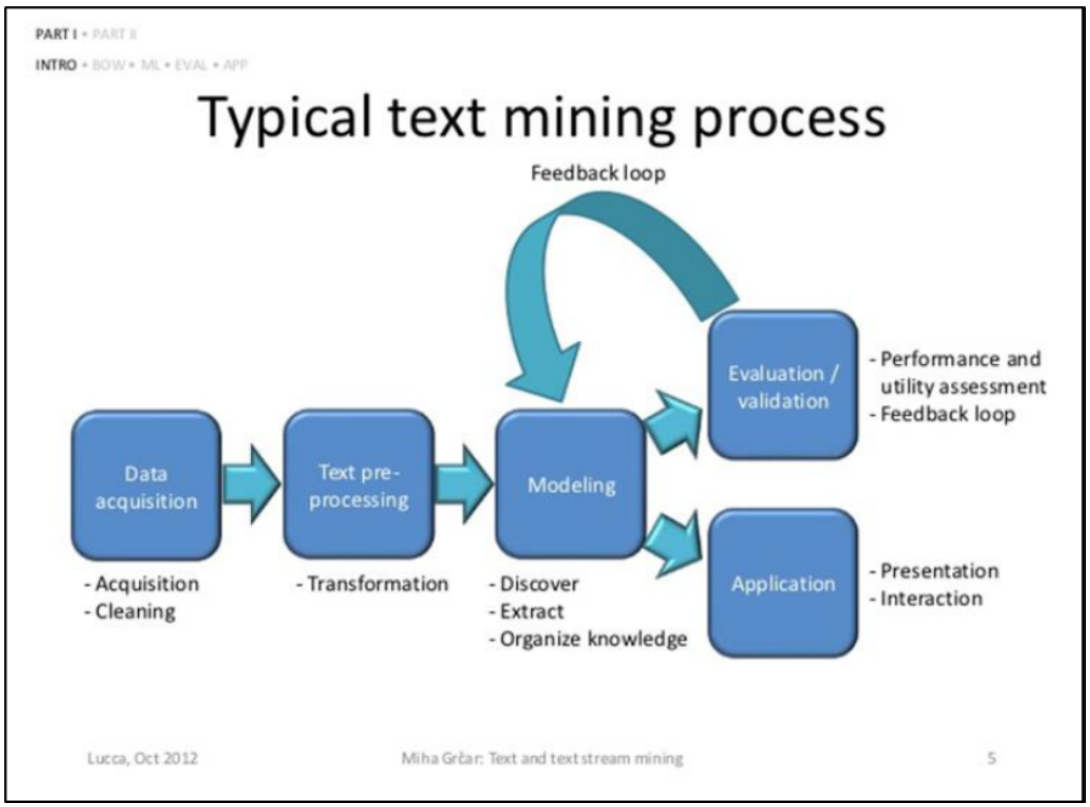


FIGURE 3.1: Illustration of the text mining work-flow
Source:Grcar, 2012

3.5.2 Text pre-processing

The flow cytometry report, contained in the variable 'comments' was used for the expert-driven information extraction and topic modelling. Pre-processing is a crucial step in TM. In this study, the pre-processing took several iterative rounds to build a corpus best suited to both expert-driven categorisation and the unsupervised learning. Such an iterative approach of continuous corrective feedback ensured the best possible knowledge model (Sun et al., 2018).

Pre-processing encompasses tokenisation, lemmatisation, and removal of punctuation, new lines and stopwords. The exact pre-processing techniques and steps depend on the task at hand. In this study all these steps were undertaken to reduce noise in the data. Characters not necessary for the project tasks were replaced with white spaces whilst those of technical significance were retained to facilitate the topic modelling component. Unnecessary text was dropped using RegEx. Stopwords are words that although in natural language are necessary, do not play a role in understanding the meaning of text such as 'and', 'it', 'he', 'she' and 'their' amongst others. They increase the size of the text (Navlani, 2019), and may introduce noise in the model. NLTK has a corpus of english stopwords that can be called. The stopwords were adapted to remove words that were necessary for the tasks, and add those that were contextually irrelevant. Removal of stopwords reduces the feature space and enables text analysis to be performed on meaningful words. Lemmatisation was only done on the topic modelling dataframe to convert words to their base forms in a meaningful way because there was benefit in reducing the feature space. Lemmatisation was preferred to stemming because morphological analysis results in words that are valid in the language (Divya et al., n.d.). It was advantageous for IE using RegEx to maintain the terms as they were in the reports.

The raw data set contained duplicated sentences in the reports, delimited by commas. Duplicate sentences were removed using the set method. Although this had the drawback of splitting some sentences, it allowed for removal of duplicate sentences. In tokenisation, text can be split in a number of ways such sentences, by white space or punctuation to create sequences of characters that are useful semantic units for processing. Sentence tokenisation was used to pre-process the raw dataset.

3.5.3 Information extraction

IE is the automated extraction of useful information from unstructured or semi-structured data, into a structured form. The extraction of leukaemia information was done in two parts: an expert-driven classification of reports into five main categories using RegEx, and topic modelling using LDA and BERTopic algorithms. N-grams analysis was used to complement the feature extraction by checking that useful sequences were in the text from which information was to be extracted. N-grams can be used to identify more informative features by retaining some context around words in the feature space. The use of n-grams has been said to be useful for short texts where the presence or absence of words is more significant (Schonlau, Guenther, and Sucholutsky, 2017). N-grams may also improve the interpretability of the outputs in topic modelling.

Extraction of leukaemia information: Expert-driven regular expressions

It is important to identify the right patterns or RegEx to ensure that the extractor delivers high-quality results (Phande and Khandelwal, 2014). A set of expert-driven terms was refined throughout the project in an iterative process, highlights of which are outlined below. RegEx allowed the expressions to be written concisely and also catered for spelling variations. It is important to format and evaluate RegEx robustly as they are error prone (Zheng et al., 2021). The goal was to identify incident (newly diagnosed) reports with a leukaemia finding, and assign each report into one of the five categories. The refinement of the terms was guided by the observations on the corpus listed below.

- The term 'leukaemia' was not specified because not all the pathologists' reports specifically mentioned the term when describing a malignancy. Terms were thus used to identify and exclude reports where there was no leukaemia.
- There was a set of reports that neither had a negative outcome nor a diagnosis. These mostly had findings of cell dyscrasia, a blood cell composition disorder that is not leukaemia. In addition, there were multiple myeloma diagnoses. The terms 'cell dyscrasia' and 'multiple myeloma' were added to the 'negative terms' to identify reports that did not have a finding of interest.
- At validation, additional expressions were added to exclude reports of flow cytometry performed on pleural fluid or peripheral blood. Another expression to match expressions in some follow up reports where the letters were followed by a '0' (the number zero), and at times the capital letter 'O'.
- In cases where both the acute and chronic forms of leukaemia were being picked up, the terms transformation and evolution were used to identify cases and diagnosis was assigned to the acute form of leukaemia.

- Terms were added, and adjusted in the four main categories as work progressed and the outcome of the algorithm was reviewed. Changes were made to improve the proportion of reports that the algorithm could classify, and accuracy of the classification.
- Elimination of differential diagnoses from the corpus was part of the iterative process of cleaning during the project. In some of the reports, the differential diagnosis terms were resulting in the report being placed in more than one leukaemia category.

The terms and RegEx used to categorise reports into the main categories of leukaemia are listed in Table 3.1.

TABLE 3.1: Expert terms and regular expressions for leukaemia categories

Leukaemia Category	**Expert Terms**	**Regular Expressions**
ALL	acute lymphoblastic leukaemia, burkitt's leukaemia, burkitt's lymphoma, precursor b cell or b-cell, pre-b, acute t-cell or acute t cell, b cell lymphoblastic leukaemia, acute lymphocytic leukaemia	acu\w+\s*lymphob\w+|acu\w+\s*lymphoc\w+|burk\w+|acu\w+\s*t\s*cel\w+|prec\w+\s*b\s*cel\w+|pre\s*b|b\s*cel\w+\s*lymphob\w+
AML	acute myeloid leukaemia	acu\w+\s*mye\w+|monob\w+
CLL	chronic lymphocytic leukaemia, chronic lymphoid leukaemia, hairy cell leukaemia, chronic t-cell or chronic t cell, mature b cell lymphoproliferative disorder, sezary syndrome, follicular lymphoma, prolymphocytic leukaemia,htlv-1	chr\w+\s*lymphoc\w+|chr\w+\s*lymphoi\w+|hairyladu\w+\s*t\s*ce\w+|chr\w+\s*t\s*ce\w+|mature\s*b\s*cell\s*lymphop\w+|sez\w+|follicu\w+|prolymph\w+|hltv
CML	chronic myeloid leukaemia, chronic myelomonocytic leukaemia or cmml, chronic myelocytic leukaemia, chronic eosinophilic leukaemia, chronic neutrophilic leukaemia, chronic myeloproliferative disease	chr\w+\s*mye\w+|chr\w+\s*eo\w+|chr\w+\ s*neu\w+
Undifferentiated	Undifferentiated	undif\w+

The other terms and their regular expressions in the algorithm are shown in Table 3.2

TABLE 3.2: Other terms and their regular expressions

Output Reports	**Expert Terms**	**Regular Expression**
Bone marrow aspirate	bone marrow aspirate	bon.*\W*mar.+ \W*asp.+
Follow-up	follow up, residual, relapse, presentation	follow|resid\w+|relap\w+|present\w+|[a-z]{2}\d{3,}
Cell dyscrasia and multiple myeloma	cell dyscrasia, multiple myeloma	dyscr\w+|mult\w+\s*myelom\w+
No leukaemia	not consistent with, no overt evidence, no evidence	not\s*consis\w+|no\s*ove\w+\s*evidence|no\s*evid\w+|no\s*ove\w+
transformation or evolution	transformation, evolution	transform\w+|evol\w+

Topic Modelling: Latent Dirichlet Allocation

The LDA model was fitted to the cleaned flow cytometry results corpus. This corpus excluded follow up reports, and those where no leukaemia or other malignancy had been found. The model was also fitted on the corpus with the reports where no malignancy finding had been made, that is, the whole corpus of new work ups. The objective was to explore whether topic modelling could be effectively applied to cluster the text corpus into useful and meaningful topics, and offer similar or additional knowledge discovery compared to the expert-driven process. One advantage of LDA is that

the topics are not known beforehand, and thus it is possible that previously unknown information can be uncovered. LDA is a statistical model, and the output topics are sometimes difficult for humans to understand. The pre-processing of text and feature engineering are thus important aspects of text preparation to ensure interpretable output. The various pre-processing and feature engineering steps taken are outlined below.

- The stopwords corpus was modified to add back the words 'not', and 'with' as these would be important to distinguish reports describing a malignancy from those where no adverse findings were made.
- The number of topics used was chosen using a plot of the c_v coherence score against the number of topics. The elbow technique was used used to narrow down the best possible number of topics. Coherence score increases with the number of topics (k). The elbow technique is a trade off between coherence and k, at the point where the declining increase in coherence does not provide any further information in the form of new topics.
- A number of iterations and passes were tested to identify the best model.

Outcome topics and top words were visualised using PyLDAvis. The selection of the number of topics was guided by the plot of topic coherence against the number of topics, and examination of the PyLDAvis plots to identify those in which the clusters representing the topics were distinct.

Topic Modelling: BERTopic

The BERTopic model was trained on the same corpus to which LDA was applied. This enabled a comparison of the two approaches when applied to the flow cytometry reports corpus. Uniform Manifold Approximation and Projection (UMAP) and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithms were incorporated into the BERTopic model. UMAP was used for dimensionality reduction and visualisation of high-dimensional data to provide more concise representation of the document embeddings. The algorithm parameters were set as follows: n_neighbors (15) and n_components (10). N_neighbors controlled the number of neighbours considered for each point whilst n_components set the dimensionality of the output space. The distance metric used was 'cosine'. It is a measure of the cosine similarity between document embeddings. Lastly, low memory was set to false. HDBSCAN was incorporated for clustering purposes. HDBSCAN is a density-based clustering algorithm useful for handling clusters of varying densities and noise. The parameters of this algorithm were set as follow: min_cluster_size, which defines the minimum number of points required to form a cluster was set at 40, and the distance metric for clustering was euclidean.

3.6 Epidemiology of leukaemia: Age Standardised Incidence Rates of cases worked up at CMJAH Laboratory

The extracted leukaemia information was used to calculate the ASIR reported per $100{,}000$ of the population as per accepted practice (Boniol and Heanue, 2007). National South African census data were used to for the population numbers in the age groups. The Segi World Standard Population was used to standardise the age rates. The World Standard Population was developed by Segi in 1960 and

modified by Doll in 1966. The population is based on the age structure from pooling of populations from 46 countries (Boniol and Heanue, 2007). Age standardisation is preferred over crude rates, as comparisons of incidence rates need to be independent of the effects of age when comparisons between populations are made (Boniol and Heanue, 2007).

The ASIRs were calculated and plotted using the Jointpoint Regression Program (National Cancer Institute, 2024). Joinpoint regression analysis uses SEER*Stat software (National Cancer Institute, n.d.) to calculate rates. The program is commonly used in health research to identify and analyse changes in trends of rates, over time.

3.7 Evaluation of the text mining

3.7.1 Evaluation of the regex based algorithm

An unseen dataset was used to evaluate the expert-driven algorithm. The reports on the evaluation dataset were coded, categorised as leukaemia or not, and for those that were leukaemia, assigned to one of the five main categories by an expert. The algorithm was implemented on the labelled unseen dataset. Performance was evaluated using the metrics precision, recall and the F1 score. A confusion matrix was drawn up for leukaemia positive or negative as illustrated in Table 3.3, as well as the categorisation of reports.

TABLE 3.3: Schematic representation of the confusion matrix

Algorithm	Gold Positive	Gold Negative	Performance Metrics
Positive	True Positives (a)	False Positives (b)	Precision=a/a+b
Negative	False Negatives (c)	True Negatives (d)	Recall=a/a+c
Overall			F1 Score=2*(Precision*Recall)/ (Precision+Recall)

3.7.2 Evaluation of the topic modelling

Topic models can be evaluated in a number of ways including how they respond to new previously unseen data, coherence of topics, and interpretability (Meaney et al., 2023; Lund et al., 2020). In LDA models, the pyLDAvis visualisation can also be used. In the visualisation, the further apart the topics are, the more distinct they are. The larger the circle representing a topic, the more it makes up the corpus. In the ideal model, the topics are represented by distinct circles, and the key words are semantically closely related, making the topic interpretable.

The choice of evaluation will often be informed by the intended use case of the model. In this study, the goal was to explore the application of topic modelling. Human evaluation with assessment of coherence scores and visualisation of topic clustering were used to evaluate the performance of models. The objective of the evaluation was to review topics generated by the model and assess the topics' distinctiveness, coherence and semantic meaningfulness. The c_v metric was used to evaluate topic coherence. The c_v score uses cosine similarity and normalised pointwise mutual information based on co-occurrence of words. The score is popular, and this has been attributed by some to its use as the default measure in the Gensim library. Other measures of topic coherence are:

- UMass is a measure of how often two words appear together in the corpus relative to the first one appearing alone. Thus, the greater the metric, the higher the score. The final score is calculated as the average pairwise values of the top words of the topic.
- The UCI metric uses sliding windows of pointwise mutual information of all the topic's top words to work out co-occurence. Co-occurence is thus, when the two words appear together in the sliding window. The size of the window can be set to any number.
- The word2vec approach is based on intra- and inter-topic similarity. The goal is to maximise intra-topic similarity whilst minimising the inter-topic similarity. The similarity is calculated as cosine similarity between words represented by word2vec embedding.

Human evaluation remains an important aspect of assessing models as the creation of topics that are understandable and meaningful to humans is a key consideration (Hoyle et al., 2021). Evaluators are often individuals with subject matter expertise or familiarity with the dataset's content. They should have a good understanding of the topics' domain and the ability to assess the quality and coherence of topics (Newman et al., 2010). Evaluators review the topics using the top words or other criteria relevant to the task (Newman et al., 2010). Evaluators may also provide qualitative feedback. The scores and feedback may be used to refine the model, re-generate topics, and conduct additional rounds of evaluation until satisfactory results are achieved.

The reviewer for this study was requested to review the key words for each topic, for both the LDA and BERTopic algorithms. For each model, the reviewer was asked to indicate if the topics were distinct from each other, and whether they could infer what the topic was about based on the top words. The reviewer is an expert in the field of haematology with decades long experience, and they are familiar with the contents of flow cytometry reports.

RESULTS

4.1 Introduction

This chapter presents the findings from the study. It begins with a description of the dataset features, and results of the pre-processing steps. The outputs from the application of text mining techniques to extract leukaemia data are presented, followed by the ASIR for the City of Johannesburg, calculated from data extracted using the RegEx based algorithm. Lastly, results of performance evaluation of the TM techniques are presented.

4.2 Application of text mining techniques to extract leukaemia data

4.2.1 Data exploration

The raw dataset contained 22 463 records of which 13 517 were reports from bone marrow aspirate samples. The dataset of bone marrow aspirate samples was explored. Ninety-five percent of reports in the dataset were for patients from Gauteng Province, followed by Mpumalanga ($3,6\%$), North West (0.6%) and Eastern Cape (0.2%) provinces. Table 4.1 shows the numbers of reports from each province.

TABLE 4.1: Number of reports by province

Province	**Number of reports**
Gauteng	12 885
Mpumalanga	486
North West	75
Eastern Cape	52
Limpopo	15
KwaZulu Natal	3
Unknown	1

A small proportion, 1.4%, of the dataset was missing patient gender. The remainder of the reports were of nearly equal proportions of females and males at 50.7% and 48.6% respectively. Nearly all the 193 reports with missing data on patient gender were from Gauteng, and distributed across the age ranges as shown in Figure 4.1.

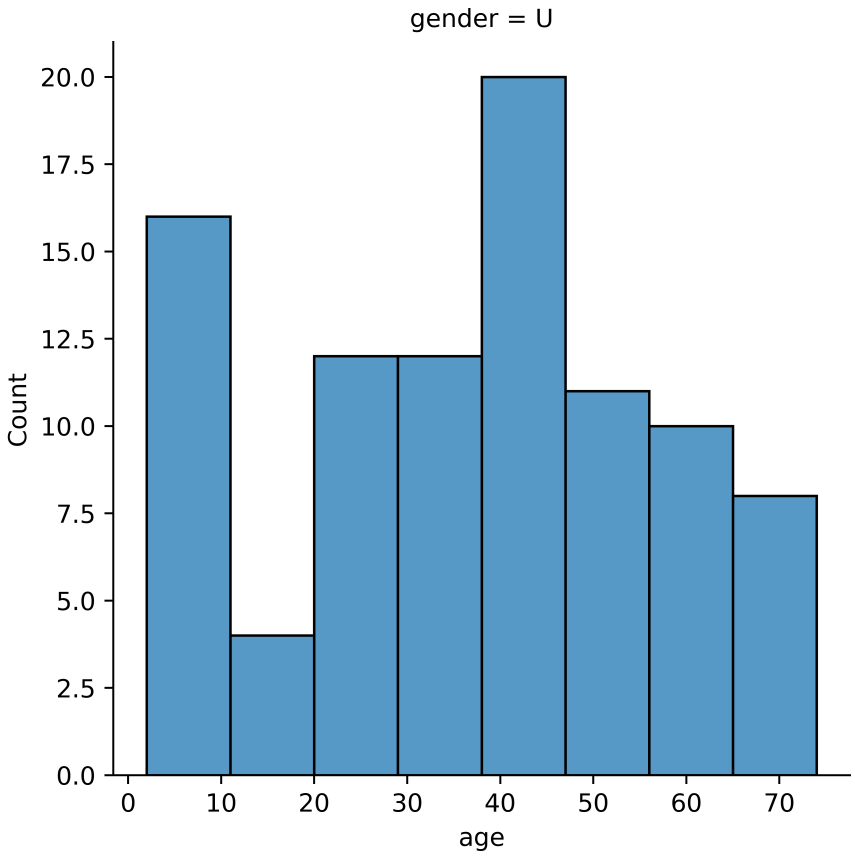


FIGURE 4.1: Age distribution of records with missing patient gender

Figure 4.2 depicts the bimodal distribution of the patients' ages, with peaks in the 0-4 and 30-34, 35-39 and 40-44 years groups. The bimodal pattern was the same in males and females as illustrated in Figure 4.3. The few cases from outside Gauteng were in the middle age ranges.

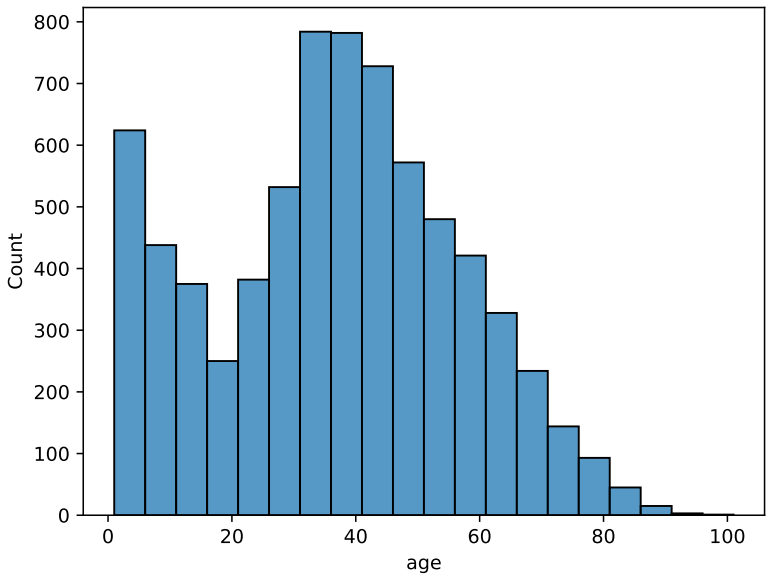


FIGURE 4.2: Distribution of patients' ages

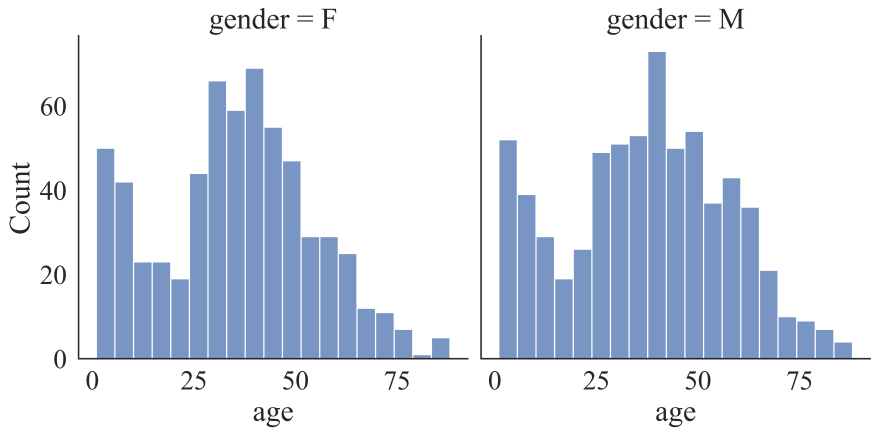


FIGURE 4.3: Age distribution of patients' ages by sex

4.2.2 Effect of cleaning on the corpus

Highest frequency words: Word clouds and bar charts

The full raw, and bone marrow corpus before and after cleaning were all different. Figure 4.4 is a wordcloud of the full raw corpus as received. Figure 4.5 is a comparison of the word clouds of the raw and cleaned corpus of flow cytometry reports performed on bone marrow aspirate samples.

