An Evaluation Of Learning Employing Natural Language Processing And Cognitive Load Assessment

Mrunal Tipari
*Dublin Institute of Technology*

Follow this and additional works at: [https://arrow.dit.ie/scschcomdis](https://arrow.dit.ie/scschcomdis)

Part of the [Artificial Intelligence and Robotics Commons](https://arrow.dit.ie/scschcomdis) and the [Computer Engineering Commons](https://arrow.dit.ie/scschcomdis)

**Recommended Citation**
[https://arrow.dit.ie/scschcomdis/172](https://arrow.dit.ie/scschcomdis/172)

This Dissertation is brought to you for free and open access by the School of Computing at ARROW@TU Dublin. It has been accepted for inclusion in Dissertations by an authorized administrator of ARROW@TU Dublin. For more information, please contact yvonne.desmond@dit.ie, arrow.admin@dit.ie, brian.widdis@dit.ie.
An Evaluation Of Learning Employing Natural Language Processing And Cognitive Load Assessment

Mrunal Tipari

A dissertation submitted in partial fulfilment of the requirements of Dublin Institute of Technology for the degree of M.Sc. in Computing (Data Analytics)

Date: January 2019
Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institutes guidelines for ethics in research.

Signed: Mrunal Tipari

Date: 4 January, 2019
Abstract

One of the key goals of Pedagogy is to assess learning. Various paradigms exist and one of this is Cognitivism. It essentially sees a human learner as an information processor and the mind as a black box with limited capacity that should be understood and studied. With respect to this, an approach is to employ the construct of cognitive load to assess a learner’s experience and in turn design instructions better aligned to the human mind. However, cognitive load assessment is not an easy activity, especially in a traditional classroom setting. This research proposes a novel method for evaluating learning both employing subjective cognitive load assessment and natural language processing. It makes use of primary, empirical and deductive methods. In details, on one hand, cognitive load assessment is performed using well-known self-reporting instruments, borrowed from Human Factors, namely the Nasa Task Load Index and the Workload Profile. On the other hand, Natural Language Processing techniques, borrowed from Artificial Intelligence, are employed to calculate semantic similarity of textual information, provided by learners after attending a typical third-level classroom, and the content of the classroom itself. Subsequently, an investigation of the relationship of cognitive load assessment and textual similarity is performed to assess learning.

Keywords: Cognitive Load, Natural Language Processing, Semantic Similarity, Nasa Task Load Index, Workload Profile.
Acknowledgments

Firstly, I wish to express my profound gratitude to my family for the continuous encouragement and unconditional support and love.

I would like to offer a special thanks to my supervisor, Dr. Luca Longo for guiding me time to time and providing valuable feedback which helped me to successfully complete my dissertation.

I would like to thank everyone who is somehow involved in the development of this thesis for their critical opinions on many aspects of the research.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Declaration</td>
<td>I</td>
</tr>
<tr>
<td>Abstract</td>
<td>II</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>III</td>
</tr>
<tr>
<td>Contents</td>
<td>IV</td>
</tr>
<tr>
<td>List of Figures</td>
<td>VII</td>
</tr>
<tr>
<td>List of Tables</td>
<td>XIII</td>
</tr>
<tr>
<td>List of Acronyms</td>
<td>XV</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Background</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Research Problem</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Research Objectives</td>
<td>2</td>
</tr>
<tr>
<td>1.4 Research Methodologies</td>
<td>3</td>
</tr>
<tr>
<td>1.5 Scope and Limitations</td>
<td>3</td>
</tr>
<tr>
<td>1.6 Document Outline</td>
<td>4</td>
</tr>
<tr>
<td>2 Literature Review and related work</td>
<td>5</td>
</tr>
<tr>
<td>2.1 Instructional Design</td>
<td>5</td>
</tr>
<tr>
<td>2.1.1 Cognitive Load Theory</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Mental Workload</td>
<td>7</td>
</tr>
</tbody>
</table>

IV
List of Figures

2.1  2-D Vector Space Model  .................................................. 18
3.1  Data Collection process .................................................... 24
3.2  Corpora  ........................................................................... 26
3.3  Hypothesis Definition .......................................................... 30
3.4  Data Preparation steps for numerical data ................................. 31
3.5  Data preparation of textual data  ............................................. 34
3.6  Model Design  .................................................................... 37
4.1  Science Corpus Wordcloud .................................................... 46
4.2  Frequency of Top 25 most common tokens in Science Corpus ...... 47
4.3  Scientific Method Corpus Wordcloud ....................................... 48
4.4  Frequency of Top 25 most common tokens in Scientific Method Corpus 48
4.5  Planning Research Corpus Wordcloud ...................................... 49
4.6  Frequency of Top 25 most common tokens in Planning Research Corpus 49
4.7  Literature Review Corpus Wordcloud ....................................... 50
4.8  Frequency of Top 25 most common tokens in Literature Review Corpus 50
4.9  Science + Scientific Method Corpus Wordcloud ......................... 51
4.10 Frequency of Top 25 most common tokens in Science + Scientific Method Corpus 51
4.11 Science + Scientific Method + Planning Research Corpus Wordcloud 52
4.12 Frequency of Top 25 most common tokens in Sc + SM + PR Corpus 52
4.13 Science + Scientific Method + Planning Research + Literature Review

Corpus Wordcloud ................................................. 53

4.14 Frequency of Top 25 most common tokens in Sc + SM + PR + LR

Corpus ................................................................. 53

4.15 Assessment of Normality for NASA-TLX Science .............................. 58

(a) Histogram ......................................................... 58
(b) Boxplot ............................................................. 58

4.16 NASA-TLX Post task: Assessment of Normality ................................. 59

(a) Histogram ......................................................... 59
(b) Boxplot ............................................................. 59
(c) Scatterplot ......................................................... 59

4.17 NASA-TLX Post task: Assessment of Normality (log transformation) ........ 59

(a) Histogram ......................................................... 59
(b) Boxplot ............................................................. 59
(c) Scatterplot ......................................................... 59

4.18 NASA-TLX Post-Pre task: Assessment of Normality ............................. 60

(a) Histogram ......................................................... 60
(b) Boxplot ............................................................. 60
(c) Scatterplot ......................................................... 60

4.19 NASA-TLX Workload score: Assessment of Normality ......................... 61

(a) Histogram ......................................................... 61
(b) Boxplot ............................................................. 61

4.20 NASA-TLX post task for Scientific Method: Assessment of Normality ...... 62

for NASA-TLX Scientific Method ................................... 62

(a) Histogram ......................................................... 62
(b) Boxplot ............................................................. 62
(c) Scatterplot ......................................................... 62

4.21 NASA-TLX Post-Pre Scientific Method: Assessment of Normality ......... 63

(a) Histogram ......................................................... 63
(b) Boxplot ............................................................. 63
4.22 NASA-TLX Planning Research: Assessment of Normality
(a) Histogram ....................................................... 64
(b) Boxplot ......................................................... 64
4.23 NASA-TLX Planning Research: Assessment of Normality
(a) Histogram ....................................................... 64
(b) Boxplot ......................................................... 64
(c) Scatterplot ..................................................... 64
4.24 NASA-TLX Post-Pre Planning Research: Assessment of Normality
(a) Histogram ....................................................... 65
(b) Boxplot ......................................................... 65
(c) Scatterplot ..................................................... 65
4.25 NASA-TLX Workload Score Literature Review: Assessment of Normality
(a) Histogram ....................................................... 66
(b) Boxplot ......................................................... 66
4.26 NASA-TLX Post task Literature Review: Assessment of Normality
(a) Histogram ....................................................... 67
(b) Boxplot ......................................................... 67
(c) Scatterplot ..................................................... 67
4.27 Assessment of Normality for NASA-TLX Post-Pre Literature Review
(a) Histogram ....................................................... 68
(b) Boxplot ......................................................... 68
(c) Scatterplot ..................................................... 68
4.28 NASA-TLX Post task Science + Scientific Method: Assessment of Normality
(a) Histogram ....................................................... 69
(b) Boxplot ......................................................... 69
(c) Scatterplot ..................................................... 69
4.29 NASA-TLX Post-Pre Science + Scientific Method: Assessment of Normality

IX
A.4 WP Workload score Scientific Method: Assessment of Normality
(a) Histogram .............................................. 96
(b) Boxplot ................................................... 96

A.5 WP post task for Scientific Method: Assessment of Normality for WP
Scientific Method .............................................. 97
(a) Histogram .............................................. 97
(b) Boxplot ................................................... 97
(c) Scatterplot .............................................. 97

A.6 WP Post-Pre for Scientific Method: Assessment of Normality
(a) Histogram .............................................. 97
(b) Boxplot ................................................... 97
(c) Scatterplot .............................................. 97

A.7 WP Planning Research: Assessment of Normality .............................................. 97
(a) Histogram .............................................. 97
(b) Boxplot ................................................... 97

A.8 WP post Planning Research: Assessment of Normality .............................................. 98
(a) Histogram .............................................. 98
(b) Boxplot ................................................... 98
(c) Scatterplot .............................................. 98

A.9 WP Post-Pre Planning Research: Assessment of Normality .............................................. 98
(a) Histogram .............................................. 98
(b) Boxplot ................................................... 98
(c) Scatterplot .............................................. 98

A.10 WP Workload Score Literature Review: Assessment of Normality .............................................. 98
(a) Histogram .............................................. 98
(b) Boxplot ................................................... 98

A.11 WP Post task Literature Review: Assessment of Normality .............................................. 99
(a) Histogram .............................................. 99
(b) Boxplot ................................................... 99
A.12 Assessment of Normality for WP Post-Pre Literature Review

(a) Histogram

(b) Boxplot

(c) Scatterplot

A.13 WP Post task Science + Scientific Method: Assessment of Normality

(a) Histogram

(b) Boxplot

(c) Scatterplot

A.14 WP Post-Pre Science + Scientific Method: Assessment of Normality

(a) Histogram

(b) Boxplot

(c) Scatterplot

A.15 Assessment of Normality for WP Sc + SM + PR Method

(a) Histogram

(b) Boxplot

(c) Scatterplot

A.16 Assessment of Normality for WP Post-Pre Sc + SM + PR Method

(a) Histogram

(b) Boxplot

(c) Scatterplot

A.17 WP post-task Sc+SM+PR+LR: Assessment of Normality

(a) Histogram

(b) Boxplot

(c) Scatterplot

A.18 WP post-pre task Sc+SM+PR+LR: Assessment of Normality

(a) Histogram

(b) Boxplot

(c) Scatterplot
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Descriptive Statistics of NASA-TLX Dataset</td>
<td>43</td>
</tr>
<tr>
<td>4.2</td>
<td>Frequency count of Keywords in NASA-TLX Dataset</td>
<td>44</td>
</tr>
<tr>
<td>4.3</td>
<td>Descriptive Statistics of Workload Profile Dataset</td>
<td>45</td>
</tr>
<tr>
<td>4.4</td>
<td>Frequency count of Keywords in Workload Profile Dataset</td>
<td>45</td>
</tr>
<tr>
<td>4.5</td>
<td>Data Understanding of the Corpus</td>
<td>46</td>
</tr>
<tr>
<td>4.6</td>
<td>Descriptive Statistics of the Keywords of NASA-TLX pre-task and post-task</td>
<td>55</td>
</tr>
<tr>
<td>4.7</td>
<td>Descriptive Statistics of the Keywords of WP pre-task and post-task</td>
<td>56</td>
</tr>
<tr>
<td>4.8</td>
<td>NASA-TLX Workload Score for Science: Kolmogorov-Smirnov Test of Normality</td>
<td>58</td>
</tr>
<tr>
<td>4.9</td>
<td>NASA-TLX Post task: Kolmogorov-Smirnov Test of Normality</td>
<td>59</td>
</tr>
<tr>
<td>4.10</td>
<td>NASA-TLX Post task: Kolmogorov-Smirnov Test of Normality (log transformation)</td>
<td>60</td>
</tr>
<tr>
<td>4.11</td>
<td>NASA-TLX Post-Pre task: Kolmogorov-Smirnov Test of Normality</td>
<td>60</td>
</tr>
<tr>
<td>4.12</td>
<td>NASA-TLX Workload score Scientific Method: Kolmogorov-Smirnov Test of Normality</td>
<td>61</td>
</tr>
<tr>
<td>4.13</td>
<td>NASA-TLX post task for Scientific Method: Kolmogorov-Smirnov Test of Normality</td>
<td>62</td>
</tr>
<tr>
<td>4.14</td>
<td>NASA-TLX Post-Pre Scientific Method: Kolmogorov-Smirnov Test of Normality</td>
<td>63</td>
</tr>
<tr>
<td>4.15</td>
<td>NASA-TLX Planning Research: Kolmogorov-Smirnov Test of Normality</td>
<td>64</td>
</tr>
<tr>
<td>4.16</td>
<td>NASA-TLX Planning Research: Kolmogorov-Smirnov Test of Normality</td>
<td>65</td>
</tr>
</tbody>
</table>
4.17 Kolmogorov-Smirnov Test of Normality ........................................ 65
4.18 NASA-TLX Literature Review: Kolmogorov-Smirnov Test of Normality 66
4.19 NASA-TLX Post task Literature Review: Kolmogorov-Smirnov Test of Normality ............................................................ 67
4.20 NASA-TLX Post-Pre Literature Review: Kolmogorov-Smirnov Test of Normality ............................................................... 68
4.21 NASA-TLX Post task Science + Scientific Method: Kolmogorov-Smirnov Test of Normality ................................................................. 69
4.22 -TLX Post-Pre Science + Scientific Method: Kolmogorov-Smirnov Test of Normality ............................................................. 70
4.23 Kolmogorov-Smirnov Test of Normality for Sc + SM + PR .............. 71
4.24 Kolmogorov-Smirnov Test of Normality for Sc + SM + PR .............. 71
4.25 NASA-TLX post-task Sc+SM+PR+LR: Kolmogorov-Smirnov Test of Normality ................................................................. 72
4.26 NASA-TLX post-pre task Sc+SM+PR+LR: Kolmogorov-Smirnov Test of Normality .............................................................. 73
4.27 Kolmogorov-Smirnov Test of Normality for WP ............................. 74
5.1 NASA-TLX Workload Score and Cosine Similarity for Post-Task: Table of Correlations ................................................................. 76
5.2 NASA-TLX and Cosine Similarity for Post-Pre Task: Table of Correlations ................................................................................. 77
5.3 WP and Cosine Similarity for Post-Task: Table of Correlations .......... 79
5.4 WP and Cosine Similarity for Post-Pre Task: Table of Correlations .... 79
List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLT</td>
<td>Cognitive Load Theory</td>
</tr>
<tr>
<td>MWL</td>
<td>Mental Workload</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>NASA-TLX</td>
<td>NASA Task Load Index</td>
</tr>
<tr>
<td>WP</td>
<td>Workload Profile</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Background

To improve one’s learning, it is essential to assess their learning experience. Various paradigms exist and one of this is Cognitivism. It focuses on mental processes, which helps to know how humans perceive, think, remember, learn, solve problems, and direct their attention to one stimulus rather than another. It essentially sees a human learner as an information processor and the mind as a black box with limited capacity that should be understood and studied. This research is aimed at investigating Learning by employing Cognitive Load Assessment and Natural Language Processing techniques in the field of education. Mental Workload (MWL) has found its application in many fields such as Ergonomics (Fallahi et al. 2016; Doebler, Ryan, & Maguire, 2017; Boele-Vos & Twisk, 2017), Human Computer Interaction (Longo 2018a; 2015b; Longo & Dondio, 2015; Longo, Rusconi, Noce, & Barrett, 2012; Longo, 2011) and Machine Learning (Moustafa, Luz, & Longo, 2017) but Natural Language Processing techniques are rarely used in conjunction to Mental workload to assess learning.

Longo (2016) says that mental overload or underload can negatively affect ones performance. There are factors which affect the working memory which in turn depends on the way in which the information is presented (F. Paas, Renkl, & Sweller, 2003). While planning instructional design, the manner in which material is presented and the learning activities required to impart knowledge should be considered (Van Mer-
1.2 Research Problem

The need to assess and analyze Learning using new and different approaches seems necessary to improve the instructional design. The use of new approaches is increasing and that opens the door to a new research problem if Natural Language Processing can make a fair contribution to the measurement of cognitive load to evaluate learning in the field of education. Learning in education should be enhanced which needs proper planning of instructional design. The ways of presenting information can have a positive or negative impact on an individual’s performance depending on the instructional design plan [Longo 2016]. In this study, the content present in an instrument design is translated into text and keywords are processed using Natural Language Processing techniques. The output of that is then used to analyze the relationship between learning and MWL measures which leads to the research question:

“To what extent do self-reporting measures of Mental Workload correlate to the semantic similarity of textual information provided by learners and a lecturer, and computed using Natural Language Processing techniques?”

