Using Semantic Richness for Metaphor Detection using Deep Learning

By

Ghadi Alnafesah

A thesis submitted to
the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
College of Engineering and Physical Sciences
University of Birmingham
Oct 2023

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

© Copyright by GHADI ALNAFESAH, 2023

All Rights Reserved

ABSTRACT

The Natural Language Processing (NLP) encounters difficulties with metaphors, known for their creative and non-literal usage. Metaphors involve using words or phrases from one context in entirely different contexts, making the meaning less clear and requiring human interpretation for understanding.

This dissertation places its focus on the semantic richness elements derived from the perceptual part of the semantic network. These elements serve as the main linguistic features integrated into vector representations. By extracting the semantic information encompassing concreteness, imageability, sensory experience, sentiment, and embodiment, this study seeks to explore the feasibility of detecting metaphors using deep learning models.

The investigation is conducted using two experimental structures: sentence-level classification for the categorisation of entire sentences and word-level classification for individual words. These models are assessed across three metaphorical datasets: VUAMC, MOH-X, and TroFi. The main objective is to evaluate the impact of these semantic elements on the metaphor detection task, with the potential for enhancing model performance.

DEDICATION

I wish I could thank everyone who contributed to the completion of this dissertation, whether on an academic, educational, or personal level. Unfortunately, this is not possible.

However, I would like to express my gratitude to my parents, who encouraged me to pursue education and provided me with various forms of support, both emotional and financial. To my twin sister, Dr. Ghada, who has been part of my journey from the beginning—we are in this together, sharing joy and tears with each step over the past eight years. She cried with me when my model did not run, and I cried with her when her RNA extraction failed. To my friends, my second support system, who share our youthful dreams and our future. To my Twitter friends, especially during the COVID-19 lockdown, thank you for your comfort and support.

My greatest gratitude goes to Dr. Mark Lee, who was an outstanding supervisor. Thank you for the enjoyable discussions on various topics, for your help, advice, and encouragement over the last years. I also want to express my appreciation to Dr. Philip Smith for your amazing assistance and advice, and to Dr. Harish Tayyar Madabushi for teaching me so much about research. Lastly, I extend my thanks to everyone at the University of Birmingham faculty and staff who has assisted, supported, or been there for me throughout the years. I am truly grateful to Qassim University in my country for providing me with this invaluable opportunity. It is through the university's support and commitment to education that I was able to embark on this academic journey.

This marks the conclusion of one adventure and the start of another.

ACKNOWLEDGMENTS

I acknowledge the people who helped me.

Contents

Page

1 Introduction **1**

1.1 Introduction 1

1.2 Metaphor and Metaphor Theories 3

1.3 Metaphor Applications 6

1.4 Semantic Richness and Psycholinguisic 7

1.5 Semantic Richness and Metaphor Detection 8

1.5.1 Strength of the Semantic Network Elements and Metaphor Detection 9

1.5.2 Perceptual Part of the Semantic Richness 11

1.6 Definition of Metaphor 13

1.7 Theories 14

1.8 Experiments Set Up 16

1.8.1 Neural Network Models 17

1.8.2 Metaphorical Datasets 24

1.8.3 Embeddings 28

1.8.4 Lexical Datasets 30

1.8.5 Experimental Procedure 32

1.9 The Research Questions and Contributions 34

1.10 Structure for the Rest of this Dissertation 36

1.11 Conclusion 37

2 Concreteness **39**

2.1 Introduction . . . 39

2.2 The Concrete-Abstract Score . . . 40

2.3 Concreteness and Metaphoricity . . . 40

2.3.1 Related Work . . . 41

2.4 Metaphor Detection Experiments . . . 43

2.4.1 Datasets . . . 43

2.4.2 Experiment Set-up . . . 46

2.4.3 Classification . . . 47

2.4.4 Sequence Labelling . . . 50

2.4.5 Discussion . . . 53

2.5 Conclusion . . . 60

3 Sentiment **62**

3.1 Introduction . . . 62

3.2 Sentiment and Figurative Language . . . 63

3.3 Related Work . . . 64

3.4 Metaphor Detection Experiments . . . 65

3.4.1 Datasets . . . 66

3.5 Experiment Set-Up . . . 68

3.5.1 Step 1 . . . 69

3.5.2 Step 2 . . . 74

3.5.3 Step 3 . . . 75

3.5.4 Classification . . . 75

3.5.5 Sequence Labelling . . . 86

3.5.6 Discussion . . . 95

3.6 Conclusion . . . 103

4 Sensory Experience and Imageability **105**

4.1 Introduction . . . 105

4.2 What are Sensory Experience and Imageability? . . . 106

4.3 Imageability, Sensory Experience and Metaphor Theories . . . 106

4.4 Related Work . . . 108

4.5 Metaphor Detection Experiment . . . 110

4.5.1 Dataset . . . 110

4.5.2 Experiments Set-Up . . . 112

4.5.3 Classification . . . 115

4.5.4 Sequence Labelling . . . 117

4.5.5 Discussion . . . 119

4.6 Conclusion . . . 126

5 Body-Object Interaction **131**

5.1 Introduction . . . 131

5.2 What is Body-Object Interaction? . . . 132

5.3 Body-Object Interaction and Metaphor . . . 133

5.4 Metaphor Detection Experiments . . . 134

5.4.1 Dataset . . . 135

5.4.2 Experiment Set-Up . . . 136

5.4.3 Classification . . . 138

5.4.4 Sequence Labelling . . . 139

5.4.5 Discussion . . . 141

5.5 Conclusion . . . 146

6 Integrating the Different Semantic Elements **149**

6.1 Introduction . . . 149

6.2 Lexical Resources and Deep Learning in NLP . . . 150

6.3 Experiments and Results . . . 151
6.3.1 Set 1 . . . 153
6.3.2 Set 2 . . . 154
6.3.3 Set 3 . . . 155
6.3.4 Set 4 . . . 155
6.3.5 Set 5 . . . 156
6.3.6 Set 6 . . . 157
6.4 Discussion and Analysis . . . 157
6.5 Conclusion . . . 164

7 Conclusion **169**
7.1 Introduction . . . 169
7.2 Semantic Richness . . . 170
7.3 Research Questions . . . 171
7.4 Models and Datasets . . . 173
7.5 The Experiments . . . 173
7.5.1 Concreteness . . . 174
7.5.2 Sentiment . . . 176
7.5.3 Sensory Experience and Imageability . . . 179
7.5.4 Body-Object Interaction . . . 182
7.5.5 Integrating the Different Semantic Elements . . . 184
7.5.6 Summary . . . 186
7.6 Observations, Limitations and Future Work . . . 187
7.6.1 Observations and Limitations . . . 188
7.6.2 Recommendations for Future Work . . . 193
7.7 Future Work Summery . . . 199
7.8 Large Language Models (LLM) and this Work . . . 202

7.9 Conclusions 203

References **205**

List of Tables

1.1 The features that measure the semantic richness as defined by Findlay and Carrol (2018). 8
1.2 The classification reported results for each of the metaphorical datasets. . . . 19
1.3 The classification baselines for each of the metaphorical datasets and the hyperparameters used to achieve these results. 19
1.4 The sequence model (Gao et al., 2018) baseline used on the VUAMC_20 datasets and the hyperparameters used to achieve the result. 21
1.5 The sequence labelling model, RNN_HG, reported results for each of the metaphorical datasets. 23
1.6 The sequence labelling model (Mao et al., 2019) baselines for each of the metaphorical datasets and the hyperparameters used to achieve these results. 24
1.7 The breakdown of the metaphorical datasets includes the number of sentences, tokens, metaphors, and literals. These datasets are used with both models for metaphor detection. CL is used with classification, and SQ is used with sequence labeling. 26

2.1 The mean, maximum, and minimum F1 scores for each Support Vector Machine (SVM) trained on the concreteness-abstractness datasets. 47

2.2 The results are compared to the reported results and the baselines built on the classification model by Gao et al. (2018) for each dataset, along with their respective hyperparameters. These experiments test the VUAMC, MOH-X, and TroFi datasets using the concreteness score. The 10-fold cross-validation procedure was applied to obtain the results for MOH-X and TroFi. Bold font indicates the best results. 49

2.3 The results are based on the VUAMC_2020 dataset provided Leong et al. (2020) for the Metaphor SharedTask competition. The results from both the single rating form and the probability distribution rating form are compared to the baselines, along with their respective hyperparameters. The bold font indicates the best results. 51

2.4 The results are compared with the reported results and the baselines based on Mao et al. (2019) using the sequence model (Seq19) for each dataset. These experiments assess the performance of the VUAMC_18, MOH-X, and TroFi datasets using the concreteness score, along with their respective hyperparameters. The 10-fold cross-validation procedure was utilised to derive the for MOH-X and TroFi. The bold font indicates the best results. 52

2.5 The details of the metaphorical datasets used. 54

2.6 Sample of the prepositions in the VUAMC_20 dataset and their counts of True-Positive and False-Negative. 57

2.7 Samples of words that appeared only once, along with their sentences, labels, the model's predicted labels, and their concreteness ratings. 59

3.1 The breakdown of the NRC words per emotion, and the metaphorical datasets' words per emotions. 71

3.2 The mean, max. and min. F1 scores for each SVM trained on the NRC list for polarity and emotions 72

3.3 The precision, recall and F1 metrics for the trained BERT model on sentences rated for emotions. 74
3.4 The breakdown of the NRC words per emotion, and the metaphorical datasets words per emotions. 74
3.5 A sample of trials showing some of the effect results of step 1 experiments on the VUAMC dataset with the single-digit SVM prediction. 77
3.6 A sample of trials showing some of the effect results of step 1 experiments on the VUAMC dataset with the probability distribution's SVM prediction. . . 78
3.7 A sample of trials showing some of the effect results of step 1 experiments on the MOH-X dataset with the single-digit SVM prediction. The hyperparametres used to reach these result are recorded. 79
3.8 A sample of trials showing some of the effect results of step 1 experiments on the MOH-X dataset with the SVM's probability distribution predictions. The hyperparametres used to reach these result are recorded. 80
3.9 A sample of trials showing some of the effect results of step 1 experiments on the TroFi dataset with the single-digit predictions. The hyperparametres used to reach these result are recorded. 81
3.10 A sample of trials showing some of the effect results of step 1 experiments on the TroFi dataset with the SVM's probability distribution predictions. The hyperparametres used to reach these result are recorded. 82
3.11 The BERT emotion with the classification model on the three metaphorical datasets and their hyperperemeters. 83
3.12 A sample of trials showing some of the effect results for this step on the three dataset and their hyperparameters. 84
3.13 The BERT emotion with the classification model as sentiment or non-sentiment testing. The table provide the hyperparameters. 86

3.14 The list of the successful sequence labelling experiments, their results and their hyperparameters for the VUAMC dataset. 88
3.15 The list of the successful sequence labelling experiments, their results and hyperparameters for the VUAMC dataset using the probability distribution. 89
3.16 A sample of trials showing some of the effect results of step 1 experiments on the MOH-X dataset with the single-digit SVM prediction. 90
3.17 A sample of trials displaying some of the effect results for the NRC sentiment experiments on the MOH-X dataset with SVM's probability distribution predictions and their hyperparameters. 91
3.18 A sample of trials showing some of the effect results from experiments with NRC sentiment on the TroFi dataset with the single-digit predictions and the hyperparameters. 92
3.19 A sample of trials showing some of the effect results from experiments with NRC sentiment on the TroFi dataset with the probability distribution predictions. The hyperparameters used to reach these results are recorded. 93
3.20 The BERT emotion with the sequence labelling model on the three datasets with their hyperparameters. 93
3.21 The BERT emotion with the sequence labelling model testing as sentiment and non-sentiment. The hyperparameters are recorded. 95

4.1 The breakdown of the total tokens for each dataset and the number covered by the imageability and sensory experience lists. 114
4.2 The mean, maximum, and minimum F1 scores for each SVM trained on the imageability and sensory experience datasets. 114

4.3 The table displays the baseline and the outcomes generated by the classification model utilised in the imageability experiments. The results related to the single-digit ratings presented in two formats, along with the results derived from probability distributions. Additionally, the hyperparameters employed in these experiments are presented. Bold font highlights the best results when compared to the baseline. 127

4.4 The table displays the baseline and the results generated by the classification model utilised in the sensory experience experiments. The results encompass single-digit ratings in two distinct formats, alongside the results for probability distributions. Furthermore, the table presents details about the hyperparameters employed in these experiments. The bold font indicates the best results compared to the baseline. 128

4.5 The table shows the baseline and the outcomes produced by the sequence labelling model employed in the imageability experiments. The results are for single-digit ratings in two distinct formats, as well as the results for probability distributions. The table also provides details about the hyperparameters utilised in these experiments. The bold indicates the best results compared to the baseline. 129

4.6 The table shows the baseline and outcomes produced by the sequence labelling model employed in the sensory experience experiments. The results are for single-digit ratings in two distinct formats as well as for probability distributions. The table also provides details regarding the hyperparameters used in these experiments. The bold font indicates the best results compared to the baseline. 130

5.1 The breakdown of total tokens for each dataset and the number that the body-object interaction list cover. 138

5.2 The mean, maximum, and minimum F1 scores for each SVM trained on the body-object interaction datasets. 138

5.3 The table shows the baseline and the outcomes produced by the classification model employed in the body-object interaction experiments. The results are for single-digit ratings in two distinct formats, as well as the results for probability distributions. The table also provides details about the hyperparameters utilised in these experiments. The bold indicates the best results compared to the baseline. 147

5.4 The table shows the baseline and the outcomes produced by the sequence labelling model employed in the body-object interaction experiments. The results are for single-digit ratings in two distinct formats, as well as the results for probability distributions. The table also provides details about the hyperparameters utilised in these experiments. The bold indicates the best results compared to the baseline. 148

6.1 The total number of token in each metaphorical dataset and how many got covered by each lexical list, with these list total number of tokens. 152

6.2 The results are compared to the reported results and the baselines built on Gao et al. (2018) classification model for each dataset. These experiments test the VUAMC, MOH-X, and TroFi with all the semantic richness elements based on the score's source: the lexical list or the SVM/BERT prediction. The 10-fold cross-validation procedure was applied to obtain the MOH-X and TroFi results. 158

6.3 These hyperparameters were the configuration used to generate the results of the classification experiments presented in Table 6.2 across the three datasets. The step indicated in bold highlights the parameters employed in that particular step that led to the best performance on each respective dataset. 166

6.4 The results are compared to the reported results and baselines built on Mao et al. (2019) sequence labelling model for each dataset. These experiments test the VUAMC, MOH-X, and TroFi with all the semantic richness elements based on the score's source: the lexical list or the SVM/BERT prediction. The 10-fold cross-validation procedure was applied to get the MOH-X and TroFi results. 167

6.5 These hyperparameters were the configuration used to generate the results of the sequence labelling experiments presented in Table 6.4 across the three datasets. The step indicated in bold highlights the parameters employed in that particular step that led to the best performance on each respective dataset. 168

Chapter One

Introduction

1.1 Introduction

A metaphor is considered a distinct type of communication, where the world it describes is expressed in a manner that can be true or false, similar to plain language. The difference lies in the unusual, clever, or artful attributes that disguise the message delivered by the metaphor (Davidson, 1978). This description of metaphor is one of meany, as researchers have explored many aspects to *what* is a metaphor, *why* it is used, *how* its meaning is understood. It is a fascinating linguistic phenomenon, as its understanding is dependent on the stage at which it is defined.