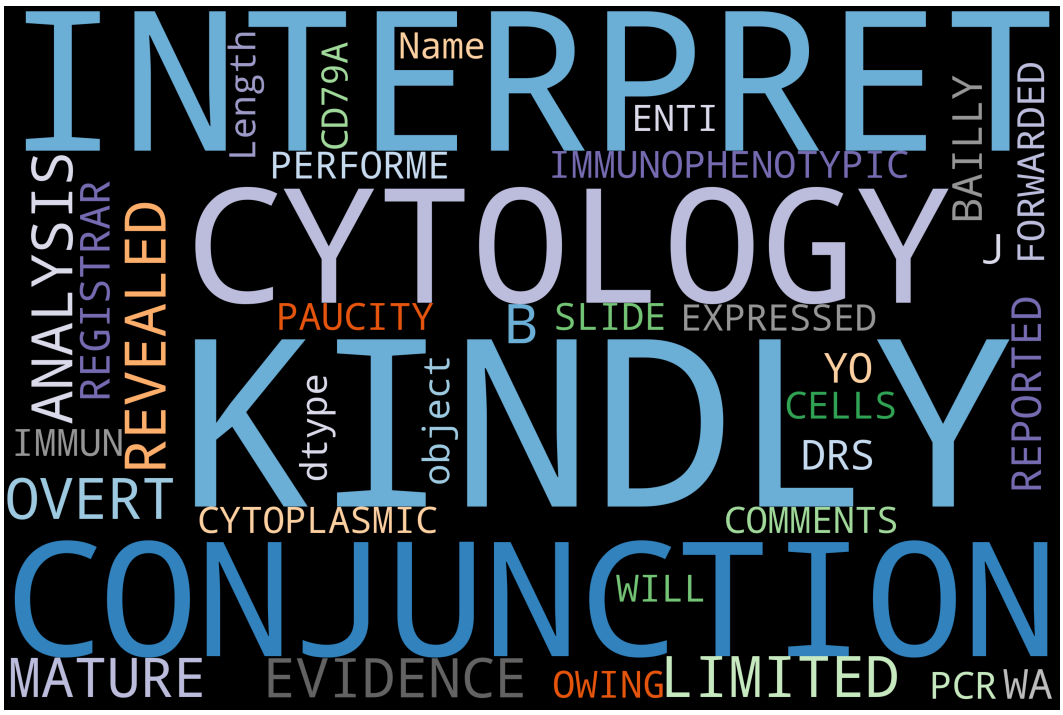
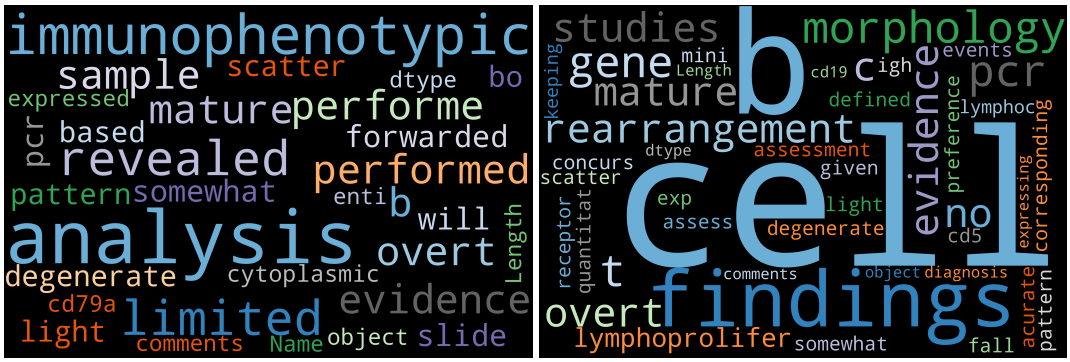


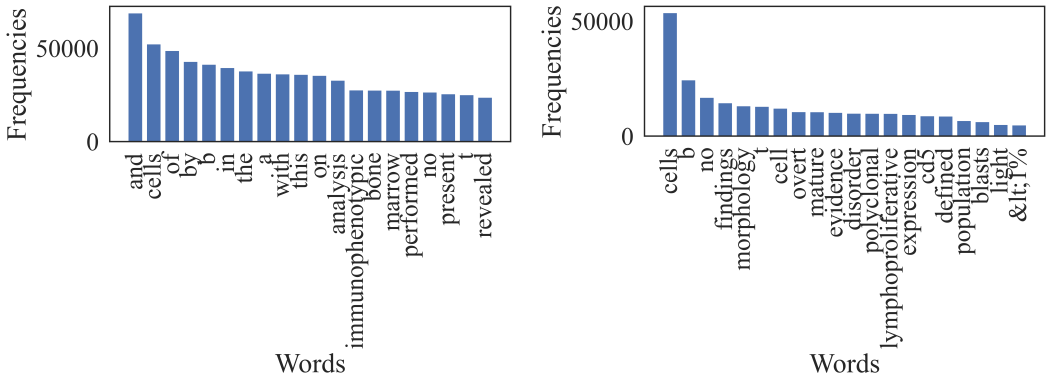
FIGURE 4.4: Word Cloud of the full corpus of reports as received



(A) Word cloud of bone marrow corpus before cleaning (B) Word cloud of bone marrow corpus after cleaning

FIGURE 4.5: Word Clouds of the bone marrow corpus before and after cleaning

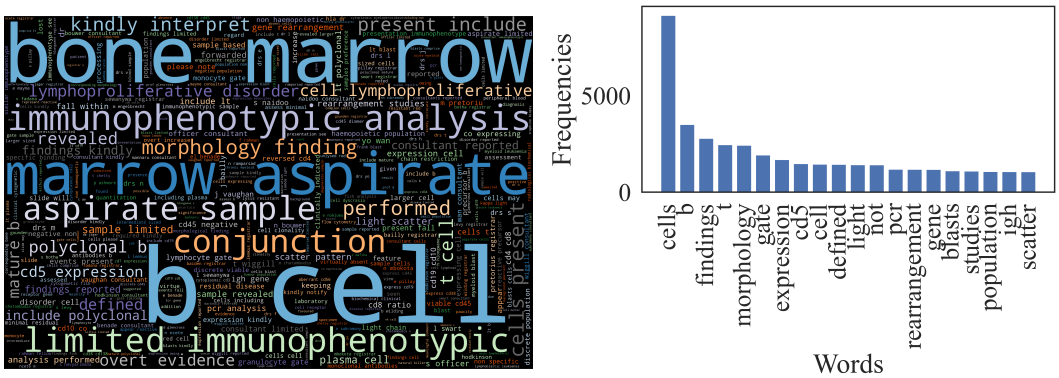
The word cloud of the bone marrow corpus before cleaning shows that stopwords were the most frequent. This is further illustrated in Figure 4.6, which is a plot of the frequencies of the top 20 words in the corpus before and after cleaning. The most frequent word, before cleaning, was 'and', and it appeared nearly 70 000 times. The words 'analysis', 'immunophenotypic', 'bone marrow', 'performed' and 'revealed' were present at nearly equal frequencies of approximately 30 000. The cleaning removed words not needed for TM, leaving relevant words and phrases with meaning in a flow cytometry immunophenotyping report as the most frequent as shown in Figure 4.6b.



(A) Top twenty words before cleaning (B) Top twenty words after cleaning

FIGURE 4.6: Top twenty words in the bone marrow corpus before and after cleaning

The final corpus used to identify reports going into the different categories of leukaemia comprised only reports of initial work-ups in which a haematological malignancy finding was made. The words in this corpus are shown in Figure 4.7 below.



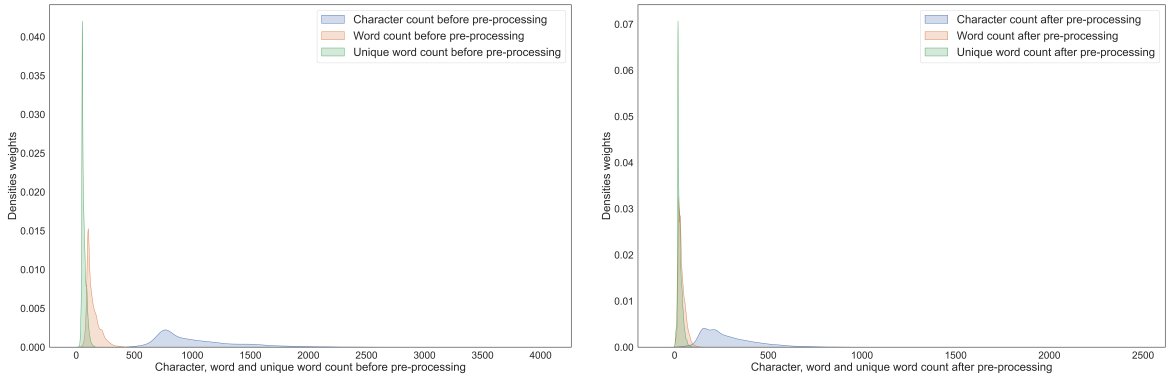
(A) Word cloud of corpus of incident haematological malignancy reports

(B) Top twenty words in corpus of incident haematological malignancy reports

FIGURE 4.7: Representation of the corpus of incident haematological malignancy reports

Reduction in the size of the feature space

Pre-processing reduced the size of the feature space. The reductions in numbers of unique characters and word counts are shown in Figure 4.8. The greatest change was in the character counts that decreased from between 500 and 2 000 in the raw corpus to between 50 and just over 500 after cleaning. The total word count of reduced from between 20 and 450 to zero to 100, nearly matching that of unique words.



(A) Text species before cleaning

(B) Text species after cleaning

FIGURE 4.8: Representation of text species in corpus of incident haematological malignancy reports before and after cleaning

Figure 4.9 shows the effect of cleaning on a follow-up report. The left side is the report before cleaning, and the processed version is on the right.

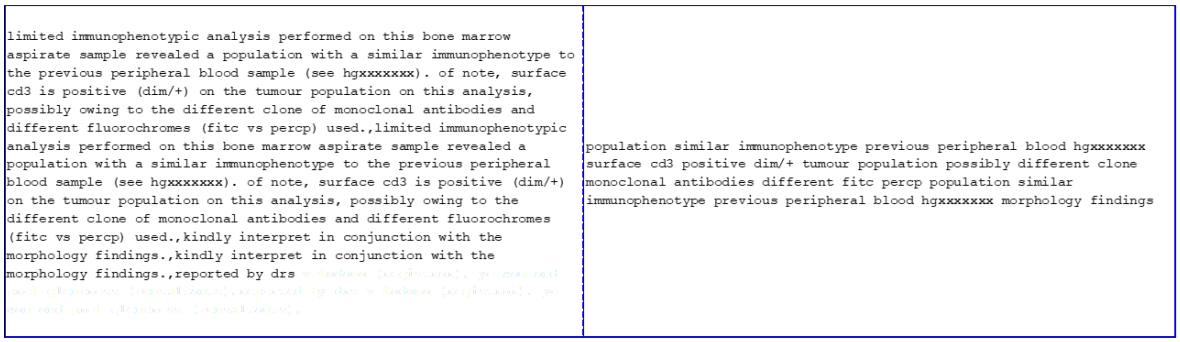


FIGURE 4.9: A comparison of a report before and after cleaning showing the effect of pre-processing

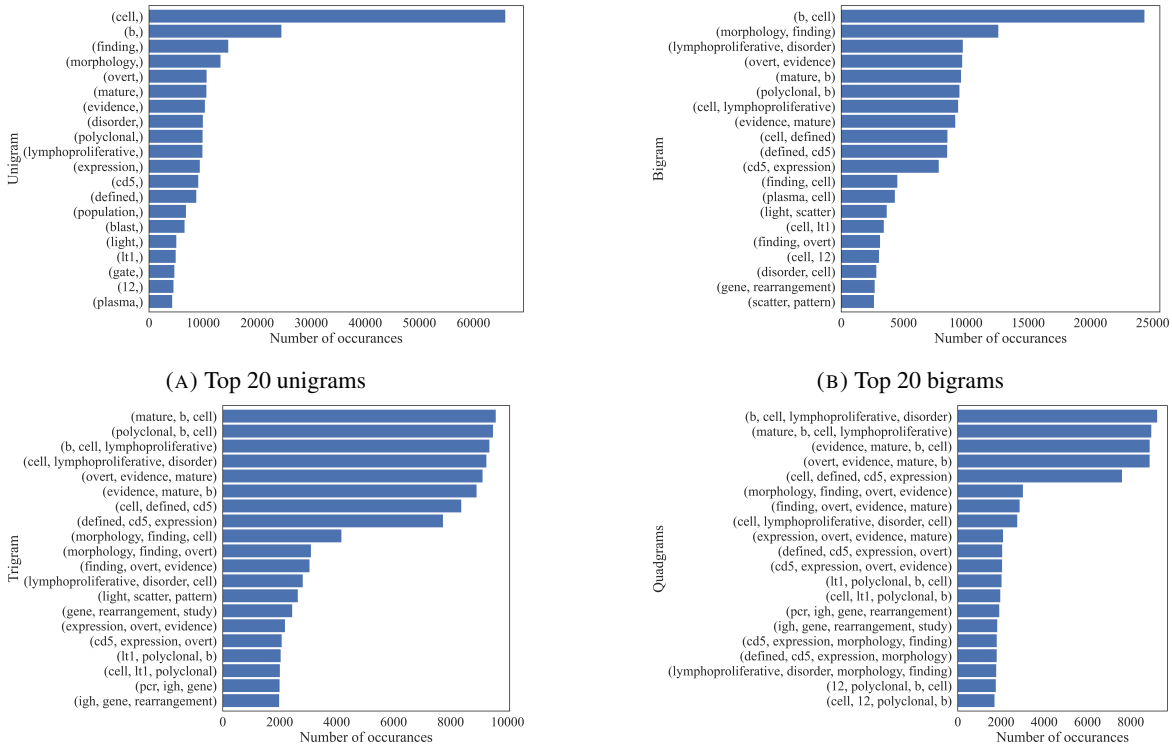
This report was subsequently dropped from the corpus by the regular expression:

`r'follow|resid\w+|relap\w+|present\w+|[a-z]{2}d{3,}'`

The expression was written to identify and exclude follow up reports by matching the phrases: 'follow' as in follow up, 'residual' as in 'residual disease', 'relapse', 'presentation' as in 'seen at presentation', and any reference to previous testing.

N-grams in the cleaned corpus

N-grams in the pre-processed corpus showed patterns that were meaningful. The n-grams included common phrases, expressions, and technical terms that would be expected in the corpus of immunophenotyping flow cytometry reports from bone marrow aspirates. The top 20 uni-, bi-, tri-, and quadgrams are shown in Figure 4.10.



(A) Top 20 unigrams

(B) Top 20 bigrams

(C) Top 20 trigrams

(D) Top 20 quadgrams

FIGURE 4.10: Top n-grams in the pre-processed corpus

4.3 Extraction of leukaemia information

4.3.1 Expert driven approach: Application of regular expressions in a rule based algorithm

The rule-based algorithm selected different reports at each step of implementation. The first step selected reports where the flow cytometry was performed on a bone marrow aspirate sample. The corpus was cleaned, and in the third step follow up reports were excluded. Lastly, the reports were put into the main categories of haematological malignancies. Figure 4.11 is an illustration of the algorithm's main steps and outcomes.

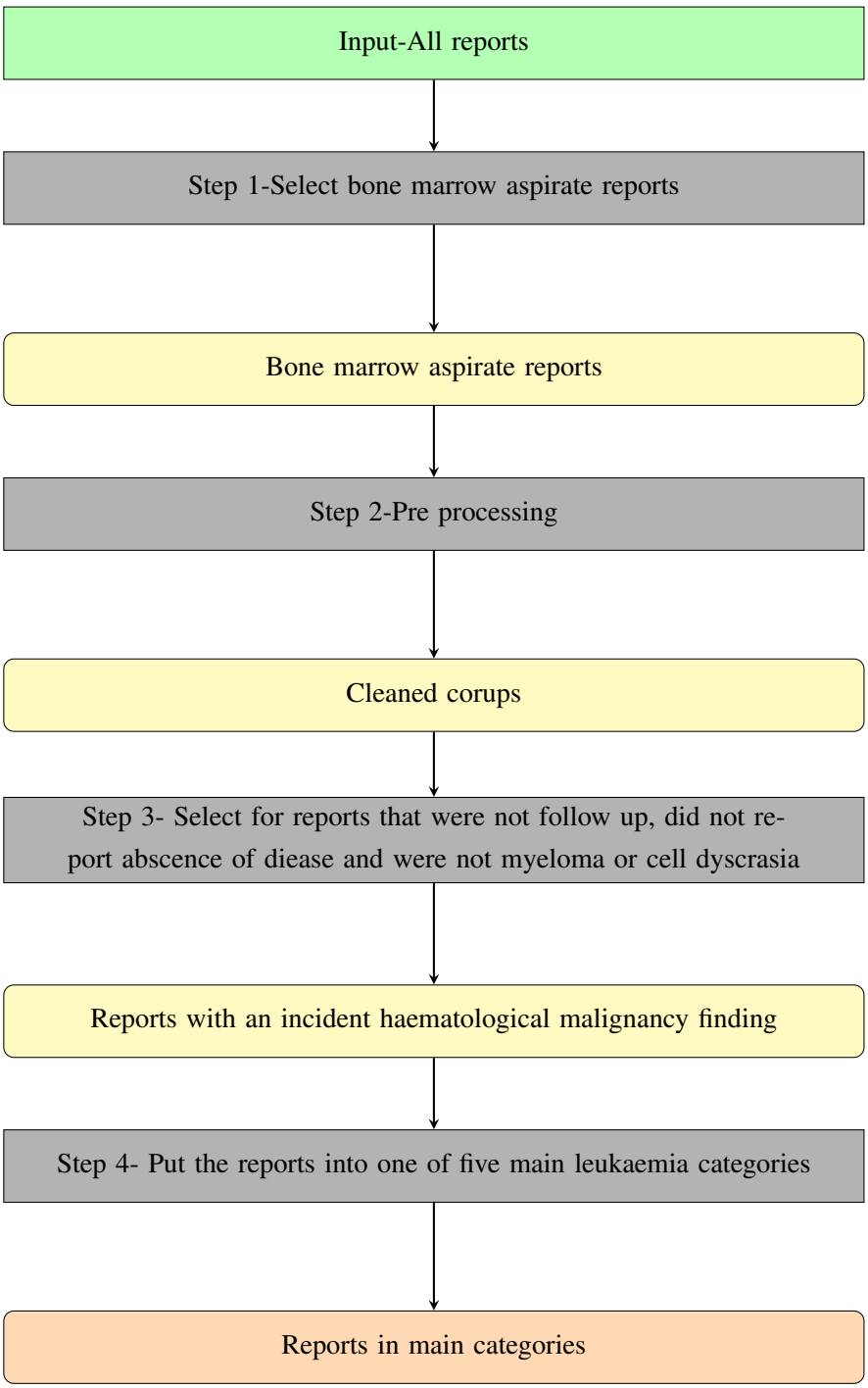


FIGURE 4.11: An illustration of the steps in the regular expressions rule based algorithm

Implementation of the algorithm identified 411 leukaemia positive, incident reports, of which 407 could be allocated a diagnosis, that is, fell into one neat category. Four reports fell into two categories, and the algorithm thus put them into the category of 'other malignant haematological disease'. The 'other malignant haematological disease' category, with 2 104 items, had other reports that were not placed into any of the four major categories. The 2 104 reports had expressions that did not match regular expressions of the expert terms. Table 4.2 provides descriptions of the corpus, and number of reports at different stages of the algorithm.

TABLE 4.2: Descriptions of the corpus and numbers of reports at different stages of the algorithm

Step	Description	Number of reports (cases)
1	Reports of bone marrow aspirate	13 517
2	New diagnoses, dropping all follow ups	10 477
3	Malignant finding, including but not limited to leukaemia	2511
4	Type of malignancy	ALL (n=200), AML (n=29),CLL (n=61), CML (n=117), other malignant haematological disease (n=2 104)

4.3.2 Application of an unsupervised learning algorithm: Latent Dirichlet Allocation implementation

Modelling the corpus of reports with an incident positive haematological malignancy finding

The features with top twenty tf-idf scores in the corpus with incident positive haematological malignancy findings were terms that would be useful in understanding the contents of the document. The terms with the top five tf-idf scores were 'gate', 'define', 'assess', 'clonality', and 'monoclonal'. Figure 4.12 below shows terms that had the top twenty tf-idf scores.

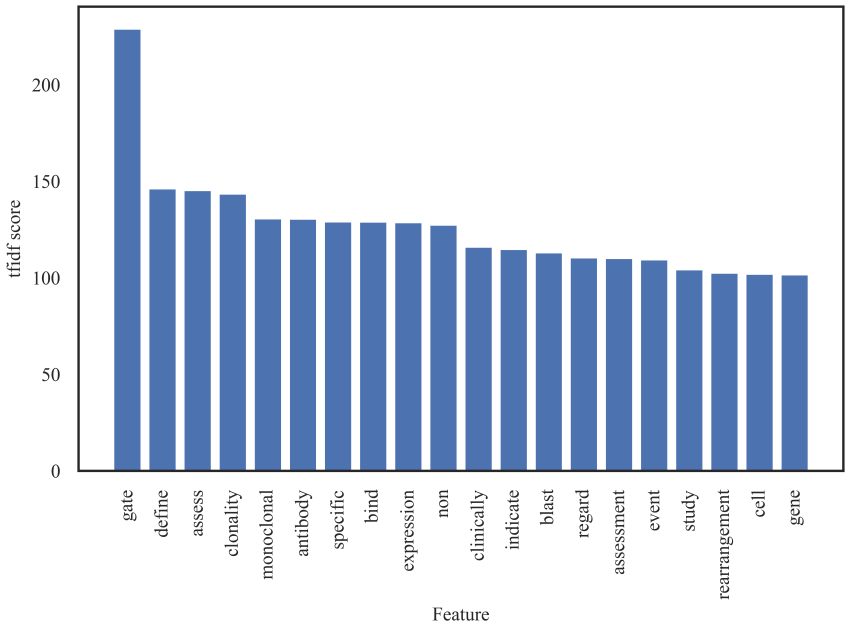


FIGURE 4.12: Features with the top twenty tf-idf scores in the corpus of reports with an incident positive haematological malignancy finding

Ten topics were identified from the corpus of reports with a newly diagnosed haematological malignancy. Identification of the ten topics was guided by the plot of coherence score c_v and the number of topics, k. The variation in coherence with the number of topics is shown in Figure 4.13. Coherence

was lowest at 0.63 for 3 topics, thereafter rising with the value of k, and peaking at 0.73 for eight topics.