1.3 Research Objectives

The key objective of this research is to evaluate the learning of the students during the third level classes by quantifying it using Natural Language Processing techniques and the cognitive load. To achieve the same, data is gathered from the students using NASA-TLX and WP questionnaires along with the textual data about the topics learned. To evaluate learning and assess the load, literature about Cognitive Load Theory and its load types are explored along with methods to assess it. Various Natural Language Processing techniques for semantic similarity were also researched.
CHAPTER 1. INTRODUCTION

From the literature, it is seen that the application of Natural Language Processing in conjunction to MWL seems to rarely happen in the discipline of Education. Therefore, an experiment is designed to test the null hypothesis which is that there is no strong correlation between Mental Workload and the difference of the cosine similarity value calculated between the keywords obtained from students before and after the third level class and the content taught by the lecturer. The Word2Vec model is built to calculate the semantic similarity of textual information, provided by learners after attending a typical third-level classroom, and the content of the classroom itself by converting words into vectors. The results are evaluated using appropriate statistical techniques which help to either reject or accept the hypothesis defined and thus answer the research question.

1.4 Research Methodologies

The aim of this research is to quantify the learning using Natural Language Processing technique to analyze the relationship between Mental Workload and cosine similarity values calculated from the keywords collected from pre-tasks and post-tasks of MWL activities conducted in third level classes.

Mixed research method is used which is concurrent triangulation by objective as the qualitative and quantitative data is collected and analyzed separately but is integrated during interpretation. The research makes use of primary, empirical and deductive methods for investigation. It is a primary research as the data used is collected by Dr. Luca Longo of Dublin Institute of Technology.

1.5 Scope and Limitations

The data used in the experiment is gathered from the students taking third level education in the university from the year 2015 to 2018 to conduct Mental Workload experiments. During those experiments, different topics were taught by the Lecturer to measure their Mental Workload using measurement methods such as NASA-TLX and
WP by providing pre-defined questionnaires along with keywords. The data collected has 224 and 213 records respectively, which is relatively small. Another concern is the size of the content taught by the Lecturer. With the small size of the input content, it is difficult to obtain statistically significant results using Natural Language Processing technique.

1.6 Document Outline

The research document includes the chapters namely, Literature Review, Design and Methodology, Implementation and Results, Evaluation and Conclusion. An overview of contents are as follows:

- Chapter 2 (Literature Review) gives an overview about the literature and related work in Cognitive Load Theory, Mental Workload and Natural Language Processing to get a better understanding of the research which helps to find the gaps and formulate a research question.

- Chapter 3 (Design and Methodology) includes the definition of the hypothesis necessary to answer the research question. It talks about the design plan of the experiment which follows the CRISP-DM model.

- Chapter 4 (Implementation and Results) gives the details about the results achieved by following the designed approach. It shows the implementation of Word2Vec model built to analyze the semantic similarity.

- Chapter 5 (Evaluation and Analysis) will give a detailed analysis of the experiment and based on the result a decision regarding the acceptance and rejection of the proposed hypothesis will be made. This chapters also outlines the strength and limitations of the findings.

- Chapter 6 (Conclusion) will summarize the working and finding of the research undertaken during this thesis work, which includes problem definition, design, implementation, and evaluation of the finding and limitation for further work.
Chapter 2

Literature Review and related work

In this chapter, Cognitive Load Theory and its load types are introduced. It then describes the Mental Workload measures in Ergonomics which helps to assess the cognitive load \cite{OrruGobboOSullivanLongo2018}. Subsequently, it focuses on a description of two self-reporting mental workload assessment techniques, these being used in the primary research study. Also, brief information about Natural Language Processing technique is given and its study in relation with MWL is also discussed. In the end, gaps are identified which lead to the formation of a research question.

2.1 Instructional Design

Instructional design is an innovation for the improvement of learning experiences and situations which help students acquire specific knowledge and skills. It consolidates known and verified learning techniques into instructional encounters which make the acquiring of knowledge and skill increasingly effective, viable, and engaging. The studies related to cognitive load framework has focused on the structure of innovative instructional designs that proficiently utilize the working memory limit which in turn helps to increase the learning.
2.1.1 Cognitive Load Theory

Cognitive load theory (CLT) was developed in the 1980s and since then the research in this area is increasing (Sweller, Van Merrienboer, & Paas, 1998). It is a learning theory which helps to investigate working memory characteristics and instructional designs by providing a framework (Longo & Barrett, 2010). Theories about human memory architecture draw a line between long term memory and short term memory. A large amount of information is stored permanently in long term memory whereas small amounts of information are stored in short term memory (Cowan, 2001; Miller, 1956) for a very short period of time (Dosher, 2003). The short term memory is commonly known as working memory as it processes information.

Cognitive load is a total amount of effort being used in working memory. Working memory is like a memory buffer which is used to manipulate the task at hand. The working memory is limited and as learning occurs it becomes overloaded which reduces the amount of information that can be moved to long term memory. CLT tells that there are many ways to utilize long term memory storage and ways to reduce the cognitive load which results in more space in working memory and easy learning. There are three different types of Cognitive load differentiated by CLT:

1. Extraneous Cognitive Load: It is based on how a material is presented (using diagrams or worked out examples)

2. Intrinsic Cognitive Load: It is based on the complexity of material which can be reduced by splitting the task using informal previous knowledge (Ayres, 2006; Seufert, Jnen, & Braken, 2007).

3. Germane Cognitive Load: It is based on building new schemas (F. G. W. C. Paas & Van Merrienboer, 1993) and refers to the load caused because of learning processes.

Understanding of CLT plays an important role in education for better instructional design, so that students are able to take most out of the lecture. The topic of how to measure the multidimensional construct of the subjective load has demonstrated
troublesome for scientists. Following F. G. Paas, Van Merrinboer, and Adam (1994) model, it is seen that cognitive load can be assessed by estimating mental load, mental effort, and performance. Therefore well-known measures of the mental workload from the discipline of Ergonomics (Human Factors) are used to assess the load.

2.2 Mental Workload

2.2.1 Foundations

Workload is thought of as a mental construct or a latent variable which is also applied to CLT (Longo & Leva, 2017; Wickens, 2017; Guastello, Marra, Correro, Michels, & Schimmel, 2017; Smith, 2017; Hancock, 2017). Sometimes, workload is also considered as an intervening variable (Gopher & Donchin, 1986). It reflects on interactions of mental demands imposed by tasks they attend to. Workload is thought to be multidimensional and multifaceted (Longo, 2015a). It results from the aggregation of many different demands. Because of these reasons, it is difficult to define mental workload uniquely. Since workload cannot be directly observed, it must be inferred from observations of different behaviors and measurement of psychological and physiological processes. The capabilities and effort of the operators in the context of specific situations all moderate the workload experienced by the operator.

There is no clear definition of MWL that is widely accepted, however, Moustafa et al. (2017) defines it as total cognitive load needed to accomplish a specific task under a finite period. Few others are, “Mental workload may be viewed as the difference between the capacities of the information processing system that are required for task performance to satisfy performance expectations and the capacity available at any given time” (Gopher & Donchin, 1986), “the mental effort that the human operator devotes to control or supervision relative to his capacity to expend mental effort” (Curry, Jex, Levison, & Stassen, 1979), etc. This concept is brought up when the necessity to know the complexity of tasks is felt. In Education, the goal to predict operator and system performance is the primary reason behind quantifying the mental cost of performing tasks (Cain, 2007).
Mental workload derives from the operators meta-controller activities: the cognitive “device” that directs attention, copes with interacting goals, selects strategies, adjusts to task complexity, sets performance tolerances, etc (Jex 1988). Alternatively, an operator faced with a task is fully engaged until the task is done, then is idle or engages in another task (Wierwille 1988). Workload is frequently described by terms such as mental strain and emotional strain. Both stress and workload involve environmental demands. The ability of the operator to cope with those demands depends on these environments.

In summary, a commonly accepted definition of mental workload still does not exist. Workload can be seen as a mental construct that reflects the mental strain resulting from performing a task under specific environmental and operational conditions, coupled with the capability of the operator to respond to those demands (Balfe, Crowley, Smith, & Longo 2017). Operational definitions will likely continue to be proposed and tested, but unless an imperative need arises for a universal definition, each field and perhaps each investigator will continue with their “culturally preferred” definition of workload.

### 2.2.2 Measurement Methods

In the last four decades, there are many techniques proposed to assess the mental workload and researchers in the applied studies prefer the use of a variety of measures rather than using only one. This behavior is acceptable given the multidimensional property of the Mental workload. The measurement techniques of MWL can be broadly classified into three main categories (Rizzo, Dondio, Delany, & Longo 2016; Cain 2007; Tsang & Vidulich 2006; M. S. Young & Stanton 2004):

1. **Self-assessment measure:** It is also known as self-reporting or subjective measure.
   
   It involves the analysis of subjective data collected from the participants interacting with the task and system. NASA Task Load Index (NASA-TLX) (Hart & Staveland 1988) and Workload Profile (WP) (Tsang & Velazquez 1996) are the most known methods. Both these techniques are multi-dimensional and are
often compared (Rizzo & Longo, 2017). Rating scales are subjective measures and have a long history for measuring feelings of workload, effort, mood, fatigue, etc. Self-assessments involve rating demands on numerical or graphical scales, typically anchored either at one or two extrema per scale. Some subjective techniques use scales that are categorical. Other techniques use an open-ended rating with a standard reference task as an anchor and subjects rate other tasks relative to the reference task. Subjective measures such as NASA-TLX and WP questionnaires are discussed briefly in the next section as they were used to collect the data during this experiment.

2. Task performance measures: It comprises of primary and secondary task measures. Primary task is aimed at directly quantifying the operators capacity to complete a task while the secondary task is based upon indirect quantification of MWL. Performance measures of workload can be classified into two major types: primary task measures and secondary task measures. Performance of primary task will always be of interest. In secondary task methods, performance of the secondary task itself may have no practical importance and serves only to load or measure the load of the operator.

Primary task measures attempt to assess the operators performance on the task of interest directly, and this is useful where the demands exceed the operators capacity such that performance degrades from baseline or ideal levels. Speed, accuracy, reaction or response times, and error rates are often used to assess primary task performance.

Secondary task measures provide an index of the remaining operator capacity while performing primary tasks. Secondary task measures are more diagnostic than primary task measures alone. The features of secondary task are used to infer the interaction between the primary and secondary tasks. The secondary task measure can be further classified into Auxiliary Task and Loading Task methodologies.

3. Physiological measures: It involves the analysis of physiological responses of the
operators body.

2.2.3 Subjective Measure

NASA-TLX

The NASA Task load Index is one of the most widely used mental workload assessment tool because of its multi-dimensional feature and easy in administration (Hart, 2006). It is built on an analogue scale which describes six independent sub-scales used to measure cognitive workload (Longo, 2017) and are as follows:

- Mental demand
- Physical demand
- Temporal demand
- Performance level
- Effort level
- Frustration level

After many years of research, it was found that these six dimension features are the reason behind the changes in the subjective workload and different types of task (Hart & Staveland, 1988). It also has pairwise comparisons features which are used to generate a weighted aggregate score from these dimensions. Each dimension \( d \) is associated with a weight \( w \) provided by the subject and a final overall workload rating is computed which is between 0 to 20 (in the experiment mentioned in this thesis) and is calculated as (Rizzo & Longo, 2018):

\[
NASA - TLX_{MWL} = \left( \sum_{i=1}^{6} d_i^*w_i \right) \left( \frac{1}{15} \right)
\]  

(2.1)

Workload Profile

Workload profile is a subjective workload assessment technique based on Wickens (2008) Multiple Resource Theory. This measurement technique is based on eight
dimensions: perceptual or central processing, response selection and execution, spatial processing, verbal processing, visual processing, auditory processing, manual output and speech output.

Based on the dimensions rating gathered from subjects, the workload profile is calculated using equation as follows:

$$WP_{MWL} = \left( \sum_{i=1}^{8} d_i \right)$$ (2.2)

2.2.4 Criteria for evaluating Measures

If workload is being measured in an experimental setting, the measurement options are generally wider than those for an operational setting. The workload measurement techniques that are available can be used successfully to differentiate among empirical conditions, and perhaps even produce interval or ration measures. There are major concerns regarding workload measures based on laboratory studies. Few of those are: lack of ecological validity/ context complexity, lack of subject acceptance, commitment or expertise, lack of assessment of the effect of strategy shifts both on performance, scheduling and on the workload measurements themselves (Longo, 2018b). There were several criteria that were laid to guide the selection or development of mental workload measurement techniques (Longo, 2014):

1. Sensitivity: The method chosen must be sensible to task difficulty changes and there should be significant variations in workload.

2. Diagnosticity: There should be a indication of the source of workload variation and quantify contribution by the type or resource demand. Thus, the method should be diagnostic.

3. Intrusiveness: The method should not be a significant source of workload by interfering with the performance of the operators task.

4. Acceptability: The method should be acceptable to the subjects.
5. Requirements: The method should comprise of minimal equipment so that it does not impair the subjects performance.

6. Bandwidth and Reliability: The method should be reliable and sufficiently rapid so that it can capture workload changes.

7. Selectivity: The method should be selectively sensitive to differences in capacity demand.

2.3 Natural Language Processing

Natural Language Processing (NLP) is a sub-field of Artificial Intelligence which can help computers with present generation computation speed to understand natural languages just the way human beings understand it (Goel, 2017). Natural language can be of any language, genre, mode, oral, written, etc. The main goal of a natural language processing system is to produce a language comprehension and production theory to such details that a normal programmer is able to write a computer program which can understand natural languages. Thus, a well-functioning NLP system will be able to rephrase an input text, answer questions about the contexts of the text, draw inferences about the text and translate the input text into another desired language. The foundations of NLP lie in a number of disciplines, like computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc. (Chowdhury, 2003). Applications of NLP include a number of fields of studies, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, artificial intelligence and expert systems, and so on. There are various tasks which a NLP system tries to solve:

1. Text Classification: It is the process of dividing the set of documents into two or more classes where each document belongs to one or multiple classes (Kaur & Saini, 2015). The process of classifying texts can be automated which increases
speed and efficiency rather than manual classification which requires more time and high accuracy (Kaur & Saini 2015).

2. Word Sequence: It is the process of assigning a set of labels to the entire sentence (T. Young, Hazarika, Poria, & Cambria 2017). Many NLP tasks can be cast into word sequence task such as part-of-speech tagging, text chunking, and named entity recognition (Zhang, Chen, Zhao, Liu, & Yin 2018).

3. Text Meaning: It is process of converting the whole set of documents into word vectors (Navigli 2018). This process involves finding similar words, sentence embeddings, topic modeling, search, and question answering.

4. Sequence to Sequence:

   It is process of machine translation, email summarization, simplification, and QA systems. Such systems are characterized by encoders and decoders (Chen, Qiu, Liu, & Huang 2018). This system works in finding hidden representation of the text.

2.3.1 Text Pre-Processing

In the area of text mining, in order to extract important, useful and non-trivial knowledge from the unstructured data, data pre-processing is a very critical step (Vijayarani, Ilamathi, & Nithya 2015). Pre-processing of text means cleaning of noise in the data. Noise can be of various forms like stop words, punctuation and words that do not carry enough weightage in the context of the text (Kalra & Aggarwal 2017). These text pre-processing methods are as follow:

1. Tokenizing: Initially, general text data is a set of characters. All processes in NLP requires analysis of words in the document. Thus, there is a need to tokenize every word in the document. Tokenizers provides the reliability of the documents. The process of dividing up text into meaningful words, or tokens, is a very important step in almost all applications of natural language processing.
Standard approach is single word tokenization where the input string is split into individual words using various delimiters (Manning & Schtze, 1999).

2. Stop-words removal: Optimization of processes for Information Retrieval, Text Summarization, Text and Data Analytic Systems becomes utmost important. Hence, in order to achieve high accuracy, redundant words that has low or no semantic meaning should be removed (Raulji & Saini, 2016). These words are called stop words. It is often assumed that topic models benefit from the use of a manually curated stop-word list (Schofield, Magnusson, & Mimno, 2017). Constructing this list can have many disadvantages like it can be time-consuming and it can be subject to user judgement about what kind of words will be important for user model. The words such as 'the', 'a', 'of', etc. are called as stop words.

3. Lemmatizing and Stemming: The increase in the size of the data and information collections over the past couple of years made it necessary for tools to be developed in order to access information with much ease. Stemming is a procedure to reduce all words with the same stem to a common form (Balakrishnan & Ethel, 2014). Lemmatization removes inflectional endings and returns the base or dictionary form of a word (Balakrishnan & Ethel, 2014). A stemming algorithm aims at obtaining the stem of the word by cleaning the affixes that carry grammatical or lexical information about the word (Manning & Schtze, 1999). The suffix removal algorithms in stemming cannot stem the alternate inflection of a word. This alternate inflection could have been declared in a different verb tense. Algorithms for lemmatization finds the lemma of the word which corresponds to collection of all words that have the same meaning.