The meaning of the metaphorical sentence is implicit, it is concealed within the *literal* meaning of its parts. The concept of literal meaning, as defined by Searle et al. (1980), refers to the definition of a particular word in isolation from the context, and it is reduced to its most basic sense, accompanied by some linguistic semantic features such as imagery and concreteness. Translating the parts of the metaphorical expression to their literal counterparts will produce an incomprehensible statement. Thus, the meaning is inferred from understanding the joint players of the metaphorical elements, which may involve recognis-

ing the underlying analogy or conceptual mapping between the *source domain* (i.e., the metaphorical word) and the *target domain* (i.e., the discussed subject).

For example, "*heart of stone*" and "*heart of iron*" are both metaphors, sharing a similar structure. However, they convey different meanings. Despite both using inanimate objects to describe the *heart*, they actually convey opposite statements at the sub-level interpretation. The first statement, which uses the *stone* to describe the *heart*, holds a negative connotation, suggesting emotional coldness and lack of compassion. On the other hand, the second statement suggests a positive connotation, as the *iron* symbolises resilience, courage, and the ability to remain strong and unbending.

The process of detecting the metaphor and understanding it requires a deep comprehension of the sub-level literal interpretation of the words used figuratively and the context in which they appeared. Due to these complexities, the field of Natural Language Processing (NLP) faces a challenging task in automatically detecting metaphors in text.

Many efforts were made to enhance the detection task. Among these efforts, some relied on metaphor and language theories as the foundational framework for developing detection models. Others used linguistic and/or semantic features to distinguish metaphors from literal expressions. In addition, more strong, effective, and capable tools were involved in the process such as leveraging neural network models and their capabilities to enhance detection performance.

This dissertation presents the findings of experiments involving the addition of semantic information as a feature to deep learning models, aiming to enhance their capabilities in detecting metaphor in text. This semantic information is an essential part of the word's literal meaning and encompasses a wide range of concepts. One of these concepts is the collection of elements known as *semantic richness* (defined in Section 1.4), which is divided into two parts, as defined by Findlay and Carrol (2018): a part refers to the strength of the

network connections, and the other part refers to the strength of the perceptual aspect.

The experiments with the detection models will primarily focus on the later part, *the perceptual aspect*. This part consists of several elements: *concreteness-abstractness* measure, *sentiment*, *imageability*, *sensory experience*, and *body-object interaction*. Each of these elements will be introduced, defined, and their potential links to the concept of metaphor explained. The dissertation will present and analyse the results obtained from these experiments. Furthermore, the dissertation will also discuss the observed limitations of the experiments and provide recommendations to address these limitations in future work. Some of these elements have never been tested before in the context of metaphor detection. By investigating these untested elements, the research seeks to fill gaps in the existing knowledge and open the way for new insights and advancements in the field of metaphor detection.

1.2 Metaphor and Metaphor Theories

Through time, many theories has been developed that attempt to explain metaphor. According to Gibbs Jr (1992), the appropriate answer to what is a metaphor is achieved by considering the question in terms of *when* metaphor is. This statement suggests that the understanding of metaphors involves a dynamic and context-dependent process. Instead of viewing metaphors as static linguistic expressions, Gibbs emphasises their temporal and situational aspects. This viewpoint resulted in the introduction of varying theories regarding the concept of metaphor. These theories could be grouped to their area of focus: the *comprehension*, *recognition*, *interpretation*, and *appreciation*. The *comprehension*-related theories describe the initial moment when the reader/hearer processes the read/heard utterances. The *recognition* is the moment when the processed, comprehended product of the initial stage is perceived as one of the linguistic types (i.e., literal, metaphorical, ironic, and so on).

Followed by the *interpretation* that goes a step further to determine the exact meaning of that processed, comprehended product. The last stage, the *appreciation*, refers to the point in which the evaluation process of that product involves subjective judgment based on personal preferences and cultural influences. The author argued further that the understanding of a metaphor does not necessarily involve following these exact steps. For example, evaluating the suitability or appropriateness of the metaphor is not crucial for its meaning to be understood.

Various theories utilised by many fields (e.g., philosophy, literary and so on) were built on the process/product notion. In this study, the focus will primarily be on the field of computer science, particularly Natural Language Processing (NLP), as it is the area of interest. In this field, the focus is on the product, therefore, the metaphor-related tasks are categorised into into two sub-tasks: detection and interpretation. The identification is the process of making the decision between categorising the target expression as *metaphorical* or *literal*. The interpretation, on the other hand, involves phrasing the intended meaning behind the metaphorical expression (Shutova, 2010). The design, evaluation, and implementation of systems and applications are grounded by one or more theory of metaphor.

In the domain of metaphor research, numerous theories have been developed. Several of these theories have served as the foundational framework for various NLP models through time.

Some theories that focus on the linguistic aspects, involving establishing sets of rules and formulas where any deviation observed in the target expression is regarded as metaphorical. One of these rules is based on the notion that if a metaphor is literally interpreted, the result is a grammatically, semantically, and/or conceptually incorrect statement (Gibbs Jr, 1992). One of these rules is the idea of "*sentences where the semantic preferences are broken*" by Wilks (1968) and Wilks (1978), who explains metaphor as a selectional preference

violation (SPV). This rule is constructed based on the uniformity of semantics in the expression parts, where any unusual usage is considered a violation and therefore identified as a metaphor. For instance, in the sentence "*she poured her heart out to me*", the verb *pour* typically relates to a noun with a liquid character. In this context, the usage of the verb is a violation, indicating a metaphorical expression.

This rule has its limitations. What if the selectional preference has been met, yet the expression is considered a metaphor, as in the example "*Her voice is music to my ears*"? The music aligns with the voices that ears hear, which did not violate the selectional preference. However, this sentence is considered a metaphor because, to the human mind, the usual sense of voice (through the normal act of speaking) does not usually associate with music (which is a deliberate act of using instruments to produce melodic sounds). In this case, Wilks et al. (2013) suggests that when the selectional preference is not violated, and the sense of the first word conforms to the sense of the second, but the sense is of lower frequency of usage, it is most likely a metaphor.

Other theories go beyond the linguistic aspects of the literal/metaphorical expression to encompass worldly experiences and knowledge. According to Lakoff and Johnson (1980), metaphors are conceptual, wherein one domain of knowledge is defined by another domain. This theory is called the Conceptual Metaphor Theory (CMT). For example, in the expression "*he shot down my best argument*", the structure follows **ARGUMENT** is **WAR**. It views criticism/discussions, the *target domain* to be understood, as the act of killing/attracting in wars, the *source domain* from which words/concepts are being borrowed. Interestingly, one *target domain* is not necessarily mapped with a specific *source domain*, but with as many as could be pulled from one's knowledge pool. **ARGUMENT** could be mapped with **CONSTRUCTION** as in "*the foundation of my argument*", and so on (Gibbs Jr, 1992). Furthermore, this theory can be used to distinguish between types of metaphors. The mappings that are derived from typical, widely experienced areas of knowl-

edge are called *conventional metaphors*, while the opposite is referred to as *novel metaphors* (Tong et al., 2021).

The previous analysis of various theories surrounding metaphors highlights the undeniable reality of the challenges they present for NLP. The ambiguity and complexity of the metaphorical expressions, which cannot be simply interpreted by referencing the literal meaning of the words, nor can they be thoroughly investigated through step-by-step rule-based approaches, have pushed the development of metaphor processing systems. This continuous development is driven by the goal of better understanding metaphors and, by extension, enhancing automated detection in natural text.

1.3 Metaphor Applications

Before delving into the idea of metaphor and its theories, it is crucial to elaborate on its significance in research across many fields of NLP. As previously mentioned, metaphors are not limited to artistic expression restricted to specific situations or contexts. On the contrary, they form an important part of communication, naturally appearing in everyday language usage. Therefore, numerous NLP applications reliant on natural language for their tasks must possess the ability to effectively handle metaphors.

One of the most useful tools integral to many applications used daily by individuals or businesses is Machine Translation (MT). However, most traditional and statistical machine translation tools lack translation accuracy due to their inability to extract important contextual information (Alkhatib and Shaalan, 2018; J. Liu et al., 2020). Context is important in many cases, as discussed above, to decide whither the expression is meant literally or metaphorically (Jang et al., 2015; Gao et al., 2018).

Sentiment analysis is one of the NLP fields that has connections with metaphor. Many researchers have studied the sentiment of figurative language, especially metaphors (Dankers et al., 2019). Rentoumi et al. (2012) detected the sentiment of metaphorical expressions by leveraging the contextual information in the sentence, demonstrating that metaphors in sentences can influence their polarity.

In addition to these applications, other areas also incorporate metaphor detection tasks into their frameworks, including Health (Shi et al., 2021), Information Extraction (Do Dinh et al., 2018; Le et al., 2020), Politics (Prabhakaran et al., 2021), and Risk Communication (Hilton et al., 2022).

1.4 Semantic Richness and Psycholinguisic

The concept of semantic richness comes from the theory of semantic representation. This concept is crucial in psycholinguistics, where language acquisition, comprehension, and production, whether oral or written, are studied (Muñoz Martín, 2012). The network of concepts contains interconnected information and features that give each concept element its *semantic meaning* in a *semantic space*. The denser a network is, the more developed it is considered, and therefore the *richer* it becomes; hence, semantic richness is defined (Pexman et al., 2007; Findlay and Carrol, 2018).

The idea of high-dimension semantic space is derived from the assumption that every token (i.e., word, sentence or text) is interconnected with other token in vast network of knowledge (Kintsch and Bowles, 2002). The token does not stand isolated but exists in a complex web of relationships and associations with other linguistic units and concepts. The *meaning* of any token is represented as a vector, which signifies the position of that token relative to all that is known in the space. Devising the meaning in this format enabled the

Strength of the connection of the semantic network	The perceptual strength of a given concept
Number of features	Concreteness-Abstractness
Neighbourhood density	Imageability
Number of senses	Body-object interaction
Number of first associates	Sensory experience rating
	Emotional valence

Table 1.1: The features that measure the semantic richness as defined by Findlay and Carrol (2018).

calculation of the distance between any two points in the space, which is considered the degree of their semantic relationship.

Words, sentences, or texts that possess richer semantic information facilitate faster and more efficient lexical processing (Pexman et al., 2002; Pexman et al., 2008; Taler et al., 2013). The measure of richness uses sets of features categorised by Findlay and Carrol (2018). There are two types of features in Table 1.1. The first type refers to the strength of the network connections (Recchia and Jones, 2012; Taler et al., 2013; Pexman et al., 2008; Ferreira et al., 2015), while the second type refers to the strength of the perceptual aspect.

The last feature is contextual dispersion (diversity), which looks into the distribution of a word's occurrences across content areas. referring to the number of times a word appears in different contexts.

1.5 Semantic Richness and Metaphor Detection

The elements of semantic richness from both sides in the Table 1.1 are not new in the field of NLP. The field has incorporated these elements through various experiments, such as analysing the number of word senses, their frequency, and other statistically based information. Concerning metaphor detection, research papers have reported their results by focusing

on the *semantic network connections* part in Table 1.1 as the main focus of the experiments, while the *perceptual* part is not as frequently examined.

1.5.1 Strength of the Semantic Network Elements and Metaphor Detection

The elements that measure the strength of the network (number of features, neighbourhood density, number of senses, and number of first associates) have been extensively studied in the field of NLP. These elements have been essential components of many valuable tools and methods that are still widely used in advanced research across various fields. They continue to play a crucial role in enhancing the understanding of semantic relationships and improving the performance of computational models in various language-related tasks. The next discussion covers studies that used these concepts in research about metaphor detection.

The first element, the sense, is the foundation of the theory of selectional preference violation (Wilks, 1968; Wilks, 1978) discussed in Section 1.2. Building on that, CorMet (Mason, 2004) relies on processing and extracting information directly from large corpora for detecting verb metaphors. Metaphor does not always violate this principle; however, as suggested by Wilks et al. (2013). In the study by Mohler et al. (2013), they thought of relying on *metaphor signatures*, which are defined as sets of highly interlinked senses from known metaphors. These signatures are used as a basis for comparison. If a new sentence's senses match one of these signatures, it is likely to be a metaphor. In their study, Mao et al. (2018) tested two hypotheses related to the sense of word-level metaphors. The first hypothesis suggests that the metaphorical word's sense and its literal sense in the context are from different domains, following the selectional preference theory. The second hypothesis suggests that senses frequently found in corpora are likely to be literal senses, rather than metaphorical ones.

The semantic neighbourhood, serves as a distribution measure of the target word's neighbours by calculating the average closeness between that word and its neighbours within the word's semantic space (Danguecan and Buchanan, 2016).

When combining the idea of the semantic neighbourhood, it can be inferred that neighboring words tend to have similar vector representations, which results in them being close in meaning. The strength of these vectors lies in their ability to enable the *generalization power* (Goldberg, 2017). Take two words as an example, *dog* and *cat*. Assuming that the former was repeatedly observed in the corpus during the training phase, while the latter was not observed as much or not at all. This occurrence of the two words will not indicate anything about each other. However, the dense vector representation of the *dog*, which occurred more frequently, holds the potential to share its statistical strength with *cat's* vector representation due to the similarity between these vectors. This argument presumes that *cat* was seen enough times so that its representation is similar to *dog*. This also describes the *distributional hypothesis* (Sahlgren, 2008), a linguistic theory that, as the example assumes, words with similar contexts will have similar meanings. This theory were used to produce several useful tools used in meany metaphor detection tasks. Word2Vec (Mikolov et al., 2013) is an algorithm for generating word embeddings. The Global Vectors for Word Representation (GloVe) (Pennington et al., 2014) extends Word2Vec. In addition, machine and deep learning models utilise word embeddings in the detection task that produced comparable results with other models (Do Dinh and Gurevych, 2016; Stemle and Onysko, 2018). Others focused on adjective-noun pairs (Bizzoni et al., 2017), cross-lingual experiments (Tsvetkov et al., 2013), combine sense and embedding to produce multi-sense embedding (Köper and Walde, 2017a).

The last two elements of the semantic richness under the strength of the network are number of associates and number of features. The number of associates is related to the idea that words are associated with other words in context (Duñabeitia et al., 2008).

In contrast, the Number of Features (NoF) is associated with polysemous words and the number of synonyms. *Polysemy* is the phenomenon of "*one word, multiple meanings*". If a concept can be described using different words, then these words are *synonyms*. On the other hand, when one word holds multiple meanings, it is *Polysemous* (Ravin and Leacock, 2000). The word *bank*, for example is considered a polysemous word because it holds many meaning, therefore, it has richer semantic representations. Similarly, even those that are not as polysemous would have richer semantic representation if they were associated with different descriptions. For example, the word *book* has the following features: *has_ authors*, *has_ page_ numbers*, *used_ for_ learning*, etc. In other words, the number of semantic features is the traits, properties, or characteristics that describe a word's meaning (Pexman et al., 2002). A word with a high number of features has a richer semantic representation and vice versa. Although these two are often tested in cognitive science, they are not as popular in computational linguistics. The number of senses, word embedding, and other static resources could be used instead of these concepts.

1.5.2 Perceptual Part of the Semantic Richness

The first section elements of the semantic richness are extensively studied and utilised not only in NLP in general but also specifically in the context of metaphor detection tasks. While some of the perceptual part elements may not be as extensively studied (e.g., *sensory experience* and *body-object interaction*), some concepts within this category are well-established and widely recognised (e.g., *concreteness*, *imageability*, and *sentiment*).

The first element, the concreteness-abstractness measure (introduced in Chapter 2), is a well-known concept in the field of NLP. According to Brysbaert et al. (2014), the concept of *concreteness* refers to the idea of describing something that is tangible, can be seen or sensed. Objects such as books, pens, cars, etc are highly concrete because they can be

perceived by one or more of human senses. On the other hand, ideas, beliefs and concepts are highly abstract. Alnafesah et al. (2020) presented the results of this concept used in detecting metaphors using deep learning models as part of of the Metaphor Shared-Task (Leong et al., 2020).