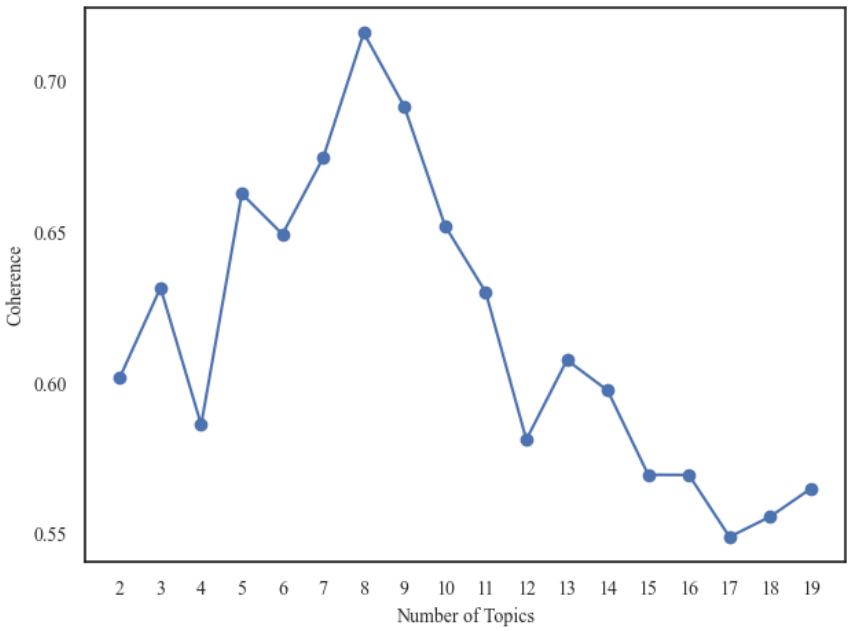


FIGURE 4.13: Plot of coherence scores against number of topics from the corpus of reports with an incident positive haematological malignancy finding

The model with nine topics was considered because the coherence score was the second highest in the plot at 0.69, and it allowed discovery of as much information as possible in the form of distinct topics spread-out over the four quadrants of the visualisation. The model with ten topics (Figure 4.14) was chosen over the one with nine topics following expert review of topics. Appendix 3 shows the top words in each of the ten topics.

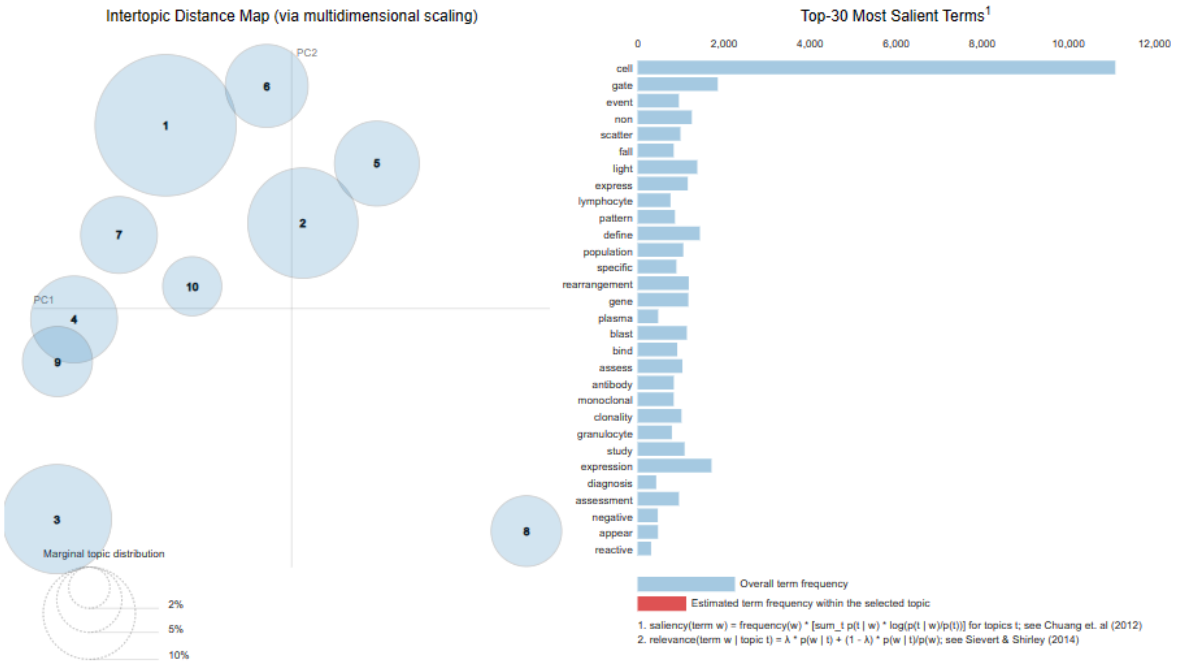


FIGURE 4.14: Visualisation of the ten topics in the chosen model of the corpus of reports with an incident positive haematological malignancy finding

Modelling the corpus of reports of all new investigations

This corpus included reports where no haematological malignancy was reported. The features with top twenty tf-idf scores in this full corpus are shown in Figure 4.15.

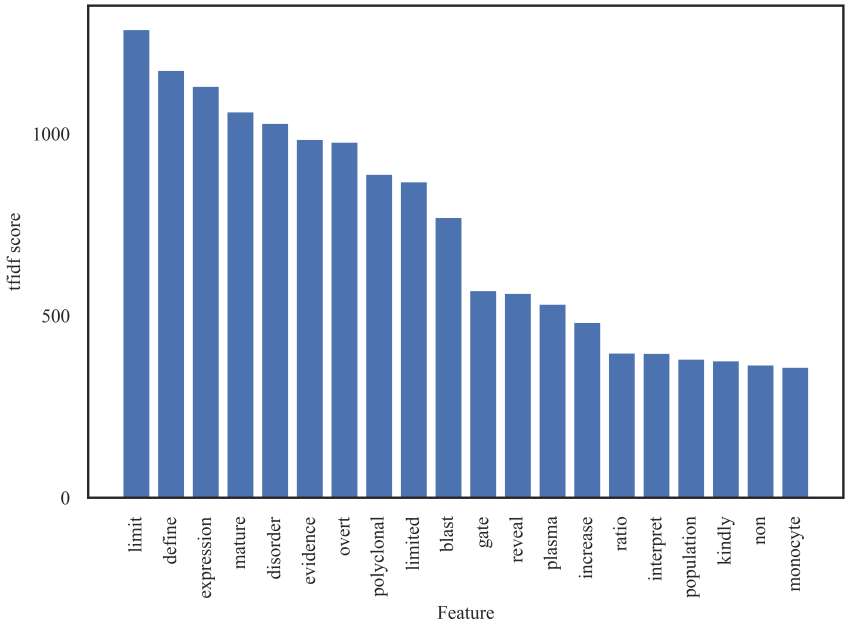


FIGURE 4.15: Features with the top twenty tf-idf scores in the corpus of reports of all new investigations

The plot of topic coherence score against the number of topics showed a peak coherence of 0.63 at 13 topics (Figure 4.16). However, pyLDAvis of the model with 12 topics showed more distinct clusters,

and this model was chosen. The model with 12 topics had a coherence score of 0.62. Appendix 4 shows the top words in each topic. Topic 12, appeared to be an outlier with no key words. On expert review, topic 12 top words were spelling errors and words that would not be useful in distinguishing a report.

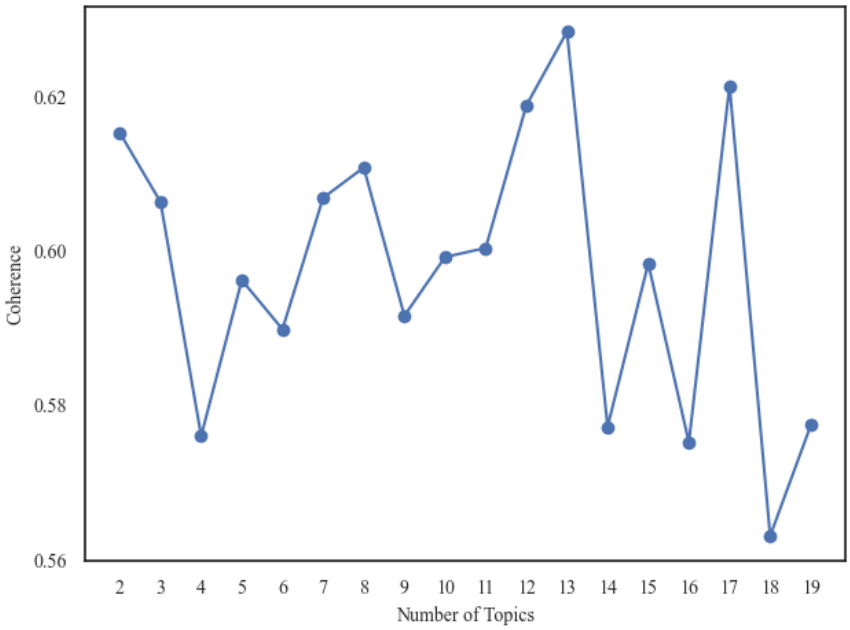


FIGURE 4.16: Plot of coherence scores against number of topics for the corpus of reports of all new investigations

The 12-topic model allowed as much information as possible to be discovered, while maintaining some distinction between the topics (Figure 4.17).

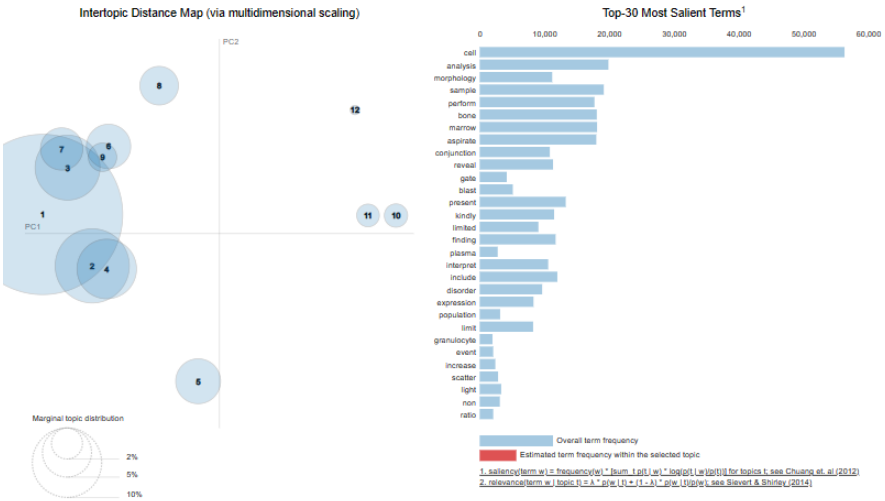


FIGURE 4.17: Visualisation of the 12 topics from chosen LDA model of the corpus of reports of all new investigations

4.3.3 Application of an unsupervised learning algorithm: BERTopic implementation

Modelling the corpus of reports with an incident positive haematological malignancy finding

The model identified 12 topics on this corpus, contrasting with the ten topics discovered from the LDA implementation. The topics and their top words are shown in Figure 4.18.

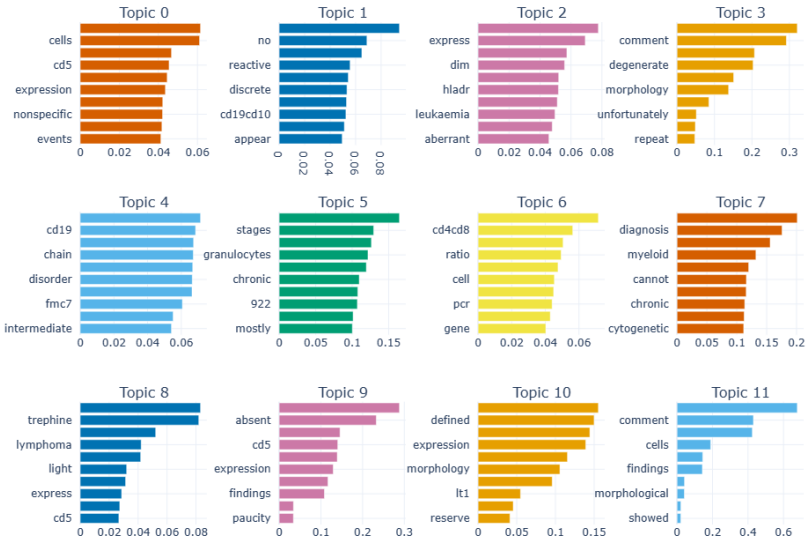


FIGURE 4.18: Barcharts of top words in topics from chosen model of the corpus of reports with an incident positive haematological malignancy finding

The topics were in clear clusters as can be seen in Figure 4.19

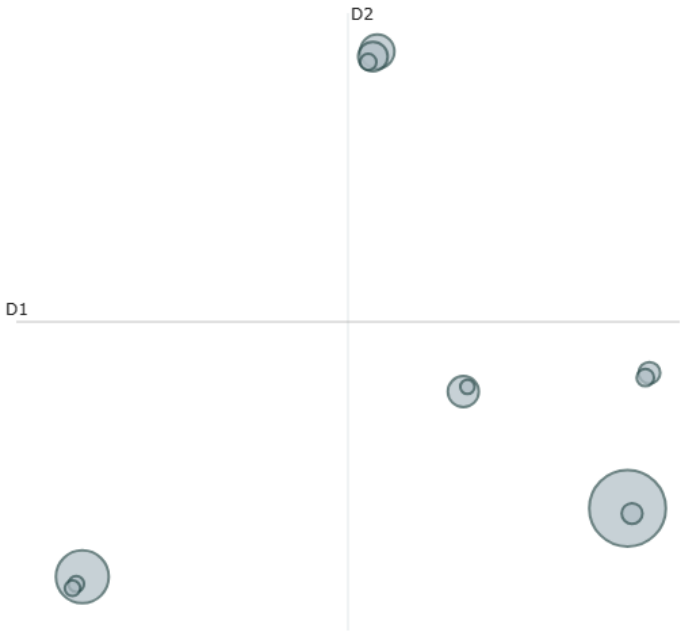


FIGURE 4.19: Intertopic distance map of topics from the chosen model of the corpus of reports with an incident positive haematological malignancy finding

Modelling the corpus of reports of all new investigations

The model identified 13 topics compared with 12 topics from the LDA implementation, on the corpus that included reports of new investigations in which no haematological malignancy was found. These topics and the top words in them are shown in Figure 4.20.



FIGURE 4.20: Barcharts of top words in topics from BERTopic model of the corpus of reports of all new investigations

The topics were, again in clear clusters as can be seen in Figure 4.21



FIGURE 4.21: Intertopic distance map of topics from the corpus of reports of all new investigations

4.3.4 Age specific incidence rates of the main leukaemia types in reports from CMJAH

The ASIR ranged from 0.54 to 0.87 per $100{,}000$ population per year, for the years 2015 to 2019. The observed and modelled age standardised incidence rates for the five years are shown in Table 4.3.

TABLE 4.3: Observed and Modelled Age Standardised Incidence Rates for 2015 to 2019

Diagnosis Year	**Age-Adjusted Rate**	**Model**	**Standard Error**
2015	0.75	0.78	0.13
2016	0.68	0.74	0.12
2017	0.87	0.69	0.15
2018	0.54	0.65	0.11
2019	0.61	0.61	0.10

Figure 4.22 is a plot of the ASIR and years when diagnoses were made.

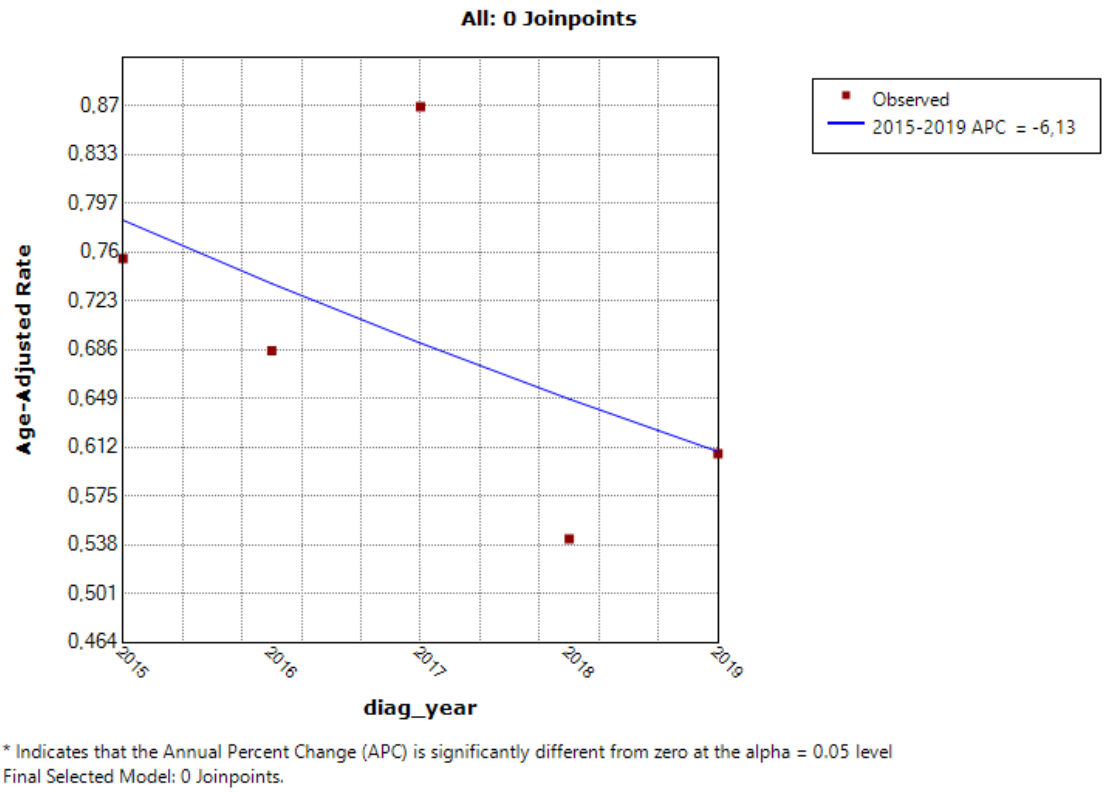


FIGURE 4.22: The leukaemia ASIR from flow cytometry data from CMJAH from 2015 to 2019

4.4 Evaluation of the performance of the text mining

Performance of the regex based algorithm

The algorithm successfully identified 93% of all reports with a leukaemia diagnosis as labelled by the experts. The algorithm, however, was only able to correctly predict 61% of reports as leukaemia positive. With regard to the leukaemia negative (other heamatological malignancy) reports, the algorithm detected 88% of reports. However, it successfully predicted as negative 98% of the reports. This translated to an F1 score of 0.74 and 0.93 for the leukaemia and other heamatological malignancy classes, respectively. These metrics are presented in Table 4.4, including support. Support is the number of positive instances in each class. The numbers are based on the evaluation dataset, after exclusion of follow up reports.

TABLE 4.4: Performance matrix of the expert driven algorithm on the binary task

Class	Precision	Recall	F1 Score	Support
leukaemia	0.61	0.93	0.74	41
other malignancy	0.98	0.88	0.93	192

The confusion matrix for the binary task of identifying the reports as leukaemia positive or negative is shown in Figure 4.23.

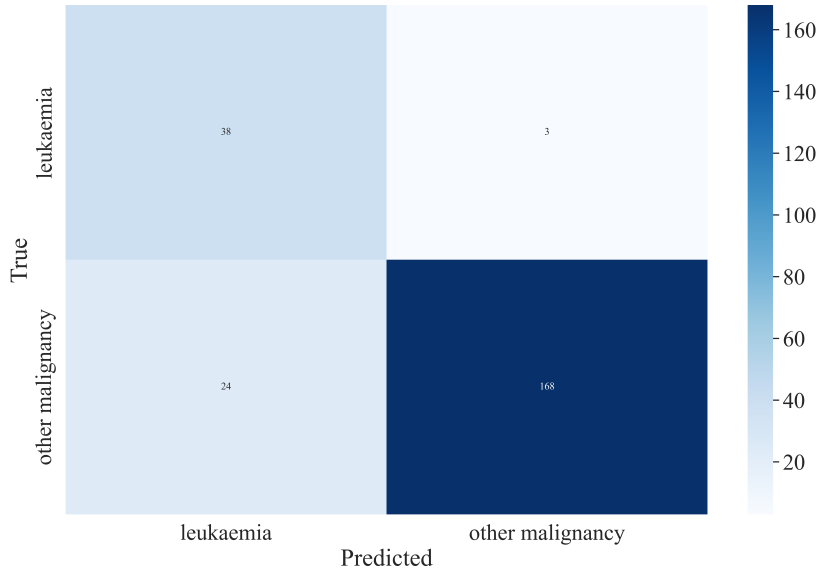


FIGURE 4.23: The confusion matrix for the binary task at evaluation with a labelled dataset

The performance of the algorithm varied for the different leukaemia classes as shown in Table 4.5. F1 scores for the leukaemia classes ranged from 0.50 to 0.92. The algorithm performed the best for CML , AML and ALL with F1 scores of 0.92, 0.75 and 0.69 respectively.

TABLE 4.5: Performance matrix of the expert driven algorithm on the multi class task

Class	Precision	Recall	F1 Score	Support
ALL	0.53	1.00	0.69	21
AML	1.00	0.60	0.75	5
CLL	0.40	0.67	0.50	3
CML	0.86	1.00	0.92	12
other malignancy	0.98	0.88	0.93	192

The confusion matrix for the multi-class task is shown in Figure 4.24.