2.3.2 Semantic Textual Similarity

Semantic Similarity is a metric which is defined for a set of documents or terms. It measures the distance between them where the distance represents the likeness of their meaning or semantic content as opposed to similarity. It is used to measure
the common characteristics between set of documents or set of words/terms (Slimani, 2013). Semantic similarity extends across numerous fields such as natural language processing, artificial intelligence and psychology. In order to perform tasks like document clustering, information retrieval and synonym extraction, accurate measurement between words is quite essential (Bollegala, Matsuo, & Ishizuka, 2009). If the two concepts in the document are not similar, then the two elements can share few of their semantic constitutive properties, semantic similarity, etc. A subset of these relatedness is used to evaluate the semantic interaction based on “is-a” hierarchy. Semantic similarity is used for biomedical informatics which compares genes and proteins based on the similarity of their functions, geo-informatics which measures similarity between concepts stored in geographic feature type ontologies, computational linguistics which constructs lexical database of English words, and natural language processing. There are various similarity measures:

1. Edge-based: It measures the similarity according to the number of semantic links separating the two concepts in the ontology. Edge-counting measures are able to provide reasonably accurate results when a detailed and taxonomically homogeneous ontology is used. This method just evaluates the shortest taxonomical path between concept pairs as evidences of distances. Compared to other ontology-based paradigms, edge-based counting usually provides the lowest accuracy.

2. Node-based: It measures the similarity by using graph nodes and their properties as the main data source. It compares the properties of the terms involved which can be related to the terms themselves, their ancestors or descendants (Albacete, Calle, Castro, & Cuadra, 2012). The most common method in node-based similarity measure is of information content (IC) which provides information about how specific and informative a term is. The most common node-based similarity measures are Resnik, Lin, Maguitman, Jiang and Conrath, DiShln, Align, Disambiguate and Walk.

3. Node-and-Relational-Content-based: It measures similarity by considering prop-
properties/content of nodes and consider types of relations between those nodes. It is applied in various fields which involves ontology. The node and relational content-based algorithms are based on eTVSM and Resniks similarity.

4. Pairwise: It is a similarity measure which compares term sets. Every term in the direct annotation set A is compared against every term in the direct annotation set B in pairwise approach (Kang & Gong, 2017). Thus, semantic similarity is considered by every pairwise combinations of terms from the two terms. These various combinations can be average, maximum or the sum of terms between two sets. Finally, only the best matching pair for each term is taken into consideration.

5. GroupWise: It is a similarity measure which compares term sets. It calculates the similarity by set, graph, or vector (Kang & Gong, 2017). Set approaches are not widely used since they only consider the direct annotations that would lose a lot of information. Based on set similarity techniques, graph approaches represent entities as the subgraphs of the whole annotations and calculate the similarity using graph matching techniques. Vector approaches compact the information in vector space (VS) as binary fingerprints which are more convenient for comparison. Various examples of group wise similarity measures are Jaccard index, simGIC, simLP, and simUI.

**Statistical Similarity**

In this kind of similarity, a model is built first and then similarity is estimated. Statistical similarity is learned from data which is a collection of written or spoken text. There are several models that are proposed to use along with statistical similarity (Majumder, Pakray, Gelbukh, & Pinto, 2016). Few of those are:

1. Latent Semantic Analysis: In this model, the contextual information of words has been extracted and represented from a large corpus of text. In the first step, text is represented as a matrix. This matrix has rows and columns which represents various unique words and text segments respectively. The frequency count of the word is represented by each entry which appears in the text.
2. Generalized Latent Semantic Analysis: One disadvantage of LSA is that its performance reduces because it generates word vectors from text corpus which is heterogeneous in nature. Advantage of GLSA over LSA is that it finds the terms and document vectors which are based on semantically motivated pair-wise term similarities \cite{Majumder et al., 2016}. In order for gaining more accuracy of the model, dimensionality reduction algorithms are also applied.

3. Explicit Semantic Analysis: In this method, meaning of any text is represented as a weighted vector of Wikipedia based concepts. Representation of vectors has been done over a high-dimensional space by using machine-learning algorithms. Wikipedia is used because it is a collection of largest encyclopedia which is defined by humans and can be easily explained.

4. Pointwise Mutual Information: This algorithm uses any search engine to issue a search query. It then analyzes the query result to find a synonym word. Thus, it is an unsupervised learning algorithm which recognizes the synonym of a problem word from a set of alternative words. The performance of this algorithm depends on two main things: power of the search engine query language and indexing of the search engine.

2.3.3 Word2Vec

Word2Vec is an extension of Latent Semantic Analysis, discussed in previous section. Word2Vec produces word embeddings by grouping related models which are shallow, two-layer neural networks that are trained to reconstruct linguistic construct of words. This neural network takes a large corpus of text as its input. It produces a vector space which consists of several hundred dimensions. Each unique word in the corpus corresponds to the vector in the space. Words vectors that share close relationship between each other share nearby positions in the space.

The purpose and usefulness of Word2Vec is to group the vectors of words that are similar together in vector space \cite{Mikolov, Chen, Corrado, & Dean, 2013}. It detects similarities between various words mathematically. Word2vec considers nu-
merical representations of word features and creates vectors. These features can be the context of individual words. It can do this work efficiently without human intervention. Word2vec can make accurate guesses about a word’s meaning based on past appearances with enough data, usage and contexts. These guesses can be used to create an association with other similar words. It can also be used to create a cluster of documents containing words that have similar meaning and to classify all the topics based on similarity. These clusters can then be used for various applications such as search, sentiment analysis and recommendations which can be used in various fields like scientific research, customer relationship management, etc. Consider for example, literature equals literature will give a similarity value of 1 whereas when compared with review, it gives a value of 0.8601. This is done by measuring cosine similarity. When the two words are not similar, they are expressed as a 90° angle while words that have perfect similarity will have a 0° angle between them or complete overlap. An example of vector space model compressed to 2-dimensional is seen in figure 2.1 produced using Literature Review corpus. It shows that the words such as ’bibliography’, ’reference’ and ’apa6’ are placed together as they appear close by in the corpus of text.

The results of Word2Vec can be sensitive to parametrisation. Few of the important parameters in Word2Vec are as follows (Caselles-Dupr, Lesaint, & Royo-Letelier, 2018):

1. Training algorithm: A model can be trained by using hierarchical softmax and/or
negative sampling. Hierarchical softmax works better for infrequent words which uses a Huffman tree to reduce calculation of the conditional log-likelihood a model seeks to maximize. While a negative sampling can be used for frequent words with low dimensional vectors which uses log-likelihood of sampled negative instances. One thing to keep in mind is that, there is a inverse relation between training epoch and hierarchical softmax. With increase in training epoch, hierarchical softmax stops being useful.

2. Dimensionality: In order to achieve high accurate results, the dimensionality of vectors is set to between 100 and 1,000. There is a direct relationship between quality of word and dimensionality. Quality of word embedding increases with higher dimensionality. This has a drawback as well. After reaching a threshold point, marginal gain will diminish.

3. Context window: How many words before and after a given word is decided by the size of the context window. This set of words will be included as the context words of the given word.

4. Sub-sampling: It is not good to have a word with a frequency above a threshold value. This is because words with high frequency provides little information. Thus, words with a frequency above a threshold value must be sub-sampled. This can have a huge impact on the training speed as well.

2.3.4 Cosine Similarity

Cosine similarity is an appropriate measure to calculate similarity between the word vectors produced by Word2Vec model. Cosine Similarity is used as a similarity measure which measures the cosine of an angle between two non-zero n-dimensional vectors in an n-dimensional space. The cosine of 0° is 1, thus two vectors with same orientation has cosine similarity of 1, while two vectors making a 90° angle have a similarity of 0. The cosine similarity is mostly used in high-dimensional positive space in information retrieval and text mining, where the outcome is between 0 to 1. It is a widely
implemented measure used for information retrieval and related studies. It models a text as term vectors. The similarity between two words or sentences can be found by calculating the cosine values between the two words term vectors\cite{Rahutomo, Kitasuka, Aritsugi, 2012}. It can be applied to any two texts such as words, sentences, paragraphs or whole documents.

It effectively calculates dot-product of two normalized vectors. Given two N dimension vectors $A$ and $B$, the cosine similarity between them is calculated as follows:

$$\text{Similarity} = \cos(\theta) = \frac{A \cdot B}{||A|| ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$  \hspace{1cm} (2.3)

The higher similarity score between two texts term vectors, more relevancy between them. Thus, it is a useful measure of how similar two texts are in terms of their subject matter.

### 2.4 Related Work on Mental Workload and Natural Language Processing

There are a lot of studies done to measure MWL under instructional designs to increase learning, related to Ergonomics based application \cite{Fallahi et al., 2016, Doebler et al., 2017, Zhang et al., 2018, Boele-Vos & Twisk, 2017}, human factors \cite{Wickens, 2008}, machine learning \cite{Zhang et al., 2018, Moustafa et al., 2017, Gmyzin, 2017} and other diverse areas. The application of Natural Language Processing in conjunction to MWL seems to rarely happen in the discipline of Education. Also, there is a lot of complexity involved in measuring MWL, thus giving the opportunity to explore new techniques like \cite{Moustafa et al., 2017} tried to find relationship between the indexes and actual class assigned in the MWL activity. \cite{Contreras, 2018} applied NLP to the subjective measures of the MWL. \cite{Gmyzin, 2017} used subjective techniques to compare the performances of theory-driven measures and supervised machine learning models that were trained using NASA-TLX and WP factors as features to predict class. \cite{Ott et al., 2016} built a supervised machine learning model by proposing a measure for MWL which used multi-modal metrics and unstructured linguistics.
2.5 Summary

2.5.1 Gaps in literature

A lot of work to increase learning by measuring Mental Workload in the fields of ergonomics, human factors, and Machine Learning has been done. From the literature, it is seen that there is not much work done in applying NLP techniques to MWL in the field of education. From the study of Contreras (2018) on MWL using NLP and from the gaps of the studies of Moustafa et al. (2017) and Gmyzin (2017), it is assuring that different types of NLP techniques can be applied on MWL. In both latter cases, the research was aimed to find the relationship between MWL scores and a class using correlation techniques. Contreras (2018) applied Natural Language Processing technique to get the similarity index from the text data which helps to measure MWL. The study gave the opportunity to try different NLP techniques.

2.5.2 Research Question

New and different approaches to assess and analyze Learning seems necessary to improve the instructional design. However, cognitive load assessment is not an easy activity, especially in a traditional classroom setting. The use of new approaches is increasing and that opens the door to a new research problem if Natural Language Processing can make a fair contribution to the measurement of cognitive load to evaluate learning in the field of education. This research proposes a novel method for evaluating learning both employing subjective cognitive load assessment and natural language processing. In this study, the content present in an instrument design is translated into text and keywords are processed using Natural Language Processing techniques. The output of that is then used to analyze the relationship between learning and MWL measures which leads to the research question. In details, on one hand, cognitive load assessment is performed using well-known self-reporting instruments, borrowed from Human Factors, namely the Nasa Task Load Index and the Workload Profile. On the other hand, Natural Language Processing techniques, borrowed from
Artificial Intelligence, are employed to calculate semantic similarity of textual information, provided by learners after attending a typical third-level classroom, and the content of the classroom itself. Accordingly, it will help to determine, “to what extent do self-reporting measures of Mental Workload correlate to semantic similarity of textual information provided by learners and a lecturer, and computed using Natural Language Processing techniques?”
Chapter 3

Experiment design and methodology

The aim of this chapter is to provide a definition of the hypotheses necessary to answer the research question. It also involves software selection, data understanding, data preparation, model design, evaluation and hypothesis testing and strengths and limitations of the design approach.

3.1 Software

Different tools are used to tackle various problems that arose during different steps of the experiment. Google sheets is used to correct the spelling mistakes that occurred during manual data entry. IBM SPSS statistics is being used for data exploration.

The Python Programming language has been used throughout all the phases of CRISP-DM model. The reason for selecting Python is, as it provides a number of packages and libraries to perform text analysis using Natural Language Processing techniques. In data understanding phase, python is used to analyze the corpus and student keywords by creating visualizations. Imputation is performed on the numerical data present in the dataset and text data is cleaned by removing stop-words and performing lemmatization before feeding it to the model. In the modelling phase, Word2Vec model has been built using gensim library of Python.
IBM SPSS Statistics software is used to explore the numerical data present in the NASA-TLX and WP datasets. The cosine similarity values calculated for further evaluation are also analyzed using SPSS. It is selected as it provides a range of easy techniques to assess the statistical properties of the data such as descriptive statistics and plots, normality and other statistical tests.

3.2 Data Understanding

In this phase, two datasets namely, NASA-TLX and Workload Profile are collected and loaded into the python environment and initial exploratory analysis is being done to get an idea about the datasets.

3.2.1 Datasets

![Data Collection process](image)

The data is gathered by conducting Mental workload experiments on the students attending third level classes in an university. The typical class schedule during the experiment is, first the lecturer tells the students what topic he is going to teach in the class. Before the teaching session, each and every student gets either a NASA-TLX questionnaire paper or the Workload Profile. All the students are expected to fill out the questionnaire which helps to find out the Mental Workload each and every student has before the lecture. Along with that, all the participants are asked to fill out at
most 15 concepts they are expecting to learn on the basis of the topic announced by the lecturer. A specific amount of time is given to the students to complete the questionnaire. As this occurs before the task of teaching, it is referred as pre-task. Lecturer teaches the class about the topic and then again a questionnaire is hand out to students either NASA-TLX or WP to note down the Mental Workload of the students after the lecture. Along with that they are asked to write at most 15 concepts they have learned from the lecture they attended. As this occurs after the task of teaching, it is referred as post-task. A specific amount of time is given to fill it out.

The electronic copy of the answers from the questionnaire of each and every student is maintained by doing manual data entry in Google Sheets. This is then used for further analysis.

**NASA-TLX dataset**

The table gives the description about NASA-TLX data set which consists of 65 columns and 224 records. It gives information about the topics taught, instructional design used by the professor to teach those topics, the scores of NASA-TLX features, and keywords taken from student before and after the lecture.

The NASA-TLX dataset is divided into 4 subsets of data on the basis of topics taught and those are Science, scientific method, literature review and planning research, to analyze each of them separately. The student keywords collected during pre and post tasks are explored by finding the frequency of them.

**Workload Profile dataset**

The table gives the description about the Workload Profile dataset which consists of 52 columns and 213 records which gives information about the topics taught, instructional design used by the professor to teach those topic, the scores of Workload Profile features, and keywords taken from student before and after the lecture.

The Workload Profile dataset is divided into 4 subsets of data on the basis of topics taught and those are Science, scientific method, literature review and planning research, to analyze each of them separately. The student keywords collected during
pre and post tasks are explored by finding the frequency of them.

Algorithm 1: Data Understanding of the NASA-TLX and WP datasets

1. Get the number of records for each topic
2. Get the column names and its data type
3. Get the descriptive statistics
4. Get the count of missing values in each column

3.2.2 Corpus

The data was collected by Dr. Luca Longo to analyze the effect of Mental Workload on learning by teaching four different topics in third-level classes, namely, research methods and computer science, scientific method, literature review and planning research. The topics were taught in the order mentioned. The lectures conducted were transformed into a transcript in a story telling format for further analysis and is considered as the core text or the corpus.

Learning is often an iterative process. It is expected that students develop a strong learning which involves some refinement each time a new topic is taught by lecturer. In the experiment conducted, the lecturer followed a specific order of topics to be taught (see figure 3.2). The order is science, scientific method, literature review and planning research. Three more corpora are created by forming a combination of the original four. This combination will help to determine if there is any temporal aspect of learning. The temporal aspect is increase in learning of student over a period of time.
by combining the learnings from previous classes. The science corpus was combined with scientific method and formed one big corpus which was tested on the keywords gathered for the topic taught latter i.e. scientific method. Same is done with the other corpora.

**Algorithm 2: Size of the corpus**

1. **Import libraries** (*NLTK*, `sent_tokenize`, `word_tokenize`)
2. `textfile = open('corpus.txt', "utf-8")`
3. `string = textfile.read()`
4. `SentList = sent_tokenize(string)`
5. `SentenceCount = length(SentList)`
6. `tokens = word_tokenize(string)`
7. `WordCount = length(tokens)`

1. **Science**

Science is a corpus having 52 sentences and 1941 tokens. The topic covers the information about the definition of science, its origin, types of sciences and scientists who has made a contribution to it. It also covers about research and the difference between research and knowledge. The discussion about science versus engineering then leads to the computer science and talks about its definitions and its mathematical and engineering origins.