Sentiment, introduced in Chapter 3, is considered one of the elements of semantic richness; however, it is a complex concept. Certain words have the ability to evoke emotions, whether described in terms of *polarity* as positive or negative, or associated with specific *emotions* like happiness or sadness. Metaphors, in particular, have been shown to elicit sentiments or emotions (Dankers et al., 2019). The main objective of this chapter is to evaluate the performance of metaphor detection models when the only additional information provided is sentiment. Both forms of sentiment, polarity, and emotions, will be tested to assess their impact on the task.

Imageability refers to the ability if some words to evoke vivid mental images when read/heard. The sensory experience refers to words that extended the mental image to involve other senses (Juhasz and Yap, 2013). For example, "*The booming sound of the fireworks filled the night sky*", the word *booming* could evoke the *audition* sense, where you can hear this strong voice of the fireworks. In addition, you can *see* the colorful flashes of the lights filling the night sky. Imageability is a well-established concept in the field of NLP and is frequently utilised in various metaphor detection tasks. As of the date of writing this dissertation, no papers have presented the concept of sensory experience in the context of metaphor detection using deep learning models. The paper by Alnafesah et al. (2023) is the first to present the results of incorporating this concept into deep learning models for metaphor detection. This work will be reported in Chapter 4.

The final concept, the body-object interaction presented in Chapter 5, is a well-studied concept in cognitive science and refers to the degree of ease of physically manipulating a

concept (Siakaluk et al., 2008). Words with high body-object interaction scores are associated with bodily actions. For example, *phone* is considered a high-body-object interaction word due to its ease of physical manipulation, such as holding, dialing, and touching. On the other hand, *river* is seen to have a lower-body-object interaction score. Although it is possible to swim in a river, other forms of physical manipulation are less straightforward. The concept of body-object interaction is a new feature that will be integrated into deep learning models for metaphor detection and will be presented in this dissertation. The paper by Alnafesah et al. (2023) presents the results of this chapter, showing the outcomes of incorporating body-object interaction as an additional feature in the metaphor detection experiments.

In Chapter 6, all the individually tested elements are combined with the deep learning models to assess their collective impact on the models' performance in detecting metaphors.

These chapters encompass a literature review presenting related works that utilised these elements for metaphor detection. Additionally, each element's individual dataset is presented and described. The procedure for setting up and executing the experiments is detailed. The results, including incorrectly and correctly predicted files, are analysed, and observations and limitations are discussed. The research questions are addressed and answered. Chapter 7 provides a summary of these observations and limitations while proposing future work recommendations to address these limitations.

1.6 Definition of Metaphor

This section to grounds this research within a solid theoretical framework by defining what constitutes a metaphor in of this study. According to the conceptual metaphor theory (CMT), metaphor is defined as follows: it is a cognitive mechanism where one conceptual domain (the source domain) is understood and expressed in terms of another conceptual

domain (the target domain). This mapping between domains allows us to comprehend abstract concepts through more concrete experiences (Lakoff and Johnson, 1980). Therefore, metaphor could be viewed as a linguistic expression which its source domain posses high perceptual semantic richness (i.e., more concrete, evoke clearer imagery and sensory and etc.) mapped into a target domain that typically lacks such rich semantic attribute.

1.7 Theories

After providing a general overview of metaphor, including its theories, the concept of semantic richness was introduced. The first part covered the use of NLP in the metaphor detection task. Following that, the second part of the semantic richness concept was introduced. This section establishes the general theory of utilising perceptual elements in the detection task using deep learning models.

However, this study does not aim to solve the question of what a metaphor is or build a theory to explain it. The goal is to utilise established models and datasets for metaphor and enhance the procedure by testing the impact of the semantic richness concept. In other words, the objective is to evaluate the semantic aspect of the concept of metaphor that is related to the semantic richness perceptual part and determine whether the introduced changes are beneficial to the neural network's automated detection of metaphor in natural text, ultimately improving its performance.

The central idea is that semantics play a crucial role in metaphors, as evidenced by the concept of selectional preference violation (Wilks, 1968; Wilks, 1978) and the significance of context, especially when there is compatibility of preference in the metaphorical expression (Barnden, 2007; Wilks et al., 2013).

Jang et al. (2015) and Jang et al. (2016), in both studies, focused on the incoherence between the *global context* topic of the text and the *local context* topic of the sentence. This transition between the global and local contexts indicated a metaphorical usage. Rai et al. (2016) detected novel metaphors using the Conditional Random Field (CRF) supervised classification by leveraging information extracted from word embedding-based features (syntactic, conceptual, and contextual). This experiment demonstrated the usefulness of providing a broad context window, which enabled the classifier to improve the detection task.

Some studies have examined the effectiveness of word embeddings in capturing context to aid various NLP and machine learning models in distinguishing between literal and metaphorical words (Stemle and Onysko, 2018; Swarnkar and Singh, 2018; C. Wu et al., 2018; Pramanick et al., 2018). These works demonstrated that both context and the model choice were crucial for accurate metaphor detection.

Neural networks have shown great capability in handling temporal text signals and preserving contextual information during different tasks. For example, Ghosh and Veale (2016) utilised a neural model for semantic representation in sarcasm detection, achieving state-of-the-art results. Furthermore, Zheng et al. (2021) constructed a multi-layer semantic representation of sentences using Bidirectional Long-Short Term Memory (Bi-LSTM) with a multiple-attention mechanism. Additionally, Karthiga et al. (2021) employed Convolutional Neural Networks (CNN) with double-LSTM for determining semantic similarity between sentences.

To summarise, the general theory behind using the semantic richness perceptual part in metaphor detection with deep learning model is that by explicitly incorporating certain semantic information as lexical linguistic features in the neural network detection process, hidden information in the context of metaphorical expressions can be enhanced and leveraged

effectively, ultimately leading to improved detection performance.

Based on the related works mentioned above, the fundamental plan for the experiments is to utilise neural networks, especially Bi-directional neural networks, for metaphor detection. The ability of these networks to handle large amounts of data efficiently is a crucial aspect. The bidirectional processing, along with LSTM, allows these models to effectively incorporate contextual information in natural language, which is a key element in semantic processing. Therefore, two types of neural models for metaphor detection will be employed to conduct the experiments: a classification neural network for sentence classification and a sequence labelling model for word classification. Both models are discussed in detail in Section 1.8.1.

The experiments will be conducted on three distinct metaphorical datasets (the VUAMC, MOH-X and TroFi), each containing diverse and varied examples of metaphors. These datasets will provide valuable insights into the models' ability to generalise across different data sources. Detailed analyses of the three datasets are presented in Section 1.8.2.

1.8 Experiments Set Up

This section provides a comprehensive overview of the experimental setup and process employed in each chapter. It covers the neural network models used in the experiments, the metaphorical datasets on which the experiments were conducted, and the additional lexical dataset used as features, which constitute the main test subjects. Additionally, the two-step procedure followed in each chapter is explained.

1.8.1 Neural Network Models

As discussed in Section 1.7, the neural networks have demonstrated remarkable success across various tasks in NLP. Their ability to effectively process natural language has made them a preferred candidate for exploring the impact of augmenting semantic richness in metaphor detection. The tests are done on two types of metaphor based on their application: the sentence-level metaphor and the word-level metaphor. Therefore, two model architectures are used to handle the different types of metaphors effectively: the classification model for sentence classification and the sequence labelling for word classification. The performance metrics are determined using Precision, Recall, and F1 scores. The precision, recall, and F1 metrics are computed by specifically emphasising the detection of metaphors, rather than evaluating the performance in predicting both non-metaphorical and metaphorical instances.

The Classification Model

The classification model designed for sentence-level metaphors assigns a metaphoricity prediction to the entire sentence, taking into account both the target word and its surrounding context. In this model, the 1 label indicates that the sentence is a *metaphor*, while the 0 label indicates that it is a *non-metaphor*. To establish this classification experiments, the model baseline is constructed based on the proposed model introduced by Gao et al. (2018).

The Architecture

The classification model introduced by Gao et al. (ibid.) uses a Bi-LSTM network with an attention layer (Vaswani et al., 2017) to predicts the metaphoricity of the target word $[x]$ as follows:

1. **Input Representation:** Each token in the sentence is represented as concatenations of $[w_i; e_i; n_i]$ forming total 1374-embedding dimension where:
   - w_i is a pre-trained GloVe word embedding, with 300-dimensions.
   - e_i is a contextualised ELMo vectors, with 1024-dimensions.
   - n_i is the indicators of whether the word x_i is the target verb or not, with 50-dimensions.
2. **Encoding Layer:** h_i is the contextualised representation for each token produced by passing the concatenated input $[w_i; e_i; n_i]$ through Bi-LSTM. A Bi-LSTM-single-layer has 300 units in each direction (in each experiments is shown with a table summarising the number of layer for each result).
3. **Attention Layer:** The feed-forward neural network computes the attention weight a_i for each token x_i as:
   $$a_i = SoftMax_i(w_a h_i + b_a)$$
   where w_a and b_a are learned parameters.
4. **Sentence Representation:** The sentence representation c is the product of the weighted sum of the LSTM output states h_i, where
   $$c = \sum_{i=1}^{n} a_i h_i$$
   .
5. **Classification Layer:** c then is passed through the feed-forward neural network to calculate the scores of the metaphor (1) or non-metaphor (0) labels l_i for each target verb. Number of nodes is 2.

The baseline results show slight differences compared to the reported results in Gao et al. (2018). In their README file, they mentioned that running the provided script is

expected to yield **lower** performance than the reported numbers. This discrepancy is due to the fact that the reported results were achieved through early stopping and additional training with smaller learning rates, which were not included in the script, in the paper or made available when requested. Therefore, the variations in the results can be attributed to the differences in training approaches and hyperparameter settings between the baseline and the reported experiments. Table 1.2 shows the reported results for the classification model in Gao et al. (2018). MaF1 for the VUAMC dataset represents the performance of metaphorical predictions, calculated based on the correct identification of metaphors.

Exp.	**P.**	**R.**	**F1**	**MaF1**
VUAMC	53.4%	65.6%	58.9%	**53.4%**
MOH-X	75.3%	84.3%	79.1%	-
TroFi	69.661%	73.04%	71.088%	-

Table 1.2: The classification reported results for each of the metaphorical datasets.

The baseline was chosen after running the provided script on each metaphorical dataset with the provided embeddings. Table 1.3 provides the results used as the baseline for each dataset and the hyperparameter adjustments used to achieve these results. The results for the VUAMC dataset were obtained by using three file segments: training, validation, and testing. On the other hand, the results for the MOH-X and TroFi datasets were obtained by employing 10-fold cross-validation. Figure 1.1 provides a detailed diagram of the model, illustrating its components and the flow of data.

Exp.	**P.**	**R.**	**F1**	**H_size**	**Lr**	**Drop1**	**Drop3**	**B_size**	**Layer**	**Epch**
VUAMC	56.28%	51.107%	53.57%	128	0.01	0.3	0.2	64	1	20
MOH-X	75.44%	77.393%	76%	300	0.001	0.2	0.2	10	1	30
TroFi	69.661%	73.04%	71.088%	300	0.001	0.2	0	10	1	15

Table 1.3: The classification baselines for each of the metaphorical datasets and the hyperparameters used to achieve these results.

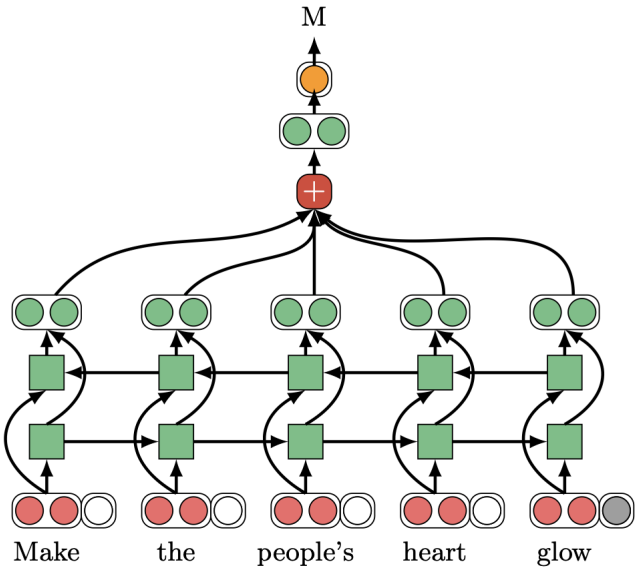

Figure 1.1: The classification model architecture (Gao et al., 2018).

The Sequence Labelling Model

The sequence labeling model, in contrast, classifies the metaphoricity of individual words in a sentence. It assigns a label of 1 for words predicted as *metaphor* and 0 for words predicted *non-metaphor*. The model determinations these predictions by analysing the surrounding context of each word in the sentence.

In this dissertation, two sequence models are used. One of these models is build on the proposed by Gao et al. (2018) and is specifically employed in Chapter 2 to test the concreteness-abstractness measure. This model is utilised with the VUAMC_20 dataset provided by Le et al. (2020) for the Metaphor Shared-Task competition, and the results were published in Alnafesah et al. (2020).

Gao et al. (2018) Architecture

The sequence labelling model based on Gao et al. (ibid.) is a vanilla BiLSTM-based sequence tagging model:

1. **Input Representation:** The token x_i is a concatenation of:
   - A pre-trained word embedding w_i from GloVe, with 300-dimension.
   - An ELMo embeddinge for the contextualise representation e_i, with 1024-dimension.

   The resulted input vector for each token is $[w_i; e_i]$
2. **Encoding Layer:** It is a Bi-LSTM with 300 units for each direction. The input vector $[w_i; e_i]$ is passed through this layer to produce a contextualised representation h_i for each token.
3. **Prediction Layer:** The feed-forward takes the resulted h_i vector as the input to produce the output l_i for each token x_i as either 0, non-metaphor, or 1, metaphor.

Figure 1.2 provides a detailed diagram of the model, illustrating its components and the flow of data. The baseline results, precision, recall, and F1 scores, were obtained using the evaluation model that was made available during the competition. Table 1.4 shows the baseline results achieved by running the script on the VUAMC_20 dataset, with the hyperparameter adjustments.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop3	B_size	Layer	Epch
VUAMC_20	64.9%	48.9%	55.8%	300	0.005	0.3	0.2	64	1	10

Table 1.4: The sequence model (Gao et al., 2018) baseline used on the VUAMC_20 datasets and the hyperparameters used to achieve the result.

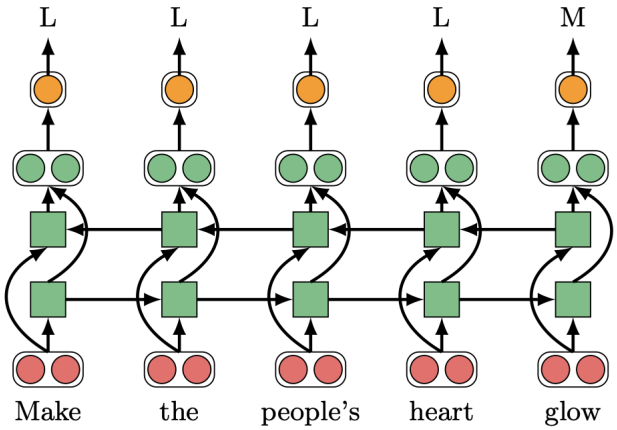

Figure 1.2: The sequence labelling model architecture (Gao et al., 2018).

The rest of the chapters, Chapters 3, 4, 5 and 6, in the dissertation employ a different baseline for the sequence labeling model, named RNN_HG, introduced by Mao et al. (2019).