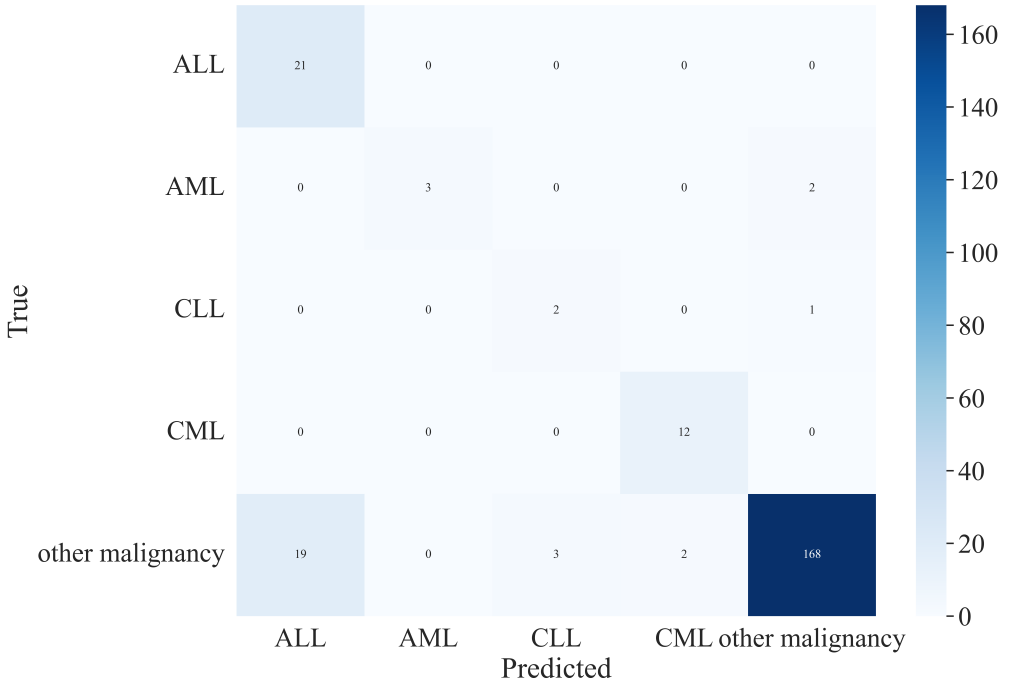


FIGURE 4.24: The confusion matrix for the different leukaemia classes at evaluation with a labelled dataset

Performance of the topic models

The coherence scores of the LDA and BERTopic models were comparable at 0.69 and 0.65, respectively, on the corpus with heamatological findings. The scores were different on the full corpus where BERTOPIC had a higher coherence of 0.73 compared to 0.62 for the LDA model. The scores are tabulated for ease of comparison in Table 4.6 below.

TABLE 4.6: Comparison of coherence scores of the different topic models

	LDA	**BERTopic**
Incident with malignancy		
Number of Topics	10	12
Coherence	0.69	0.65
All incident		
Number of Topics	12	13
Coherence	0.62	0.73

The domain expert could more readily identify the topics from the BERTopic algorithm. The topics identified on the models from the corpus of reports with an incident positive haematological malignancy finding are shown in Table 4.7.

TABLE 4.7: Topics identified by the domain expert from models of the corpus of reports with an incident positive haematological finding

Latent Dirichlet Allocation		BERTopic	
Topic	**Theme**	**Topic**	**Theme**
1	No disease	1	CLL work up
2	b lymphoproliferative disorders	2	Paediatric work up
3	no evidence of light	3	AML or CML
4	b lymphoproliferative disorders, picks up a rare, specific type of leukaemia	4	Poor sample quality
5	Cluster of leukaemias (AML,, t-ALL,pre-B, common ALL)	5	b lymphoproliferative including CLL work up but CLL not found
6	CLL	6	CML
7	Likely no disease but not sure	7	Likely no disease
8	CML	8	CLL work up
9	b lymphoproliferative disorders (Hodgkins)	9	CLL work up
10	multiple myeloma	10	CLL work up where it was not found
		11	Bad quality sample

The topics identified from the corpus of reports of all new investigations are shown in Table 4.8

TABLE 4.8: Topics identified by the domain expert from models of the corpus of reports of all new investigations

Latent Dirichlet Allocation		**BERTopic**	
Topic	**Theme**	**Topic**	**Theme**
1	No disease	1	No disease
2	No disease	2	AML
3	No disease	3	CLL
4	B lymphoproliferative disorders (Hodgkin's)	4	Common ALL or B-ALL
5	Cluster of leukemias (AML, T-ALL, pre-B, common ALL)	5	Multiple myeloma
6	CLL	6	CLL
7	Samples with paucity of cells	7	CLL
8	Cluster of leukemias (CML, ALL, common ALL)	8	Bad quality sample
9	CML	9	B lymphoproliferative
10	Peripheral blood specimens (should have been excluded)	10	B lymphoproliferative (probably CLL)
11	Bad samples	11	B lymphoproliferative (in Hodgkin's)
12	Bad quality sample and spelling errors	12	No evidence of CLL
		13	Peripheral blood specimens (should have been excluded)

Discussion

5.1 Introduction

This study applied a rule-based TM algorithm, and explored application of unsupervised ML to flow cytometry diagnostic pathology text reports. This chapter discusses the study findings in light of the objectives and documented knowledge in the literature. The discussion begins with a consideration of outcomes and performance of the text mining tasks, comparing the regex based algorithm with the topic modelling. The performance of the two topic modelling algorithms is also compared. The ASIR findings are discussed, and compared to statistics published by the NCR. Lastly, the limitations of the study are highlighted.

5.2 Applicability of text mining techniques to flow cytometry narrative reports

5.2.1 Outcomes and performance of text mining techniques

The rule-based RegEx algorithm achieved an F1 score of 0.74 for the leukaemia, and 0.93 for the other haematological malignancy classes respectively. In the multi class task, the F1 scores ranged from 0.50 to 0.92. A 2014 review found that IE systems for simple tasks were highly accurate (Spasić et al., 2014). Rule-based IE systems applied to pathology corpora have yielded high accuracies ranging between 0.7 and 0.9 (Weegar, Nygård, and Dalianis, 2017). Lower scores are not unheard of. Applications to extract data from Norwegian prostate cancer pathology reports have achieved F1-scores between 0.24 to 0.94 (Dahl, Özkan, and Dalianis, 2016). Weegar and Dalianis, 2015 reported F1-scores between 0.67 and 1.0 from breast cancer reports. A study that extracted information from paper based laboratory records reported an F1 score of 0.86 on a test dataset of 153 reports (Ma et al., 2023). The information extraction module was applied to laboratory reports digitised using optical character recognition techniques. It is thus not only possible to extract diagnostic and other data from laboratory text, but it can also be done accurately.

The acceptability of a TM F1-score depends on a number of considerations. These include the complexity of the data, the nature of the task, and application requirements. Pathology reports containing specialised domain terminology, or diverse writing styles may present greater challenges (O'Mara-Eves et al., 2015b). TM tasks can be expected to have varying levels of intricacy and difficulty, and this needs to be taken into consideration. The intended application of the text mining results is also taken into consideration. In clinical decision support systems for example, higher scores would be required. The F1-scores from this study were comparable to these reported in the literature, ranging between 0.50 and 0.93. This score validates the potential for application of TM techniques to

leukaemia flow cytometry reports for the extraction of structured data. In the study by Ma et al., 2023, the F1 score of 0.86 was deemed sufficient for validation of feasibility of their TM pipeline (Ma et al., 2023).

Rule-based systems have been the most commonly applied to pathology reports (Kreimeyer et al., 2017; Savova et al., 2019), and they find application possibly due to the semi structured nature of the reports. Another feature seen in reviews of clinical NLP systems is that they are tailored to specific tasks (Kreimeyer et al., 2017), and it was the same in this study for the rule-based algorithm.

This study has demonstrated that unsupervised topic models have potential to facilitate understanding of themes from the textual resources on leukaemia. The themes or topics discovered were distinct, and could be interpreted by human experts at evaluation. The topics from the BERTopic models were much more easily interpreted by the expert. There is sometimes no agreement between coherence and other automated performance measures, and human interpretability (Fan, Doshi-Velez, and Miratrix, 2019; Rijcken et al., 2022). Nonetheless, this human interpretability was complemented with some automated evaluation, the assessment of coherence scores, and this was useful in setting the number of topics. It will be useful to evaluate other measures such as perplexity when ML models are further refined to provide some measure of performance assurance in routine use. Topic modelling is widely used to uncover themes from textual resources and has been applied in a variety of fields, with discovery of new or additional information in some cases.

In this study, the topic modelling enabled the discovery of additional information compared to the rule-based algorithm. The topic modelling picked up subtypes of the main classes of leukaemia. The expert review of the topics has also highlighted stop words, for a corpus of this type of pathology reports for leukaemia that need to be added to the stopwords corpus at pre-processing. These are words that are common in the reports and, or do not add any value to the leukaemia work up or disease data in the report. Inclusion of these words is expected to improve the performance of the models, and quality of topics as some of the top words in topics from the LDA algorithm were these words, that did not differentiate the reports meaningfully.

This study, as far as the authors know, presents initial efforts directed at applying TM techniques to increase leukaemia data representation in South Africa. It is significant as it applied TM to access textual data towards increasing the data representation, and use the data to determine ASIR of leukaemia. The first article on leukaemia and TM on the citation database Scopus reportedly appeared in 1987 (Bhosale et al., 2021). The majority of these articles were from China and the United States of America (Bhosale et al., 2021). The activity on AI related work in developing countries is lower than that in middle and high income countries (Savova et al., 2019). This project contributes to the AI work in developing countries demonstrating application in South Africa.

It is important to pursue efforts to increase leukaemia data representation in South Africa given the bottlenecks to the availability of this data from clinical testing platforms as highlighted in the introductory chapters. These efforts could eventually contribute to efforts on a platform similar to the United States Food and Drug Administration (FDA) and the Centers for Disease Control and Prevention (CDC) 'Natural Language Processing (NLP) Web Service for Structuring and Standardizing Unstructured Clinical Information'. This platform is intended to be a repository of systems and tools that can be applied to clinical text to extract structured data (Kreimeyer et al., 2017). A drawback to application of ML based NLP systems is the unavailability of large labelled datasets. Indeed we had

to annotate the evaluation dataset for this study. The 2017 review of clinical NLP systems, albeit with inherent limitations, found only four purely ML based systems out of 71 (Kreimeyer et al., 2017).

5.2.2 Data pre-processing and feature engineering

Pre-processing is an important step in AI applications, and it is often informed by state of the input data and the objectives of the task. In the case of regular expressions that utilise pattern matching to extract information, the pre-processing was kept to a minimum to improve the performance of the algorithm. The input text data in this study had duplicated sentences, line breaks and other characters. The pre-processing enabled a reduction of the corpus size before rules were applied to extract diagnosis information by matching patterns frequently used by pathologists in reporting. There is still more work required to refine the regular expressions to capture more reports into one of the main categories of leukaemia, or possible more categories of haematological malignancy. This is particularly important given that not all reports mention the word leukaemia.

The pre-processing of the corpus for the topic modelling is possibly a more challenging task and needs to be refined further. There are some words that probably do not add any value and could be added to the stopwords corpus depending on the purpose of the task. In this project, the stopwords corpus was kept to a minimum as the goal was to assess how much useful knowledge could be discovered from the corpus. This enabled the discovery of stopwords, relevant to this particular corpus, that can be included in further work. The implementation of the topic modelling in Gensim could have led to some loss of useful clinical information. The pre-processing in Gensim trimmed numbers from the corpus. These could have been useful in the clustering process of the LDA model, as according to the domain expert, these numbers were useful information. Their exclusion thus likely had a negative impact on the output from the models. There might be value in exploring existing cNLP platforms with a view to customise and test them on the data. Alternatively, implementation of the topic modelling in the sklearn library, with customised pre-processing to retain as much of the text data might also improve the topics and their quality. Some other approaches to topic modelling have been reported to outperform LDA in some contexts (Egger and Yu, 2022). The current work in which BERTopic appears to give more interpretable topics, illustrates this.

N-grams are important features of a corpus. In this project, they were used to assess the state of the corpus during the pre-processing to ensure that phrases were now useful for pattern matching. Examining top n-grams can be a way to validate pre-processing steps. In some tasks such as sentiment analysis, topic modelling, or text classification, identifying informative n-grams can be crucial for feature selection (Zengul et al., 2021). In the case of the rule- based algorithm, some n-grams were expected to be useful indicators of belonging to a category.

There were some 'noisy' or irrelevant n-grams in the topic modelling but these were retained nonetheless to enable assessment of their impact on topic quality. Generally, noisy n-grams might indicate a need to refine pre-processing steps. N-grams larger than necessary for the task may increase sparsity of features and hamper model performance, rather than improve it. Sparse data is not easily modelled, and bi-grams were an effort to strike a balance on a corpus with a limited number of tokens (the pathologists tend to use the same standard terms in the reports), with high frequency. Some authors recommend that for corpora with rich vocabulary, with lower frequency per token, a lower order

n-gram model would give better results, striking a balance between over, and under-fitting. In addition to interpretability, bi-grams can be expected to have lower perplexity on held out data because there is more context to predict the next word. Topics from a bi-gram topic model, however were not interpretable by experts and this was not pursued. This could have been due to increased sparsity introduced into the data.

In this study where the goal was to extract diagnosis information from incident reports, the identification of follow up reports for exclusion, was a challenge. The algorithm has been built to include the steps to identify and exclude follow up cases. This is important and should be pursued and considered as a quality control step on the platform because the identification of incident leukaemia cases is critical as the other processes derive from it. Although a somewhat simpler challenge, this illustrates the challenges faced in handling temporality in NLP systems (Kreimeyer et al., 2017).

5.2.3 Development of accurate regular expressions algorithm and topic models

The RegEx used in this study are specific to flow cytometry and leukaemia. The focus on a specific use is common for NLP systems (Kreimeyer et al., 2017). Systems accurately extract data with very specific rules but are not scalable. The rule-based algorithm and topics models in this study are intended to address specific needs in a specific context. Scalability to other domains or contexts is likely to be limited. It is important, however, that the systems meet needs accurately in their routine implementation as that is the value add that can be realised with current technologies (Kreimeyer et al., 2017).

In the development of RegEx for pattern matching, spelling errors were a challenge but they were minimal. Elimination of differential diagnoses from the corpus was part of the iterative process of cleaning during the project. In some of the reports, the differential diagnosis terms resulted in the report being placed in more than one category. Future work is required to handle the issue of differential diagnoses as it has practical implication in the clinical setting. A ML approach could probably be useful, or rules can be written to handle this, or a combination. The current system is quite simplistic in this regard.

In an evaluation of the suitability of a topic model trained on one corpus for application to another (Shaikina and Funkner, 2020), found that this did not always yield good results. The corpora needed different word level processing and had specific sets of descriptions, which limits the use of predictive models for some diseases (Shaikina and Funkner, 2020). Style and structure of EMR can differs between health centres, limiting the transferrabiity of models (Shaikina and Funkner, 2020). This might impact the rule-based algorithm if data from other laboratories were added, necessitating validation of the models on the additional reports. The validation can provide guidance on any pre-processing that may be necessary.

5.3 Age Standardised Incidence Rates of Leukaemia in Johannesburg from 2015 to 2019

This project has shown that it is possible to apply smart tools for a more granular epidemiological picture to be developed for South Africa using data from flow cytometry immunophenotying text reports. The ASIR for the main categories of leukaemia can thus potentially be described. Methodological

rigour is critical to ensure that data used for epidemiological studies is accurate. NLP generated variables data was recently shown to be reliable and simulated errors did not significantly affect the epidemiological interpretations of relationships between NLP generated predictor variables and outcomes data (Lan and Turchin, 2023).

The leukaemia ASIR calculated from the data extracted was between 0.54 and 0.87 per 100 000 population per year for the years 2015 to 2019. This was lower than the national reported figures for the same period. The national figures ranged from 0.82 to 0.95, and 1.22 to 1.61 per 100 000 population among females and males respectively. The calculated ASIR showed a declining trend over the years 2015 to 2019. This trend is also seen in the data reported by the NCR (Table 5.1).

TABLE 5.1: ASIR reported by the NCR for histologically diagnosed leukaemia in 2015 to 2019

Year	**Female**	**Male**
2015	0.95	1.61
2016	0.83	1.22
2017	0.82	1.52
2018	0.82	1.31
2019	0.92	1.26

The ASIR figures may change when the performance of the algorithm is refined, and data extraction improves. The population figures used were for the City of Johannesburg as this was the closest approximation of the population served by CMJAH. The difference in rates may also be due to possibly, higher testing rates in private health care. The NCR figures include private health care data as it became mandatory from 2011 for all laboratories to report cancer diagnoses to the NCR.

In addition to limitations imposed by the accuracy of the algorithm, epidemiological analysis of extracted data in this project may be biased due to the social and economic factors that influence the samples received and analysed at CMJAH. The CMJAH laboratory is a state facility and largely serves the City of Johannesburg and the findings may be biased towards this catchment. The contents of the narrative reports are entirely dependent on the opinion of the attending pathologist. Reporting styles and attention to detail in reporting vary by pathologist. These factors can affect text mining efficiency as alluded to by (Zhu et al., 2013).

5.4 Limitations of the study

The data in the study was mostly from Gauteng Province, most likely the City of Johannesburg patients and this could have introduced bias due to the social and economic factors that influence the samples received and analysed at CMJAH. One such factor could be the use of private testing facilities. Determinants of leukaemia incidence peculiar to Gauteng might also have had an influence. In the future, heterogeneity might increase, and there will be a need to consider ML approaches and also the rules might get complicated, and the process time consuming. ML based methods would then be preferable (Pironet et al., 2021).

Conclusions and Recommendations

6.1 Introduction

This chapter presents the final conclusions and recommendations towards application of smart techniques to improve leukaemia data representation.

6.2 Text mining techniques can improve data representivity for leukaemia

This project has demonstrated that it is possible to apply expert-driven terms to extract information from flow cytometry reports and improve leukaemia data representation in the following ways:

- Diagnostic data for the four main categories, and undifferentiated leukaemia could be extracted, enabling the reports to be placed in categories. The extracted data was more granular than the current leukaemia category, reported in the cancer registry reports. Currently, all the different types of leukaemia are aggregated into one category of leukaemia.
- Socio-economic data for the epidemiological analysis can also be extracted and used to generate leukaemia category specific ASIR.

There is, however, still a need to improve both the precision and recall of the algorithm because it is important that accurate information is captured for epidemiological studies, possible automation of cancer registry updates and surveillance. The precision in this study for the leukaemia class was quite low at 0.61. Kunz et al., 2018 extracted diagnostic data from bone marrow biopsy reports of patients diagnosed with myeloid neoplasms with overall precision of 0.9117 and recall of 0.7951. The low precision represents a high number of false positives as the algorithm places the reports in the different leukaemia classes. It is recommended that for the next steps domain experts further review the terms, rules and steps of the algorithm to improve precision.

The project has also shown that unsupervised ML (topic modelling) can be applied to extract information and nuanced knowledge. This study has shown that distinct topics can be extracted. One downstream application might be using unsupervised learning for the assignment of reports to topics which could be diagnoses. An immediate build on to the current work might be to explore the corpus to investigate whether there is any other information other than the diagnoses that can be gleaned from the corpus of text reports, and also whether there are any linkages that might exist between the flow cytometry, genetics, histology and possibly clinical notes.

6.3 Recommendations for further work towards routine implementation of smart tools

A key requisite for smart tools to add value to health systems is that they are able to accurately meet practical needs and enhance patient and health system outcomes (Kreimeyer et al., 2017). It is recommended that the following be considered for work towards routine implementation of the smart tools.

- Immediate follow on work to focus on refining the rule-based algorithm to improve its performance. In addition, for routine implementation, the system needs to be able to flag reports where there is ambiguity and a human expert needs to review and code. In the prostate cancer tumour information extracting study, the system identified ambiguous reports that needed to be reviewed by experts (Weegar, Nygård, and Dalianis, 2017). This and other quality assurance check will need to be built in to the algorithm. Systems incorporating smart techniques need to reliably satisfy unmet needs or support real-life use cases on a routine basis, to enable full realisation of their impact.
- More data should be sourced and labelled to evaluate the final rule-based algorithms and topic model. It would also add value if the data sourcing and labelling project focusses on leukaemia flow cytometry data across the country. This will allow a national epidemiological study, that would report granular findings and enable further interrogation of variables such as province, gender, race which are currently seen as possible confounders of representation from pathology reports. For the finalisation of the current rule-based system, reports in the LIS from 2020 to date could be used.
- Development of the ML models should continue as they hold promise for wider application, and discovery of more knowledge including across domains.
- In the absence of standardised terminology for the reports, a keywords check be built into the report approval systems for the pathologists to address the challenges introduced by terminology inconsistencies and reporting styles. Engagement with the haematology expert on the study indicated that this is an intervention that can potentially be built into the reporting system.
- Epidemiological findings need to be disseminated, in addition to the tools as they have potential to contribute to health planning and access. The dissemination activities would form part of the framework for routine implementation of the rule-based algorithm and granular reporting of the epidemiological trends. This can be using pipelines that link to a dashboard showing data disaggregated by geographic location.This would be a medium to long term project that will also require stakeholder mobilisation.
- When the algorithm has achieved higher precision levels, it can be considered for automated reporting by the NCR.
- Additional efforts to improve leukaemia data and equity could investigate the application of ML based cNLP and to other pathology reports and expand to clinical notes, discharge summaries especially in communities with limited access to healthcare and pathology services.

Bibliography

Achilonu, Okechinyere J. et al. (2021). "A text mining approach in the classification of free-text cancer pathology reports from the south african national health laboratory services". In: *Information (Switzerland)* 12.11. ISSN: 20782489. DOI: 10.3390/info12110451.

ACRN (2021). *South Africa National Cancer Registry (NCR-SA)*. URL: https://afcrn.org/index.php/membership/membership-list/87-ncrsa (visited on 04/22/2021).

Adeline Abbe, Cyril Grouin, Pierre Zweigenbaum, Bruno Falissard (2016). "Text mining applications in psychiatry: a systematic literature review". In: *International Journal of Methods in Psychiatric Research* 2, pp. 86–100. DOI: 10.1002/mpr.1481.

Appugliese, Carlo, Paco Nathan, and William S Roberts (2020). *Agile AI: A Practical Guide to Building AI Applications and Teams*. Tech. rep. Sebastopol. URL: https://www.oreilly.com/library/view/agile-ai/9781492074984/.

Arber, Daniel A et al. (2016). "The 2016 revision to the World Health Organization classifi cation of myeloid neoplasms and acute leukemia". In: *Blood* 127.20, pp. 2391–2405. DOI: 10.1182/blood-2016-03-643544.

Arnold, Corey and William Speier (2012). "A topic model of clinical reports". In: *SIGIR'12 - Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval* 90024, pp. 1031–1032. DOI: 10.1145/2348283.2348454.

Bauer, Chris et al. (2021). "Large-scale literature mining to assess the relation between anti-cancer drugs and cancer types". In: *Journal of Translational Medicine* 19.1, pp. 1–13. ISSN: 14795876. DOI: 10.1186/s12967-021-02941-z. URL: https://doi.org/10.1186/s12967-021-02941-z.