2. **Scientific Method**

Scientific Method is a corpus having 66 sentences and 2397 tokens. The topic gives information about origins, history, definition, elements of scientific method, testing hypothesis, inductivism, falsification and steps in defining a scientific method.

3. **Planning Research**

Planning Research is a corpus having 34 sentences with 899 corpus. The topic starts with the ways to develop a research question by asking WH type questions. It also covers the fish model which helps in planning research.

4. **Literature Review**
Literature Review is a corpus having 128 sentences and 2458 tokens. It gives overview of literature review and its sub-parts and how to write it, which leads to formulation of research question. Furthermore, types of research question and its connection to the development research hypothesis and its types are discussed. The referencing style, plagiarism and its type and measures to avoid plagiarism are further mentioned.

5. **Science + Scientific Method** The Science + Scientific Method corpus is formed by combining Science and Scientific method corpora already mentioned. It is also referred as Sc + SM in the report and has 118 sentences and 4240 tokens.

6. **Science + Scientific Method + Planning Research** The Science + Scientific Method + Planning Research corpus is formed by combining Science, Scientific method and planning research corpora already mentioned. It is also referred as Sc + SM + PR in the report and has 152 sentences and 5127 tokens.

7. **Science + Scientific Method + Planning Research + Literature Review** The Science + Scientific Method + Planning Research + Literature Review corpus is formed by combining Science, Scientific method, planning research and literature review corpora already mentioned. It is also referred as Sc + SM + PR + LR in the report and has 280 sentences and 7514 tokens.

**Exploratory analysis of Corpus**

The number of sentences and tokens for each corpus are found to get an idea about the size of the dataset. Wordcloud is created to visualize and see which words have more importance in terms of frequency. The frequency of all the words in the corpus is determined.
Algorithm 3: Data Understanding of the corpus

1. `tokenCount = length(tokens)`
2. `distinctWords = length(set(tokens))`
3. `lexicalRichness = \frac{tokenCount}{distinctWords} \times 100`
4. `Frequency = FreqDist(tokens)`
5. `Top25 = Frequency.plot(25)`
6. Create `stopWordsList` (user defined list)
7. Create `punctuationList` (user defined list)
8. `for` `words in tokens` `do`
9. `if` `word.lower()` `not in stopWordsList` `then`
10. `Add the word to the tokensWithoutStopwords list` `end`
11. `end`
12. `countTokensWOStopwords = length(tokensWithoutStopwords)`
13. `for` `words in tokens` `do`
14. `if` `word.lower()` `not in punctuationList` `then`
15. `tokensWithoutPunct.append(word)` `end`
16. `end`
17. `countTokensWOPunct = length(tokensWithoutPunct)`
18. `for` `words in tokens` `do`
19. `if` `word.lower()` `not in stopWordsList` `and word.lower()` `not in punctuationList` `then`
20. `uniqueTokens.append(word)` `end`
21. `end`
22. `countUniqueTokens = length(uniqueTokens)`
23. `percentageStopwords = \frac{countTokensWOStopwords - tokenCount}{tokenCount} \times 100`
24. `percentageStopwords = \frac{countTokensWOPunct - tokenCount}{tokenCount} \times 100`
25. `percentageEffectiveTokens = \frac{countUniqueTokens}{tokenCount} \times 100`

3.3 Hypothesis Definition

A survey was conducted by Dr. Luca Longo in the third level classes of Dublin Institute of Technology (DIT) between 2015 to 2018.
HA: There is a strong correlation between Mental Workload and the difference of the cosine similarity value calculated between the keywords obtained from students before and after the third level class ($S2 - S1$) and the content taught by Lecturer (corpus).

The hypothesis is based on the assumption that cosine similarity values calculated from the keywords collected during the experiments of MWL at third level sessions and the topics in form of text data provide insights of the MWL activity. Hence, the difference of the cosine similarity of the keywords after the lecture and the cosine similarity of the keywords before the lecture ($S2-S1$) is analyzed and the strength of its relationship is evaluated against MWL.

### 3.4 Data Preparation

In this step of data preparation, the data is pre-processed to bring it to the most suitable form. It involves dealing with the data quality problems such as missing values, outliers in the subjective measures of Mental Workload and spelling mistakes, abbreviations check and other text cleaning processes on the set of students keywords present in NASA-TLX and Workload Profile datasets.
3.4.1 Datasets

NASA-TLX and Workload Profile (WP)

NASA-TLX and Workload Profile dataset’s numerical features go through some pre-processing before calculating MWL (see figure 3.4).

**Imputation**

For some of the dimensions in the NASA-TLX and Workload Profiles features, students have failed to note down the scores in the questionnaires during pre and post tasks. These missing values might be because of lack of understanding of the dimension by the students (see algorithm 4). K-nearest neighbour (K-NN) imputation technique is used to impute the missing values. The reason behind using K-NN is that the value for a variable are approximated by the values of points that are closest to it, based on other variables.

**Algorithm 4: Imputation**

```python
1  dataframe = read(MWL.csv)
2  Create a subset of original dataframe with columns useful for calculation
3  Convert all the columns to numeric data type
4  Replace all the missing values with NaN
5  Import library(fancyImpute)
6  Impute the missing values using K nearest neighbour algorithm
7  Round off the imputed values to bring it the uniformity in the column
8  Convert all the columns to integer datatype and store it as a dataframe
```
Worload Score Calculation

Algorithm 5: Data Preparation for NASA-TLX

1. Imputedf = Get the imputed dataframe from Algorithm 4
2. Create an object for each of the six dimension feature (d1,d2,...d6)
3. Create an object for each of the 15 pairwise comparison feature (pc1, pc2,...pc15)
4. Initialize error = False
5. for i in range of 0 to length of the d1: do
   6. if the values in d1,d2..d6 does not lie between 0 to 20 scale then
      7. error = True
   8. end
   9. if the values in pc1,pc2..pc15 is outside the range of 0 and 1 then
      10. error = True
   11. end
   12. if error = True then
      13. return i which is row number
      14. break
   15. end
   16. else
      17. weights(w1,w2...w6) are declared and initialized to zero for each of the dimension feature(d1,d2...d6)
      18. weights are incremented by one if the pc1,pc2..pc15 is associated with the corresponding d1,d2...d6
      19. Calculate the NASA-TLX workload using the formula, \( (d1[i]* w1 + ... + d6[i] * w6)/15 \) Append the score to the list. Reset the weights.
   20. end
21. end
22. Convert list to the dataframe. Merge the dataframe with NASA-TLX score with the imputed dataframe from Algorithm 4

The Mental Workload is calculated from the NASA-TLX and Workload Profile ratings using a mathematical equation discussed in previous chapter. NASA-TLX dimension feature scores and Pairwise comparison scores which are considered as weights are used to calculate the Mental Workload (see algorithm 5), whereas for the Workload Profile only dimension features are used (see algorithm 6).
Algorithm 6: Data Preparation for WP

1. Get the imputed dataframe from Algorithm 4
2. Create an object for each of the WP dimension features (d1,d2,...,d8)
3. Initialize error = False
4. for i in range of 0 to length of the d1: do
   5.  if the values in d1,d2..d6 does not lie between 0 to 20 scale then
      6.     error = True
   7.  end
   8.  if error = True then
      9.     return i which is row number
     10.    break
   11.  end
   12. else
   13.    Calculate the WP workload using the formula, \( (d1+d2+...d8) \) Append the score to the list.
   14. end
15. end
16. Convert list to the dataframe
17. Merge the dataframe with NASA-TLX score with the imputed dataframe from Algorithm 4

Keywords

While conducting Mental Workload experiments, students were asked to write 15 concepts each containing at most five keywords related to topic. During the pre task they were asked to write about what they are expecting to learn from the class by just knowing the topic and during the post task they were asked to write what they have learned from the class.

Both the questionnaires were collected on the same day before and after the class. Therefore, NASA-TLX and WP datasets contains 15 features each having keywords for pre and post tasks. It is possible that the some of the students left empty spaces due to various reasons such as students running out of time to complete the questionnaire, not interested in the topic or not able to recall and write 15 concepts learned from the session. To deal with this, the blank spaces are replaced by the term ‘unknown’.
Data understanding will be performed on the new set of dataframes. This analysis comprises of finding out the total number of tokens, unique tokens, percentage of stopwords, percentage of punctuation symbols, total number of effective tokens present, the number of missing values for pre and post list of each dataset.

A specific flow is being followed for the data preparation phase for keywords. Firstly, the keywords are inspected for spelling mistakes and are corrected using the Spell Checker tool of Google sheets. Using Python programming language, all the keywords present as a sentence are cleaned by converting them to lower case then performing sentence and word tokenisation, removing the stop-words and punctuation symbols, lemmatize it to the base form by considering the part of speech of the words.
in terms of noun, verb, adjective and adverbs (see algorithm 7).

**Algorithm 7**: Data Preparation of Keywords and Corpus

1. Create user defined list of stop words and punctuation symbols
2. **Function** sentence_to_wordlist(tokenized words)
3. Create an empty list list 1.
4. Keep the tokens having letters, numbers and percentage symbol using regular expression
5. **for** i in range of 0 to length of the tokens: **do**
   6. if lowercase(words) not in stop words list **then**
   7. Append words to the list 1
   8. **end**
   **end**
10. return list created
11. **End function**
12. **Function** readKeywords(keywordsList)
13. Create an empty list list 2
14. **for** each list in the nestedkeywordsList **do**
   15. Create an empty list list 3
   16. **for** each item in list **do**
   17. **end**
   18. if length of an item is > 0 **then**
   19. Save the results returned from the **Function** sentence_to_wordlist(item)
   20. Lemmatized the returned tokens
   21. Append the results to list 3
   **end**
23. Append the tokenized result to list 2
24. **end**

The preprocessed tokens are checked for any abbreviations that the students might have used and are replaced with the common terminology to bring uniformity. A private dictionary has been created with abbreviation as the key and the words to replace them as value. The words with maximum of three letters are checked if they are an abbreviation by inspecting the words adjacent to it and if found, are replaced with the values present in the dictionary (refer algorithm 8).
Algorithm 8: Inspection and replacement of Abbreviations

1. Create user defined dictionary of key and value pair with key being Abbreviation and value is its word phrase

2. Function Check_Abbreviation(KeywordsList)

3. for each student list in KeywordsList do

4.     for each word list in student list do

5.         for each token in the word list do

6.             if length of token <= 2 then

7.                 Find the index of the token and check if it is greater than or equal to 0 and if yes get the previous token

8.                 Find the next token if present and save it

9.                 Print the current, previous and next token and take user input, 1 for token is abbreviation else 0

10.                if user input is 1 then

11.                    if token is present in the dictionary then

12.                        Print the suggestion from the dictionary and take user input, 1 for accepting the suggestion update the list else 0

13.                    end

14.                end

15.            end

16.        end

17.    end

18. return updated list

3.4.2 Corpus

The same procedure is followed for the corpora of all the four topics and three newly created corpora. As the model takes the input as the nested list, the corpora is converted into that form. To achieve that, first the sentence tokenization is performed on corpora, creating a tokenized list of sentences. The words present in those sentences are then tokenized. These lists are then cleaned by removing the stop-words and punctuation symbols and are lemmatized to the base form in terms of noun, verb, adjective and adverb.
3.5 Modelling

In this phase, seven Word2Vec models are built on seven corpora of which there are four original story-telling documents and three more corpora are created by combining the original four. The aim of the model is to find semantic similarity between the corpora and the student keywords. This is done by creating word embedding. Refer figure 3.6 for the flow of the modelling phase.

**Algorithm 9: Modelling**

1. Build Word2Vec Model on the preprocessed corpus from the Algorithm 7 by choosing appropriate parameters
2. Explore the model
3. Save the model in .w2v format to use it anytime
4. Compress the word vectors into 2D space
5. Plot the graphs to explore

Algorithm 9 is followed to build the model and using algorithm 10 average word vectors are found for each sentence called sentence vectors to further calculate semantic similarity.
CHAPTER 3. EXPERIMENT DESIGN AND METHODOLOGY

Algorithm 10: Sentence vectors for corpus and student keywords

1 Function `sent_vectorizer(sentence,model)`
2 Create an empty numpy vector(`sent_vect`) with the same number of dimensions of the vectors in the model
3 Initialize a variable for keeping the word count in a sentence to zero
4 for `word` in a `sentence` do
5     Append the vector for the word from the model and append it to the `sent_vect`
6     Increment the word count variable by one
7 end
8 if `word count` > 0 then
9     Divide the `sent_vect` by the word count to get the average
10 end

Semantic Analysis of corpus and student keywords

The Word2Vec takes the corpus as input and produces vector space with a number of dimensions, with each unique word in the corpus assigned to a corresponding vector in the vector space. The words that share common context or are closer to one another in a sentence are placed closed to each other in the vector space giving it a semantic meaning.

To find the semantic similarity between the content taught by professor called the corpus and the students keywords, the corpus and the keywords written by each student is converted into word vectors using the model built (refer Algorithm 10), giving the same semantic meaning to the keywords written by students. There are at most 15 sentences written by each student are converted to vectors which are then stored in a list. The average of those vectors are taken called sentence vectors to compute the cosine similarity (refer Algorithm 11) between the sentence vector of the corpus to tell how similar students learning is with the Lecturer’s teaching. Using the most suitable measure of central tendency, cosine similarity signifying learning of each student is then calculated for further analysis.
3.6 Evaluation and Hypothesis Testing

The relationship between the MWL measures obtained from NASA-TLX and WP datasets and cosine similarity computed between corpus and student keywords is calculated using correlation. The Pearson correlation, a parametric technique and Spearman and Kendall’s Tau correlation which are non parametric techniques are selected to explore the relationship.

The Pearson Correlation Coefficient denoted by $r$ measures the strength of linear relationship between two variables. There are few assumptions that needs to be satisfied by the variables before proceeding with the Pearson Correlation, such as
the variables should be either interval or ratio and should be normally distributed. Also, the outliers should be kept minimum and presence of homoscedasticity should be checked. If these assumptions fail, then Spearman ($r_s$) or Kendall’s Tau correlation should be performed. The correlation value can range from +1 to -1. When there is no relationship present, the correlation gives value of 0(zero). The closer the value towards +1 or -1 the stronger the relationship. Positive correlation value means that if value of one variable increases, the value of other increases whereas, negative correlation value means that if one value increases, the other decreases.

Accepting or rejecting Hypothesis

The alternate hypothesis will be rejected if the correlation coefficients, $r$, $r_s$ or $Tb$ has a non significant value ($p > 0.05$), stating that there is no statistical significant relationship between the MWL and difference of the cosine similarity values obtained from the keywords during the post-task and pre-task ($S2-S1$) under different corpora.

3.7 Strength and limitations of designed approach

3.7.1 Strengths

The first major strength is that the approach outlined in this chapter that is the use of word embedding(Word2Vec), a natural language processing technique in conjunction to Mental Workload assessment technique is novel. A lot of algorithms are described to understand the corpora of topics by finding the number of tokens, its lexical richness, distribution of words across the corpora. Same analysis is done on the keywords gathered from the students during the pre-task and post-task. The algorithms also describes the steps to pre-process the tokens by removing the stop words and punctuation symbols, lemmatization, identifying and replacing abbreviations, replace missing values, calculate the NASA-TLX and WP workload scores, identify and impute the missing values present in the numerical data and textual data, thus maintaining the number of records as the dataset is not too big.
Another strength of this design is that along with the learning of the students related to the topic, the temporal aspect of learning is also analyzed. To test those aspects, wide range of records are evaluated as the datasets are based on NASA-TLX and WP which has both numerical and textual data.

The research conducted gives scope and base to future work that can be conducted on MWL using different Natural language processing techniques.

3.7.2 Limitations

The main limitation to this research is the size of the corpora and the size of data sets. The small size of the corpora creates an issue as the Word2Vec model built on it covers words present only in the corpora. If the students write synonyms of the words present in the corpus, the model will not understand as it tries to maintain the semantic similarity between the words by creating the vocabulary of words only present in the corpus. Although the NASA-TLX and WP datasets were gathered from 2015 to 2018, the number of records are not optimal as they were further divided on the basis of the topic taught for analysis.

Another major limitation is the presence of missing values in the keywords section of dataset. The reason for those can be student’s handwriting is not clear while filling the data manually to create electronic version of data, student gave up writing due to some reason or the time to fill up the questionnaire was over. As the analysis of the learning is based on the keywords, lesser the number of keywords, less accurate is the analysis. It creates issues such as higher value of cosine similarity even if the student wrote about only one concept out of 15.