Mao et al. (2019) Architecture

This model is based on the Metaphor Identification Procedure (MIP) (Pragglejaz-Group, 2007). RNN_HG incorporates a multi-channel CNN layer to integrate different contextualised representation embeddings, GloVe and ELMo. In addition, the model employs the *hierarchical contextualised representation* procedure to improve its predictions. The model's architecture, shown in Figure 1.3, is as follows:

1. **Input Representation:** Each word w_i is represented with a 300-dimensional GloVe embedding g_i as the word representation and a 1024-dimensional ELMo embedding e_i as the contextualised representation.

2. **Encoding Layer:** A Bi-LSTM layer encodes the contextual information for each

token.

3. **Contrastive Mechanism:** The model applies the Metaphor Identification Procedure (MIP) by contrasting the literal and contextual meanings of words. In this step, the comparison is made between the literal meaning provided by GloVe and the contextual meaning provided by the Bi-LSTM hidden states h_i for each word.

4. **Hidden State Comparison:** The identification of discrepancies between the literal and contextual senses of the target words is performed by comparing the hidden states from the Bi-LSTM with the original GloVe embeddings.

5. **Prediction Layer:** The feed-forward network takes the contrasted representations and produces a 0 for non-metaphor or a 1 for metaphor for each word in the input sequence.

The model applies 10-fold cross-validation on the MOH-X and TroFi datasets, while utilising three separate segment files for the VUAMC dataset: training, validation, and testing files. Table 1.5 provides the repotred results for RNN_HG in Mao et al. (2019), while Table 1.6 provides the baselines for the three metaphorical datasets along with their corresponding hyperparameter adjustments.

Exp.	P.	R.	F1
VUAMC	71.8%	76.3%	74.0%
MOH-X	79.7%	79.8%	79.8%
TroFi	67.4%	77.8%	72.2%

Table 1.5: The sequence labelling model, RNN_HG, reported results for each of the metaphorical datasets.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop3	B_size	Layer	Epch
VUAMC	79.58%	64.45%	71.22%	256	0.2	0.5	0.2	2	1	29
MOH-X	76.408%	81.63%	78.46%	256	0.001	0.5	0.1	2	1	20
TroFi	68.98%	74.489%	71.575%	256	0.2	0.5	0.1	2	1	20

Table 1.6: The sequence labelling model (Mao et al., 2019) baselines for each of the metaphorical datasets and the hyperparameters used to achieve these results.

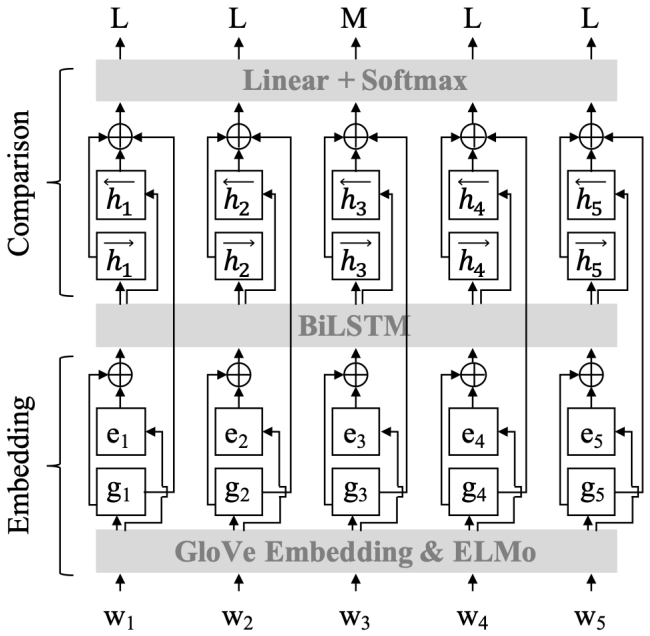

Figure 1.3: The sequence labelling model architecture named RNN_HG (Mao et al., 2019).

1.8.2 Metaphorical Datasets

The goal of this study is to evaluate the impact of the semantic richness of perceptual part elements on the performance of metaphor detection neural networks. These elements are added as lexical features to the vector representation. The experiments are conducted on three different metaphorical datasets: the VUAMC, MOH-X, and TroFi. Each one of these datasets was constructed differently and covers various genres, conversations, interviews, and

more. These datasets are diverse in contexts and domains, which ensures that the resulting models are robust and capable of handling metaphorical language use in a wide range of linguistic situations.

The VUAMC

The VU Amsterdam Metaphor Corpus (VUAMC) (Tighe, 2010) is the largest corpus manually annotated for metaphors. This dataset covers diverse contexts and domains, including different linguistic texts such as news, fiction, and academic writings. The Metaphor Identification Procedure (MIP) (Pragglejaz-Group, 2007) was followed to identify and construct metaphors in the dataset, which involves, for each lexical unit: checking its context, comparing its different dictionary meanings, assessing whether the word has a more basic meaning. If the contextual meaning is different from the basic meaning, then this lexical unit is considered a metaphor (Stowe, 2019). Three versions of the VUAMC dataset are used in the experiments, statistics in Table 1.7:

1. **The VUAMC 2018 CL** dataset, provided by Gao et al. (2018) and Mao et al. (2019), is annotated for metaphoricity at the sentence level. It contains a total of 23,113 sentences, out of which 10,489 are unique (non-repeated). The dataset contains a total of 603,555 words, with 18,695 unique (non-repeated) tokens. Of the 23,113 total sentences, 6,554 sentences are metaphorical, while 16,559 sentences are non-metaphorically expressed. Throughout this dissertation, the VUAMC 2018 dataset is used in every lexical feature of semantic richness experiments related to metaphor detection.

2. **The VUAMC 2018 SEQ** dataset for word-level, as provided by Gao et al. (2018) and Mao et al. (2019), is annotated for individual words in sentences. The dataset consists of 10,567 sentences in total, with a total count of 205,425 tokens. Among these tokens, 18,657 are unique (unrepeated). Within the dataset, 181,644 words are

labeled as non-metaphors, while 23,781 words are labeled as metaphors. Throughout the dissertation, this dataset is utilised with every lexical feature of semantic richness to conduct the experiments related to metaphor detection.

3. **The VUAMC 2020 SEQ** dataset for word-level is provided by Leong et al. (2020) specifically for the Metaphor Shared-Task competition. This dataset is annotated for individual words, indicating whether they are used metaphorically or not. The dataset contains a total of 16,943 sentences, and out of the 244,947 total tokens, there are 20,014 unique tokens. In Chapter 2, this dataset was used exclusively with the concreteness-abstractness measure.

Dataset	# Sent.	# Token	# Unique Token	# Metaphor	# Literal
VUAMC_18 CL	23,113	603,555	18,695	6,554	16,559
VUAMC_18 SQ_19	10,567	205,425	18,657	23,781	181,644
VUAMC_20 SQ_18	16,943	24,4947	20,014	41,31	15,505
MOH-X	647	5,176	1,677	192	195
TroFi	3,737	10,5949	13,738	50	50

Table 1.7: The breakdown of the metaphorical datasets includes the number of sentences, tokens, metaphors, and literals. These datasets are used with both models for metaphor detection. CL is used with classification, and SQ is used with sequence labeling.

MOH-X

The Metaphor in Ontologies and Hubs (MOH-X) (Gao et al., 2018; Mao et al., 2019) is a dataset produced by S. Mohammad et al. (2016) and refined by Shutova et al. (2016). The MOH-X dataset shares similarities with the VUAMC, as its annotation is based on sentences, with a focus on a single verb in each sentence. Notably, the sentences in MOH-X are simpler and shorter compared to other datasets. The metaphor identification process in MOH-X

followed the Selectional Preference Violation (SVP) procedure, which involves determining if the verb in each sentence violated the semantic restrictions associated with it.

The statistics of the dataset are presented in Table 1.7. According to the table, there are 647 sentences with a total of 1,639 tokens. Among these tokens, there are 214 verbs. Out of the 214 verbs, 192 are metaphorical, while 195 are non-metaphorical. For the sequence labelling model, the dataset was processed by assigning the label 1 to only the annotated verb if it is metaphorical, and 0 otherwise and for all other words. Due to this format, the metaphoricity of every single word in the sentence is unspecified. To conduct experiments on the MOH-X dataset for metaphor detection, a 10-fold cross-validation approach was followed.

TroFi

Similar to MOH-X, the Tropes and Figures Corpus (TroFi) (Birke and Sarkar, 2006) comprises simpler sentences. It follows the MIP approach for metaphor identification. Like MOH-X, TroFi contains sentences, and each sentence has one verb rated for metaphor or non-metaphor. The dataset covers 50 verbs from the Wall Street Journal. Table 1.7 shows that the total number of sentences in the dataset is 3,737, with 13,738 tokens. There are 50 verbs labeled for metaphor and non-metaphor. Similar to MOH-X, only the verbs are labelled in the TroFi dataset, and the metaphoricity of the words in the sentence is unspecified. Consequently, when testing with the sequence labelling model, the verb with a label of 1 is identified as metaphorical, while all the other words receive a label of 0. The metaphor detection experiments in TroFi adopt a 10-fold cross-validation approach.

1.8.3 Embeddings

As discussed in previous sections about the important of the context and the information it carry that is a strong determiner for metaphor or non-metaphor decision to be made about the expression. The providers are two types of contextualise representations: GloVe and ELMo.

GloVe stands for Global Vectors for Word Representation (Pennington et al., 2014). It is a 300-dimensional vector space that utilise statistical methods applied to a words co-occurrence matrix within a given corpus. As briefly mentioned in Section 1.5.1, GloVe is an extension of Word2Vec (Mikolov et al., 2013), a neural network-based approach designed to capture semantic associations among words (referred to as local context). Word2Vec itself is derived from the Latent Semantic Analysis (LSA) (Landauer et al., 1998), which revolves around a matrix showing the relationships between documents and their terms, known as the global context. The strength of GloVe lies in that it integrates both local and global contextual information from both techniques. The GloVe vectors serve as the provider of word representations for each word within the three metaphorical datasets throughout the metaphor detection experiments.

Another embedding utilised in the metaphor detection experiments is ELMo, short for Embeddings from Language Models (Peters et al., 2018). ELMo is a 1024-dimensional vectors generated from a 2-layer bi-directional LSTM that is trained on an large corpus. This model excels at capturing the contextual information of words, which is crucial for addressing polysemy. ELMo is employed to supply the context of the sentence by assigning a specialised vector representation to each word within the sentence.

Both vectors are employed in the main metaphor detection task, utilising the classification and sequence[1] labeling models, aiming to test the impact of incorporating semantic

[1]GloVe alone was utilised in the published work by Alnafesah et al. (2020), employing the sequence

richness elements. The total length is a 1324-dimension vector for each word within the metaphorical datasets.

Another embedding that is used during the pre-processing stage is the BERT pre-trained embedding. The Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) is a language model that utilises a 12-layer bidirectional transformer encoder to generate deep contextual representations with a 768-dimensional vector. It differs from ELMo in that the encoding is conducted on both left and right contexts across all layers. BERT can be fine-tuned to suit various NLP tasks, resulting in numerous state-of-the-art outcomes.

The pre-trained BERT model is employed in the pre-processing stage of the metaphor detection experiments, which is explained further on in the next section, Section 1.8.5. It serves as the word representation for each word in the lexical lists, Section 1.8.4, associated with the semantic richness elements that are being tested. These representations are then used as input to train Support Vector Machine (SVM). Furthermore, BERT pre-trained vectors are used to represent words in the metaphorical dataset, allowing the SVMs to predict scores for each tested element. BERT is utilised at this stage due to its contextualised nature, stemming from being trained on extensive unlabeled corpora. Its tokenisation capabilities are also crucial, as they guarantee a representation for each word. BERT employs a technique in which *unknown* words are broken down into smaller word tokens, and their combined weights form a representation for the entire word (Clark et al., 2019).

labeling model based on the model by Gao et al. (2018) known as SEQ_18. This model was used on the VUAMC_2020 dataset provided by Leong et al. (2020) for the Metaphor Shared-Task competition, focusing on evaluating concreteness-abstractness measures.

1.8.4 Lexical Datasets

This study aims to evaluate the impact of incorporating the elements of the perceptual part of the semantic richness, concreteness-abstractness, sentiment, imageability, sensory experience, and body-object interaction, into neural network models used for metaphor detection. These elements have been widely studied in the domains of psycholinguistics and cognitive science. However, some of these elements have not been examined as main features for detecting metaphors with neural networks.

The lexical datasets are extensively covered in their corresponding chapters. Here, a brief overview is provided. These lists primarily consist of English language unigrams, meaning they include individual words presented in isolation, without any surrounding context to influence their interpretation. Each list employs its own scoring scale that varies in range, which is built upon the framework developed by researchers and derived from human experiments.

The concreteness-abstractness measure is considered the most explored semantic richness element in metaphor detection. The lexical list, produced by Brysbaert et al. (2014), consists of 39,955 English words, each assigned a concreteness-abstractness rating ranging from 0 to 5. A lower score indicates more abstractness, while a higher score indicates more concreteness. To prepare for the SVM training, the original scoring scale is transformed into integers within a standardised range. The new scale ranges from 1 (more abstract) to 4 (more concrete). This list is covered in detailed in Chapter 2 Section 2.4.1.

The next element, which has an entire domain within NLP, is sentiment, Chapter 3. In the context of metaphor detection, sentiment is integrated into neural networks. There are two aspects of sentiment examined: *polarity*, distinguishing between *positive* and *negative*, and a collection of distinct *emotions* like *joy* and *sadness*. The latter form of sentiment

is investigated independently, not linked to whether it's *positive* or *negative*. The goal is to assess the potential of evaluating each set of emotions individually as an indicator of metaphoricity. For example, if the words in a specific metaphorical expression are associated with the emotion of *anger*, can the model detect these cues and extract sufficient information for accurate predictions? Two sources of sentiment are employed for this study. The first source is a dataset from S. Mohammad and Turney (2010), which covers over 14,000 English words rated across 10 sentiment categories: *positive*, *negative* (for polarity), as well as specific emotions that covers *anger, anticipation, disgust, fear, joy, sadness, surprise*, and *trust*. The second source involves a BERT sentiment analysis, fine-tuned proposed by Garbas (2019). This BERT model classifies sentences into one of five emotions: *joy, sadness, anger, fear*, and *neutral*. The BERT model is trained on three emotion-rated sentence datasets (Ghazi et al., 2015; Y. Li et al., 2017; Scherer and Wallbott, 1994).

Chapter 4 examines two elements: imageability and sensory experience, both elements have similar definitions. The imageability dataset consists of approximately 18,156 words that have been rated for imageability on a scale of 4 (Ljubešić et al., 2018). On the other hand, the sensory experience dataset consists of 5,000 English words rated on a scale of 7 (Juhasz and Yap, 2013). Similar to the approach used for the concreteness scale, the imageability and sensory experience scales are rearranged into a standardised format to facilitate SVM training and score assignment. As a result, imageability scores are distributed across a scale of 3, where 1 indicates low imageability and 3 indicates high imageability. Similarly, the sensory experience scores are rearranged on a scale ranging from 1 to 6, where words with scores closer to 1 signify low sensory experience, while higher numbers indicate a greater degree of sensory experience.

The final element is a new concept that has not been used in metaphor detection before, which is the body-object interaction. The list of words used, produced by Pexman et al. (2019), consists of over 9,000 English words rated for their *ease of body interaction*

on a scale of 1 to 7. Words with scores closer to 1 have low ease of body interaction, while words with higher scores indicate a greater degree of ease of body interaction. This concept is discussed in detail in Chapter 5.