Bhattacharya, Moumita et al. (2016). "Topic Models of Electronic Medical Records". In: pp. 466–469.

Bhosale, Rohanraje et al. (2021). "Bibliometric Analysis of Machine Learning and Text Mining Algorithms for Diagnosis of Leukemia". In: *Library Philosophy and Practice*, pp. 1–10. ISSN: 15220222.

Blei, David M., Andrew Y. Ng, and Michael T. Jordan (2003). "Latent dirichlet allocation". In: *Advances in Neural Information Processing Systems* 3, pp. 993–1022. ISSN: 10495258.

Boggs, Jennifer M. and Julie M. Kafka (2022). "A Critical Review of Text Mining Applications for Suicide Research". In: *Current Epidemiology Reports* 9.3, pp. 126–134. ISSN: 2196-2995. DOI: 10.1007/s40471-022-00293-w. URL: https://doi.org/10.1007/s40471-022-00293-w.

Boniol, M and M Heanue (2007). "Chapter 7: Age-standardisation and denominators". In: *Cancer Incidence in Five Continents Vol. IX*, pp. 99–101.

Bui, Duy Duc An and Qing Zeng-Treitler (2014). "Learning regular expressions for clinical text classification". In: *Journal of the American Medical Informatics Association* 21.5, pp. 850–857. ISSN: 1527974X. DOI: 10.1136/amiajnl-2013-002411.

CANSA (2021). "Fact Sheet on Adult Chronic Lymphoblastic Leukaemia (CLL)". In: pp. 1–13.

Cassim, Naseem et al. (2021). "Using text mining techniques to extract prostate cancer predictive information (Gleason score) from semi-structured narrative laboratory reports in the Gauteng province, South Africa". In: *BMC Medical Informatics and Decision Making* 21.1, pp. 1–11. ISSN: 14726947. DOI: 10.1186/s12911-021-01697-2. URL: https://doi.org/10.1186/s12911-021-01697-2.

Chan, Katherine Redfield et al. (2013). "An empirical analysis of topic modeling for mining cancer clinical notes". In: *Proceedings - IEEE 13th International Conference on Data Mining Workshops, ICDMW 2013*, pp. 56–63. DOI: 10.1109/ICDMW.2013.91. URL: https://www.biorxiv.org/content/10.1101/062307v1.full.pdf.

Cho, Hae-Wol (2019). "Osong public health and research perspectives". In: *Osong Public Health and Research Perspectives* 10.3, pp. 115–116. ISSN: 22336052. DOI: 10.24171/J.PHRP.2021.12.1.01.

Cho, Kyoung Won, Shine Young Kim, and Young Woon Woo (2019). "Osong Public Health and Research Perspectives Analysis of Women ' s Health Online News Articles Using". In: 10.3, pp. 158–169.

Conceição, Sofia I.R. and Francisco M. Couto (2021). "Text mining for building biomedical networks using cancer as a case study". In: *Biomolecules* 11.10. ISSN: 2218273X. DOI: 10.3390/biom11101430.

Currie, Anne Marie et al. (2006). "Automated extraction of free-text from pathology reports." In: *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, p. 899.

Dahl, Anders, Atilla Özkan, and Hercules Dalianis (2016). "Pathology text mining - On Norwegian prostate cancer reports". In: *2016 IEEE 32nd International Conference on Data Engineering Workshops, ICDEW 2016*, pp. 84–87. DOI: 10.1109/ICDEW.2016.7495622.

Dhokotera, Tafadzwa et al. (2019). "The burden of cancers associated with HIV in the South African public health sector, 2004-2014: A record linkage study". In: *Infectious Agents and Cancer* 14.1, pp. 1–12. ISSN: 17509378. DOI: 10.1186/s13027-019-0228-7.

Divya, Khyani et al. (n.d.). "An Interpretation of Lemmatization and Stemming in Natural Language Processing". In: 22.10 (), pp. 350–357.

Egger, Roman and Joanne Yu (2022). "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts". In: *Frontiers in Sociology* 7.May, pp. 1–16. ISSN: 22977775. DOI: 10.3389/fsoc.2022.886498.

El Alaoui, Yousra et al. (2022). "A Review of Artificial Intelligence Applications in Hematology Management: Current Practices and Future Prospects." eng. In: *Journal of medical Internet research* 24.7, e36490. ISSN: 1438-8871 (Electronic). DOI: 10.2196/36490.

Fan, Angela, Finale Doshi-Velez, and Luke Miratrix (2019). "Assessing topic model relevance: Evaluation and informative priors". In: *Statistical Analysis and Data Mining* 12.3, pp. 210–222. ISSN: 19321872. DOI: 10.1002/sam.11415.

Fleming, Kenneth A et al. (2018). "High-Quality Diagnosis: An Essential Pathology Package". In: *Disease Control Priorities: Improving Health and Reducing Poverty. 3rd edition.* Ed. by Dean T

Jamison et al. Third. Washington, DC: World Bank. Chap. 11, pp. 215–234. ISBN: 9781464805271. DOI: 10.1596/978-1-4648-0527-1.

Fleuren, Wilco W.M. and Wynand Alkema (2015). "Application of text mining in the biomedical domain". In: *Methods* 74, pp. 97–106. ISSN: 10959130. DOI: 10.1016/j.ymeth.2015.01.015. URL: http://dx.doi.org/10.1016/j.ymeth.2015.01.015.

Grcar, M (2012). *No Title*. URL: https://www.slideshare.net/mgrcar/text-and-text-stream-mining-tutorial-15137759.

Grootendorst, Maarten (2022). "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *ArXiv* abs/2203.0. arXiv: 2203.05794. URL: http://arxiv.org/abs/2203.05794.

Grubbs, Allison E. et al. (2023). "Use of topic modeling to assess research trends in the journal Gynecologic Oncology". In: *Gynecologic Oncology* 172, pp. 41–46. ISSN: 10956859. DOI: 10.1016/j.ygyno.2023.03.001. URL: https://doi.org/10.1016/j.ygyno.2023.03.001.

Hallek, Michael and Othman Al-Sawaf (2021). "Chronic lymphocytic leukemia: 2022 update on diagnostic and therapeutic procedures". In: *American Journal of Hematology* 96.12, pp. 1679–1705. ISSN: 10968652. DOI: 10.1002/ajh.26367.

Hanahlyn M Park, Vicky Sandhu, Paul Fearn, Kathleen Shannon Dorcy, Elihu H. Estey, Emily Silgard (2014). "Using Natural Language Processing to determine Chemotherapeutic Regimens Administered within 30 Days Prior to Death in Acute Myelogenous Leukemia patients". In: *Blood* 21, p. 1267. DOI: doi.org/10.1182/blood.V124.21.1267.1267.

Hanisch, Daniel et al. (2005). "ProMiner: Rule-based protein and gene entity recognition". In: *BMC Bioinformatics* 6.SUPPL.1, pp. 1–9. ISSN: 14712105. DOI: 10.1186/1471-2105-6-S1-S14.

Harrison, Matt and Michael Prentiss (2016). *Learning the Pandas Library: Python Tools for Data Munging, Analysis, and Visual*. Ed. by Michael Prentiss. ISBN: 9781533598240. URL: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://mcgillmsaa.wordpress.com/wp-content/uploads/2018/05/learning-the-pandas-library-python-tools-for-data-munging-analysis-and-visual.pdf.

Hauser, Ronald G. et al. (2021). "A machine learning model to successfully predict future diagnosis of chronic myelogenous leukemia with retrospective electronic health records data". In: *American Journal of Clinical Pathology* 156.6, pp. 1142–1148. ISSN: 19437722. DOI: 10.1093/ajcp/aqab086.

Heng-Yi Wu, Chien-Wei Chiang and Lang Li (2014). "Text Mining for Drug-Drug Interaction". In: *Methods in Molecular Biology* 1159, pp. 47–75. DOI: 10.1007/978-1-4939-0709-0.

Hoffbrand, A.V. and Moss, P.A.H. (2011). *Essential Hematology*. Sixth Edit. Oxford: Blackwell Publishing Limited.

Höglund, Martin, Fredrik Sandin, and Bengt Simonsson (2015). "Epidemiology of chronic myeloid leukaemia: an update". In: *Annals of Hematology* 94.2, pp. 241–247. ISSN: 14320584. DOI: 10.1007/s00277-015-2314-2.

Hoyle, Alexander et al. (2021). "Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence". In: *Advances in Neural Information Processing Systems* 3.NeurIPS, pp. 2018–2033. ISSN: 10495258. arXiv: 2107.02173.

Huilgol, Purva (2019). *Accuracy vs. F1 Score*. URL: https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2{\#}:{~}:text=Accuracyisusedwhenthe,andFalsePositivesarecrucial{\&}text=Inmostreal-lifeclassification,toevaluateourmodelon.{\%}3E (visited on 02/24/2021).

IARC (2020). *Population Fact Sheets - South Africa*. Tech. rep. URL: https://gco.iarc.fr/today/data/factsheets/populations/710-south-africa-fact-sheets.pdf.

Jones-diette, Julie S et al. (2019). "Validation of text-mining and content analysis techniques using data collected from veterinary practice management software systems in the UK". In: *Preventive Veterinary Medicine* 167.July 2018, pp. 61–67. ISSN: 0167-5877. DOI: 10.1016/j.prevetmed.2019.02.015. URL: https://doi.org/10.1016/j.prevetmed.2019.02.015.

Kho, Soon Jye et al. (2017). "A novel approach for classifying gene expression data using topic modeling". In: *ACM-BCB 2017 - Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 388–393. ISBN: 9781450347228. DOI: 10.1145/3107411.3107483.

Kim, Hyun Yong et al. (2022). "Web-Based Research Trends on Child and Adolescent Cancer Survivors over the Last 5 Years: Text Network Analysis and Topic Modeling Study". In: *Journal of Medical Internet Research* 24.2, e32309. ISSN: 14388871. DOI: 10.2196/32309.

Kreimeyer, Kory et al. (2017). "Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review". In: *Journal of Biomedical Informatics* 73, pp. 14–29. ISSN: 15320464. DOI: 10.1016/j.jbi.2017.07.012. URL: http://dx.doi.org/10.1016/j.jbi.2017.07.012.

Kunz, Isaac et al. (2018). "Extracting Diagnostic Data from Unstructured Bone Marrow Biopsy Reports of Myeloid Neoplasms Utilizing a Customized Natural Language Processing (NLP) Algorithm". In: *Blood* 132.Supplement 1, pp. 2272–2272. ISSN: 0006-4971. DOI: 10.1182/blood-2018-99-119049. URL: http://dx.doi.org/10.1182/blood-2018-99-119049.

Kyle H. Ambert, Aaron MCohen (2012). "Chapter Six - Text mining and Neuroscience". In: *International Review of Neurobiology*. Ed. by Melissa A. Haendel Elissa J. Chesler. Amsterdam: Elsevier B.V. Chap. 6, pp. 109 –132. ISBN: 978-0-12-388408-4. DOI: 10.1016/B978-0-12-388408-4.00006-X. URL: https://www.sciencedirect.com/bookseries/international-review-of-neurobiology/vol/103/suppl/C.

Lan, Zhou and Alexander Turchin (2023). "Impact of possible errors in natural language processing-derived data on downstream epidemiologic analysis". In: 6.4.

Levy_Mendelovich, Sarina et al. (2021). "Pediatric literature trends: high-level analysis using text-niming". In: *Pediatric Research* 90, pp. 212–215.

Li, Weijie (2022a). "Flow cytometry in the diagnosis of lymphomas". In: *Cytometry*. Fifth. Brisbane: Exon. Chap. 4, pp. 53–77. ISBN: 9780645332070. DOI: 10.1002/cyto.990090809.

– (2022b). "The 5th Edition of the World Health Organization Classification of Hematolymphoid Tumors". In: *Leukemia*. Fifth. Brisbane: Exon Publications. Chap. 1, pp. 1–21. ISBN: 9780645332070. DOI: 10.36255/exon-publications-leukemia-who-5th-edition-hematolymphoid-tumors.

Li, Yuelin et al. (2019). "Leveraging Latent Dirichlet Allocation in processing free-text personal goals among patients undergoing bladder cancer surgery". In: *Quality of Life Research* 28.6, pp. 1441–

1455. ISSN: 15732649. DOI: 10.1007/s11136-019-02132-w. URL: http://dx.doi.org/10.1007/s11136-019-02132-w.

Luhn, H. P. (1958). "A Business Intelligence System". In: *IBM Journal of Research and Development* 2.4, pp. 314–319. ISSN: 0018-8646. DOI: 10.1147/rd.24.0314.

Lund, Jeffrey et al. (2020). "Automatic and human evaluation of local topic quality". In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 788–796. arXiv: arXiv:1905.13126v1.

Ma, Ming Wei et al. (2023). "Extracting laboratory test information from paper-based reports". In: *BMC Medical Informatics and Decision Making* 23.1, pp. 1–11. ISSN: 14726947. DOI: 10.1186/s12911-023-02346-6. URL: https://doi.org/10.1186/s12911-023-02346-6.

Machado-Neto, João Agostinho et al. (2023). "Data mining in acute myeloid leukemia: identification of disease biomarkers, prognostic factors, novel targets, and potential drugs". In: *Translational Cancer Research* 12.9, pp. 2235–2238. ISSN: 22196803. DOI: 10.21037/tcr-23-993.

Madhusoodhan, P. Pallavi, William L. Carroll, and Teena Bhatla (2016). "Progress and Prospects in Pediatric Leukemia". In: *Current Problems in Pediatric and Adolescent Health Care* 46.7, pp. 229–241. ISSN: 15383199. DOI: 10.1016/j.cppeds.2016.04.003. URL: http://dx.doi.org/10.1016/j.cppeds.2016.04.003.

Maposa, Sibonginkosi (2025). *LeukamiaTM*. URL: https://github.com/SibonginkosiM/LeukamiaTM..

Meaney, Christopher et al. (2023). "Quality indices for topic model selection and evaluation: a literature review and case study". In: *BMC Medical Informatics and Decision Making* 23.1, pp. 1–18. ISSN: 14726947. DOI: 10.1186/s12911-023-02216-1. URL: https://doi.org/10.1186/s12911-023-02216-1.

Michael, Louis G. et al. (2019). "Regexes are hard: Decision-making, difficulties, and risks in programming regular expressions". In: *Proceedings - 2019 34th IEEE/ACM International Conference on Automated Software Engineering, ASE 2019*, pp. 415–426. DOI: 10.1109/ASE.2019.00047.

Moreno, Antonio and Teófilo Redondo (2016). "Text Analytics: the convergence of Big Data and Artificial Intelligence". In: *International Journal of Interactive Multimedia and Artificial Intelligence* 3.6, p. 57. DOI: 10.9781/ijimai.2016.369.

Musaigwa, Fungai et al. (2021). "Chronic lymphocytic leukaemia trends and features at a tertiary hospital in South Africa (2011–2016)". In: *South African Journal of Oncology* 5, pp. 1–6. ISSN: 2518-8704. DOI: 10.4102/sajo.v5i0.168.

Narasimhulu, K. et al. (2022). "A novel sampling-based visual topic models with computational intelligence for big social health data clustering". In: *Journal of Supercomputing* 78.7, pp. 9619–9641. ISSN: 15730484. DOI: 10.1007/s11227-021-04300-7. URL: https://doi.org/10.1007/s11227-021-04300-7.

National Cancer Institute (n.d.). *SEER*Stat Software*. URL: https://seer.cancer.gov/seerstat/ (visited on 04/18/2024).

– (2024). *Joinpoint Trend Analysis Software*. URL: https://surveillance.cancer.gov/joinpoint/ (visited on 04/18/2024).

National Health Laboratory Service: National Cancer Registry (2023). *Cancer In South Africa: 2019 Full Report*. Tech. rep. Johannesburg: National Health Laboratory Service: National Cancer Registry. URL: https://www.nicd.ac.za/centres/national-cancer-registry/.

NCR-SA (2022). *Childhood Cancer Registry 2018 Annual Report*. Tech. rep. Johannesburg: National Cancer Registry.

Newman, David et al. (2010). "Evaluating topic models for digital libraries". In: *Proceedings of the ACM International Conference on Digital Libraries*, pp. 215–224. DOI: 10.1145/1816123.1816156.

Nguyen, Anthony et al. (2011). "Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications". In: *Studies in Health Technology and Informatics* 168, pp. 117–124. ISSN: 18798365. DOI: 10.3233/978-1-60750-791-8-117.

NHLS (2019a). *Annual Performance Plan Fiscal Year : 2019 / 20*. Tech. rep. June. National Health Laboratory Service. URL: https://www.nhls.ac.za/wp-content/uploads/2019/11/NHLS{_}APP{_}2019.pdf.

– (2019b). *Annual Report*. Tech. rep.

NICD (2020). *National Cancer Registry*. URL: https://www.nicd.ac.za/centres/national-cancer-registry/ (visited on 07/16/2020).

Ning, Liqing et al. (2020). "Trends in disease burden of chronic myeloid leukemia at the global , regional , and national levels : a population - based epidemiologic study". In: *Experimental Hematology & Oncology* 9.29, pp. 1–14. ISSN: 2162-3619. DOI: 10.1186/s40164-020-00185-z. URL: https://doi.org/10.1186/s40164-020-00185-z.

Nyasulu, Peter S., Christine Paszko, and Nontombi Mbelle (2014). "A Narrative Review of the Laboratory Information System and Its Role in Antimicrobial Resistance Surveillance in South Africa". In: *Advances in Microbiology* 04.10, pp. 692–696. ISSN: 2165-3402. DOI: 10.4236/aim.2014.410074.

O'Mara-Eves, Alison et al. (2015a). "Erratum to: Using text mining for study identification in systematic reviews: A systematic review of current approaches. [Systematic Rev., 4, (2015) 5]". In: *Systematic Reviews* 4.1, pp. 1–2. ISSN: 20464053. DOI: 10.1186/s13643-015-0031-5. URL: ???.

– (2015b). "Erratum to: Using text mining for study identification in systematic reviews: A systematic review of current approaches. [Systematic Rev., 4, (2015) 5]". In: *Systematic Reviews* 4.1, pp. 1–22. ISSN: 20464053. DOI: 10.1186/s13643-015-0031-5.

Pan, Yuyan, Yong Zhang, and Jiaqi Liu (2018). "Text mining-based drug discovery in cutaneous squamous cell carcinoma". In: *Oncology Reports* 40.6, pp. 3830–3842. ISSN: 17912431. DOI: 10.3892/or.2018.6746.

Patel, Swayamprakash et al. (2022). "Probiotic Formulations: A Patent Landscaping Using the Text Mining Approach". In: *Current Microbiology* 79.5, pp. 1–15. ISSN: 14320991. DOI: 10.1007/s00284-022-02836-2. URL: https://doi.org/10.1007/s00284-022-02836-2.

Phande, Varsha C and A.S Khandelwal (2014). "A Survey Of Different Text Mining Techniques". In: *IBMRD's Journal of Management and Research* 3.1, pp. 125–133. DOI: 10.17697/ibmrd/2014/v3i1/46903.

Piedra, David, Antoni Ferrer, and Joaquim Gea (2014). "Text Mining and Medicine: Usefulness in Respiratory Diseases". In: *Archivos de Bronconeumología (English Edition)* 50.3, pp. 113–119. ISSN: 15792129. DOI: 10.1016/j.arbr.2014.02.008.

Pironet, Antoine et al. (2021). "Machine Learning-Based Extraction of Breast Cancer Receptor Status From Bilingual Free-Text Pathology Reports". In: *Frontiers in Digital Health* 3.August, pp. 1–11. ISSN: 2673253X. DOI: 10.3389/fdgth.2021.692077.

Porturas, Thomas and R. Andrew Taylor (2021). "Forty years of emergency medicine research: Uncovering research themes and trends through topic modeling". In: *American Journal of Emergency Medicine* 45, pp. 213–220. ISSN: 15328171. DOI: 10.1016/j.ajem.2020.08.036. URL: https://doi.org/10.1016/j.ajem.2020.08.036.

Przybyla, Piotr et al. (2016). "Text mining resources for the life sciences". In: *Database* 2016, pp. 1–30. ISSN: 17580463. DOI: 10.1093/database/baw145.

Rb-Silva, Rita; and Yulia Karimova (2021). *AMILE : APPLICATION OF TEXT MINING TO CLINICAL REPORTS OF PATIENTS WITH ACUTE MYELOID LEUKEMIA AMILE : APPLICATION OF TEXT MINING TO CLINICAL REPORTS OF PATIENTS WITH ACUTE MYELOID LEUKEMIA*. DOI: https://doi.org/10.5281/zenodo.4477657.

Renganathan, Vinaitheerthan (2017). "Text mining in biomedical domain with emphasis on document clustering". In: *Healthcare Informatics Research* 23.3, pp. 141–146. ISSN: 2093369X. DOI: 10.4258/hir.2017.23.3.141.

Rijcken, Emil et al. (2022). "Topic Modeling for Interpretable Text Classification From EHRs". In: *Frontiers in Big Data* 5.May, pp. 1–11. ISSN: 2624909X. DOI: 10.3389/fdata.2022.846930.

Rumshisky, A. et al. (2016). "Predicting early psychiatric readmission with natural language processing of narrative discharge summaries". In: *Translational Psychiatry* 6.10, e921–5. ISSN: 21583188. DOI: 10.1038/TP.2015.182. URL: http://dx.doi.org/10.1038/tp.2015.182.

Savova, Guergana K. et al. (2010). "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications". In: *Journal of the American Medical Informatics Association* 17.5, pp. 507–513. ISSN: 10675027. DOI: 10.1136/jamia.2009.001560.

Savova, Guergana K. et al. (2019). "Use of natural language processing to extract clinical cancer phenotypes from electronic medical records". In: *Cancer Research* 79.21, pp. 5463–5470. ISSN: 15387445. DOI: 10.1158/0008-5472.CAN-19-0579.

Schonfeld, Sara J et al. (2016). "Hematologic malignancies in South Africa 2000-2006: Analysis of data reported to the National Cancer Registry". In: *Cancer Medicine* 5.4, pp. 728–738. ISSN: 20457634. DOI: 10.1002/cam4.597.

Schonlau, Matthias, Nick Guenther, and Ilia Sucholutsky (2017). "Text mining with n-gram variables". In: *Stata Journal* 17.4, pp. 866–881. ISSN: 15368734. DOI: 10.1177/1536867X1801700406.

Severance, Charles R (2018). *Python for Everybody | Trinket*, p. 245. URL: https://books.trinket.io/pfe/index.html.

Shaikina, Alevtina A. and Anastasia A. Funkner (2020). "Medical Corpora Comparison Using Topic Modeling". In: *Procedia Computer Science* 178.2019, pp. 244–253. ISSN: 18770509. DOI: 10.1016/j.procs.2020.11.026. URL: https://doi.org/10.1016/j.procs.2020.11.026.