Lastly, the misspelling and abbreviations creates a challenge. The misspelling were corrected using the Google Sheets’ spell checker. Even though, an algorithm is described to replace the abbreviation with the best suitable and appropriate word phrase, it takes a lot of time to create a user defined dictionary with abbreviations and word phrase. As during the pre-task questionnaire, no one knows what students will write when they are asked what they expect to learn from the session giving a wide range of possibility of different abbreviations.
Chapter 4

Implementation and results

The aim of this chapter is to present the results obtained by implementing the designed research described in the previous chapter. The chapter covers the following topics:

- Data Understanding
- Data Preparation
- Modelling
- Strengths and Limitation of findings

4.1 Data Understanding

4.1.1 Datasets

In this step, data present in NASA-TLX and Workload Profile is inspected for data quality problems with the methods mentioned in the previous chapter.

NASA-TLX dataset

A table with descriptive statistics of the numerical variables is presented in the table [4.1]. It gives information about Dimensions and Pairwise Comparison features of the NASA-TLX such as the number of records (N), Minimum, Maximum, Mean and
CHAPTER 4. IMPLEMENTATION AND RESULTS

Standard Deviation of each variable. All the 224 participants responded to the Dimension features of NASA-TLX and their scale varies from 0 to 20. However, there are some missing values seen for the Pairwise Comparison features and the values for those variables is either 0 or 1.

Table 4.1: Descriptive Statistics of NASA-TLX Dataset

<table>
<thead>
<tr>
<th>Dimensions of NASA-TLX</th>
<th>Feature</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NASA_Mental</td>
<td>224</td>
<td>1</td>
<td>20</td>
<td>10.01</td>
<td>3.417</td>
</tr>
<tr>
<td></td>
<td>NASA_Physical</td>
<td>224</td>
<td>1</td>
<td>20</td>
<td>6.27</td>
<td>4.166</td>
</tr>
<tr>
<td></td>
<td>NASA_Temporal</td>
<td>223</td>
<td>1</td>
<td>20</td>
<td>9.05</td>
<td>3.561</td>
</tr>
<tr>
<td></td>
<td>NASA_Performance</td>
<td>224</td>
<td>2</td>
<td>17</td>
<td>8.85</td>
<td>3.608</td>
</tr>
<tr>
<td></td>
<td>NASA_Frustration</td>
<td>224</td>
<td>1</td>
<td>17</td>
<td>7.64</td>
<td>3.855</td>
</tr>
<tr>
<td></td>
<td>NASA_Effort</td>
<td>224</td>
<td>1</td>
<td>20</td>
<td>9.89</td>
<td>3.859</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pairwise Comparisons of NASA-TLX</th>
<th>Feature</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NASA_TempFrus</td>
<td>221</td>
<td>0</td>
<td>1</td>
<td>0.21</td>
<td>0.407</td>
</tr>
<tr>
<td></td>
<td>NASA_PerMen</td>
<td>220</td>
<td>0</td>
<td>1</td>
<td>0.53</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>NASA_MenPhy</td>
<td>220</td>
<td>0</td>
<td>1</td>
<td>0.11</td>
<td>0.312</td>
</tr>
<tr>
<td></td>
<td>NASA_FrusPer</td>
<td>220</td>
<td>0</td>
<td>1</td>
<td>0.80</td>
<td>0.401</td>
</tr>
<tr>
<td></td>
<td>NASA_TempEffort</td>
<td>217</td>
<td>0</td>
<td>1</td>
<td>0.62</td>
<td>0.487</td>
</tr>
<tr>
<td></td>
<td>NASA_PhyFrus</td>
<td>221</td>
<td>0</td>
<td>1</td>
<td>0.48</td>
<td>0.501</td>
</tr>
<tr>
<td></td>
<td>NASA_PerpTemp</td>
<td>222</td>
<td>0</td>
<td>1</td>
<td>0.43</td>
<td>0.497</td>
</tr>
<tr>
<td></td>
<td>NASA_MenEffort</td>
<td>220</td>
<td>0</td>
<td>1</td>
<td>0.36</td>
<td>0.481</td>
</tr>
<tr>
<td></td>
<td>NASA_PhyTemp</td>
<td>217</td>
<td>0</td>
<td>1</td>
<td>0.81</td>
<td>0.392</td>
</tr>
<tr>
<td></td>
<td>NASA_FrustEffort</td>
<td>221</td>
<td>0</td>
<td>1</td>
<td>0.80</td>
<td>0.404</td>
</tr>
<tr>
<td></td>
<td>NASA_PhyPerf</td>
<td>221</td>
<td>0</td>
<td>1</td>
<td>0.88</td>
<td>0.328</td>
</tr>
<tr>
<td></td>
<td>NASA_TempMen</td>
<td>220</td>
<td>0</td>
<td>1</td>
<td>0.70</td>
<td>0.459</td>
</tr>
<tr>
<td></td>
<td>NASA_EffortPhy</td>
<td>222</td>
<td>0</td>
<td>1</td>
<td>0.10</td>
<td>0.299</td>
</tr>
<tr>
<td></td>
<td>NASA_FrustMen</td>
<td>221</td>
<td>0</td>
<td>1</td>
<td>0.81</td>
<td>0.393</td>
</tr>
<tr>
<td></td>
<td>NASA_PerfEffort</td>
<td>221</td>
<td>0</td>
<td>1</td>
<td>0.50</td>
<td>0.501</td>
</tr>
</tbody>
</table>

A table of frequencies was created for the keywords written by the students before and after the task (refer table 4.2). Every student had to write at most 15 key concepts.
they wish to learn before the teaching session which are shown in the table with variable names having pre as the suffix and 15 concepts after the session and has post as the suffix in their variable names. The table shows that all the keywords features have missing values and it can be seen that as the number of keywords increases, the number of participants writing all the 15 concepts decreases which can be inferred as the number of missing values increases.

Table 4.2: Frequency count of Keywords in NASA-TLX Dataset

<table>
<thead>
<tr>
<th>Pre</th>
<th>Count</th>
<th>Post</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>NASA_k1.pre</td>
<td>220</td>
<td>NASA_k1.post</td>
<td>222</td>
</tr>
<tr>
<td>NASA_k2.pre</td>
<td>220</td>
<td>NASA_k2.post</td>
<td>219</td>
</tr>
<tr>
<td>NASA_k3.pre</td>
<td>214</td>
<td>NASA_k3.post</td>
<td>217</td>
</tr>
<tr>
<td>NASA_k4.pre</td>
<td>213</td>
<td>NASA_k4.post</td>
<td>217</td>
</tr>
<tr>
<td>NASA_k5.pre</td>
<td>210</td>
<td>NASA_k5.post</td>
<td>213</td>
</tr>
<tr>
<td>NASA_k6.pre</td>
<td>203</td>
<td>NASA_k6.post</td>
<td>208</td>
</tr>
<tr>
<td>NASA_k7.pre</td>
<td>201</td>
<td>NASA_k7.post</td>
<td>204</td>
</tr>
<tr>
<td>NASA_k8.pre</td>
<td>193</td>
<td>NASA_k8.post</td>
<td>195</td>
</tr>
<tr>
<td>NASA_k9.pre</td>
<td>183</td>
<td>NASA_k9.post</td>
<td>190</td>
</tr>
<tr>
<td>NASA_k10.pre</td>
<td>181</td>
<td>NASA_k10.post</td>
<td>183</td>
</tr>
<tr>
<td>NASA_k11.pre</td>
<td>168</td>
<td>NASA_k11.post</td>
<td>177</td>
</tr>
<tr>
<td>NASA_k12.pre</td>
<td>163</td>
<td>NASA_k12.post</td>
<td>167</td>
</tr>
<tr>
<td>NASA_k13.pre</td>
<td>159</td>
<td>NASA_k13.post</td>
<td>162</td>
</tr>
<tr>
<td>NASA_k14.pre</td>
<td>158</td>
<td>NASA_k14.post</td>
<td>156</td>
</tr>
<tr>
<td>NASA_k15.pre</td>
<td>153</td>
<td>NASA_k15.post</td>
<td>151</td>
</tr>
</tbody>
</table>

The Workload Profile dataset is explored by calculating descriptive statistics. There are 213 participants, who undertook workload profile assessment method of mental workload. The scale of the features of the workload profile is between 0 and 20. Table 4.3 shows information about the number of records (N), number of missing values, Minimum, Maximum, Mean and standard deviation of each Dimension feature. WP_task-space, WP_auditory_resources, WP_manual_response and WP_speech_response have missing values which will be imputed using the technique mentioned in the previous chapter.
Table 4.3: Descriptive Statistics of Workload Profile Dataset

<table>
<thead>
<tr>
<th>Feature</th>
<th>N</th>
<th>Missing</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP_solving_deciding</td>
<td>213</td>
<td>0</td>
<td>1</td>
<td>20</td>
<td>11.00</td>
<td>3.899</td>
</tr>
<tr>
<td>WP_response_selection</td>
<td>213</td>
<td>0</td>
<td>1</td>
<td>20</td>
<td>9.86</td>
<td>4.369</td>
</tr>
<tr>
<td>WP_task_space</td>
<td>212</td>
<td>3</td>
<td>1</td>
<td>20</td>
<td>8.78</td>
<td>4.582</td>
</tr>
<tr>
<td>WP_verbal_material</td>
<td>213</td>
<td>0</td>
<td>1</td>
<td>20</td>
<td>12.41</td>
<td>3.860</td>
</tr>
<tr>
<td>WP_visual_resources</td>
<td>213</td>
<td>0</td>
<td>2</td>
<td>17</td>
<td>12.26</td>
<td>3.840</td>
</tr>
<tr>
<td>WP_auditory_resources</td>
<td>212</td>
<td>1</td>
<td>1</td>
<td>20</td>
<td>12.77</td>
<td>3.780</td>
</tr>
<tr>
<td>WP_manual_response</td>
<td>213</td>
<td>2</td>
<td>0</td>
<td>20</td>
<td>9.34</td>
<td>4.913</td>
</tr>
<tr>
<td>WP_speech_response</td>
<td>213</td>
<td>2</td>
<td>0</td>
<td>20</td>
<td>9.19</td>
<td>5.011</td>
</tr>
</tbody>
</table>

Table 4.4: Frequency count of Keywords in Workload Profile Dataset

<table>
<thead>
<tr>
<th>Pre</th>
<th>Count</th>
<th>Post</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP_k1_pre</td>
<td>213</td>
<td>WP_k1_post</td>
<td>212</td>
</tr>
<tr>
<td>WP_k2_pre</td>
<td>213</td>
<td>WP_k2_post</td>
<td>211</td>
</tr>
<tr>
<td>WP_k3_pre</td>
<td>213</td>
<td>WP_k3_post</td>
<td>211</td>
</tr>
<tr>
<td>WP_k4_pre</td>
<td>211</td>
<td>WP_k4_post</td>
<td>211</td>
</tr>
<tr>
<td>WP_k5_pre</td>
<td>209</td>
<td>WP_k5_post</td>
<td>205</td>
</tr>
<tr>
<td>WP_k6_pre</td>
<td>205</td>
<td>WP_k6_post</td>
<td>203</td>
</tr>
<tr>
<td>WP_k7_pre</td>
<td>201</td>
<td>WP_k7_post</td>
<td>203</td>
</tr>
<tr>
<td>WP_k8_pre</td>
<td>192</td>
<td>WP_k8_post</td>
<td>201</td>
</tr>
<tr>
<td>WP_k9_pre</td>
<td>184</td>
<td>WP_k9_post</td>
<td>197</td>
</tr>
<tr>
<td>WP_k10_pre</td>
<td>178</td>
<td>WP_k10_post</td>
<td>192</td>
</tr>
<tr>
<td>WP_k11_pre</td>
<td>171</td>
<td>WP_k11_post</td>
<td>184</td>
</tr>
<tr>
<td>WP_k12_pre</td>
<td>168</td>
<td>WP_k12_post</td>
<td>178</td>
</tr>
<tr>
<td>WP_k13_pre</td>
<td>158</td>
<td>WP_k13_post</td>
<td>165</td>
</tr>
<tr>
<td>WP_k14_pre</td>
<td>154</td>
<td>WP_k14_post</td>
<td>159</td>
</tr>
<tr>
<td>WP_k15_pre</td>
<td>146</td>
<td>WP_k15_post</td>
<td>155</td>
</tr>
</tbody>
</table>

A table of frequencies was created for the keywords written by the students before and after the task (see table 4.4). Every student had to write at most 15 key concepts they wish to learn before the teaching session which are shown in the table with variable names having pre as the suffix and 15 concepts after the session and has

45
post as the suffix in their variable names. The table shows that all the keywords features have missing values except first three pre keywords: WP_k1.pre, WP.k2.pre and WP.k3.pre It can be seen that as the number of keywords increases, the number of participants writing all the 15 concepts decreases which can be inferred as the number of missing values increases.

4.1.2 Corpus

The basic understanding of the seven corpora is done using algorithm 2 and algorithm 3 mentioned in the previous chapter. The table 4.5 gives information about the total tokens, unique tokens, lexical richness, stop words and punctuation symbols count and effective tokens in the each corpus.

Table 4.5: Data Understanding of the Corpus

<table>
<thead>
<tr>
<th></th>
<th>Sc¹</th>
<th>SM²</th>
<th>PR³</th>
<th>LR⁴</th>
<th>(Sc + SM)</th>
<th>(Sc + SM + PR)</th>
<th>(Sc + SM + PR + LR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Tokens</td>
<td>1893</td>
<td>2347</td>
<td>886</td>
<td>2387</td>
<td>4240</td>
<td>5127</td>
<td>7514</td>
</tr>
<tr>
<td>Total Unique Tokens</td>
<td>745</td>
<td>785</td>
<td>349</td>
<td>687</td>
<td>1313</td>
<td>1510</td>
<td>1892</td>
</tr>
<tr>
<td>Lexical Richness (%)</td>
<td>39</td>
<td>33</td>
<td>39</td>
<td>29</td>
<td>31</td>
<td>29</td>
<td>25</td>
</tr>
<tr>
<td>Stop words (%)</td>
<td>28.47</td>
<td>31.79</td>
<td>37.70</td>
<td>35.57</td>
<td>30.31</td>
<td>31.58</td>
<td>32.85</td>
</tr>
<tr>
<td>Punctuation Symbols (%)</td>
<td>20.81</td>
<td>18.41</td>
<td>17.72</td>
<td>17.68</td>
<td>19.48</td>
<td>19.19</td>
<td>18.71</td>
</tr>
<tr>
<td>Total Effective Tokens</td>
<td>960</td>
<td>1169</td>
<td>395</td>
<td>1116</td>
<td>2129</td>
<td>2524</td>
<td>3640</td>
</tr>
<tr>
<td>Effective Tokens (%)</td>
<td>50.71</td>
<td>49.81</td>
<td>44.58</td>
<td>46.75</td>
<td>50.21</td>
<td>49.23</td>
<td>48.44</td>
</tr>
</tbody>
</table>

¹ Sc : Science ² SM : Scientific Method ³ PR : Planning Research ⁴ LR : Literature Review

1. Science

![Figure 4.1: Science Corpus Wordcloud](image)
The Science (Sc) corpus has 1893 total tokens of which 745 are unique. The corpus has 39% lexical richness. The higher the number, the more lexically rich the corpus is, it means the more unique words (and punctuation) it contains compared to the total number of words and punctuation it contains. Almost 29% of words are stop words and 21% are punctuation symbols. There are 960 effective tokens which is almost 51% of words in the corpus which are clean and are without any stop words or punctuation symbols.

![Frequency of Top 25 most common tokens in Science Corpus](image)

Figure 4.2: Frequency of Top 25 most common tokens in Science Corpus

The wordcloud (see figure 4.1) and figure 4.2 shows that Science is the most frequent word in the corpus followed by the word computer as the topic is about science and gives information about science and computer science as well, whereas ”,” is the highest used stop word with frequency count of approximately 150 and of is commonly used stop word in the Science corpus (see figure 4.2).

2. **Scientific Method**
The Scientific Method (SM) corpus has 2347 total tokens of which 785 are unique. The corpus has 33% of lexical richness. Almost 32% of words are stop words and 19% are punctuation symbols. There are 1169 effective tokens which is almost 50% of tokens in the corpus which are clean and are without any stop words or punctuation symbols.

The wordcloud (see figure 4.3) and figure 4.4 shows that Scientific, method, science and hypothesis are the most frequent words used in the corpus which is obvious as the topic is about scientific method, whereas ”,” (comma) is the highest used punctuation symbol with frequency count of approximately 140 and the is commonly used stop word in the Science corpus (see Figure 4.4).