1.8.5 Experimental Procedure

Exploring new concepts can potentially enhance the detection task. Therefore, the initial step involves adapting the knowledge from the more experienced field into a format suitable for the new field, resulting in laying the groundwork for future enhancements. In the context of this study, the output from the experienced domain is the lexical lists containing words rated according to the specific element being examined. For example, list of words rated for their sensory experience. The adaptation step aims to automatically replicate these ratings for new words present in the metaphorical datasets by using a machine learning model, the SVM. In sequence, these predicted rated words are integrated with word/context representations as inputs for neural network models, allowing the testing of metaphor detection performance. Finally, by analysing the outputs and drawing observations, including limitations, the path for further experimental development and future work that addresses these limitations is established.

Following the outlined framework, the experiments are carried out using a two-step procedure. The ultimate objective is to assess the performance of deep learning models for metaphor detection when contextual information is augmented by integrating semantic richness elements. These elements are provided by relatively small lexical datasets; notably, these datasets do not cover all the tokens present in the metaphorical dataset (the subject of testing). Therefore:

1. The first step is training a machine learning model, the Support Vector Machine (SVM),

which takes both the word and the rating from the provided lexical lists as inputs. This model then generates predicted ratings for all words present in the metaphorical dataset. As discussed in Section 1.8.3, where the reason behind the use of BERT pre-trained vectors was explained, these vectors are employed as the word representation for each token within the lexical lists. These vectors, in addition to the ratings, are the input for the SVM.

2. The second step involves incorporating these newly predicted ratings into the GloVe+ELMo as the word/context vector representation within the main experiments of the metaphor detection neural networks. This is done by extracting the BERT pre-trained vector for every token in the three metaphorical datasets. The outcome is a predicted score for the specific element being examined. It is worth mentioning that the assigned predicted scores are included in two formats: a single-digit score and as a probability distribution score:

    - The single-digit score represents a value within the range of the trained scale. For instance, if the element being evaluated is the concreteness rating, the predicted score for each token within these metaphorical datasets will fall within the range of 1, 2, 3, or 4. On the other hand, if the element is the body-object interaction, the predicted score will be a single number between 1 and 7. This predicted score is then integrated into the GloVe+ELMo vector representation.
    - On the contrary, the probability distribution consists of an array of predicted scores, its length aligning with the scale of scores for the specific element under examination. For instance, if the element is the concreteness measure, the length of the predicted score probability distribution will be 4. In the case of body-object interaction, the length will be 7. Within this distribution, each value signifies the probability of the word belonging to a particular score category. The objective is to assess whether incorporating this probability distribution enhances the in-

formation within the GloVe+ELMo vector, potentially benefiting the metaphor detection task.

BERT's strength lies in its high contextualisation, where the vector representation is tailored to each word in its specific context. This contrasts with GloVe vectors, which represent words with the same vector regardless of context. As mentioned previously, the lexical lists mostly consist of unigrams, individual words without context. This might suggest that BERT's usefulness is diminished due to the lack of context, leading to a literal vector representation for each word. However, the usage of BERT is justified for another reason. GloVe has limited word coverage, and not every word in the lexical lists or metaphorical datasets is found in the GloVe list, resulting in numerous out-of-vocabulary words. BERT overcomes this limitation by dividing words without pre-trained embeddings into smaller tokens, each assigned its representation. These separate parts can then be combined into one vector by adding or multiplying the vectors, ensuring that each word is given a unique vector representation that captures its meaning as accurately as possible. This ensures that the SVMs are trained on all words in the psycholinguistic lexical datasets and that all words in the metaphorical datasets are assigned ratings.

This procedure is applied to each evaluated semantic richness element for both the classification and sequence labeling models across the three metaphorical datasets. The experiments are conducted for each score format independently, and the outcomes are subsequently presented to facilitate a comparative analysis.

1.9 The Research Questions and Contributions

The previous sections covered the concept of metaphor and its theories. The discussions showed the considerable difficulty and challenges the topic of metaphor is, not only in lin-

guistic contexts but also in the realm of NLP. The deep learning models have displayed promising results across numerous NLP tasks, this study aims to utilise this technology to evaluate the significance of richer semantic information for the task of metaphor detection. This is pursued by incorporating semantic information into the context. This semantic information is a component of the perceptual aspect of the concept of semantic richness. The experiments are attempting to answer these questions at the end of each chapter:

RQ1 What is the relationship between metaphors and human-derived perceptual semantic richness models and theories? What is the link between perceptual semantic richness and the development of metaphor models based on deep learning?

RQ2 Could these resources for perceptual semantic richness be used to make and test deep-learning metaphor models, and to accurately detect metaphors in natural text?

RQ3 Lastly, what can these models tell us about the function, distribution, and use of metaphor in natural language?

By answering these questions, the study makes the following contributions:

1. The study attempts to enhance the semantic information by incorporating these elements of semantic richness into the word/context representations offered by GloVe and ELMo vectors. The tests also assess the influence of the score format, whether single-digit or probability distribution. These lexical features could enhance metaphor comprehension, which in turn could potentially lead to improved detection performance.

2. The experiments are conducted by leveraging various deep learning models. One model is designed to identify the metaphoricity of sentences, which was build on the BiLSTM classification model suggested by Gao et al. (2018). Another model is developed to

identify the metaphoricity of individual words within sentences. This task employs two distinct models: the SEQ_18 sequence labeling model introduced by Gao et al. (2018), which was utilized only with the concreteness-abstractness measure in Chapter 2, and the RNN_HG model proposed by Mao et al. (2019). Comprehensive discussion on these models is found in Section 1.8.1.

3. The assessment of the deep learning models is conducted using three distinct metaphorical corpora: the VUAMC, TroFi, and MOH-X. These corpora vary in terms of token sizes and encompass a range of linguistic domains, ranging from spoken language to formal academic text, Section 1.8.2. The diverse array of metaphors present in these corpora offers valuable insights into the models' ability to generalise effectively across disparate datasets.

1.10 Structure for the Rest of this Dissertation

The remaining chapters within this thesis present the metaphor detection experiments integrated with the element of semantic richness perceptual part. Each individual chapter focuses on one or two specific elements. Within these chapters, the element's definition is expounded upon, followed by a thorough literature review. Furthermore, the dataset employed in the experiments is outlined, accompanied by a comprehensive analysis of the outcomes resulted by the models. In addition, each chapter addresses the research questions regarding the element that being investigated. Chapter 2 starts with the first element, the concreteness-abstractness measure. Subsequently, Chapter 3 explores the sentiment, covers both *polarity* and *emotions*. Chapter 4 combined imageability and sensory experience due to their similar definitions. The concluding element the body-object interaction, which derived from concreteness and imageability, is discussed in Chapter 5. Chapter 6 combines all previously examined elements, aiming to evaluate their collective impact on the performance

of neural networks for metaphor detection. The final chapter of this thesis comprises the concluding chapter, Chapter 7. This chapter summarises the key findings from preceding chapters. Additionally, it provides a comprehensive analysis of observations and limitations derived from the conducted experiments. Furthermore, it suggests some recommendations for future endeavors aimed to addressing these limitations.

1.11 Conclusion

This chapter functions as the introduction of this thesis, which aims to enhance the task of metaphor detection through the utilisation of deep learning models. It begins by exploring the concept of metaphor and its ambiguous nature, posing challenges not only to linguists but also extending these challenges to the realm of Natural Language Processing (NLP). NLP applications greatly benefit from the detection and interpretation of metaphors in various tasks, such as machine translation. This chapter delves further into popular metaphor theories examined in the context of NLP for the purpose of metaphor detection.

Subsequently, the main focus of examination in this study is introduced: the perceptual elements of semantic richness. These elements are used to augment the existing semantic information provided by the GloVe+ELMo word/context vector representation. The aim is to evaluate the performance of deep learning models enhanced by these elements across three metaphorical datasets: VUAMC, MOH-X, and TroFi.

To evaluate the effectiveness of this approach, two types of neural network models are employed. The first model involves classification, classifying the metaphoricity of sentences based on the target verb. The second model is the sequence labelling approach, classifying the metaphoricity of individual words within the context. A comprehensive exploration of these datasets and models can be found in Section 1.8.

The final portion of this chapter outlines the research questions that this thesis endeavors to address, providing the summarisation of contributions made by the research.

Chapter Two

Concreteness

2.1 Introduction

The preceding chapter, Chapter 1, offered a thorough account of metaphors, covering their theories, applications, and challenges. It also introduced the concept of semantic richness and its divisions: the elements related to the strength of the network and the perceptual components. The chapter extensively examines the underlying theory regarding the connection between the perceptual part of semantic richness and metaphor, the methodology employed in the upcoming experiments, and the research questions addressed.

This chapter introduces the concept of concreteness and abstractness measures, used to assist in the task of metaphor detection using deep learning models. The chapter begins by presenting the relevant literature on the definition of concrete-abstract measures, their role in various metaphor-related experiments, and the theory connecting them to metaphors. The results are presented and analysed. Finally, the research questions are addressed.

2.2 The Concrete-Abstract Score

Brysbaert et al. (2014) defined *concreteness* as the concept of describing a perceivable, detectable entity. These concrete entities are easily remembered and explained (e.g., *books*, *birds*, and *mountains*), contrasting with abstract entities. Abstract concepts, such as *intuition*, *anxiety*, and *love*, are difficult to understand, unlike the perception of concrete objects by the senses, and usually, these abstract terms have to be used in context to fully understand their meaning.

2.3 Concreteness and Metaphoricity

The intuition behind using the concept of concreteness/abstractness is that metaphors are often employed to describe abstract concepts by utilising concrete ones. They enable the utilisation of a familiar concrete concept to acquire knowledge and understanding of an abstract concept (Jamrozik et al., 2016). In Harrington (2012)'s study, the author examines the communication and description of cancer patients' diagnosis, awareness, and healing process with other patients and care providers. During these processes, cancer patients frequently employ metaphorical imagery related to **MILITARY** and **JOURNEY** themes, both in positive and negative contexts. For instance, the phrase "*war on cancer*" depicts the pursuit of medical advancements as military troops advancing to combat this enemy. Regarding the case of "*cancer invades the immune system*", this metaphor portrays the body's defense against the disease. Patients often employ more journey-related metaphors in their communication, as these metaphors provide a more comprehensive account of their experience with illness. For example, when a patient describes their treatment as a "*bumpy road*", it conveys the challenges and obstacles associated with that particular course of treatment (Reisfield and Wilson, 2004).

To rephrase, when there is a shift in the concrete/abstract rating of words in the context, it is most likely indicates the use of metaphor. As mentioned earlier, concrete words used to describe abstract concepts can indicate the presence of metaphor (e.g., *won the argument*). Consequently, in the experiments conducted in this chapter, the shifts in concreteness-abstractness scores within the metaphorical datasets are utilised to assess the performance of the neural network and the impact of this semantic richness element on the metaphor detection task.

2.3.1 Related Work

The usage of the concreteness concept is not novel in the field of metaphor detection. Traditional models, which rely on approaches other than neural networks for automatic metaphor detection, incorporate the concrete-abstract rating in their experiments. In their work, Turney et al. (2011) used Coltheart (1981)'s MRC psycholinguistic dataset, which consists of a small number of English words rated for their abstractness. They employed a logistic regression model to classify metaphors in adjective-noun pairs and target verbs within the TroFi dataset. Similarly, Tsvetkov et al. (2013) used the MRC list in their experiment, focusing on subject-verb-object pairs instead. They further expanded their analysis to classify metaphors within a Russian dataset.

The study by Klebanov et al. (2014) employed the concreteness measure proposed by Brysbaert et al. (2014), along with other feature sets, to train a logistic regression model for classifying unigrams in the VUAMC. Dunn (2015) focused on investigating the mapping direction and distance of concreteness score shifts in metaphorical words, specifically examining whether these shifts occurred from more concrete to less concrete or vice versa.

Köper and Walde (2017b) reproduced concreteness scores for individual words, phrases,

and even specific word senses to enhance metaphor detection. They compared multiple supervised learning models and resources to learn and expand ratings for a broader range of words. Additionally, they generated a multi-word abstractness rating for verb-noun pairs using the method proposed by Shutova et al. (2016). This approach allowed them to obtain vector representations for these verb-noun phrases and extend the concreteness rating to these phrases. Finally, they developed distinct word representations associated with each sense of the word, enabling the assignment of different abstractness ratings for each word sense, distinct from the overall abstractness rating.

In contrast, the approach taken by Rabinovich et al. (2018) for examining abstractness differed. They utilised lexical and contextual clues to derive abstractness scores in the absence of labelled data. The hypothesis was that the immediate neighboring words, as manifested in their contextual usage, could serve as indicators of the degree of abstractness of a concept within a sentence. For instance, the morphology of a word could provide insights into its meaning. In the word "*feminism*", the "*-ism*" suffix likely indicates an abstract concept, while the word "*assistant*" suggests a concrete concept due to the "*-ant*" suffix, which typically describes a person.

The following sections detail the experiments conducted in this chapter using deep learning models. It is important to note that two sets of results are presented: one pertaining to the published results, (Alnafesah et al., 2020), which is related to the Metaphor Shared-Task competition published in the *Proceedings of the Second Workshop on Figurative Language Processing* (Leong et al., 2020), and other results (not published) are based on the same experimental setup as the other semantic richness elements that will be applied in the upcoming chapters. These two sets of experiments differ in terms of the provided datasets, the embeddings used as the representation, and the specific models employed to carry out the experiments.

2.4 Metaphor Detection Experiments

This section provides an overview of: 1) the datasets and models employed in the SVM training for assigning the concreteness-abstractness rating, 2) the deep learning models and the datasets utilised for the metaphor detection task. The objective of these experiments is to evaluate the impact of incorporating the concreteness-abstractness rating, which is a component of semantic richness perceptual part, into the metaphor detection neural network models. The architectural details of the models used in this chapter and subsequent chapters were explained in Chapter 1. However, only in this chapter, two sequence labelling models are presented, along with two versions of the VUAMC dataset[1]. One sequence model[2] is proposed by Gao et al. (2018), which was employed in the *Metaphor Shared-Task* competition using the VUAMC[3] dataset provided by them. It should be noted that this dataset differs from the VUAMC[4] dataset provided by Gao et al. (ibid.) and used with the sentence classification model and the second sequence[5] labelling model proposed by Mao et al. (2019), the RNN_HG .

2.4.1 Datasets

Concreteness-Abstractness Resource

A concreteness-abstractness rating is assigned to every token in the metaphorical datasets. This rating is the result of SVM predictions, after it was trained a list of words rated for

[1]To make the reading of this chapter easier, the two sequence models and the two versions of the VUAMC will be renamed.

[2]Seq18

[3]VUAMC_20

[4]VUAMC_18

[5]Seq19

their concreteness-abstractness. Whether the experiments are conducted in sentence-level classification or word-level classification, the pre-processing stage remains the same. The objective is to incorporate this lexical resource as an additional agent to the word/context representation for the metaphor detection experiments, in order to assess its impact. The word representation is obtained by using GloVe exclusively for words or by adding ELMo for contextual information.

The training dataset for the SVM consists of a list of 39,955 English words, as provided by Brysbaert et al. (2014). Each token in the dataset is assigned a rating between 1 (indicating more abstract) and 5 (indicating more concrete). For instance, the word '*meatball*" has a rating of 4.93, indicating high concreteness. On the other hand, the word "*lethal*" has a rating of 3.04, suggesting a moderately abstract nature. Similarly, the word '*compatible*" has a rating of 1.64, indicating a relatively more abstract concept.

In order to standardise the ratings for all words in the list, the ratings were adjusted to fit within a scale of 1 to 4 is the first step. As a result, "*meatball*" has now a 4, "*lethal*" has a 3, and "*compatible*" has a 1. The subsequent step involves extracting pre-trained BERT word representations for each token in the list. These embeddings serve as the input for training the SVM. The final step is to assign predicted concreteness-abstractness ratings to all tokens in the metaphorical datasets. This is achieved by extracting BERT pre-trained embeddings for each token. These assigned predictions are either in a single-digit score or in a 4-digit probability distribution.