Shelly Soffer , Eyal Klang, Noam Tau, Roni Zemet, Shomron Ben-Horin, Yiftach Barash and Uri Kopylov (2020). "Evolution of colorectal cancer screening research in the past 25 years: text-mining analysis of publication trends and topics". In: *Therapeutic Advances in Gastroenterology* 13, pp. 1–12. DOI: 10.1177/https://doi.org/10.1177/1756284.

Simmons, Michael, Ayush Singhal, and Zhiyong Lu (2016). "HHS Public Access". In: *Advances in Experimental Medicine and Biology* 939.January 2015, pp. 139–166. DOI: 10.1007/978-981-

10-1503-8. URL: https://0-www-ncbi-nlm-nih-gov.innopac.wits.ac.za/pmc/articles/PMC5931382/pdf/nihms952504.pdf.

Singh, Elvira et al. (2014). "Registry : an update". In: *The Lancet Oncology* 15.9, e363. DOI: 10.1016/S1470-2045(14)70310-9.

Singh, Elvira et al. (2016). "Establishment of a cancer surveillance programme: the South African experience". In: *The Lancet Oncology* 16.8, e414–e421. DOI: 10.1016/S1470-2045(15)00162-X.Establishment.

Sivakumar, Muthuramalingam, Sudhaman Parthasarathy, and Thiyagarajan Padmapriya (2024). "Trade-off between training and testing ratio in machine learning for medical image processing". In: *PeerJ Computer Science* 10, e2245. ISSN: 23765992. DOI: 10.7717/peerj-cs.2245.

Song-Quan Ong, Maisarah Binti Mohamed Pauzi, Keng Hoon Gan (2022). "Text mining in mosquito-borne disease: A systematic review". In: *Acta Tropica* 231, p. 106447.

Spasić, Irena et al. (2014). "Text mining of cancer-related information: Review of current status and future directions". In: *International Journal of Medical Informatics* 83.9, pp. 605–623. ISSN: 18728243. DOI: 10.1016/j.ijmedinf.2014.06.009.

Spishak, Eric, Werner Dietl, and Michael D. Ernst (2012). "A type system for regular expressions". In: *Proceedings for FTfJP 2012: The 14th Workshop on Formal Techniques for Java-Like Programs - Co-located with ECOOP 2012 and PLDI 2012, Papers Presented at the Workshop*, pp. 20–26. DOI: 10.1145/2318202.2318207.

Stevens, Wendy S et al. (2016). "Cloud-Based Surveillance, Connectivity, and Distribution of the GeneXpert Analyzers for Diagnosis of Tuberculosis (TB) and Multiple-Drug-Resistant TB in South Africa". In: *Molecular Microbiology: Diagnostic Principles and Practice*. Ed. by David H Persing et al. Third. February. Washington DC: American Society for Microbiology Press. Chap. 49, pp. 707–718. ISBN: 9781555819071. DOI: 10.1128/9781555819071.ch49.

Sun, Wencheng et al. (2018). "Data processing and text mining technologies on electronic medical records: A review". In: *Journal of Healthcare Engineering* 2018. ISSN: 20402309. DOI: 10.1155/2018/4302425.

Sun, Xiao Hong et al. (2023). "Identifying a prognostic model and screening of potential natural compounds for acute myeloid leukemia". In: *Translational Cancer Research* 12.6, pp. 1535–1551. ISSN: 22196803. DOI: 10.21037/tcr-22-2500.

Tran, Bach Xuan et al. (2019). "Modeling research topics for artificial intelligence applications in medicine: Latent dirichlet allocation application study". In: *Journal of Medical Internet Research* 21.11, pp. 1–13. ISSN: 14388871. DOI: 10.2196/15511.

Trevethan, Robert (2017). "Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice". In: *Frontiers in Public Health* 5.November, pp. 1–7. DOI: 10.3389/fpubh.2017.00307.

Trindade, Fábio et al. (2021). "Automatic text-mining as an unbiased approach to uncover molecular associations between periodontitis and coronary artery disease". In: *Biomarkers* 26.5, pp. 385–394. ISSN: 13665804. DOI: 10.1080/1354750X.2021.1904002.

Wagland R, Recio-Saucedo A, Simon M, Bracher M, Hunt K, Foster C, Downing A, Glaser A, Corner J (2016). "Development and testing of a text-mining approach to analyse patients' comments on their experiences of colorectal cancer care". In: *BMJ Quality and Safety* 25.8, pp. 604–14. DOI: doi.org/10.1136/bmjqs-2015-004063.

Walter, Wencke et al. (2021). "How artificial intelligence might disrupt diagnostics in hematology in the near future". In: *Oncogene* 40.25, pp. 4271–4280. ISSN: 14765594. DOI: 10.1038/s41388-021-01861-y. URL: http://dx.doi.org/10.1038/s41388-021-01861-y.

Wang, Liqin et al. (2019). "Development and Validation of a Deep Learning Algorithm for Mortality Prediction in Selecting Patients with Dementia for Earlier Palliative Care Interventions". In: *JAMA Network Open* 2.7, pp. 1–14. ISSN: 25743805. DOI: 10.1001/jamanetworkopen.2019.6972.

Wang, Lucy Lu and Kyle Lo (2021). "Text mining approaches for dealing with the rapidly expanding literature on COVID-19". In: *Briefings in Bioinformatics* 22.2, pp. 781–799. ISSN: 14774054. DOI: 10.1093/bib/bbaa296.

Warrer, Pernille et al. (2012). "Using text-mining techniques in electronic patient records to identify ADRs from medicine use". In: *British Journal of Clinical Pharmacology* 73.5, pp. 674–684. ISSN: 03065251. DOI: 10.1111/j.1365-2125.2011.04153.x.

Weegar, Rebecka and Hercules Dalianis (2015). "Creating a rule based system for text mining of Norwegian breast cancer pathology reports". In: *EMNLP 2015 - 6th International Workshop on Health Text Mining and Information Analysis, LOUHI 2015 - Proceedings of the Workshop* September, pp. 73–78. DOI: 10.18653/v1/w15-2609.

Weegar, Rebecka, Jan F. Nygård, and Hercules Dalianis (2017). "Efficient encoding of pathology reports using natural language processing". In: *International Conference Recent Advances in Natural Language Processing, RANLP* 2017-Septe.2011, pp. 778–783. ISSN: 13138502. DOI: 10.26615/978-954-452-049-6-100.

Wong, Hwee Bee and Gek Hsiang Lim (2011). "Measures of Diagnostic Accuracy : Sensitivity , Specificity , PPV and NPV". In: *Proceedings of Singapore Healthcare* 20.4, pp. 316–318. DOI: 10.1177/201010581102000411.

Yuan Luo, Gregory Riedlinger and Peter Szolovits (2014). "Cancer Informatics". In: *Cancer Informatics* 13.S1, pp. 69–79. DOI: :10.4137/CIn.s13874.

Zengul, Ayse G. et al. (2021). "Identifying research themes and trends in the top 20 cancer journals through textual analysis". In: *Journal of Cancer Policy* 30.January, p. 100313. ISSN: 22135383. DOI: 10.1016/j.jcpo.2021.100313. URL: https://doi.org/10.1016/j.jcpo.2021.100313.

Zhang, Yu, Mengdong Chen, and Lianzhong Liu (2015). "A review on text mining". In: *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS* 2015-Novem, pp. 681–685. ISSN: 23270594. DOI: 10.1109/ICSESS.2015.7339149.

Zheng, Li Xiao et al. (2021). "Ensuring the Correctness of Regular Expressions: A Review". In: *International Journal of Automation and Computing* 18.4, pp. 521–535. ISSN: 17518520. DOI: 10.1007/s11633-021-1301-4.

Zheng, Si et al. (2019). "Text Mining for Drug Discovery". In: *Bioinformatics and Drug Discovery*. Ed. by Richard S Larson and Tudor I Oprea. New York, NY: Springer New York, pp. 231–252. ISBN: 978-1-4939-9089-4. DOI: 10.1007/978-1-4939-9089-4_13. URL: https://doi.org/10.1007/978-1-4939-9089-4{_}13.

Zhu, Fei et al. (2013). "Biomedical text mining and its applications in cancer research". In: *Journal of Biomedical Informatics* 46.2, pp. 200–211. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2012.10.007. URL: http://dx.doi.org/10.1016/j.jbi.2012.10.007.

Zuckerman, Tsila and Jacob M Rowe (2015). "Leukaemia". In: *UICC Manual of Oncology*. Ed. by Shao Hui Huang Brian O'Sullivan, James D. Brierley, Anil D'Cruz, Martin Fey, Raphael E. Pollock, Jan Vermorken. Ninth. Chicester: John Wiley and Sons Incorporated. Chap. 34, pp. 427–448. ISBN: 9781444332445. URL: https://0-ebookcentral-proquest-com.innopac.wits.ac.za/lib/wits-ebooks/detail.action?docID=4040526.

Zweigenbaum, Pierre et al. (2007). "Frontiers of biomedical text mining: current progress". In: *Briefings in Bioinformatics* 8.5, pp. 358–375. ISSN: 14675463. DOI: 10.1093/bib/bbm045.

Appendix 1

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Sibonginkosi Maposa (Student number: 2293644) am a student registered for the degree of MSc Epidemiology in the academic year 2024.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: Maposa Date: 19 August 2024

Appendix 2

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG	HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

Office of the Deputy Vice-Chancellor (Research and Innovation)

TO: Ms S Maposa
School of Public Health
Division of Epidemiology and Biostatistics
Medical School
University

E-mail: 2293644@students.wits.ac.za

CC: Supervisor: Messrs M Mapundu & N Cassim; Prof DK Glencross
<Michael.Mapundu@wits.ac.za>
and <HREC-Medical Research Office@wits.ac.za>

FROM: Mr Iain Burns
Human Research Ethics Committee (Medical)
Tel: 011 717 1252

E-mail: Iain.Burns@wits.ac.za

DATE: 2021/11/24

REF: R14/49

PROTOCOL NO: **M210812** (This is your ethics application reference number. Please quote it in all enquiries, oral or written, relating to this study.)

PROJECT TITLE: *Application of text mining techniques to extract clinical and epidemiological information from narrative leukaemia reports concluded between 2010 and 2019 at the Charlotte Maxeke Johnnesburg Academic Hospital laboratory, South Africa*

Please find attached the Clearance Certificate for the above project. I hope it goes well and that an article in a recognized publication comes out of it. This will reflect well on your professional standing and contribute to Government funding of the University.



MSWorks2000/Iain0007/Clearscan.wps

Appendix 3

C.1 Top words in topics from the corpus of reports with haematological malignancy findings

This Appendix contains visualisation of the model of reports with haematological malignancy findings, highlighting the tops words in each topic.