3. Planning Research
The Planning Research (PR) corpus has 886 total tokens of which 349 are unique. The corpus has 39% of lexical richness. Almost 38% of words are stop words and 18% are punctuation symbols. There are 395 effective tokens which is almost 45% of tokens in the corpus which are clean and are without any stop words or punctuation symbols.

The wordcloud (see Figure 4.5) and Figure 4.6 shows that research and problem are the most frequent words used in the corpus as the topic tries to solve plan the research by solving the research problem, whereas ”-” (hyphen) is the highest used punctuation symbol with frequency count of approximately 50 and to is commonly used stop word in the Planning Research corpus (see Figure 4.6).

4. Literature Review
The Literature Review (LR) corpus has 2387 total tokens of which 687 are unique. The corpus has 29% of lexical richness. Almost 36% of words are stop words and 18% are punctuation symbols. There are 1116 effective tokens which is almost 47% of tokens in the corpus which are clean and are without any stop words or punctuation symbols.

The wordcloud (see Figure 4.7) and Figure 4.8 shows that Research, review, question and literature are the most frequent words used in the corpus which is obvious as the topic is about literature review, whereas ”?” (question mark) is the highest used stop word with frequency count of approximately 140 and the is commonly used stop word and most common token as well in the Literature Review corpus (see Figure 4.8).
Science + Scientific Method Corpus

Science + Scientific Method (Sc+SM) corpus is formed by combining Science and Scientific Method corpora. Therefore, the total number of tokens are 4240 of which 1313 tokens are unique which gives 31% lexical richness. The corpus comprises of almost 30% of stop words and 20% punctuation symbols. There are 2129 effective tokens which is 50% of the corpus.

The wordcloud (see Figure 4.9 and Figure 4.8) shows that science, scientific, method and hypothesis are the most frequent words used in the corpus which is obvious as the corpus is formed combining science and scientific method, whereas “,” (comma) is the highest used punctuation symbol with frequency count of
approximately 290 and *of* and *the* is commonly used stop word in the science +
scientific method corpus (see Figure 4.10).

6. **Science + Scientific Method + Planning Research Corpus**

![Figure 4.11: Science + Scientific Method + Planning Research Corpus Wordcloud](image)

Science + Scientific Method + Planning Research (Sc+SM+PR) corpus is formed
by combining Science, Scientific Method and Planning research corpora. There-
fore, the total number of tokens are 5127 of which 1510 tokens are unique which
gives 29% lexical richness. The corpus comprises of almost 32% of stop words
and 19% punctuation symbols. There are 2524 effective tokens which is 49% of
the corpus.

![Figure 4.12: Frequency of Top 25 most common tokens in Sc + SM + PR Corpus](image)

The wordcloud (see Figure 4.11) and Figure 4.12 shows that *science*, *scientific*
and *method* are the most frequent words used in the corpus which is obvious as
the corpus is formed combining science, scientific method and planning research,
whereas ”,” (comma) is the highest used punctuation symbol with frequency count of approximately 310 and of and the is commonly used stop word in the Science + Scientific Method + Planning Research corpus.

7. Science + Scientific Method + Planning Research + Literature Review Corpus

Science + Scientific Method + Planning Research + Literature Review (Sc+SM+PR+LR) corpus is formed by combining Science, Scientific Method, Planning research and Literature Review corpora. Therefore, the total number of tokens are 7514 of which 1892 tokens are unique which gives 25% lexical richness. The corpus comprises of approximately 33% of stop words and 19% punctuation symbols. There are 3640 effective tokens which is 48% of the corpus.
The wordcloud (see Figure 4.13) and Figure 4.14 shows that research, science and scientific are the most frequent words used in the corpus as the corpus is formed combining all the four topics, whereas ’,’ (comma) is the highest used punctuation symbol with frequency count of approximately 400 and the is commonly used stop word in the Science + Scientific Method + Planning Research + Literature Review corpus (see Figure 4.14)

4.2 Data Preparation

The data preparation phase involves solving the data quality problems associated with the numerical as well as textual data present in both the datasets: NASA-TLX and Workload Profile.

4.2.1 Datasets

The datasets were explored and it was seen that datasets has some missing values in both numerical and textual data. Also, to proceed with modelling, mental workload was calculated using subjective assessment techniques like NASA-TLX and Workload Profile.

NASA-TLX and Workload Profile

During the data understanding technique, Pairwise Comparison features were found to have some missing values. Instead of using measures of central tendency to replace the Nan, K-nearest neighbour algorithm was used as the values of the variables are either 0 and 1. Missing values were imputed with the value closest to the $k$ nearest cases to the input case having missing value. Same procedure is followed for the features in the Workload Profile dataset. Algorithm 4 was followed to impute the values.

After imputation, the updated dataset was used to calculate the workload score. It is a two step process which uses the Dimension features which are rating and Pairwise comparison which are weights. Algorithm 5 is used for the same from the previous
chapter. Workload from the Workload Profile dataset is calculated using the Algorithm 6. The workload score is calculated separately for each of the seven corpora making it easy for the modelling step.

Keywords

All the keywords collected along with the NASA-TLX and WP assessment techniques were pre-processed using the Algorithm 7 by converting all the tokens to lower case, removing stopwords and punctuation symbols then lemmatizing the effective tokens to the base form. Those tokens are then checked for abbreviations to avoid any loss of information. The tokens abbreviated such as 'x' or 'vs' were replaced by 'versus', 'cs' by 'computer science', 'sm' by 'scientific method', etc.

The Table 4.6 gives information about the descriptive statistics of the NASA-TLX keywords from pre-task and post-task. It is seen that the pre-task keywords collected before the lecture of Science topic has the highest number of total tokens, unique tokens and thus the lexical richness amongst all the topics taught, however Literature topic has the highest percentage of the effective tokens during the pre-task. Science topic has the highest number of total tokens during the post-task as well but the topic planning research has the high lexical richness and high percentage of effective tokens than the other three topic’s post-task keywords.

Table 4.6: Descriptive Statistics of the Keywords of NASA-TLX pre-task and post-task

<table>
<thead>
<tr>
<th>Nasa Pre</th>
<th>Nasa Post</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sc</td>
</tr>
<tr>
<td>Total Tokens</td>
<td>2840</td>
</tr>
<tr>
<td>Total Unique Tokens</td>
<td>1423</td>
</tr>
<tr>
<td>Lexical Richness (%)</td>
<td>49.26</td>
</tr>
<tr>
<td>Stop words (%)</td>
<td>9.20</td>
</tr>
<tr>
<td>Punctuation Symbols (%)</td>
<td>30.27</td>
</tr>
<tr>
<td>Total Effective Tokens</td>
<td>1717</td>
</tr>
<tr>
<td>Effective Tokens (%)</td>
<td>60.53</td>
</tr>
</tbody>
</table>

Sc : Science
SM : Scientific Method
PR : Planning Research
LR : Literature Review
The table 4.7 gives information about the descriptive statistics of the WP keywords from pre-task and post-task. It is seen that the pre-task keywords collected before the lecture of Science topic has the highest number of total tokens, unique tokens and thus the lexical richness amongst all the topics taught. Science and Literature topics has the highest percentage of the effective tokens of almost 60% during the pre-task. Science topic has the highest number of total tokens during the post-task as well but the topic planning research has the high lexical richness. The Scientific method topic has the high percentage of effective tokens than the other three topic’s post-task keywords who has almost similar percentage of effective tokens.

| Table 4.7: Descriptive Statistics of the Keywords of WP pre-task and post-task |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                                 | WP Pre          | WP Post         |                 |                 |                 |                 |                 |
|                                 | Sc  | SM  | PR  | LR  | Sc  | SM  | PR  | LR  |
| Total Tokens                    | 2728| 2853| 2305| 2186| 3048| 2847| 2417| 2414|
| Total Unique Tokens             | 1373| 1426| 1190| 1139| 1413| 1450| 1352| 1222|
| Lexical Richness (%)            | 50.07| 50.05| 51.37| 51.94| 46.32| 50.75| 55.16| 51.15|
| Stop words (%)                  | 9.52| 11.40| 9.63| 7.44| 13.76| 8.00| 12.06| 11.21|
| Punctuation Symbols(%)          | 30.45| 31.82| 31.08| 32.26| 28.69| 31.60| 30.04| 29.82|
| Total Effective Tokens          | 1629| 1614| 1367| 1314| 1748| 1727| 1402| 1407|
| Effective Tokens(%)             | 60.03| 56.78| 59.29| 60.19| 57.56| 60.39| 57.90| 58.97|
| Sc     : Science                | SM  : Scientific Method |
| PR     : Planning Research      | LR  : Literature Review  |

4.2.2 Corpus

The Corpus is pre-processed to feed it to the model in the most suitable form. It starts with sentence tokenization. The semantic meaning of the sentence is preserved by keeping the words of a sentence in the list. Thus, a nested list is created having tokenized sentences. The words inside each sentence is then tokenized by performing word tokenization. Those tokens are checked for any stop words or punctuation symbols from the manually created private list of stop words and are removed if present in the list. The leftover tokens are lemmatized to bring to the base form making it easier to compare with the lemmatized keywords.
4.3 Modelling

The modelling phase is aimed to determine the semantic similarity between the corpora of the topics taught by the lecturer and the keywords written by the students before and after the lecture to determine learning. This is done by calculating the cosine similarity between the average of vector of words in a sentence of the corpus with the average vector of sentences of the keywords written by students.

To achieve this, Word2Vec models are created on the seven corpora to create word embedding. The dimensionality of the resulting word vectors known as size is set to 300 as greater number of dimensions gives more accurate results. The minimum word count (min_count) is set to 1 as this parameter ignores the words that does not satisfy the min_count and the corpora are very short and even the word having frequency of one is important to achieve the goal. The context size is kept to 5. The threshold for the occurrence of the word given as sample is usually set between 0 to 1e-5 to downsample frequent words. Seven models are built by setting it the parameters mentioned and seven corpora are passed to create word vectors.

First, the sentence vectors are calculated for the corpus using the Word2Vec model for the corresponding topic and then the sentence vectors are calculated for the student keywords using Algorithm 10. The cosine similarity is calculated between each of the sentence vector for a particular student with all the sentence vectors present for a corpus. Maximum cosine similarity value is then retained considering it is the most representative similarity value for that sentence with the corpus. Using the most appropriate measure of central tendency, one similarity value was chosen to proceed with evaluation. In this case, median was used as the distribution of the values may or may not be normal. Same procedure is followed for the pre-task and post-task of both the datasets of NASA-TLX and WP of all the seven corpora.

All those cosine similarity values were then saved in a .csv file as further process of evaluation is done on SPSS.
4.3.1 NASA-TLX

1. Science

(a) NASA-TLX Workload Score

The distribution of NASA-TLX score is analyzed for the Science topic calculated using Algorithm 5 based on the figures and normality test performed in SPSS.

![Histogram and Boxplot](image.png)

Figure 4.15: Assessment of Normality for NASA-TLX Science

<table>
<thead>
<tr>
<th>Simi_Nasa_Tlx</th>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.079</td>
<td>58</td>
<td>0.200</td>
</tr>
</tbody>
</table>

From the Figure 4.15a, it is seen that the mean of the NASA-TLX score of the 58 participants is 8.51, standard deviation is 2.671 with a normal distribution. Figure 4.15b suggests that there are no outliers and the Kolmogorov-Smirnov test (see Table 4.8) confirms that the NASA-TLX data for Science topic is normally distributed because of the non-significant result ($p > 0.05$).

(b) NASA-TLX Post Task

The distribution, linearity and homoscedasticity of the Cosine Similarity values obtained for the keywords written by 58 participants in post-task.
are analyzed along with the test of Kolmogorov-Smirnov test of normality to check if it satisfies the assumptions for further evaluation.

![Histogram](image1)
![Boxplot](image2)
![Scatterplot](image3)

(a) Histogram  
(b) Boxplot  
(c) Scatterplot

Figure 4.16: NASA-TLX Post task: Assessment of Normality

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simi_Nasa_Post</td>
<td>0.308</td>
<td>58</td>
</tr>
</tbody>
</table>

Table 4.9: NASA-TLX Post task: Kolmogorov-Smirnov Test of Normality

From the Figure 4.16 it is seen that because of the outlier present in the data, neither a normal distribution curve is seen nor a linear graph. The Kolmogorov-Smirnov test of normality also gives statistical significant results thus showing that the data is not normally distributed. To solve the issue of normality, log transformation is performed on the data.

![Histogram](image4)
![Boxplot](image5)
![Scatterplot](image6)

(a) Histogram  
(b) Boxplot  
(c) Scatterplot

Figure 4.17: NASA-TLX Post task: Assessment of Normality (log transformation)
After log transformation, it is seen that there are no outliers present and there is also a normal distribution curve, but there is no linear relationship (see figure 4.17). The test of normality shows non significant results \((p>0.05)\), thus telling that the data is normal (see Table 4.10).

(c) NASA-TLX Post-Pre Task

The cosine similarity values obtained by taking the difference of the cosine similarity values of the keywords from the pre-task and from the post task is analyzed. There is an outlier present which cannot be removed because the behaviour of that student seems unusual and is important which needs to be inspected.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simi_log_Nasa_Post</td>
<td>0.114</td>
<td>58</td>
</tr>
</tbody>
</table>

The Table confirms that the data is not normal as the results given by the test are statistically significant.

2. Scientific Method
(a) NASA-TLX Workload Score

The distribution of NASA-TLX score is analyzed for the Scientific Method topic calculated using Algorithm 5 based on the figures and normality test performed in SPSS.

![Histogram and Boxplot](image)

Figure 4.19: NASA-TLX Workload score: Assessment of Normality

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simi_Nasa_Tlx</td>
<td>0.049</td>
<td>61</td>
</tr>
</tbody>
</table>

From the Figure 4.19a it is seen that the mean of the NASA-TLX score of the 61 participants is 10.11, standard deviation is 2.282 with a normal distribution. Figure 4.19b suggests that there are no outliers and the Kolmogorov-Smirnov test (see Table ??) confirms that the NASA-TLX data for Scientific Method topic is normally distributed because of the non-significant result ($p > 0.05$).

(b) NASA-TLX Post Task

The distribution, linearity and homoscedasticity of the Cosine Similarity values obtained for the keywords written by 61 participants in post-task are analyzed along with the test of Kolmogorov-Smirnov test of normality to check if it satisfies the assumptions for further evaluation.
Figure 4.20: NASA-TLX post task for Scientific Method: Assessment of Normality for NASA-TLX

Scientific Method

Table 4.13: NASA-TLX post task for Scientific Method: Kolmogorov-Smirnov Test of Normality

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simi_Nasa_Post</td>
<td>0.089</td>
<td>61</td>
</tr>
</tbody>
</table>

From the Figure 4.20b, it is seen that there is an outlier present in the data. The case 43 which acts as an outlier cannot be removed as it shows that it has the highest similarity with the corpus. The Kolmogorov-Smirnov test of normality also gives non statistical significant results thus showing that the data is normally distributed.

(c) NASA-TLX Post-Pre Task

The cosine similarity values obtained by taking the difference of the cosine similarity values of the keywords from the pre-task and from the post task is analyzed. There is an outlier present which cannot be removed because the behaviour of that student seems unusual and is important which needs to be inspected.
The Figure 4.21a shows a normal distribution curve and no outliers are seen from the Figure 4.21b but no linear relationship is observed from the scatterplot (see Figure 4.21c).

Table 4.14: NASA-TLX Post-Pre Scientific Method: Kolmogorov-Smirnov Test of Normality

<table>
<thead>
<tr>
<th>Simi_Nasa_PostPre</th>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.054</td>
<td>61</td>
<td>0.200</td>
</tr>
</tbody>
</table>

The test of normality shows non significant results thus stating that the data is normally distributed.

3. Planning Research

(a) NASA-TLX Workload Score

The distribution of NASA-TLX Workload score is analyzed for the Planning Research topic calculated using Algorithm 5 based on the figures and normality test performed in SPSS.
CHAPTER 4. IMPLEMENTATION AND RESULTS

Figure 4.22: NASA-TLX Planning Research: Assessment of Normality

![Histogram and Boxplot](image)

(a) Histogram  
(b) Boxplot

Table 4.15: NASA-TLX Planning Research: Kolmogorov-Smirnov Test of Normality

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simi_Nasa_Tlx</td>
<td>53</td>
<td>0.060</td>
</tr>
</tbody>
</table>

From the Figure 4.22a, a normal distribution curve is seen and there are no outliers as well (see Figure 4.22b). The Table 4.16 also shows that the data is normally distributed as the \( p-value > 0.05 \).

(b) NASA-TLX Post Task

The distribution, linearity and homoscedasticity of the Cosine Similarity values obtained for the keywords written by 61 participants in post-task are analyzed using Figure 4.23 along with the test of Kolmogorov-Smirnov test of normality to check if it satisfies the assumptions for further evaluation.