Metaphorical Datasets

As mentioned previously, three models were utilised for the experiments. Two of these models apply the sequence classification approach build on the work of Gao et al. (2018) (Seq18), the other based on the work of Mao et al. (2019) (Seq19). The third model is a sentence

classification model build on Gao et al. (2018). The Seq19 model and the classification model both employed the same pre-processed VUAMC dataset (VUAMC_18) in addition to MOH-X and TroFi datasets. However, the Seq18 model used a different version of the VUAMC provided by Leong et al. (2020) (VUAMC_20). Furthermore, the last model solely employed GloVe embeddings as the word representation, whereas the first two models utilised both GloVe and ELMo for contextual representation.

As previously mentioned, one of the published models is a classification model proposed by Gao et al. (2018). This model classifies a sentence as either a metaphor (assigned 1) or a non-metaphor (assigned 0) by considering the target verb and the surrounding context. The published results only show for the VUAMC_18 dataset, as the tests on the other two datasets (MOH-X and TroFi) were conducted after the publication. The Seq18 sequence labelling model also proposed by Gao et al. (ibid.) classified each word into metaphor, labelled 1, or non-metaphor, labelled 0, based on the surrounding context. The additional unpublished results presented in this chapter are associated with the Seq19 model proposed by Mao et al. (2019). Similar to the previous Seq18 model, it classifies each word. However, the key difference is that Seq19 compares the input's literal representation by the embedding vector with its contextual representation in the hidden state using a BiLSTM.

The three models employed in the study underwent the same pre-processing step, which involved assigning concreteness-abstractness predicted scores to the tokens using an SVM. The subsequent section provides comprehensive details on how this step was performed, including the results obtained.

2.4.2 Experiment Set-up

The objective is to evaluate the impact of incorporating concreteness information on the overall performance of the metaphor detection models. These rating are the sole additional information to the words in the context. Furthermore, the tests explore the feasibility of employing a generalisation approach, where the rating is predicted by a machine learning model (SVM) trained on a list of rated words (Brysbaert et al., 2014). The neural networks used in the chapter, the classification and the two sequence labeling, share a same initial step for word/context representation. The words are assigned static GloVe vectors of 300-dimensions, which represent the basic meaning of the words regardless of the intended senses. These vectors are combined with ELMo vectors of 1024-dimensions, which provide contextual information. An extension is made to this step, the overall vector representation is concatenated with additional dimensions that hold the SVM predictions for the concreteness-abstractness rating. The predicted scores are in two forms:

Single-Rating Concreteness Score

This part of the experiment aims to evaluate the effectiveness of incorporating a single rating into the existing 1324-dimensional vectors of GloVe+ELMo, for each token in the three metaphorical datasets. The single rating, predicted within the range of 1, 2, 3, or 4, is concatenated with the vector representation as an additional dimension in the overall GloVe-ELMo vector. Table 2.1 shows the training result of the single SVM.

Probability Distribution Concreteness Score

In this step, the SVM provides a four-digit prediction that represents a probability distribution for the concreteness rating. Each digit corresponds to a score on the rating scale. This

step aims to compare the impact of using a single-rating approach versus a multi-dimensional probability distribution rating. There is a possibility that the format of a probability distribution could enable the incorporation of additional information that might assist in the metaphor detection task. The probability indicates that this word probably falls under the second rating of abstractness, while it is less likely to be under the rest of the ratings. These digits are concatenated with GloVe-ELMo vector. Table 2.1 shows the training result of the probability distribution SVM.

SVM	F1 Mean	F1 MAX	F1 MIN
Conc-Single-Digit	58.3%	65.1%	52.2%
Conc-Prob. Dist.	58.9%	67.1%	49.2%

Table 2.1: The mean, maximum, and minimum F1 scores for each Support Vector Machine (SVM) trained on the concreteness-abstractness datasets.

2.4.3 Classification

The metaphor classification task predicts whether a sentence is a metaphor (assigned 1) or literal (assigned 0). The rating forms discussed above are added with each run for each dataset used to examine the effect of the concreteness score on the metaphor detection task. The results related to the VUAMC_20 dataset were published in Alnafesah et al. (2020). Initially, the experiments were built upon the code provided by Gao et al. (2018) for the classification model, utilising GloVe and ELMo. However, as discussed in the previous chapter, due to the absence of certain information such as hyperparameter tuning and other information, the results did not match the published report, as it was noted in their README file. Consequently, the baseline was selected based on the best run after hyperparameter adjustment. The selected baselines serve as a reference for evaluating the subsequent experiments conducted on all three datasets. For the VUAMC datasets, a train-validation-testing approach was employed, while the MOH-X and TroFi datasets were tested using a 10-fold

cross-validation approach.

The baseline model predicts the metaphoricity of the sentence by extracting relevant information from the GloVe+ELMo vectors, considering the target verb and its context. The subsequent results were built upon this baseline model by incorporating SVM-predicted scores for concreteness rating. After each run, the results were recorded, and further hyperparameters were adjusted, tested, and then recorded. The performance evaluation scales are for precision, recall, and F1. For this model, the F1 denotes the "*metaphor prediction performance*" as opposed to the total "*literal & metaphor*" performance. Meaning, the model calculated the information related to "*metaphor prediction*" instead of "*literal prediction and metaphor prediction*" together. Table 2.2 summarises the results of the three forms of concreteness experiments. It shows the results of incorporating the concreteness rating in different forms with the metaphor detection on the sentence-level classification. The table also provides the hyperparameters used to acquire the best run for each rating format.

In the VUAMC_18 dataset, the single-digit slightly improved precision, but it had a negative impact on recall and F1 score. The recall is necessary because it shows how many correct results are correctly predicted (i.e., how many correct predictions are correctly predicted metaphors). While precision indicates how many correct prediction are correct predictions, literal and metaphorical, the recall reflects the ability of the model to identify all relevant instances of metaphors. Thus, while improved precision might indicate fewer false positives, the decrease in recall means that a significant number of true metaphors are being missed, ultimately lowering the overall effectiveness of the model as represented by the F1 score. The average 10-fold cross validation for MOH-X and TroFi experiments with the classification model also exhibits slight improvements in recall and F1, with slight decrease in precision for the single-digit rating form. This shows that the added information may have given more valuable clues to the classification model that helped correctly predict the metaphors in the testing datasets.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC_18											
Gao et al., 2018 Reported	53.4%	65.6%	53.4%								
Baseline	*56.28%*	*51.107%*	*53.57%*	128	0.01	0.3	0.0	0.2	64	1	20
SD. SVM	**57.93%**	44.57%	50.38%	300	0.01	0.3	0.0	0.2	64	1	20
Prob. Dist.	54.21%	**62.46%**	**58.04%**	128	0.01	0.5	0.0	0.2	16	1	20
MOH-X											
Gao et al., 2018 Reported	75.3%	84.3%	79.1%								
Baseline	**75.44%**	*77.393%*	*76%*	300	0.001	0.2	0.0	0.2	10	1	30
SD. SVM	75.182%	**79.432%**	76.83%	300	0.001	0.2	0.0	0.2	10	1	15
Prob. Dist.	**76.6%**	78.524%	**77.175%**	256	0.001	0.2	0.0	0.2	10	1	15
TroFi											
Gao et al., 2018 Reported	69.661%	73.04%	71.088%								
Baseline	*69.661%*	*73.04%*	*71.088%*	300	0.001	0.2	0.0	0.1	10	1	15
SD. SVM	69.101%	**74.98%**	**71.823%**	256	0.001	0.2	0.0	0.2	10	1	15
Prob. Dist.	**70.157%**	73.374%	71.56%	128	0.001	0.2	0.0	0.2	10	1	15

Table 2.2: The results are compared to the reported results and the baselines built on the classification model by Gao et al. (2018) for each dataset, along with their respective hyperparameters. These experiments test the VUAMC, MOH-X, and TroFi datasets using the concreteness score. The 10-fold cross-validation procedure was applied to obtain the results for MOH-X and TroFi. Bold font indicates the best results.

The second form of concreteness format is the probability distribution rating. The overall produced results were better for all three datasets. For the VUAMC_18 dataset, the recall improved to reach 62.46%, and F1 increased reaching 58.04% with a slight decrease in precision. As for MOH-X and TroFi datasets, all three metrics were improved. The recall improvement shows the model classified more metaphors correctly as metaphors. The probability distribution may have facilitated better knowledge gains than the other forms of ratings for the neural network that improved overall performance.

2.4.4 Sequence Labelling

The sequence labeling model is employed to classify each word as either a metaphor (labelled as 1) or a non-metaphor (labelled as 0) based on the context. Two different models for sequence labeling were utilised in the study:

1. Seq18: proposed by Gao et al. (2018), was used to compete in the Metaphor Shared-Task using the VUAMC_20 dataset version provided by the organisers Leong et al. (2020). This model only uses GloVe word representations. Initially, the model was run on the provided dataset using GloVe, and the output file was submitted to the competition website's machine evaluator to obtain the baseline results (precision, recall, and F1). Following the classification model, the same procedures were applied to obtain the SVM rating assignment for each token in the dataset. Table 2.3 shows the results of word-level experiment on the VUAMC 2020 using Seq18 and their hyperparameters.

2. Seq19 (RNN_HG): proposed by Mao et al. (2019), uses GloVe+ELMo as its representation. This model was evaluated on the three datasets provided by Gao et al. (2018) and Mao et al. (2019) in their experiments (VUAMC_18, MOH-X and TroFi). The results of the Seq19 model on the three metaphorical datasets are presented in Table 2.4. Further adjustments to hyperparameters were performed for each dataset to obtain the best possible results.

As discussed in the previous sections above, each concreteness rating is an SVM prediction, trained on Brysbaert et al. (2014) reformed list of 39,955 English words. This prediction was concatenated with the GloVe vector in Seq18 sequence model, and with GloVe-ELMo vectors in Seq19 sequence model. The Seq18 model show the results tested on the VUAM_20 used for the Shared-Task competition (Leong et al., 2020), while the Seq19 model shows the results of testing on three metaphorical datasets: VUAMC_18, MOH-X

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC_2020											
Baseline	*64.9%*	*48.9%*	*55.8%*	128	0.01	0.3	0.0	0.2	64	1	20
SD. SVM	60.3%	53.7%	56.8%	300	0.005	0.5	0.0	0.1	64	1	10
Prob. Dist.	**65.3%**	**54.8%**	**59.6%**	300	0.005	0.5	0.1	0.1	64	3	20

Table 2.3: The results are based on the VUAMC_2020 dataset provided Leong et al. (2020) for the Metaphor SharedTask competition. The results from both the single rating form and the probability distribution rating form are compared to the baselines, along with their respective hyperparameters. The bold font indicates the best results.

and TroFi.

The experiment's prediction for the Seq18 was produced as an "*answer.txt*" file, which was then submitted for acquiring the scores. The machine evaluator on the competition website calculated the precision, recall, and F1 scores. The results are shown in Table 2.3. The first run without including the concreteness rating produced a precision of 64.9%, recall of 48.9%, and F1 of 55.8%. After incorporating the concreteness as a single-digit rating concatenated with GloVe vectors, the produced precision decreased to 60.3%, while the recall and F1 increased to 53.7% and 56.8%, respectively. Similarly, the concatenated probability distribution to the GloVe vectors improved the results generally. The best result showed an increase in the three metrics: precision reached 65.3%, recall reached 54.8%, and F1 reached 59.6%

The experiments performed using Seq19 model showed similar pattern on the three datasets. The models use the GloVe+ELMo+concreteness vector to reach the decision about the metaphoricity of a word. The single-digit rating tested on the VUAMV_18 show increase in F1 by 3.28%. The recall showed greater increase by 10.06%. While the precision decreased to 76.38%. Similarly, the recall increase to 71.39% and the F1 to 73.96, while the precision decreased to 76.73% for the probability distribution ratings. Overall, the model benefited

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC_2018											
Mao et al., 2019 Reported	71.8%	76.3%	74.0%								
Baseline	*79.58%*	*64.45%*	*71.22%*	256	0.2	0.5	0.0	0.2	2	1	29
SD. SVM	76.38%	70.94%	73.56%	300	0.2	0.5	0.0	0.2	2	1	29
Prob. Dist.	76.73%	**71.39%**	**73.96%**	300	0.2	0.5	0.0	0.2	2	1	29
MOH-X											
Mao et al., 2019 Reported	79.7%	79.8%	79.8%								
Baseline	*76.408%*	*81.63%*	*78.46%*	256	0.001	0.5	0.0	0.2	2	1	20
SD. SVM	**79.177%**	79.944%	79.274%	256	0.001	0.5	0.0	0.1	2	1	20
Prob. Dist.	76.963%	**82.337%**	**79.437%**	300	0.001	0.5	0.0	0.1	2	1	20
TroFi											
Mao et al., 2019 Reported	67.4%	77.8%	72.2%								
Baseline	*68.98%*	*74.489%*	*71.575%*	256	0.001	0.5	0.0	0.1	2	1	20
SD. SVM	**69.101%**	**74.98%**	**71.823%**	128	0.001	0.5	0.0	0.1	2	2	20
Prob. Dist.	67.807%	77.7103%	72.341%	128	0.001	0.5	0.0	0.1	2	2	20

Table 2.4: The results are compared with the reported results and the baselines based on Mao et al. (2019) using the sequence model (Seq19) for each dataset. These experiments assess the performance of the VUAMC_18, MOH-X, and TroFi datasets using the concreteness score, along with their respective hyperparameters. The 10-fold cross-validation procedure was utilised to derive the for MOH-X and TroFi. The bold font indicates the best results.

from these addition, it helped improve the detection on the VUAMC_18 dataset.

The results for MOH-X and TroFi improved, but not to the same extent as the VUAMC experiments. The single-digit rating showed a slight increase in F1 to 79.274%, an increase in precision to 79.177%, but a slight decrease in recall to 79.944%. As for TroFi, the increase was in F1, 72.587%, and recall, 78.66%. The slight decrease was in precision, reaching 67.512%. The probability distribution rating with MOH-X improved all three metrics over the baseline, 76.963% for precision, 82.337% for recall, and 79.437% for F1. However, for TroFi, precision slightly decreased while the recall and F1 improved slightly.

Similar to the previous model, the concreteness rating in its two forms helped the word-level metaphor detection task.

2.4.5 Discussion

The classification experiment used the VUAMC_18 dataset provided by Gao et al. (2018). The dataset was divided into three files, 15516 sentence in the training file, 1724 sentences in the validation file, and testing files contains 5873 sentences. The test file is formatted a little differently. Every sentence in the file with more than one verb was repeated as the same number of verbs. The purpose is to test the model's prediction when the target is that particular verb, which means the model focuses on that verb and its context as the target for the metaphor prediction decision. To illustrate, "*It plays a key role in keeping city streets on the move and is even the object of the best modern designers' desire.*" was repeated two times because the target verbs are the underlined words *play* and *keep*. The model should consider the target verb and its context to produce the prediction each time. MOH-X and TroFi datasets were evaluated by applying the 10-fold cross validation. Their sizes is smaller when compared with the VUAMC_18, as well as the metaphor percentage is higher where TroFi has 43% and MOH-X has 49%. All datasets were tested with GloVe+ELMo representations also provided by by Gao et al. (ibid.).