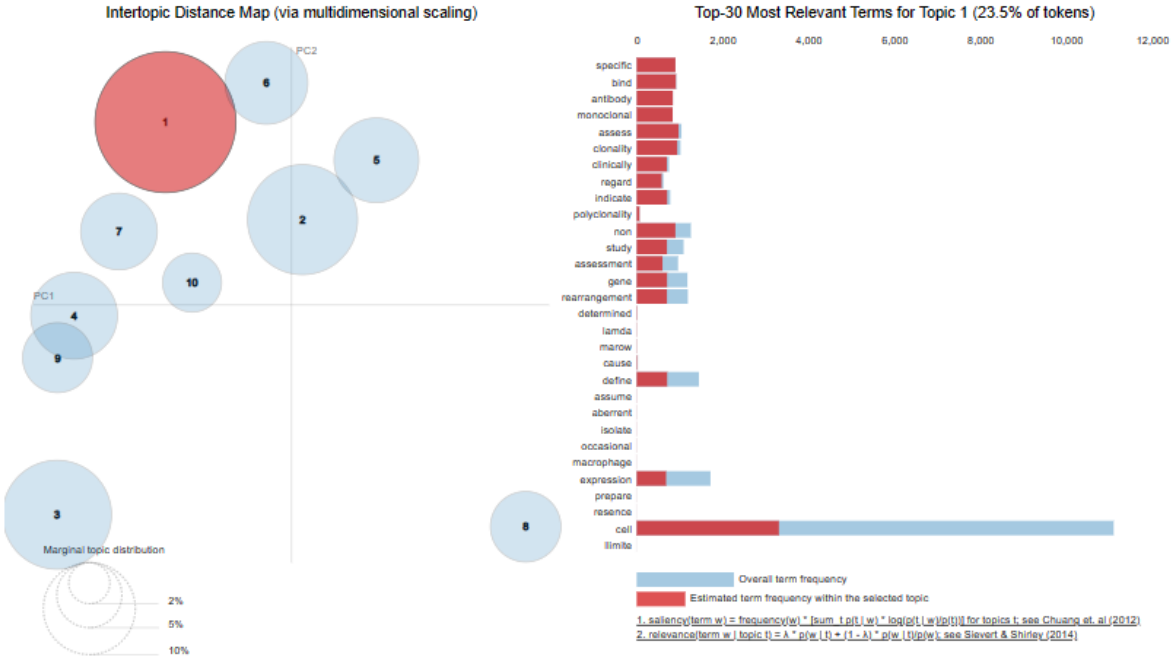


FIGURE C.1: Top words in Topic 1

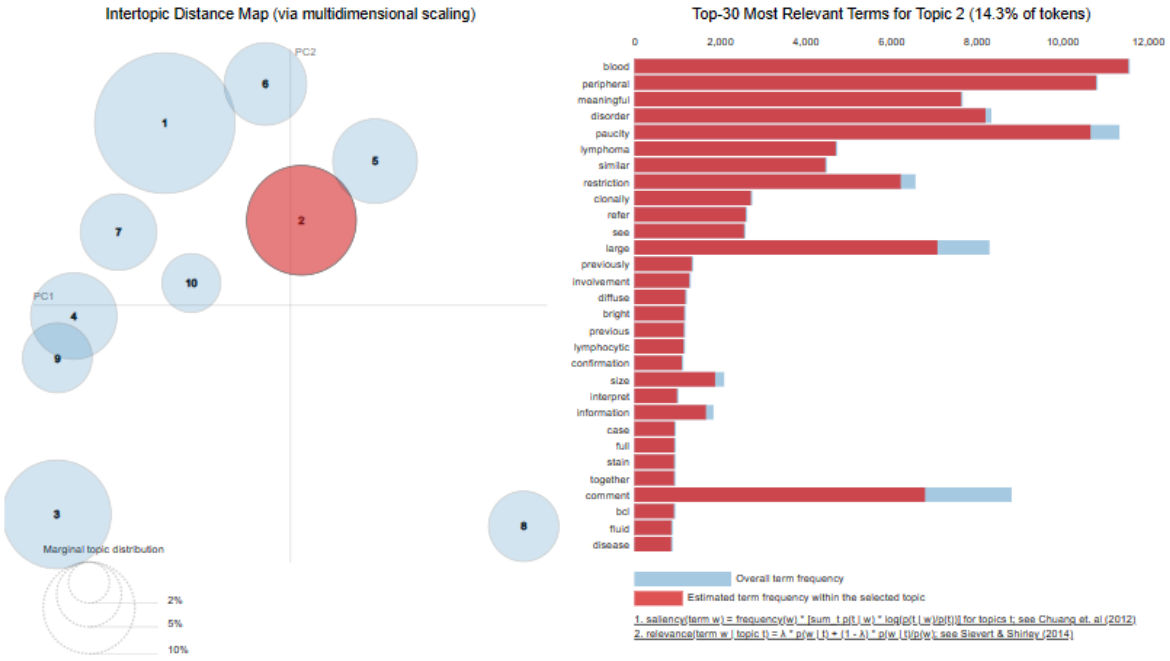


FIGURE C.2: Top words in Topic 2

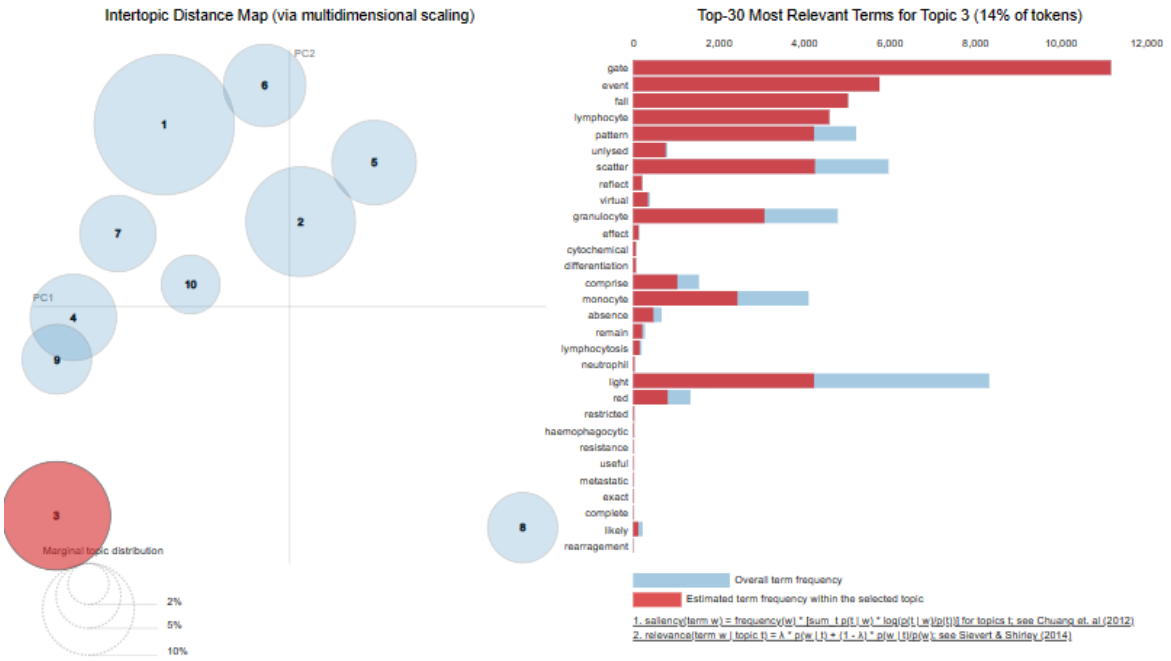


FIGURE C.3: Top words in Topic 3

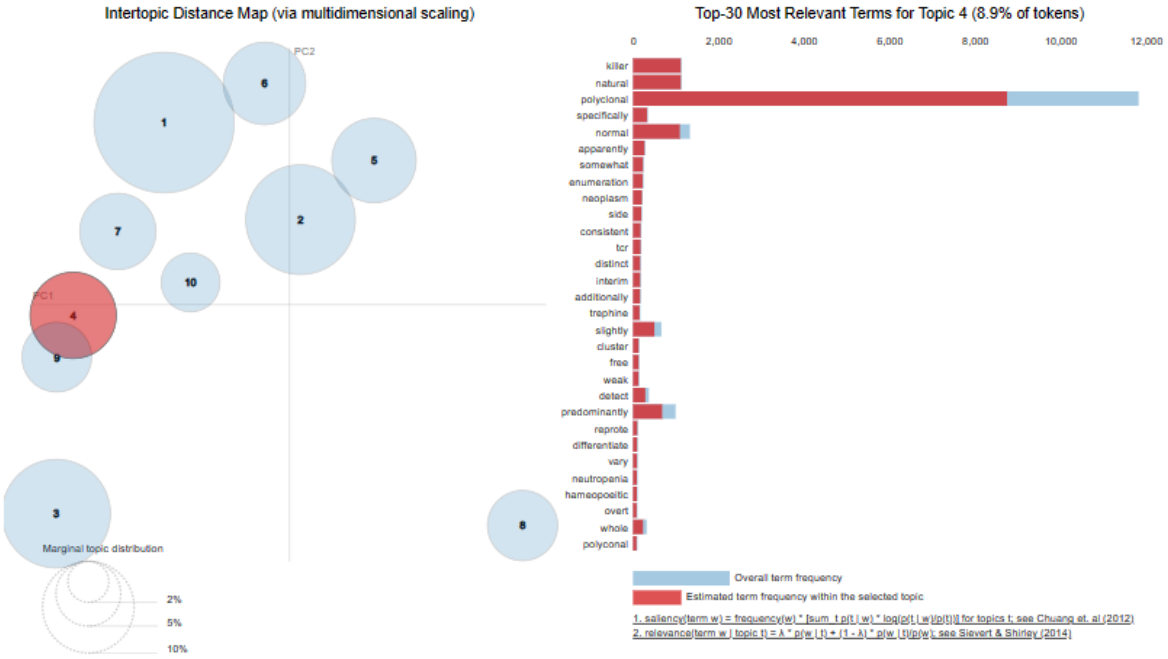


FIGURE C.4: Top words in Topic 4

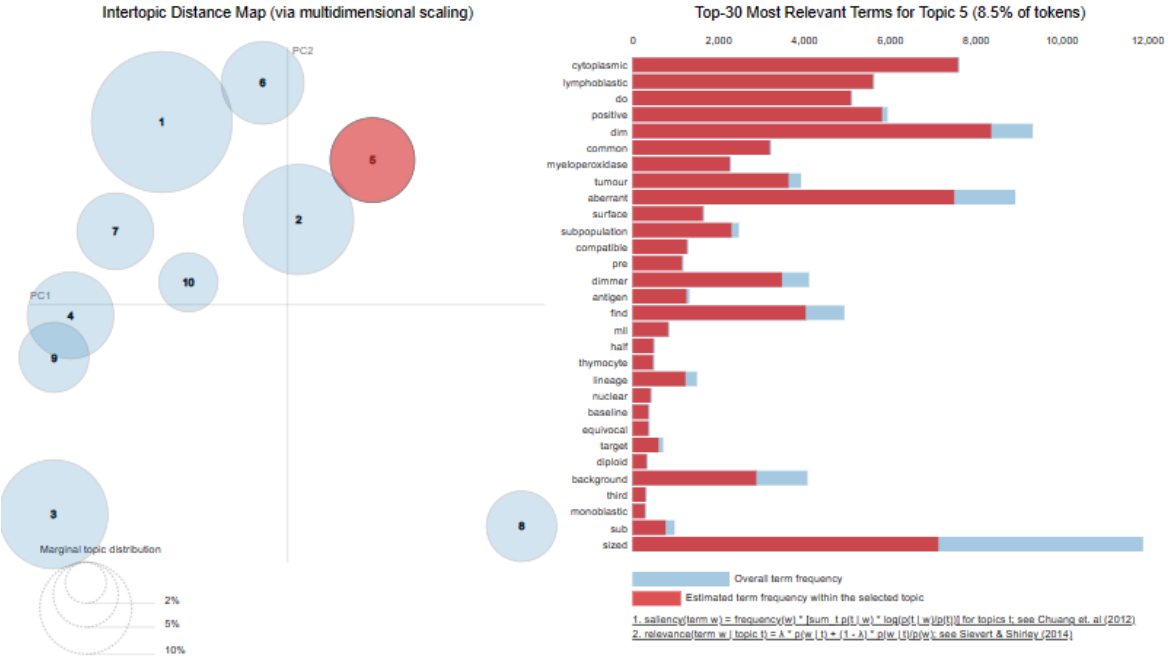


FIGURE C.5: Top words in Topic 5

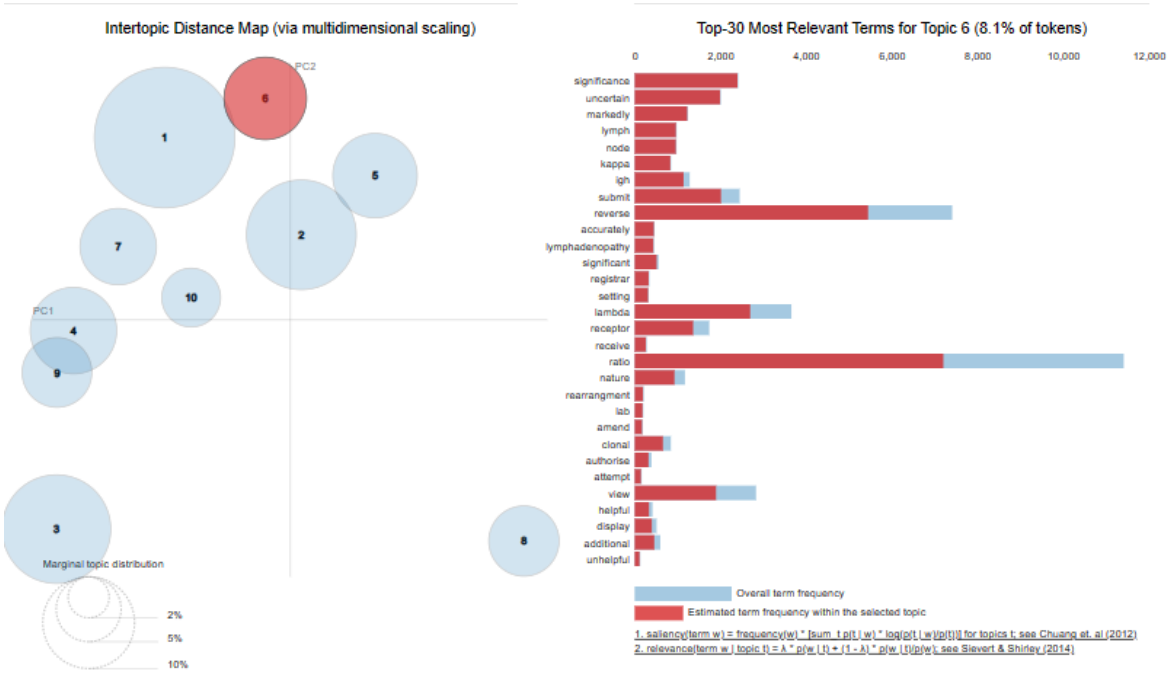


FIGURE C.6: Top words in Topic 6

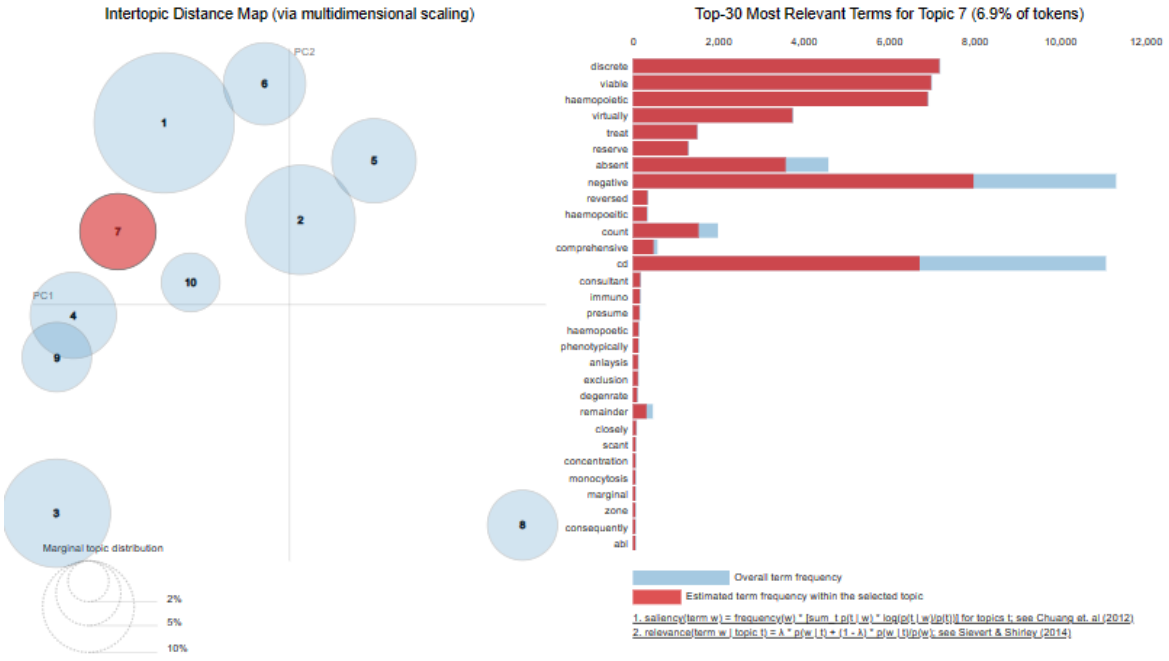


FIGURE C.7: Top words in Topic 7

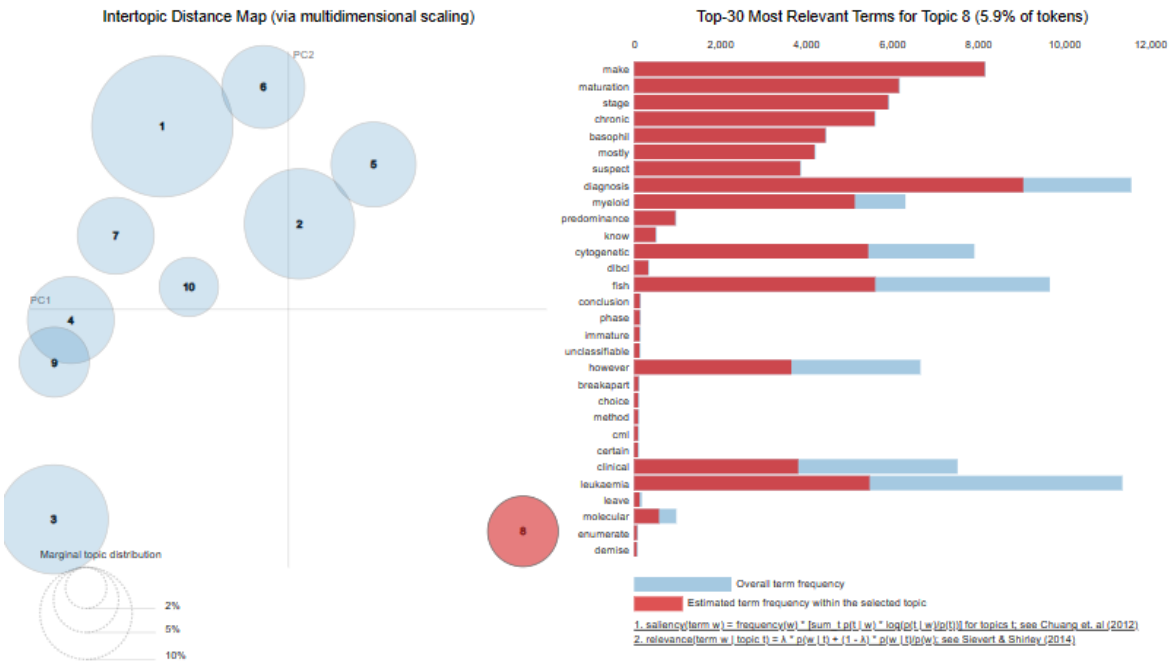


FIGURE C.8: Top words in Topic 8

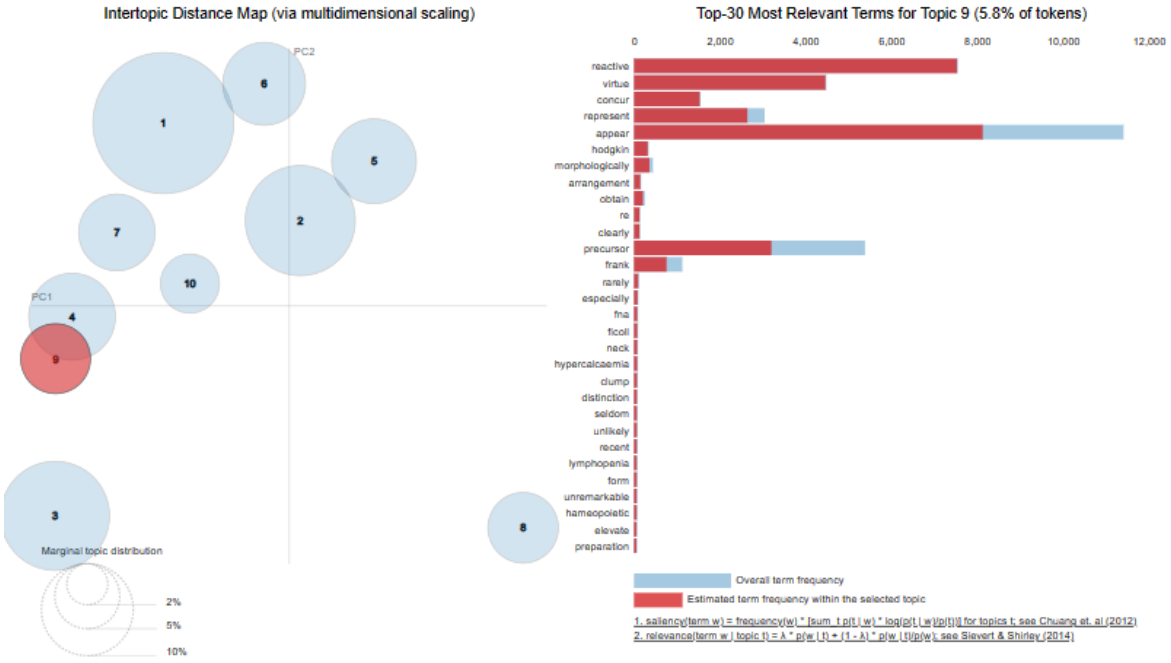


FIGURE C.9: Top words in Topic 9

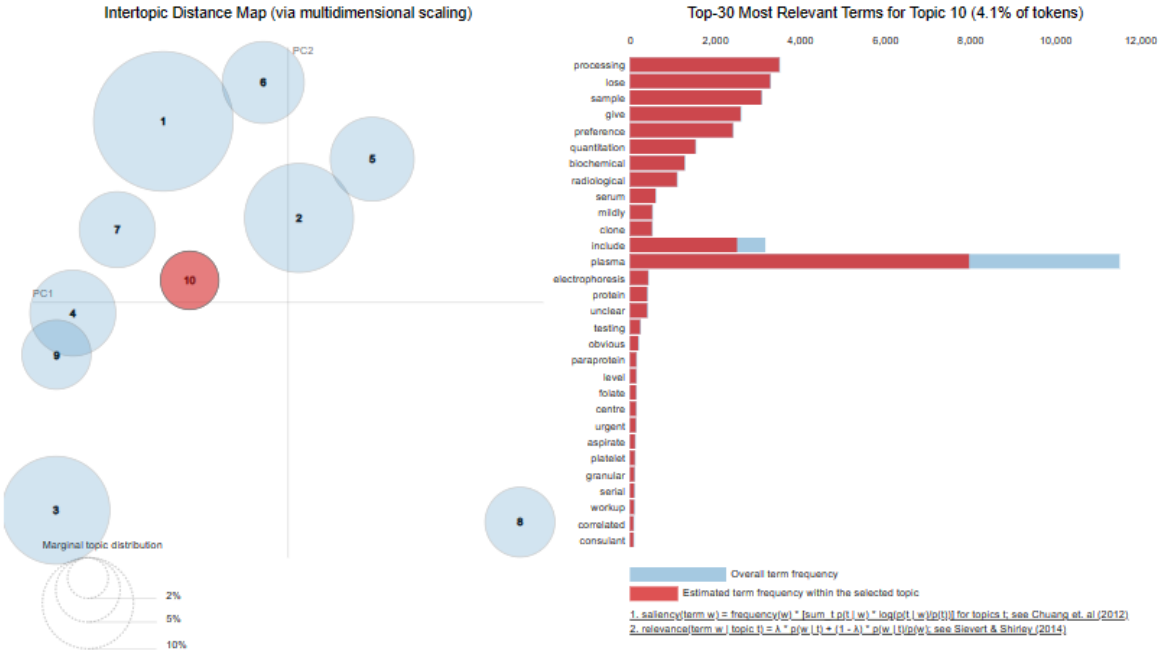


FIGURE C.10: Top words in Topic 10

Appendix 4

D.1 Top words in topics from the corpus of all reports

This Appendix contains visualisation of the model of all reports, highlighting the tops words in each topic.