![Histogram, Boxplot, and Scatterplot](image)

(a) Histogram  
(b) Boxplot  
(c) Scatterplot

Figure 4.23: NASA-TLX Planning Research: Assessment of Normality

64
Table 4.16: NASA-TLX Planning Research: Kolmogorov-Smirnov Test of Normality

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simi_Nasa_Post</td>
<td>0.109</td>
<td>53</td>
</tr>
</tbody>
</table>

From the Figure 4.23b it is seen that an outlier is present in the data and there is not linear relationship present between the NASA-TLX post cosine similarity value and NASA-TLX workload score for the planning research topic. However, a normal distribution curve is seen. The Kolmogorov-Smirnov test of normality also gives non statistical significant results thus showing that the data is normally distributed.

(c) NASA-TLX Post-Pre Task

The cosine similarity values obtained by taking the difference of the cosine similarity values of the keywords from the pre-task and from the post task is analyzed.

Figure 4.24: NASA-TLX Post-Pre Planning Research: Assessment of Normality

(a) Histogram  
(b) Boxplot  
(c) Scatterplot

Table 4.17: Kolmogorov-Smirnov Test of Normality

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simi_Nasa_PostPre</td>
<td>0.095</td>
<td>53</td>
</tr>
</tbody>
</table>

From the Figure 4.24 it is seen that there are no outliers present, there is a normal distribution curve over the data and no linear relationship is seen.
Thus, it does not satisfy all the assumptions for the evaluation. The test of normality shows that the data is normally distributed.

4. Literature Review

(a) NASA-TLX Workload Score

The distribution of NASA-TLX score is analyzed for the Literature Review topic calculated using Algorithm 5 based on the figures and normality test performed in SPSS.

From the figure 4.25 it is seen that there is a normal distribution with no outliers present. The mean of that data is 0.99. The test of normality shows the data of the 61 participants is normally distributed (see Table 4.18).

Table 4.18: NASA-TLX Literature Review: Kolmogorov-Smirnov Test of Normality

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simi_Nasa_Tlx</td>
<td>0.079</td>
<td>51</td>
</tr>
</tbody>
</table>

(b) NASA-TLX Post Task

The distribution, linearity and homoscedasticity of the Cosine Similarity values obtained for the keywords written by 61 participants in post-task are analyzed along with the test of Kolmogorov-Smirnov test of normality to check if it satisfies the assumptions for further evaluation.
CHAPTER 4. IMPLEMENTATION AND RESULTS

(a) Histogram  (b) Boxplot  (c) Scatterplot

Figure 4.26: NASA-TLX Post task Literature Review: Assessment of Normality

Table 4.19: NASA-TLX Post task Literature Review: Kolmogorov-Smirnov Test of Normality

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simi_Nasa_Post</td>
<td>61</td>
<td>0.200</td>
</tr>
</tbody>
</table>

Figure 4.26 shows that, the frequency distribution of the cosine similarity values of the post task have a normal distribution even though there is an outlier present in the data. The outlier is not removed as the data gives some information about the learning of the student. There is no linear relationship between the cosine similarity and NASA-TLX workload score for the Literature Review post task. The test of normality (see Table 4.19) shows non significant result stating that the data is normally distributed.

(c) NASA-TLX Post-Pre Task

The distribution of cosine similarity score calculated from the difference of pre task and post task is analyzed. It is seen that neither there is normal distribution of frequencies of cosine similarity nor there is a normal relationship present between the NASA-TLX workload score and the cosine similarity of post-pre task as there are outliers present in the data. The outliers are not removed as it shows that cases 50 and 51 has a huge learning difference before and after teaching session which is an useful data. Also, the difference is high, that means those two students have learnt a lot.
5. Science + Scientific Method Corpus on Scientific Method Keywords

(a) NASA-TLX Workload Score

The analysis is done between the Science + Scientific Method corpus and the keywords collected during the Scientific Method topic. The distribution of NASA-TLX score is already analyzed (see Figure 4.19 and Table 4.12).

(b) NASA-TLX Post Task

The distribution, linearity and homoscedasticity of the Cosine Similarity values obtained for the keywords written by 61 participants in post-task are analyzed along with the test of Kolmogorov-Smirnov test of normality to check if it satisfies the assumptions for further evaluation.
CHAPTER 4. IMPLEMENTATION AND RESULTS

(a) Histogram  (b) Boxplot  (c) Scatterplot

Figure 4.28: NASA-TLX Post task Science + Scientific Method: Assessment of Normality

Table 4.21: NASA-TLX Post task Science + Scientific Method: Kolmogorov-Smirnov Test of Normality

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simi_Nasa_Post</td>
<td>0.079</td>
<td>61</td>
</tr>
</tbody>
</table>

From the Figure 4.28, normal frequency distribution is seen with no outliers. No linear relationship is seen between the NASA-TLX workload score and the NASA-TLX post task cosine similarity value. However, the data is normally distributed (see Table 4.21).

(c) NASA-TLX Post-Pre Task

The difference of cosine similarity values obtained during pre-task and post-task are analyzed and their relationship with NASA-TLX workload score is explored.

(a) Histogram  (b) Boxplot  (c) Scatterplot

Figure 4.29: NASA-TLX Post-Pre Science + Scientific Method: Assessment of Normality

69
Table 4.22: -TLX Post-Pre Science + Scientific Method: Kolmogorov-Smirnov Test of Normality

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simi_Nasa_PostPre</td>
<td>61</td>
<td>0.000</td>
</tr>
</tbody>
</table>

From the Figure 4.29, non normal frequency distribution with no outliers is seen along with no linear relationship between the data. Table 4.22 shows statistically significant results which means that the data is not normally distributed.

6. Science + Scientific Method + Planning Research Corpus on Planning Research Keywords

(a) NASA-TLX Workload Score

The analysis is done between the Science + Scientific Method + Planning Research corpus and the keywords collected during the Planning Research topic. The distribution of NASA-TLX score is already analyzed (see Figure 4.23 and Table 4.16).

(b) NASA-TLX Post Task

The distribution of NASA-TLX score is analyzed for the Planning Research topic calculated using Algorithm 5 based on the figures and normality test performed in SPSS.

Figure 4.30: Assessment of Normality for NASA-TLX post task: Sc + SM + PR
Table 4.23: Kolmogorov-Smirnov Test of Normality for Sc + SM + PR

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simi_Nasa_Post</td>
<td>61</td>
<td>0.200</td>
</tr>
</tbody>
</table>

(c) NASA-TLX Post-Pre Task

The distribution of NASA-TLX score is analyzed for the Planning Research topic calculated using Algorithm 5 based on the figures and normality test performed in SPSS.

![Histogram](image1)
![Boxplot](image2)
![Scatterplot](image3)

Figure 4.31: Assessment of Normality for NASA-TLX Post-Pre: Sc + SM + PR

Table 4.24: Kolmogorov-Smirnov Test of Normality for Sc + SM + PR

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simi_Nasa_PostPre</td>
<td>61</td>
<td>0.200</td>
</tr>
</tbody>
</table>

7. Science + Scientific Method + Planning Research + Literature Review Corpus on Literature Review Keywords

(a) NASA-TLX Workload Score

The distribution of NASA-TLX score is already analyzed for the Literature review topic calculated using Algorithm 5 based on the figures and normality test performed in SPSS using Figure 4.25 and Table 4.18.

(b) NASA-TLX Post Task
Table 4.25: NASA-TLX post-task Sc+SM+PR+LR: Kolmogorov-Smirnov Test of Normality

<table>
<thead>
<tr>
<th>Statistic</th>
<th>df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simi_Nasa_Post</td>
<td>0.103</td>
<td>51</td>
</tr>
</tbody>
</table>

From the Figure 4.32, there is an outlier and no linear relationship is present. It is also seen that the test of normality shows statistically significant results, thus the data is normally distributed.

(c) NASA-TLX Post-Pre Task

The cosine similarity values of Post-pre task are analyzed for distribution, linearity, outliers etc. from Figure 4.33
There is an outlier and no linear relationship is present between the variables. The histogram (see Figure 4.33) and the test of normality shows that the data is not normally distributed (see Table 4.26).

During this section, 19 tests were conducted to assess the normality of the NASA-TLX workload score and cosine similarity values associated with pre-task and post-task of each of the seven corpora namely: Science, Scientific Method, Planning Research, Literature Review, Science + Scientific Method, Science + Scientific Method + Planning Research and Science + Scientific Method + Planning Research + Literature Review. It involved analysis of frequency distribution through histogram, checking for outliers through boxplot, check if linear relationship exists between the NASA-TLX workload profile and cosine similarity values and conducting Kilmogorov-Smirnov tests of normality to check for any statistical significant results.

The skewness is observed in the cosine similarity values obtained for post-pre task for almost all the topics. After analysis it was seen that the reason behind the skewness is presence of outliers. Those outliers were not removed as it gave important information like, the student had no knowledge prior to the teaching session hence the cosine similarity value for the post-task for higher, as the cosine similarity value for pre-task was zero.

The calculations showed that even though the data was normally distributed, there were outliers present in the data and for all the corpora no linearity was found between the NASA-TLX workload score and cosine similarity values, thus not satifying all the assumptions to conduct parametric tests such as Pearson Correlation. Hence, Non parametric tests are selected for further evaluation.
4.3.2 Workload Profile

For the Workload Profile dataset, same procedure applied to the NASA-TLX dataset in the previous section is followed as part of modelling technique. Thus, 18 tests were conducted to assess the normality of the WP and cosine similarity values on each of seven corpora. It involved analysis of frequency distribution through histogram, checking for outliers through boxplot, check if linear relationship exists between the NASA-TLX workload profile and cosine similarity values and conducting Kilmogorov-Smirnov tests of normality to check for any statistical significant results. The graphs and tables of the tests are annexed in Appendix A.1.

Table 4.27: Kolmogorov-Smirnov Test of Normality for WP

<table>
<thead>
<tr>
<th>Topic</th>
<th>df</th>
<th>MWL</th>
<th>Pre</th>
<th>Post-Pre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>47</td>
<td>0.200</td>
<td>0.200</td>
<td>0.200</td>
</tr>
<tr>
<td>Scientific Method</td>
<td>60</td>
<td>0.200</td>
<td>0.200</td>
<td>0.200</td>
</tr>
<tr>
<td>Planning Research</td>
<td>48</td>
<td>0.084</td>
<td>0.200</td>
<td>0.200</td>
</tr>
<tr>
<td>Literature Review</td>
<td>47</td>
<td>0.200</td>
<td>0.200</td>
<td>0.200</td>
</tr>
<tr>
<td>Sc + SM</td>
<td>60</td>
<td>0.200</td>
<td>0.200</td>
<td>0.200</td>
</tr>
<tr>
<td>Sc + SM + PR</td>
<td>48</td>
<td>0.200</td>
<td>0.200</td>
<td>0.200</td>
</tr>
<tr>
<td>Sc + SM + PR + LR</td>
<td>47</td>
<td>0.200</td>
<td>0.200</td>
<td>0.200</td>
</tr>
</tbody>
</table>

Sc : Science  SM : Scientific Method
PR : Planning Research  LR : Literature Review

From the Figure 4.27, it is seen that all the topics have normally distributed data. There is no linearity present between the workload score and cosine similarity values. Hence, Non-parametrics tests will also be performed as it does not satisfy the assumptions for Parametric tests.
Chapter 5

Evaluation and analysis

In this chapter, the results obtained from the previous chapter are evaluated by implementing the methodology proposed. The relationship between the MWL and cosine similarities calculated for the keywords collected during pre-task and post-task of third level teaching sessions on the topics: Science, Scientific Method, Planning Research and Literature Review. Three more corpora are created using combination of original four topics. On the basis of the evaluation, hypothesis stated during Chapter 3 will be either accepted or rejected. The strengths and limitations of the findings are also discussed in this chapter.

5.1 Hypothesis Testing

The analysis of the features involved in the dataset were performed in the previous chapter. It was seen that most of the features had outliers and none of them had linearity between the variables. Even though the data was normally distributed for most of them but due the presence of outliers non-parametric tests are selected to be performed on the data as they are not sensitive to the outliers. Spearman correlation and Kendall Tau’s correlation is performed to find the relationship between the MWL and the cosine similarities of the keywords.
CHAPTER 5. EVALUATION AND ANALYSIS

5.1.1 NASA-TLX

Accepting or rejecting Hypothesis

The relationship between the MWL and the cosine similarity values calculated between the keywords and respective corpus during the post-task of teaching sessions in the third level classes for the corpora: Science, Scientific Method, Planning Research, Literature Review, Science + Scientific Method, Science + Scientific Method + Planning Research, Science + Scientific Method + Planning Research + Literature Review was investigated using Pearson Correlation Coefficient ($r$), Spearman correlation coefficient ($r_s$) and Kendall’s Tau-b ($T_b$). Even though the data for the post-task is normally distributed, due to the presence of outliers, non parametric correlation tests are also performed.

Table 5.1: NASA-TLX Workload Score and Cosine Similarity for Post-Task: Table of Correlations

<table>
<thead>
<tr>
<th>Topic</th>
<th>$r$</th>
<th>p-value</th>
<th>$r_s$</th>
<th>p-value</th>
<th>$T_b$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>-0.154</td>
<td>0.247</td>
<td>-0.072</td>
<td>0.589</td>
<td>-0.044</td>
<td>0.629</td>
</tr>
<tr>
<td>Scientific Method</td>
<td>0.164</td>
<td>0.207</td>
<td>0.108</td>
<td>0.405</td>
<td>0.075</td>
<td>0.397</td>
</tr>
<tr>
<td>Planning Research</td>
<td>0.063</td>
<td>0.656</td>
<td>0.092</td>
<td>0.517</td>
<td>0.057</td>
<td>0.554</td>
</tr>
<tr>
<td>Literature Review</td>
<td>-0.074</td>
<td>0.606</td>
<td>-0.109</td>
<td>0.447</td>
<td>-0.065</td>
<td>0.500</td>
</tr>
<tr>
<td>Sc + SM</td>
<td>0.343</td>
<td>0.007*</td>
<td>0.317</td>
<td>0.013*</td>
<td>0.223</td>
<td>0.012*</td>
</tr>
<tr>
<td>Sc + SM + PR</td>
<td>-0.042</td>
<td>0.764</td>
<td>-0.094</td>
<td>0.505</td>
<td>-0.073</td>
<td>0.438</td>
</tr>
<tr>
<td>Sc + SM + PR + LR</td>
<td>0.024</td>
<td>0.868</td>
<td>-0.022</td>
<td>0.879</td>
<td>-0.021</td>
<td>0.836</td>
</tr>
</tbody>
</table>

Sc : Science  
SM : Scientific Method  
PR : Planning Research  
LR : Literature Review

From the Table 5.1 above, it can be observed that the correlation for the corpora such as Scientific Method, Planning Research, Science + Scientific Method and Science + Scientific Method + Planning Research + Literature Review shows positive relationship between the cosine similarity values and the MWL that means as the similarity between the keywords and corpus increases, MWL increases but from the p-values, it can be seen that none of the correlation coefficients are statistically significant. While rest of the corpora have negative relationship.
CHAPTER 5. EVALUATION AND ANALYSIS

It is seen that the Science + Scientific Method corpus has a weak positive relationship \( (r = 0.343, n = 61, p < 0.01) \) between the MWL calculated from the NASA-TLX and the cosine similarity values of the keywords gathered during the post-task with statistically significant result. This shows temporal aspect of learning. The students have used the knowledge from the previous class as well to fill up the questionnaire of keywords. The two variables that correlate \( r = 0.343 \) share only 11.76\% \( (0.343\times0.343 = 0.1176 \times 100) \) of their variance, thus, indicating the presence of an overlap between the two variables. In this sense, the cosine similarity values of the post-task helps to explain nearly 12\% of the variance in students’ scores on the Mental Workload scale (NASA-TLX).