The Seq19 model used the VUAMC_18 that has 23113 sentences with 18657 unique tokens. The number of tokens labelled as metaphor is 23781, and 181644 are literal. The same in the classification model, MOH-X and TroFi datasets were evaluated by applying the 10-fold cross validation. In addition, GloVe+ELMo representations were used as the word and contextual information sources. While the Seq18 model used the VUAMC_20 version (in the competition) has 16943 sentence with 20014 unique tokens. This model only used GloVe for the word representation. Table 2.5 provide more details about the datasets used

in this chapter's experiments.

Dataset	# Sent.	# Token	# Unique Token	# Metaphor	# Literal
VUAMC_18 CL	23113	603555	18695	6554	16559
VUAMC_18 SQ_19	10567	205425	18657	23781	181644
VUAMC_20 SQ_18	16943	244947	20014	4131	15505
MOH-X	647	5176	1677	192	195
TroFi	3737	105949	13738	50	50

Table 2.5: The details of the metaphorical datasets used.

There were a couple of observations made when analysing the prediction output files for these models. The predictions were divided into two files: correctly predicted, which contained the metaphor sentences/words that models correctly predicted them as metaphor. The second file is the incorrectly predicated file contains the models' prediction of metaphors that were labelled as non-metaphors.

In the incorrectly predicted files for the VUAMC datasets, the word *get* is one of the top words on the incorrectly predicted as literal list. Some words appeared only once, such as *offer, paint* and *distracted*. Words such as *get*, are very common and frequently used in the language. Therefore, they have many meanings/senses. In some instances, *get*, when used in a metaphorical sentence, has other meanings than the usual *catch* or *hold* or *receive*...etc. Understanding these differences is related to their extended, frequent usage over time. This situation led to the loss of its novelty and it became very familiar. For example:

1. "I said I *got* a good joke for you here, started to laugh and he said I ain't said it yet!"
2. "I've only *got* until tomorrow."

In these examples, the verb *get* is used to express very common concept. These concepts do not relate to the *physical possess*, instead, the notion of *possess* is metaphorical,

referring to abstract entities such as *a good joke* and *time*. Although these entities are not something that would be defined using one of the senses, the model incorrectly classified them as literal.

Another point to consider is the word meaning given to the model through the vector representation. Because GloVe embedding only supplies one definition (the most commonly used) regardless of the intended sense, the model judges the target word only on that meaning, which may or may not be the one intended for use in the metaphorical sentence. ELMo provides contextual information in some of the models, but it seems that these common words pose challenges for detection models. In addition, the model may have detected more occurrences of these structures in the non-metaphor tagged samples. As a result, any such case was categorised as non-metaphor by the model (i.e., more non-metaphor sentences in the training and validation files than the metaphorical ones).

Lastly, the concreteness rating might have also played a role in the points discussed earlier. The SVM predicted 2 as the concreteness score for the word *get*. It also predicted concreteness scores of 1, 1, 2, 1 for the surrounding words *a*, *good*, *some*, *until*. As mentioned previously, if there is a shift in the words' concrete-abstract ratings in a sentence, it is most likely used as a metaphor. However, in the case of the sentence *got a good joke* with the ratings 2 1 1 2, this slight degree shift may have failed to signal the model to consider this situation as a metaphor

This point is supported when analysing the examples in the correct metaphor file. Although the same meaning and frequent logic usage are applied here, the concreteness of the drastic shift observation is achieved. In the example, "*And with a miserable face like hers I'm surprised you ever get any work done at all*" the words *any word* each have 1, 3 concreteness scores. The shift from 2 in *get* to 1 and 3 for *any* and *work* is more than one degree, which could successfully have signalled the model to consider the possibility of

metaphor usage.

The MOH-X for incorrectly predicted words has 87 tokens, among the most incorrectly predicted words as literal: match, push and rotate. In the example "*She is pushing fifty.*", the word *push* in this context requires additional context to clarify its meaning. The sentence suggests that the person referred to is almost fifty years old. However, the sentence does not explicitly specify that *fifty* refers to *age*, which could also be interpreted as *something that weighs fifty.* Consequently, the context representation of the ELMo vector is somewhat lacking, making the word *push* in this sentence ambiguous without further clarification. In addition, the concrete-abstract ratings for the sentence are the word she has 4, is has 2, pushing has 3, and fifty has 3. The rating for fifty is very close to be concrete which reflects the assumption above that it could be referring to weights.

Other type of words consist most of the list incorrectly predicted which are the prepositions. Table 2.6 shows a sample of prepositions, their frequency in the VUAMC_20 dataset, and the number of true positives and false negatives. Among the 4,638 occurrences of the word *of*, labelled as a metaphor only 151 times. The word *with*, in contrast, labelled metaphor 620 out of 995 appearances, and the word *up* has 137 metaphor labelled occurrences out of 335 times. These frequent prepositions' occurrences blurred the distinction between their literal and metaphorical senses. For example, in the sentence "*Francesca Simon describes some of the pitfalls and how to avoid them.*" the model incorrectly labelled the word *of* as literal.

Moreover, prepositions form phrases that occupy a big chunk of the sentence structure, such as in *some of the*. The phrase structure raises the difficulty of correct classification because, usually, the phrasal-verb phrase as a chunk provides a meaning that differs from its parts. This difference holds true as well with the parts' concreteness ratings. Each element in the phrasal-verb fragment could have a score that differs from the supposed score

Prepositions	True Positives	False Negatives	Total
of	3	148	4638
to	538	239	3731
in	1198	285	2811
with	510	110	995
go	24	40	258

Table 2.6: Sample of the prepositions in the VUAMC_20 dataset and their counts of True-Positive and False-Negative.

of the whole. The overall meaning of the phrase is *set up* in this example "*The use of role play, incorporating previously discussed difficulties (i.e. homework assignment session 4) in real or set up situations provide an opportunity for testing these skills.*" is different from the meanings of its parts *set* and *up*. Additionally, the word *set* was correctly labelled as a metaphor, but incorrectly labelled the second part of the phrase *up*. Moreover, the information related to this fragment *real or set up* situation is as follows: The word *real* has a 2 rating as its concreteness score, labelled as literal in the dataset, was correctly classified as such; likewise for the word *or*. As for the metaphor word *set*, correctly classified as a metaphor, it has a concreteness score of 3. The case is different for the following ward *up*, concreteness rating is 3, was incorrectly labelled as literal. Lastly, the word *situations*, which has a concreteness score of 2 was incorrectly labelled as literal. It could be explained that the noticeable shift in concreteness score for the words *or* and *set* (1 to 3) has contributed to successful labelling of the word *set* as a metaphor. Failing to successfully do the same with the second part of the phrase *up* was due to the lack of a shift in concreteness score between *set* and *up*; both have 3 as the rating when the two parts are combined as a single phrase with one meaning.

There are 30 tokens in TroFi's list of incorrectly predicted words, with step being among the most misclassified. Almost all instances of misclassification for the word *step* occurred when it was used as a phrasal verb as in the example, "*Mr.Instone, 64 years old,*

stepped down in February from his posts as president and chief executive officer.". In this case, the phrase *stepped down* does not have a literal *descending* meaning but refers to *resigning from a position.* The issue stems from the fact that the word labelled as metaphor was *push* without including the rest of the fragments. Additionally, GloVe representation was not helpful in capturing the intended meaning as it focuses on the basic meaning associated with the word. Moreover, the assigned concreteness ratings based on the basic meaning further reinforced the false assumptions. The word *stepped* received a rating of 3, *down* a rating of 3, *in* a rating of 4, and *February* a rating of 3. The lack of shifts did not provide any hints regarding the possibility of metaphorical usage in this context.

When examining the words and their reparations, some words did not appeared in many sentences. Approximately 7527 unique tokens appeared exactly once. For example, *disclosed* in the sentence "*There were others , but Lucy never disclosed any of them to us*" was initially labelled as a metaphor but incorrectly classified as literal. Reading this sentence gives two interpretations. Either Lucy's knowledge of some people's identities and metaphorically points out that Lucy did not *uncover* this information when referencing *others* and *them*, or that it is a literal *uncover* of Lucy's secrets when referencing *others* and *them.* While the concreteness rating for the words in the sentences ranged between 1 and 2, and Lucy with 4, it could be said that the ratings did not help clarify the intended meaning. Similar ambiguous sentences in the same position, such as "*Bribery and corruption!*", need more context to catch the intended meaning (literal or metaphorical). The latter sentence has the following concreteness ratings: 2, 1, 2. In order to increase the probability of the successful classification of the metaphor word corruption, the model needs to be provided with more context. Table 2.7 presents a sample of sentences that have words that appear only once and shows these words' labels, predicted labels, and their concreteness ratings.

After analysing the results of these models on several metaphorical datasets, employing GloVe only or including the contextual representation ELMo, the research question

The Sentence	The Label	The Predicted Label	Concreteness Rating
And they told it without onscreen questioning , though the programme is *skilfully* structured to give it a coherence it might have lacked .	0	0	1
The burn threads a wild and inhospitable crevice of the hills , where the wind blows cold and the sense of *isolation* grows with each lonely mile .	1	0	2
Although that is the position in law , he court emphasised that as a matter of sense a tenant should first complain to the landlord before *exercising* the right to prosecute .	1	1	2

Table 2.7: Samples of words that appeared only once, along with their sentences, labels, the model's predicted labels, and their concreteness ratings.

could be answered. The first research question (RQ1) covers the idea of whether this human-driven knowledge of the concept of concreteness-abstractness contributes to understanding of metaphors and building some of their theories. Based on the previously analysed result of two models on three different datasets, this question could positively be answered. The hypothesis states that metaphors could be used in combining abstract concepts with concrete ones in order to convey the intended meaning. Typically, abstract concepts that are difficult to explain and comprehend are presented in contexts that are more concrete, tangible, and easily understood. As a result, it is more likely to see that metaphorical words tend to have higher concrete-abstract scores compared to the surrounding words. This scoring indicates that the metaphorical word is grounded in a more tangible and accessible framework, making it easier for the listener to grasp its intended meaning. As in the instance of this sentence "*solid idea*" (solid has 3 and idea has 1) where the general intention is to say that the idea is logical, well-founded and coherent. The ability of many metaphorical neural network models to utilise this observation and improve their performances can answer the RQ3 about the role of these models in explaining the function and distribution of metaphor in language.

When RQ2 asked about the resource used in this experiment, the current available resource showed that concreteness is a useful tool for metaphor detection tasks. It does, however, fall short when it comes to determining the scope of the score if it is related to a single token or a whole phrase. It would be interesting to give chunks or phrases as a whole the concrete-abstract score and see how the models do. Chapter 7 provides an in-depth analysis and observation that builds upon the previous discussions. It highlights the limitations observed in the methodology and tools employed during the research. Additionally, it offers recommended solutions and suggestions to address these limitations and improve future work. By presenting a comprehensive examination of these aspects, Chapter 7 contributes valuable insights to advance further studies in this area.

2.5 Conclusion

Chapter 2 presented an in-depth examination regarding the impact of adding concreteness-abstractness ratings to the deep learning models on their performance in the metaphor detection task. The hypothesis proposes that metaphorical sentences often involve an abstract idea expressed through more concrete words. For instance, the metaphor "*life is a rollercoaster*" represents the abstract concept of *life* through the tangible example of a thrilling amusement park ride. By perceiving the *rollercoaster* as a game with its ups and downs, the metaphor conveys the meaning of the struggles and joys of life.

To test this hypothesis, two detection approaches, sentence-level and word-level, were employed on three metaphor datasets. Each token was assigned a rating using a trained Support Vector Machine (SVM) on a list of words rated for concreteness. The models' performances demonstrated improvements in the detection task, providing insights into the role of metaphor in language and the potential application of these resources in automated

detection experiments. Additionally, the analysis identified certain limitations which will be further discussed in the concluding chapter (Chapter 7). This chapter will also provide recommended solutions for future research, extending the understanding of metaphor detection and its implications.

The next chapter covers the second component of semantic richness's perceptual part, which is the sentiment. The chapter will cover sentiment in different forms, measure as polarity (0 or 1) or as definitive emotion (happiness, sadness and others). The rating assignment process will follow the same steps as concreteness, utilising a trained SVM. Additionally, the assignment will be tested using a BERT model trained on five specific emotions. It will be interesting to observe how sentiment influences the models and to determine which rating method, either SVM or BERT, yields a more significant influence. This exploration will shed light on the role of sentiment in metaphor detection and provide valuable insights into the relative effectiveness of these two rating approaches.

Chapter Three

Sentiment

3.1 Introduction

In Chapter 2, the first element of the perceptual aspect of semantic richness, known as the concreteness-abstractness measure, was introduced. This element was incorporated into neural networks to evaluate its impact on the automated metaphor detection task. It served as the only external feature concatenated with the word-context vector representations of GloVe and ELMo. The concreteness ratings were obtained through predictions made by a trained SVM, which were then assigned to tokens from three metaphorical datasets. The experimental results displayed promising outcomes, giving improvements across all three datasets for the three models. Although some limitations were identified and briefly discussed, a more comprehensive analysis and future recommendations addressing these limitations are presented in Chapter 7.

This chapter explores the second component of semantic richness, which is sentiment (i.e. the overall attitude, opinion, and emotional disposition someone holds in regards to a particular idea or statement). Since metaphors often convey emotional intent, incorporating sentiment could be valuable for metaphor detection. Two types of sentiment are tested

and compared: polarity (*positive* or *negative*) and emotion (such as *anger*, *joy*, etc.). These sentiment ratings are obtained from SVM and BERT models trained on multiple sentiment datasets. The experiments are conducted using two neural network models: one for sentence classification and one for sequence labeling, applied to three metaphorical datasets (VUAMC, MOH-X, and TroFi). Subsequently, the results are presented and analysed, while also discussing limitations and proposing potential solutions.

3.2 Sentiment and Figurative Language

Communications became easier with the inventions and the continuous development of social media. These communications involve covering and discussing many topics and the growing exchange of viewpoints. These discussions carry feelings and perspectives, which are the focus of sentiment analysis, whether expressly stated "*that movie was exciting*" or implicitly stated "*that movie was a roller-coaster*". Even though this improved communication helped the researchers in sentiment analysis and other NLP fields by searching useful information from textual data, the implicit statements and other figurative language, such as metaphors, made the work harder. According to several academics, metaphors communicate emotions and attitudes (S. M. Mohammad, 2016). "*She is the sun of my life*", for example, was written to discuss **LOVE** when the line addressed the **SUN**. The statement, "*I attacked that first point*" depicted the passionate **ARGUMENT** similarly using **WAR** words. Even though the study of metaphor sentiment has been established, there has been little research into detecting metaphors using sentiment. An investigation into the field of sentiment as an additional feature to the neural network for the metaphor detection task was conducted.

This chapter focus on the sentiment as the primary lexical feature. The aim is to explore the feasibility and effect of combining polarity or emotions with neural networks for

metaphor detection.

3.3 Related Work

The majority of research on metaphor and polarity has focused on conventional metaphors or metaphors with expressed emotion structures. For example, "*fear is killing me*". Rentoumi et al. (2012) attempted to find the polarity, negative or positive, of metaphorical statements regardless of the structure or type. The investigation divided the study into three consecutive stages. The first stage was Word sense Disambiguation (WSD), due to the study's use of metaphorical statements. A novel metaphor has a meaning that is difficult to decipher. The second stage is sense-level polarity assignment, which involves determining the polarity of a word based on its context. Words with different contexts achieve different functions, natural or with negative/positive connotations. The last stage is assigning the polarity decision to the whole sentence.

The goal of the SemEval 2015 "The Sentiment Analysis of Figurative Language in Twitter" track is to improve figurative language sentiment scores. The goal was to obtain fine-grained sentiment scores for sarcasm, irony, metaphor, and other sentiment categories. The supervised models in Karanasou et al. (2015) employed feature selection to assign scores to tweets. They used the Senti-WordNet resource, which gives WordNet and words a sentiment score based on similarity. On the other hand, Farias et al. (2015) utilised not only polarity-related sentiment resources but also emotions and other psycholinguistic information (e.g., word frequency, POS frequency, and Affective Norms) with the regression model to assign the polarity score to tweets.