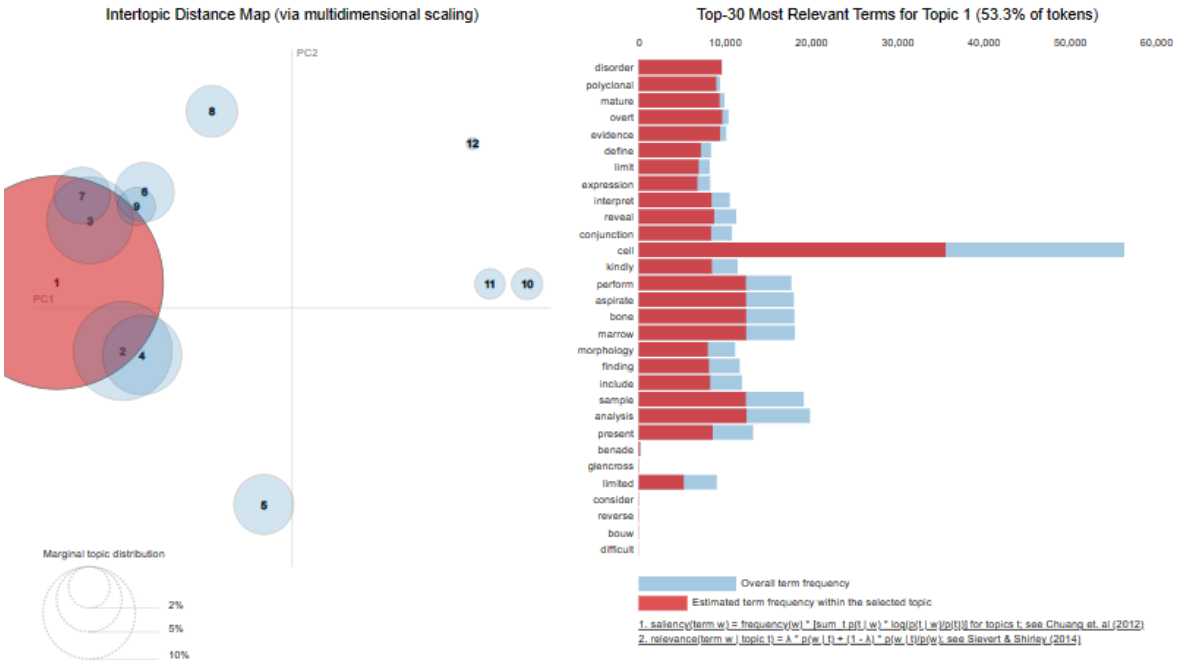


FIGURE D.1: Top words in Topic 1

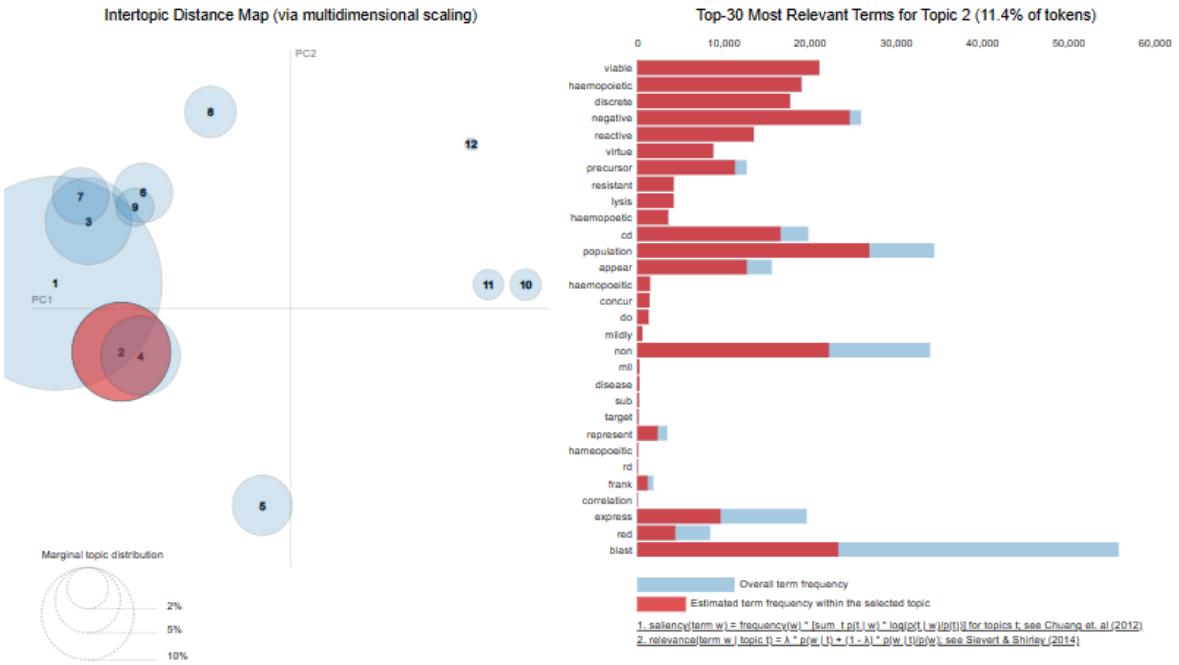


FIGURE D.2: Top words in Topic 2

FIGURE D.3: Top words in Topic 3

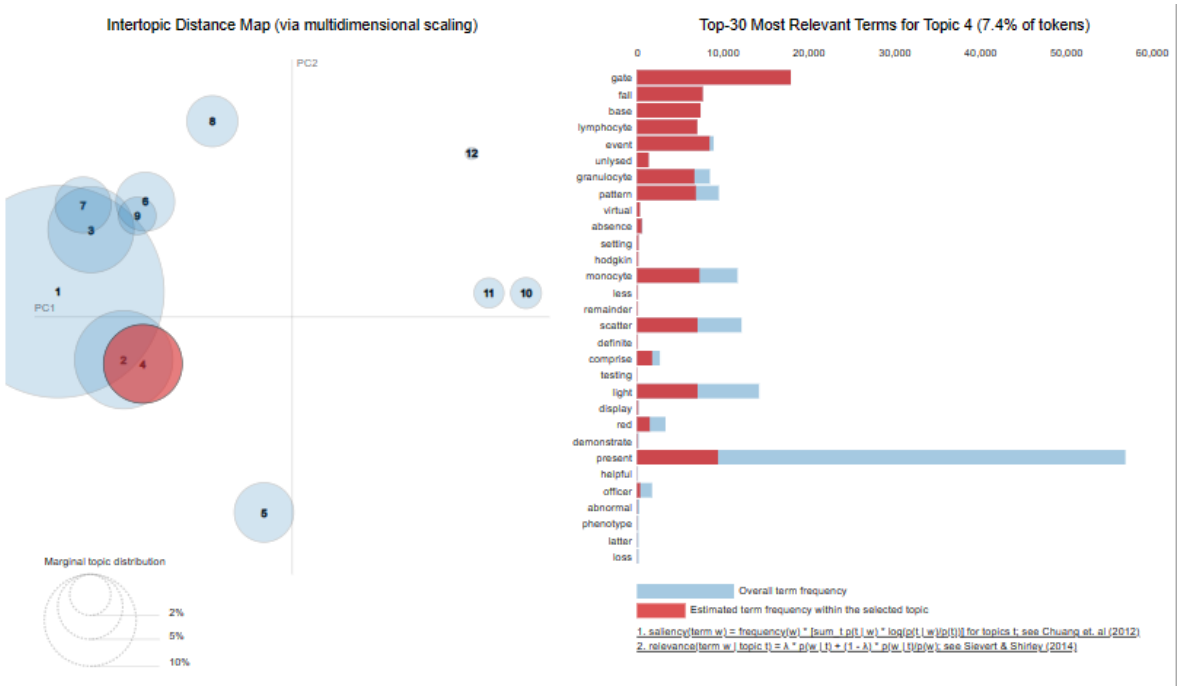


FIGURE D.4: Top words in Topic 4

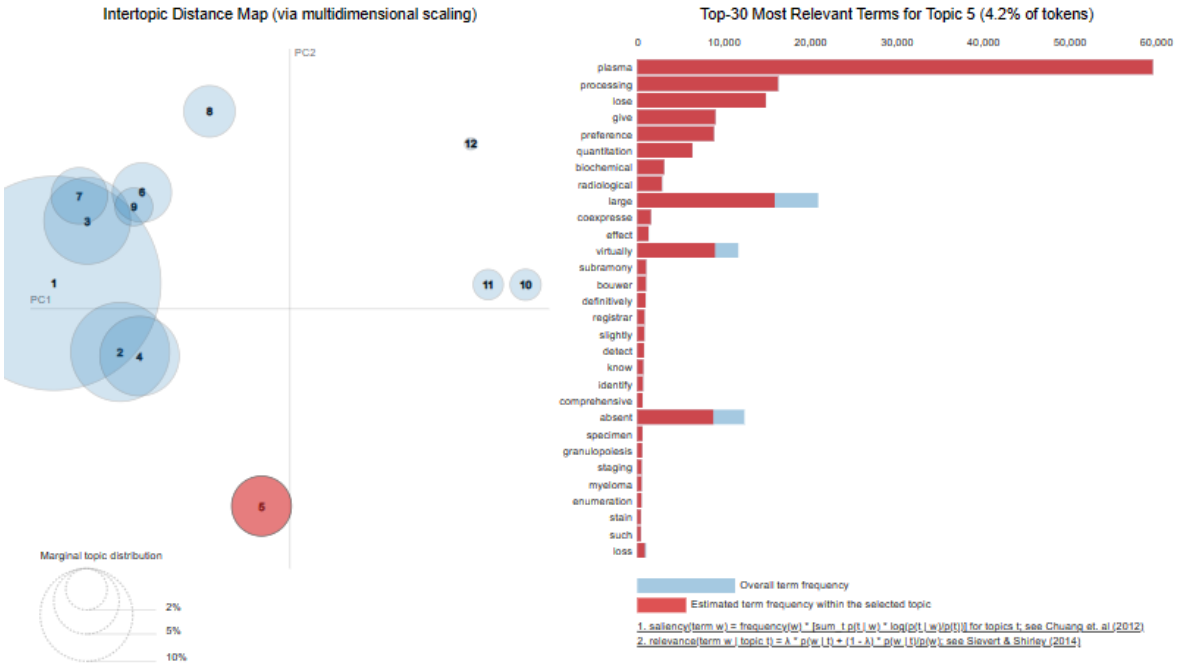


FIGURE D.5: Top words in Topic 5

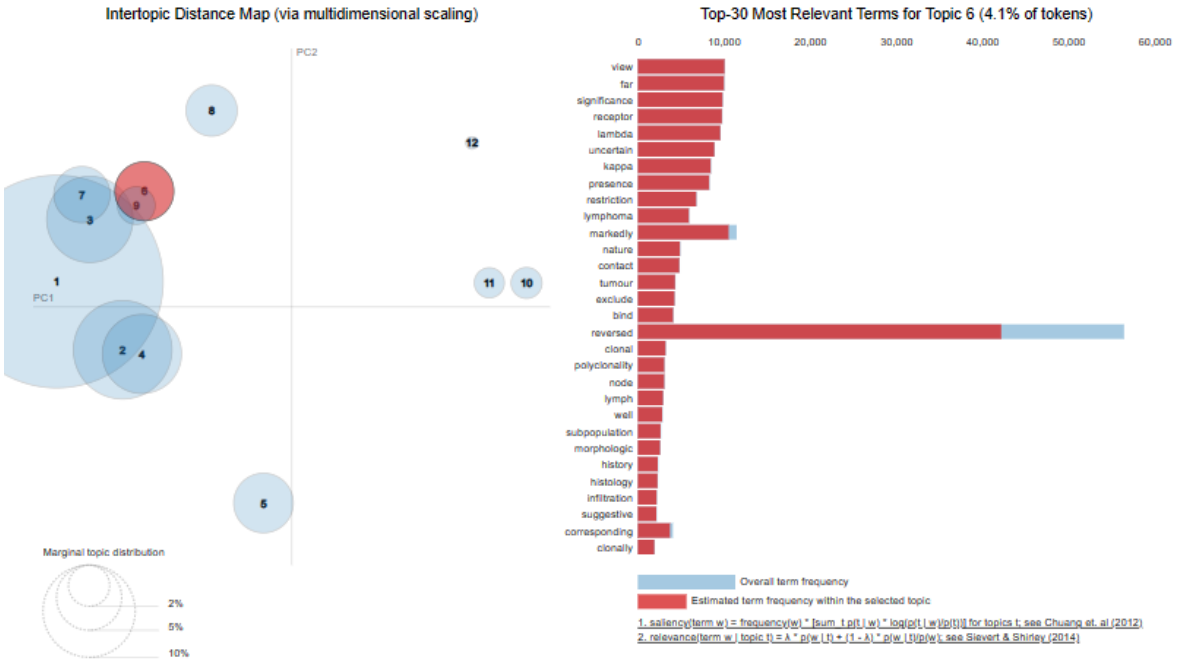


FIGURE D.6: Top words in Topic 6

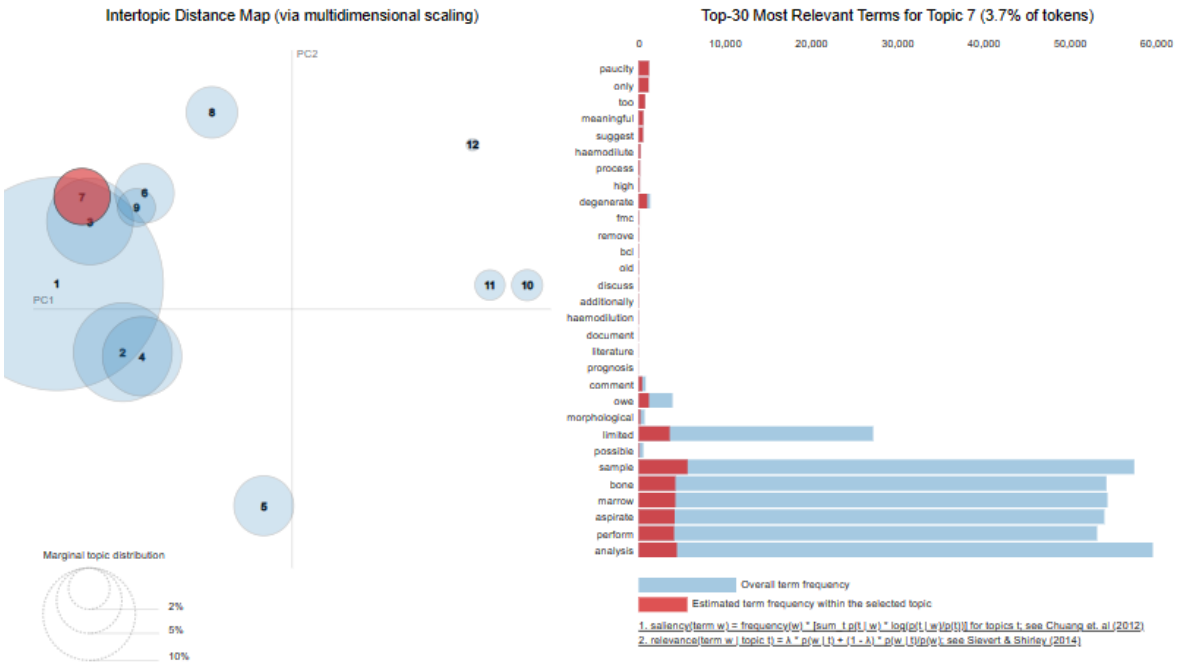


FIGURE D.7: Top words in Topic 7

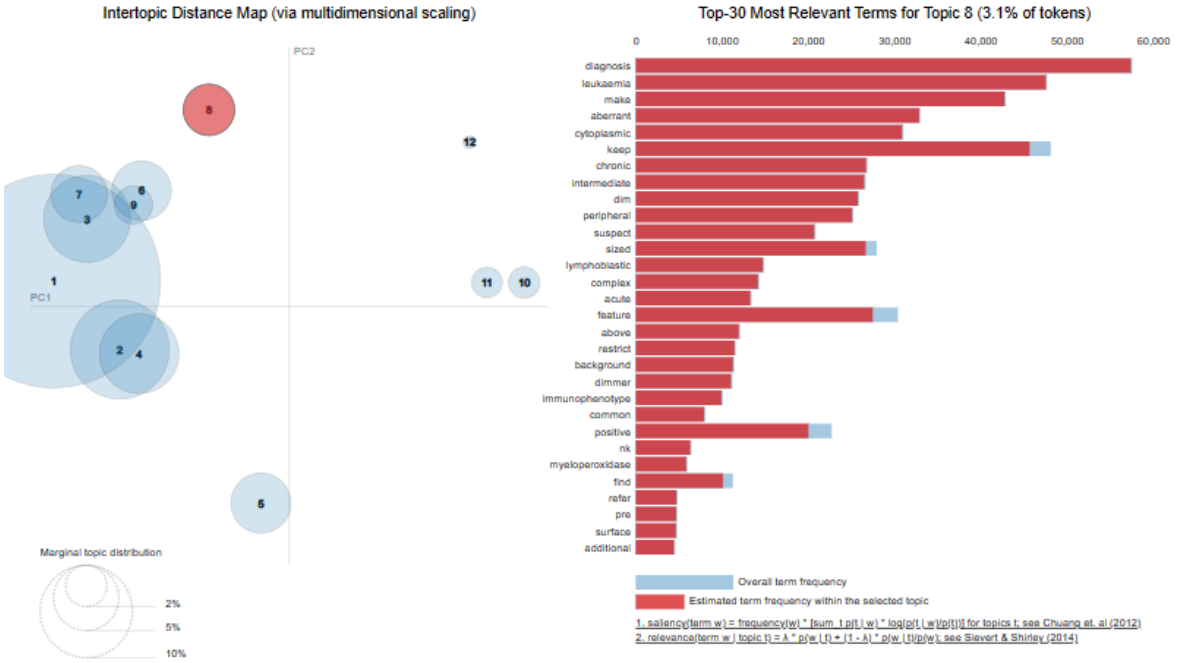


FIGURE D.8: Top words in Topic 8

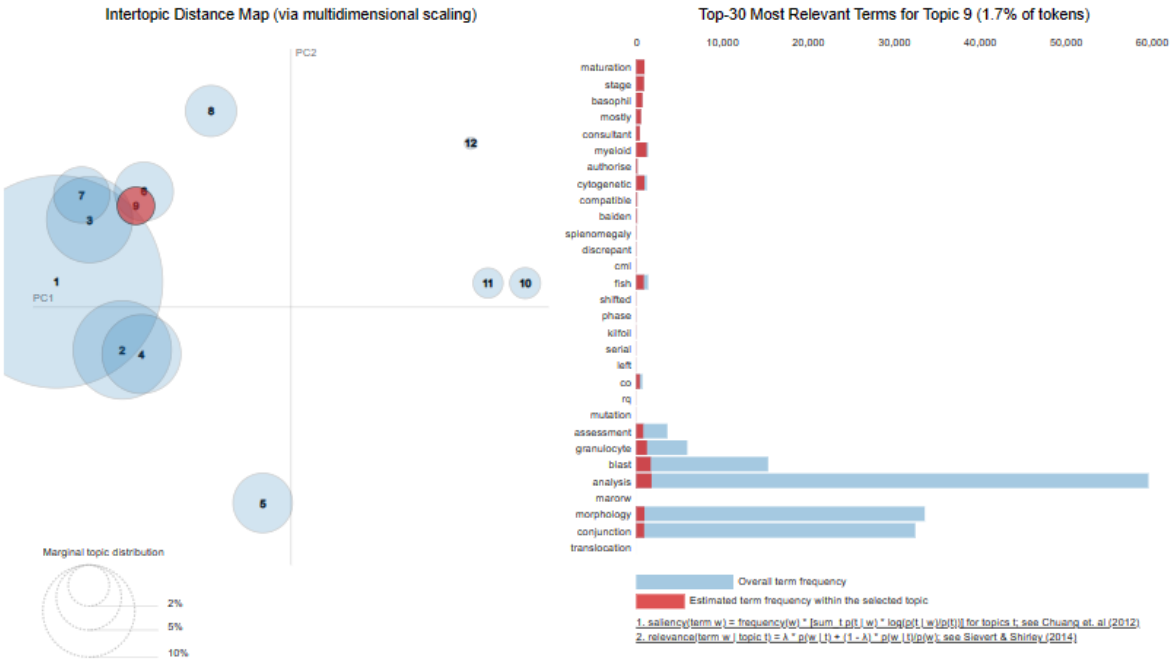


FIGURE D.9: Top words in Topic 9

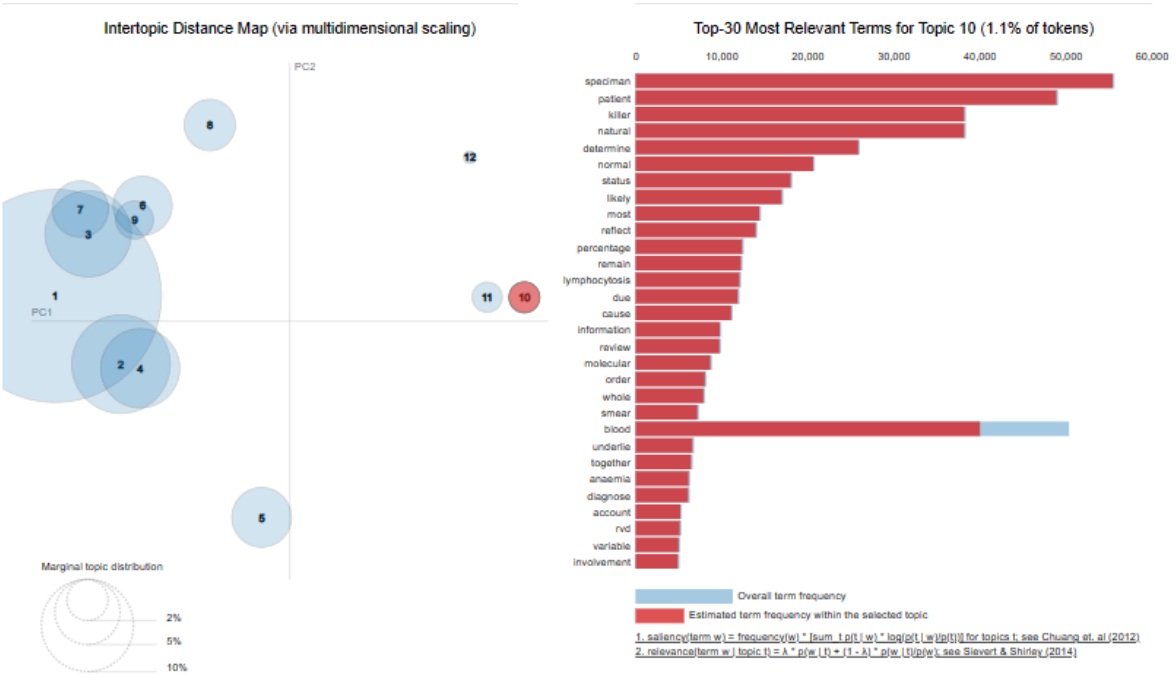


FIGURE D.10: Top words in Topic 10

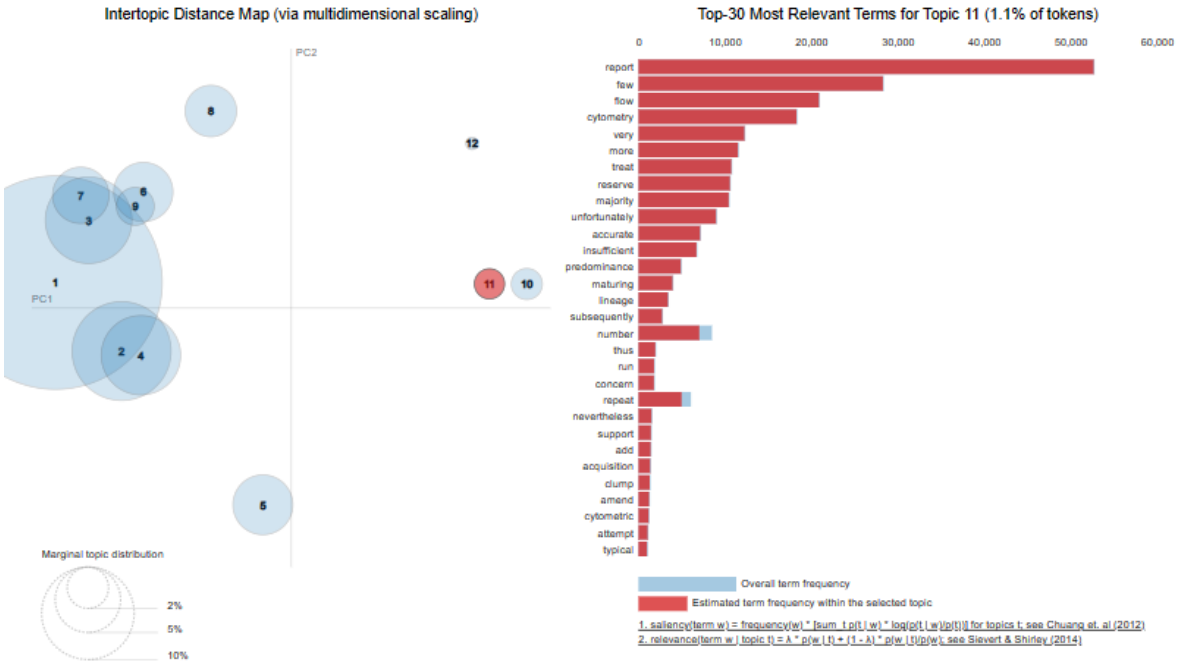


FIGURE D.11: Top words in Topic 11

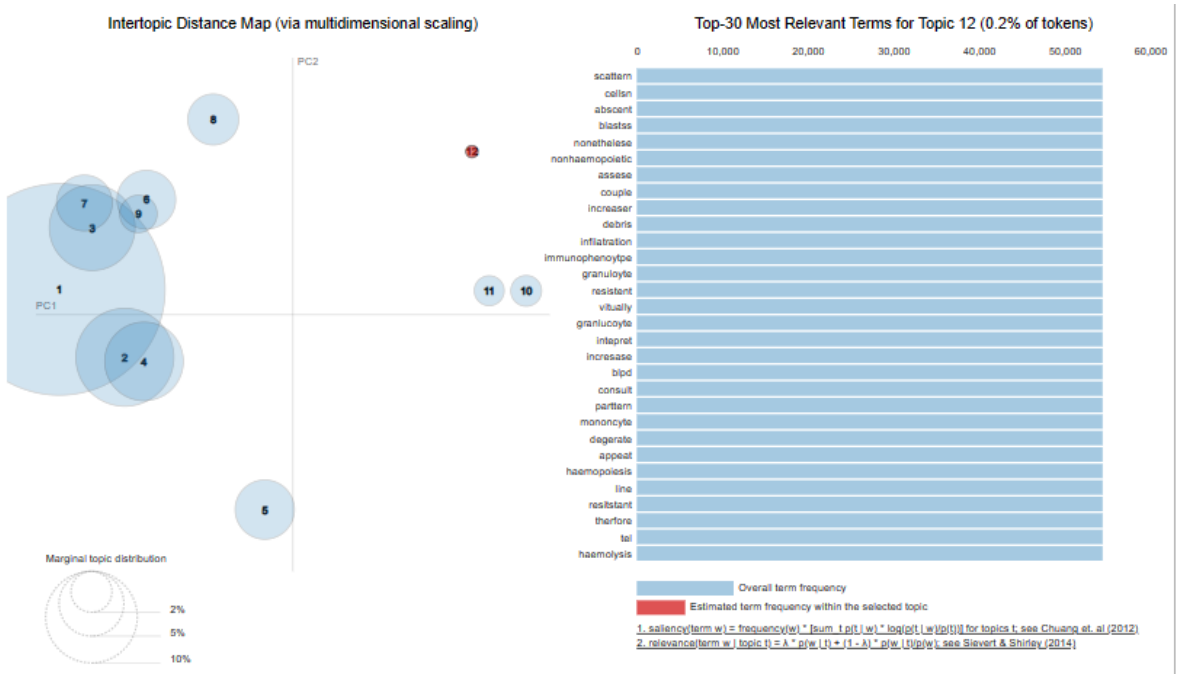


FIGURE D.12: Top words in Topic 12

Appendix 5

This Appendix is the similarity report for the write up

Turnitin Originality Report

Processed on: 19-Aug-2024 1:49 PM SAST
ID: 2434395536
Word Count: 19382
Submitted: 1

main_no_refs-3.pdf By Sibonginkosi Maposa

Similarity Index	Similarity by Source	
8%	Internet Sources:	5%
	Publications:	4%
	Student Papers:	3%

1% match (student papers from 10-Sep-2010)
Submitted to Kingston University on 2010-09-10

1% match (student papers from 04-May-2023)
Submitted to Washington University in St. Louis on 2023-05-04

< 1% match (Internet from 02-Oct-2022)
https://wiredspace.wits.ac.za/bitstream/handle/10539/31733/10-06-20%20FINAL.....RESEARCH%20PROTOCOL1887347%20%28CORRECTED%20VERSION%29%20%281%29.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 08-May-2024)
https://wiredspace.wits.ac.za/server/api/core/bitstreams/9576f42b-98cf-427f-9e40-5919f70b2da5/content

< 1% match (Internet from 04-May-2024)
https://wiredspace.wits.ac.za/server/api/core/bitstreams/05e934c2-559a-4266-9bf6-d869fb020ec7/content

< 1% match (Internet from 01-Jun-2024)
https://wiredspace.wits.ac.za/server/api/core/bitstreams/720f9dae-0d6d-45e4-b76f-e61e863f5159/content

< 1% match (Singh, Elvira, Paul Ruff, Chantal Babb, Mazvita Sengayi, Moira Beery, Lerato Khoali, Patricia Kellett, and J Michael Underwood. "Establishment of a cancer surveillance programme: the South African experience", The Lancet Oncology, 2015.)
Singh, Elvira, Paul Ruff, Chantal Babb, Mazvita Sengayi, Moira Beery, Lerato Khoali, Patricia Kellett, and J Michael Underwood, "Establishment of a cancer surveillance programme: the South African experience", The Lancet Oncology, 2015.

< 1% match ()
Wencheng Sun, Zhiping Cai, Yangyang Li, Fang Liu, Shenggun Fang, Guoyan Wang, "Data Processing and Text Mining Technologies on Electronic Medical Records: A Review", Journal of Healthcare Engineering

< 1% match ()
A Rahim, Afiq Izzudin, "Development of machine learning sentiment analyzer and quality classifier (MLSAQC) and its application in analysing hospital patient satisfaction from Facebook reviews in Malaysia", 2022

< 1% match (Curiskis, Stephan. "The Evolution of Topological Concepts in Online Social Networks", University of Technology Sydney (Australia), 2023)
Curiskis, Stephan. "The Evolution of Topological Concepts in Online Social Networks", University of Technology Sydney (Australia), 2023

< 1% match (Internet from 09-Feb-2023)
https://www.researchgate.net/figure/Graphical-representation-of-the-LDA-model-The-shaded-circle-represents-the-observed_fig1_303203779

< 1% match (Internet from 14-Dec-2022)
https://www.researchgate.net/publication/319285543_A_Novel_Approach_for_Classifying_Gene_Expression_Data_using_Topic_Mode

< 1% match (Internet from 04-Oct-2022)
https://web.archive.org/web/20180724061020if_/http:/knoesis.wright.edu/Library/export/ris/?f%5Bpublications%5D=iswc10_paper218.pdf

< 1% match (Internet from 17-May-2024)
https://dokumen.pub/introduction-to-artificial-intelligence-1nbsped-9783031259272-9783031259289.html

< 1% match (Internet from 29-Feb-2024)
https://dokumen.pub/sentiment-analysis-in-the-medical-domain-3031301862-9783031301865.html

< 1% match (Internet from 12-Aug-2024)
https://dokumen.pub/inventive-systems-and-control-proceedings-of-icisc-2023-9819916232-9789819916238.html

< 1% match (student papers from 09-Sep-2023)
Submitted to Sim University on 2023-09-09

< 1% match (student papers from 26-Oct-2022)
Submitted to Sim University on 2022-10-26