Table 5.2: NASA-TLX and Cosine Similarity for Post-Pre Task: Table of Correlations

<table>
<thead>
<tr>
<th>Topic</th>
<th>( rs )</th>
<th>p-value</th>
<th>( \text{Tau-b} )</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>-0.053</td>
<td>0.691</td>
<td>-0.036</td>
<td>0.687</td>
</tr>
<tr>
<td>Scientific Method</td>
<td>0.059</td>
<td>0.655</td>
<td>0.040</td>
<td>0.651</td>
</tr>
<tr>
<td>Planning Research</td>
<td>0.125</td>
<td>0.376</td>
<td>0.085</td>
<td>0.372</td>
</tr>
<tr>
<td>Literature Review</td>
<td>-0.126</td>
<td>0.377</td>
<td>-0.084</td>
<td>0.385</td>
</tr>
<tr>
<td>Sc + SM</td>
<td>0.338</td>
<td>0.008*</td>
<td>0.229</td>
<td>0.010*</td>
</tr>
<tr>
<td>Sc + SM + PR</td>
<td>-0.007</td>
<td>0.961</td>
<td>-0.011</td>
<td>0.908</td>
</tr>
<tr>
<td>Sc + SM + PR + LR</td>
<td>0.114</td>
<td>0.436</td>
<td>0.086</td>
<td>0.384</td>
</tr>
</tbody>
</table>

Sc : Science
SM : Scientific Method
PR : Planning Research
LR : Literature Review

From the Table 5.2 above, it can be observed that the correlation for the corpora such as Scientific Method, Planning Research, Science + Scientific Method and Science + Scientific Method + Planning Research + Literature Review shows positive relationship between the cosine similarity values and the MWL that means as the similarity between the keywords and corpus increases, MWL increases but from the p-values, it can be seen that none of the correlation coefficients are statistically significant. While rest of the corpora have negative relationship with no statistically significant results.

It is seen that the Science + Scientific Method corpus has a weak positive relationship from the Spearman correlation coefficient \( (rs = 0.338, n = 61, p < 0.01) \)
between the MWL calculated from the NASA-TLX and the cosine similarity values of the keywords gathered during the post-pre task with statistically significant result. This shows temporal aspect of learning. The students have used the knowledge from the previous class as well to fill up the questionnaire of keywords. The two variables that correlate $r = 0.338$ share only 11.42% ($0.338^2 = 0.1142 \times 100$) of their variance, thus, indicating the presence of an overlap between the two variables. In this sense, the difference of the cosine similarity values of the post-task and pre-task helps to explain nearly 11% of the variance in students’ scores on the Mental Workload scale (NASA-TLX). Based on the correlation coefficients shown in table 5.1 and table 5.2, although not all the results show statistically significant results, HA is accepted as the assumption is met when the strength of relationship between MWL and difference of cosine similarity values of post-task and pre-task is compared for Science + Scientific Method corpora. On the other hand, HA is rejected for all the other corpora as there are no statistical significant results.

5.1.2 Workload Profile

The relationship between the MWL and cosine similarity values calculated between the keywords and respective corpus of WP datasets during the post-task of teaching sessions in the third level classes for the corpora: Science, Scientific Method, Planning Research, Literature Review, Science + Scientific Method, Science + Scientific Method + Planning Research, Science + Scientific Method + Planning Research + Literature Review was investigated using Pearson Correlation Coefficient ($r$), Spearman correlation coefficient ($rs$) and Kendall’s Tau-b ($Tb$). Even though the data for the post-task is normally distributed, due to the presence of outliers, non parametric correlation tests are also performed.

From the Table 5.3 above, it can be observed that according to the Pearson Correlation Coefficient only Science corpus shows positive relationship between the cosine similarity values and the MWL that means as the similarity between the keywords and corpus increases, MWL increases but from the p-values, it can be seen that none of the correlation coefficients are statistically significant except the Science + Scien-
tific Method corpus. While rest of the corpora have negative relationship with no statistically significant results.

It is seen that the Science + Scientific Method corpus has a weak negative relationship based on the Spearman correlation coefficient ($rs = -0.252, n = 60, p < 0.05$) between the MWL calculated from the WP and the cosine similarity values of the keywords gathered during the post-pre task with statistically significant result.

Table 5.3: WP and Cosine Similarity for Post-Task: Table of Correlations

<table>
<thead>
<tr>
<th>Topic</th>
<th>r</th>
<th>p-value</th>
<th>rs</th>
<th>p-value</th>
<th>$\tau$-b</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>0.024</td>
<td>0.868</td>
<td>-0.022</td>
<td>0.879</td>
<td>-0.021</td>
<td>0.836</td>
</tr>
<tr>
<td>Scientific Method</td>
<td>-0.150</td>
<td>0.252</td>
<td>-0.096</td>
<td>0.464</td>
<td>-0.057</td>
<td>0.519</td>
</tr>
<tr>
<td>Planning Research</td>
<td>-0.193</td>
<td>0.189</td>
<td>-0.119</td>
<td>0.420</td>
<td>-0.070</td>
<td>0.482</td>
</tr>
<tr>
<td>Literature Review</td>
<td>-0.101</td>
<td>0.498</td>
<td>-0.040</td>
<td>0.790</td>
<td>-0.026</td>
<td>0.797</td>
</tr>
<tr>
<td>Sc + SM</td>
<td>-0.251</td>
<td>0.053</td>
<td>-0.252</td>
<td>0.046</td>
<td>-0.176</td>
<td>0.049*</td>
</tr>
<tr>
<td>Sc + SM + PR</td>
<td>-0.162</td>
<td>0.266</td>
<td>0.003</td>
<td>0.979</td>
<td>0.003</td>
<td>0.979</td>
</tr>
<tr>
<td>Sc + SM + PR + LR</td>
<td>-0.120</td>
<td>0.420</td>
<td>-0.038</td>
<td>0.800</td>
<td>-0.015</td>
<td>0.883</td>
</tr>
</tbody>
</table>

Sc : Science  SM : Scientific Method  PR : Planning Research  LR : Literature Review

Table 5.4: WP and Cosine Similarity for Post-Pre Task: Table of Correlations

<table>
<thead>
<tr>
<th>Topic</th>
<th>r</th>
<th>p-value</th>
<th>rs</th>
<th>p-value</th>
<th>$\tau$-b</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>0.135</td>
<td>0.356</td>
<td>0.114</td>
<td>0.436</td>
<td>0.086</td>
<td>0.384</td>
</tr>
<tr>
<td>Scientific Method</td>
<td>0.178</td>
<td>0.178</td>
<td>0.194</td>
<td>0.141</td>
<td>0.141</td>
<td>0.116</td>
</tr>
<tr>
<td>Planning Research</td>
<td>-0.089</td>
<td>0.548</td>
<td>-0.058</td>
<td>0.698</td>
<td>-0.051</td>
<td>0.612</td>
</tr>
<tr>
<td>Literature Review</td>
<td>0.095</td>
<td>0.524</td>
<td>0.100</td>
<td>0.504</td>
<td>0.076</td>
<td>0.452</td>
</tr>
<tr>
<td>Sc + SM</td>
<td>0.253</td>
<td>0.047*</td>
<td>0.152</td>
<td>0.246</td>
<td>0.095</td>
<td>0.287</td>
</tr>
<tr>
<td>Sc + SM + PR</td>
<td>-0.162</td>
<td>0.267</td>
<td>0.129</td>
<td>0.379</td>
<td>0.095</td>
<td>0.338</td>
</tr>
<tr>
<td>Sc + SM + PR + LR</td>
<td>0.092</td>
<td>0.537</td>
<td>0.009</td>
<td>0.952</td>
<td>0.017</td>
<td>0.869</td>
</tr>
</tbody>
</table>

Sc : Science  SM : Scientific Method  PR : Planning Research  LR : Literature Review

Based on the correlation coefficients shown in table 5.3 and table 5.4, although not all the results show statistically significant results, HA is accepted as the assumption is
met when the strength of relationship between MWL and difference of cosine similarity values of post-task and pre-task is compared for Science + Scientific Method corpora. On the other hand, HA is rejected for all the other corpora as there are no statistical significant results.

5.2 Strength and limitations of findings

5.2.1 Strengths

The steps and methods followed throughout the experiment were clearly identified to achieve good results. The research was fully focused on understanding and preparing the textual as well as numerical data to avoid issues that may arise while analyzing and interpreting results especially text analysis using Natural Language Processing. A new approach was proposed and tested to evaluate the learning of the students using Word2Vec model. All the records were retained during the analysis by imputing the missing values. Appropriate methods were used to get good results.

All the assumptions while implementing the selected methods were taken into consideration. The statistical techniques were selected and assessed based on the properties of the features present in the dataset. Parametric test such as Pearson correlation coefficient and Non Parametric technique such as Spearman and Kendall’s Tau-b were used. As the dataset had lot of features, it gave opportunity to test and analyze different aspects of the data.

5.2.2 Limitations

As mentioned in the Limitations section of the designed chapter, due to the relatively small size of the data set, it is difficult to get statistically significant results. Based on the evaluation, hypothesis $HA$ was accepted for Science + Scientific Method corpus because of statistically significant results and was rejected for all the other topics.

The other limitation is the number of missing values in the textual data. The missing values were imputed by 'unknown' keyword and was treated as a stop word. Thus
giving small amount of textual data to analyze. Lastly, the misspelling and abbreviations creates a challenge. The misspelling were corrected using the Google Sheets’ spell checker. Even though, an algorithm is described to replace the abbreviation with the best suitable and appropriate word phrase, it takes a lot of time to create a user defined dictionary with abbreviations and word phrase as during the pre-task of the teaching session, no one knows what student expects to learn from the session giving a wide range of possibility for different abbreviations.
Chapter 6

Conclusion

6.1 Research Overview

The research is carried out to evaluate the relationship between Mental Workload and Learning which is quantified using Natural Language Processing Technique. It started with a literature review, describing related works and gaps in the literature. On the basis of those gaps, research question is formulated. The experiment is designed and a Word2Vec model is built and evaluated to answer the research question.

The study was conducted on the students of third level classes, who answered NASA-TLX and WP questionnaire on which two data sets were built on four topics taught by Lecturer. Each data set had almost 220 records. Word2Vec, a natural language processing technique is used to quantify the text data into a score signifying learning of the student. Word vectors are generated using the Word2Vec model for the corpora and the keywords which helped to calculate sentence vectors. The cosine similarity measure was used to calculate the semantic similarity between two sentence vectors. Thus calculating the learning of the student by comparing the sentence vectors between the corpora and the keywords. The correlation between the learning computed and MWL scores is then determined.
6.2 Problem Definition

The main purpose of the current research was to show that Natural Language Process-
ing can be used to evaluate learning by quantifying the textual data and evaluate its relationship with Mental Workload. In order to answer the question, following issues were tackled:

- Investigate Cognitive Load Theory and its load types.
- Investigate Mental workload and its measurement methods.
- Identify and investigate various Natural Language Processing technique that can be used to quantify the keywords and the corpus.
- Identify most suitable semantic similarity measure that can be used to find the similarity between corpus and keywords.
- Select statistical techniques and software on which those can be performed.
- Analyze data sets features and corpus.
- Identify and solve the data quality problems present in both numerical as well as textual data.
- Prepare the data in the form which is most suitable as input to the model.
- Implement the model and evaluate the results in order to answer the research question.

6.3 Design/Experimentation, Evaluation & Results

The content taught by the lecturer called instructional designs is translated into text as storytelling format, along with the keywords written by the students and were analysed to find out how they are related to Mental Workload and check if Mental Workload affects the learning of the student. The current hypothesis was based on the assumption that the cosine similarity values calculated between the lecturer’s text
data and the keywords gathered from the student provides insights on the Mental workload activity. The Word2Vec model was built to get the word vectors and then calculate cosine similarity. Thus, the relationship between the similarity value and the MWL were analysed using Pearson correlation coefficient (r), Spearman correlation coefficient (rs) and Kendalls Tau (Tb). All the three tests were performed even though the data was normally distributed in most of the cases, as there was presence of outliers in the data. Based on the results, it was seen that out of the seven corpora, only one corpus showed weak positive relationship between the MWL(NASA-TLX) and the difference of the cosine similarity values of keywords gathered from students during post-task and pre-task in third-level sessions.

6.4 Contributions & Impact

A novel attempt was made to find a Natural Language Processing technique that can help to quantify the text data which will help to find similarity between the Lecturer’s content and the student’s keywords gathered before and after the class. The gaps from the existing literature, helped to demonstrate the application of new technologies to contribute to the analysis of theoretical approaches. This work was done to expand the existing studies done in Mental Workload.

6.5 Future Work & Recommendations

The aim of the study was to evaluate the relationship between MWL and learning through the use of Natural Language Processing technique. This study used word embeddings by converting words into vectors using Word2Vec model as an NLP technique. During the study, the main limitation was dealing with small data sets of not only students’ keywords but also Lecturer’s content which was input to the model.

Due to the time constraint, finding the accuracy of the model is not included as test data set had to be a created containing semantic and syntactic examples in “A is to B as C is to D” format. Word2vec training is an unsupervised task, there is no
defined way to evaluate the results produced by the model. The research mentioned in this project can be extended by finding the accuracy of the model and trying different similarity measures like Word Mover’s distance, etc. More data will surely help to get statistically significant results. The future work will be trying other Natural Language Processing techniques to investigate the relationship between learning and MWL subjective measures. Also, as there is a very weak correlation between mental workload assessment and the similarity scores, these could be jointly employed in a unified model towards the prediction of task performance measures, as a multiple-choice questionnaire, which is an objective form of learning.
References


REFERENCES


REFERENCES


Rizzo, L., & Longo, L. (2017). Representing and inferring mental workload via defeasible reasoning: A comparison with the NASA task load index and the workload profile. In *Proceedings of the 1st workshop on advances in argumentation in artificial intelligence co-located with XVI international conference of the italian association for artificial intelligence, a3@ai*ia 2017, bari, italy, november 16-17, 2017. (pp. 126–140).


REFERENCES


Appendix A

Additional content

A.1 Workload Profile : Assessment of Normality

1. Science

(a) WP Workload Score

![Histogram]

(b) Boxplot

Figure A.1: Assessment of Normality for WP Science

(b) WP Post Task
APPENDIX A. ADDITIONAL CONTENT

(a) Histogram
(b) Boxplot
(c) Scatterplot

Figure A.2: WP Post task Science: Assessment of Normality

(c) WP Post-Pre Task

(a) Histogram
(b) Boxplot
(c) Scatterplot

Figure A.3: WP Post-Pre task Science: Assessment of Normality

2. Scientific Method

(a) WP Workload Score

(a) Histogram
(b) Boxplot

Figure A.4: WP Workload score Scientific Method: Assessment of Normality

(b) WP Post Task

96
APPENDIX A. ADDITIONAL CONTENT

(a) Histogram  
(b) Boxplot  
(c) Scatterplot

Figure A.5: WP post task for Scientific Method: Assessment of Normality for WP Scientific Method

(c) WP Post-Pre Task

(a) Histogram  
(b) Boxplot  
(c) Scatterplot

Figure A.6: WP Post-Pre for Scientific Method: Assessment of Normality

3. Planning Research

(a) WP Workload Score

(a) Histogram  
(b) Boxplot

Figure A.7: WP Planning Research: Assessment of Normality

(b) WP Post Task
APPENDIX A. ADDITIONAL CONTENT

(a) Histogram  
(b) Boxplot  
(c) Scatterplot

Figure A.8: WP post Planning Research: Assessment of Normality

(c) WP Post-Pre Task

(a) Histogram  
(b) Boxplot  
(c) Scatterplot

Figure A.9: WP Post-Pre Planning Research: Assessment of Normality

4. Literature Review

(a) WP Workload Score

(a) Histogram  
(b) Boxplot

Figure A.10: WP Workload Score Literature Review: Assessment of Normality
APPENDIX A. ADDITIONAL CONTENT

(a) Histogram  
(b) Boxplot  
(c) Scatterplot

Figure A.11: WP Post task Literature Review: Assessment of Normality

(b) WP Post-Pre Task

(a) Histogram  
(b) Boxplot  
(c) Scatterplot

Figure A.12: Assessment of Normality for WP Post-Pre Literature Review

5. Science + Scientific Method Corpus on Scientific Method Keywords

(a) WP Post Task

(a) Histogram  
(b) Boxplot  
(c) Scatterplot

Figure A.13: WP Post task Science + Scientific Method: Assessment of Normality
APPENDIX A. ADDITIONAL CONTENT

(b) **WP Post-Pre Task**

![Histogram](image1)
![Boxplot](image2)
![Scatterplot](image3)

Figure A.14: WP Post-Pre Science + Scientific Method: Assessment of Normality

6. **Science + Scientific Method + Planning Research Corpus on Planning Research Keywords**

(a) **WP Post Task**

![Histogram](image4)
![Boxplot](image5)
![Scatterplot](image6)

Figure A.15: Assessment of Normality for WP Sc + SM + PR Method

(b) **WP Post-Pre Task**

![Histogram](image7)
![Boxplot](image8)
![Scatterplot](image9)

Figure A.16: Assessment of Normality for WP Post-Pre Sc + SM + PR Method
7. Science + Scientific Method + Planning Research + Literature Review Corpus on Literature Review Keywords

(a) WP Post Task

Figure A.17: WP post-task Sc+SM+PR+LR: Assessment of Normality

(b) WP Post-Pre Task

Figure A.18: WP post-pre task Sc+SM+PR+LR: Assessment of Normality