The study by Biddle et al. (2020) used sentiment to improve Health Mention Classification (HMC) in tweets. Research studies in the field have discovered a link between the

mentioned disease as figurative language and negative sentiment. Both the HMC and the figurative language tasks got better when they used the pre-trained contextual representation and sentiment for health-related words together.

The primary objective of these cases was to automatically detect the sentiment polarity or emotion in various text forms. Figurative language, particularly metaphor, is believed and shown to evoke some level of emotion or sentiment. Therefore, sentiment in both forms, polarity and emotion, were added to the GloVe+ELMo vector for word/context representation. The sentiment ratings were acquired from two sources: 1) a trained SVM on a list of words rated for sentiment and 2) a BERT model that is trained on sentences rated for emotions. The purpose was to compare the two models and their assignment (words vs. sentences) and test which rating output positively impacted the metaphor detection models. The experiments were done using two models: sentence classification built on Gao et al. (2018) and sequence labelling Mao et al. (2019) on three metaphorical datasets: the VUAMC, MOH-X, and TroFi provided by Gao et al. (2018). This chapter presents the incorporation of sentiment information into the word/context representation to enhance the detection of metaphors

3.4 Metaphor Detection Experiments

As previously mentioned, studies looked into detecting the sentiment of different figurative language forms. In this chapter, however, the goal is use sentiment to detect metaphors automatically using Deep Learning models. As explained in Chapter 2, the concreteness-abstractness rating were tested as a single-digit rating and as a probability distribution rating. This format (discussed in detail in section 3.5) is going to be tested in this chapter with sentiment.

3.4.1 Datasets

Several resources for sentiment information are used in these experiments because the nature of polarity and distinct emotions differ in the way they are presented. Polarity gives the entities it classifies (words, phrases, or sentences) a positive or negative score without specifying any named emotion (*joy*, *sadness*, or so on). For example, the word "*captive*" entails negative sentiment, but it does not specify under which negative emotion it falls. On the other hand, the emotion information classifies the entities into several types of named emotions.

Polarity

S. Mohammad and Turney (2010) produced a dataset that contained more than 14 thousand English words with sentiment polarity and emotions indicated. Usually, a 1 in the sentiment score annotation means the polarity is *positive*, and a 0 means the polarity is *negative*. However, S. Mohammad and Turney (ibid.) annotated their data differently because they included a score for *others*. The new annotation becomes: a word has a score of 1 if it indicates a positive or negative polarity, and 0 if it indicates neither. This means that some words have 1 for both positive and negative, and others have 0 for both. For example, the word *bookworm* was given a 1 in both the negative and positive senses. That is because the word's sentiment could refer to either of the two options depending on the context. As for words with 0 for both positive and negative, these words refer to neither and are categorised as *others*.

Emotion

S. Mohammad and Turney (2010) hypothesised that words might elicit specific emotions; therefore, they further investigated eight named emotions. Since metaphors could have *negative* or *positive* sentiments (without specifying which negative emotion (*sadness* and *anger*) or positive emotion (*joy* and *anticipation*) it is, they could also evoke certain emotions such as *sadness* and *surprise*. Two different forms of emotional resources were looked into: the NRC list (ibid.) and the BERT multi-class sentiment analysis model for assigning emotions to sentences (Garbas, 2019).

1. Each word in the NRC list (S. Mohammad and Turney, 2010) has either a 0 or 1 for each of the eight emotions. These emotions are: *anger, anticipation, disgust, fear, joy, sadness, surprise* and *trust*. Annotated words with a 1 indicate that the word evokes this emotion, whereas a 0 indicates that it does not. Some words have 1 for multiple emotions, indicating that these words express these emotions (e.g., *captive* marked 1 for both *fear* and *sadness*). In other cases, words have 0 for all emotions, meaning that these words evoke none (e.g., *distant*). Each of these emotions was looked at on its own and then combined to see how they affected the metaphor detection task.

2. The BERT model for multi-class sentiment analysis (Garbas, 2019) is a fine-tuned BERT model trained on three resources (Ghazi et al., 2015; Y. Li et al., 2017; Scherer and Wallbott, 1994) for five emotions: *joy, sadness, anger, fear*, and *neutral*. These resources are lists of annotated sentences for one of the five named emotions.

Metaphorical Datasets

The Deep Learning models performed the detection tests on three different metaphorical datasets previously mentioned in detail in the previous chapters. The VUAMC, MOH-X

and TroFi datasets are the version provided by Gao et al. (2018) and used by Mao et al. (2019) in their experiment. The verbs in the VUAMC for sentences-level are the target for metaphocicty prediction based on the context (MOH-X and TroFi also for sentences-level), while each word in the sentence in the VUAMC for words-level is target for metaphocicty prediction. Both versions were divided into three files: train, validation and testing. The other two dataset, however, were tested as 10-fold cross validation. Both model, the sentence classification and the sequence labelling, use GloVe+ELMo vectors as the word/context representation. The new predicted sentiment ratings were concatenated with these vectors and fed to the neural network models.

The next section gives an in-dept breakdown to the experiments' set up. Due to different sentiment forms and models used to acquire the ratings, the metaphor experiments will follow the several steps discussed below.

3.5 Experiment Set-Up

The current hypothesis states that if words can express negative or positive sentiment, then metaphorical statements could as well. In addition, literal statements are more neutral or have a lower sentiment degree. Therefore, it is plausible to use sentiment information to detect metaphors. These hypotheses were challenged with classification and sequence labelling models tested on the VUAMC, MOH-X and TroFi datasets combined with GloVe-ELMo vector representations.

The first stage of the setup is to obtain sentiment ratings. Two forms of sentiment will be tested: polarity and emotion. One dataset consists of unigrams rated for polarity and eight emotions will be used to train an SVM model. This model will assign two forms of rating to the metaphorical datasets' tokens: single-digit rating and probability distribution.

The other dataset consists of three lists of sentences rated for one of five emotions. This set of datasets will be used to train a BERT model used to classify the emotion of a sentence. However, this model will be used as one format, single-digit, assigned to the metaphorical datasets' tokens. The divisions in details are as follows:

1. This step used the NRC list to train the SVM using BERT pre-trained embeddings for tokens to predict the polarity and emotion for all words in the metaphorical datasets. First, its predictions were tested as a single-digit number, representing each type's sentiment. After that, the probability distribution predictions were examined.

    - Polarity: for single assignment it is either: 1, 0 (negative or positive) or -2 (neither). The portability distributing is a 3-digit rating representing each rating.
    - NRC eight emotions: each emotion is represented by a number from 0 to 7). Each emotion is tested separately, meaning the assigned prediction is 1 or 0. The portability distributing is a 2-digit rating representing each rating.
    - BERT 5 emotions: 0 to 4 represents one of the 5 emotions. The prediction is a single-digit

2. The second step observed the impact of adding the two forms of sentiment together in one experiment. The NRC polarity and emotions' best results were combined and examined.

3. The third step tests the detection when the words are being annotated as *sentiment* or *not*, regardless of what polarity or emotion they were.

3.5.1 Step 1

The NRC list by S. Mohammad and Turney (2010) has about 14 thousand words annotated for both polarity and eight emotions. If a word was assumed to evoke a particular senti-

ment type, then the sentiment type was labelled 1. The NRC word list does not cover the metaphorical datasets' vocabulary sizes. Table 3.1 provides statistics on the number of words in the NRC list per sentiment form (polarity and emotions) that exist in each metaphorical dataset. It demonstrates that the NRC list has a relatively small number of words compared to the vocabulary size of each dataset. To examine the impact of sentiment on the metaphor detection models, each token will receive a sentiment rating. This is accomplished by training SVMs on each sentiment form separately and assigning the resulting predictions to the metaphorical tokens.

The 300-dimension GloVe and the contextual 1024-dimension ELMo vectors for each of the VUAMC, MOH-X and TroFi words was extracted first. The SVM provided a prediction related to the sentiment form that was being tested, then the label was added to the vector. After that, the eight emotions were combined into one experiment for the VUAMC dataset.

Similarly, the fine-tuned BERT model assigned a prediction for one of five emotion to each token in the metaphorical dataset. This prediction then was concatenated with the GloVe+ELMo vector and fed to the metaphor detection neural models.

The step is applied using both neural network models, the classification model for sentence-level build on the proposed Gao et al. (2018) and the sequence labelling model (RNN_HG) for word-level proposed Mao et al. (2019).

Polarity

The annotation in the NRC list differs in terms of the annotation scheme. In the known polarity annotation, a value of 0 represents negative, and 1 represents positive. However, the annotation scheme used here indicates that if a word evokes a positive, negative, or both

Lexical Feature	VUAMC Cl.	VUAMC Seq.	MOH-X	TroFi	Total Lexicon Tokens
Pos_Neg	2420	2420	298	1842	5474
Anger	509	509	64	407	1247
Anticipation	452	452	371	407	839
Disgust	384	384	47	278	1058
Fear	608	608	85	535	1476
Joy	357	357	58	292	689
Sadness	515	515	66	412	1191
Surprise	246	246	38	205	534
Trust	656	656	88	553	1231
Total	17017	16979	1677	13738	_

Table 3.1: The breakdown of the NRC words per emotion, and the metaphorical datasets' words per emotions.

sentiments, it is annotated with 1, and it receives a 0 if it evokes neither, as discussed in Section 3.4.1 above. Therefore, some readjustment were needed. A word is labelled 1 is it evokes the positive polarity, labelled 0 if evokes the negative polarity and -2 for neither. The resulted list for pos_neg polarity words that used to train the SVM contains 5474 words.

The SVMs were trained on polarity and emotions in each case above, and the results were compared in this section. The SVM takes an input of pre-trained BERT embeddings for each NRC word. After training, the pre-trained BERT embedding for the VUAMC, MOH-X and TroFi tokens was extracted, and the SVM's polarity predictions were obtained. The prediction forms were either:

1. Single digit where each word gets either a 1 for the positive, a 0 for negative, or -2 for neither label. This prediction was then added to the GloVe+ELMo vector for the classification and sequence labelling models.

2. The probability distribution consists of three digits. Each digit represents the prob-

ability that this word falls under each polarity type, then concatenated with the GloVe+ELMo vector.

The Table 3.2 shows the F1 results of the trained SVM for each sentiment. The results are produced using a 10-fold cross validation.

NRC Emotion	F1 Mean	F1 MAX	F1 MIN
Single-Digit			
POS_NEG	72.3%	74.4%	69.6%
Anticipation	93.9%	94.2%	93.5%
Anger	91.5%	92.9%	90.1%
Disgust	92.3%	93%	90.9%
Fear	90.1%	91.2%	88.5%
Surprise	95.8%	96.3%	95.2%
Sadness	91.6%	92.4%	90.4%
Joy	94.6%	95.3%	93.3%
Trust	90.9%	91.9%	90.1%
Probability Distribution			
POS_NEG	84.2%	88.1%	80.8%
Anticipation	93.4%	94.1%	91.8%
Anger	91.5%	92.5%	90.5%
Disgust	92.4%	93.6%	91%
Fear	89.9%	91.5%	88.7%
Surprise	95.5%	96.1%	94.7%
Sadness	91.5%	92.5%	90.2%
Joy	94.7%	95.7%	93.5%
Trust	90.6%	91.5%	89.7%

Table 3.2: The mean, max. and min. F1 scores for each SVM trained on the NRC list for polarity and emotions

Emotions

The NRC list covered eight emotions in addition to polarity: *anger, anticipation, disgust, fear, joy, sadness, surprise* and *trust*. The annotation scheme stated that words with label 1 evoke a particular emotion, while words with 0 do not. A word can have a label of 1

for multiple emotions or have 0 for all. Therefore, the metaphor detection experiments were divided, and each emotion was separately tested. In the last experiment, they were combined to observe their collective impact. In addition to the NRC list, a BERT model that classifies the emotions of sentences was trained. Table 3.2 presents the F1 scores for the SVM of each NRC emotion. The BERT model classifies five emotions: *joy, sadness, anger, fear*, and *neutral*. BERT assigned one of these emotions to the metaphorical datasets' tokens.

- NRC emotion experiments were split based on the type of emotion. For each of the eight, only one label (1 if the word evokes that emotion, 0 if not) was added to the GloVe+ELMo vector.

- The combined eight emotions were different. The concatenated labels formed a one-hot encoding vector. If a word evoked that emotion, then 1 was assigned under its designated index, and 0 if not. The rating vector was concatenated with GloVe+ELMo vector. This is tested on the VUAMC dataset.

- For the BERT case, the model was trained on sentences and classified sentences, but the trained model here was used to classify the metaphorical datasets' words. Therefore, the BERT was not divided into smaller sections for each emotion. The goal was to observe the ability of BERT against the NRC emotions list. The number of labels added was from 0 to 4, five labels in total. Table 3.4 shows the number of tokens in each metaphorical dataset under each BERT emotion. The model classifies sentences, but it was used to classify words. Table 3.3 shows the result of the BERT model in terms of precision, recall and F1 for each individual emotion.

Emotion	Precision	Recall	F1
Joy	0.85	0.85	0.85
Sadness	0.79	0.81	0.80
Fear	0.84	0.85	0.85
Anger	0.81	0.78	0.80
Neutral	0.82	0.83	0.82

Table 3.3: The precision, recall and F1 metrics for the trained BERT model on sentences rated for emotions.

Lexical Feature	VUAMC Cl.	VUAMC Seq.	MOH-X	TroFi
Total words	17941	16979	3426	14881
Joy	106	76	340	438
Sadness	172	313	60	258
Fear	194	249	50	447
Anger	187	260	47	291
Neutral	17282	12890	3193	13447

Table 3.4: The breakdown of the NRC words per emotion, and the metaphorical datasets words per emotions.

3.5.2 Step 2

This step tests the impact of combining the polarity and the emotion into the same vector on the metaphor identification task. The best performance out of the polarity and two emotions with the best results (*anticipation* and *fear*) were chosen for the VUAMC dataset. The GloVe+ELMo vectors were concatenated with two additional digits. One got either 1 or 0 for positive or negative, and the second got a 2 for *anticipation* or a 4 for *fear*. The combination of *polarity* and *disgust* was done on MOH-X dataset, while TroFi combined *polarity* and *anger*. This step was tested using the classification model (Gao et al., 2018).

3.5.3 Step 3

In this step, the BERT multi-class sentiment for the VUAMC classification result was taken and changed; every word was changed to become a *sentiment* word or *non-sentiment* word. Words classified with one of the four BERT emotions got a 1, *sentiment* class. If they got the fifth, *neutral*, The words with 0, *non-sentiment* were labelled. The idea was to test the overall *sentiment-non-sentiment* situation without specifying the category.

This step was tested on the two models (Gao et al., 2018) and (Mao et al., 2019).

3.5.4 Classification

The classification model classifies the sentences into metaphors or non-metaphors. Based on the target word and its context, the model extracts the necessary information for the prediction. As discussed earlier in Chapter 1, the baseline for classification model is different from the reported result in Gao et al. (2018). Therefore, the baseline was chosen as the best run out of several with hyperparamtert adjustments. The model was applied using GloVe+ELMo vectors on the VUAMV dataset. The tests on the MOH-X and TroFi datasets were based on the 10-fold cross validation with hyperparameter adjustment, too.

The model was tested with each case in the steps explained above, and compared against the chosen baselines. The results were for precision, recall, and F1 for metaphorical prediction.

Step 1

This step involves evaluating the classification performance using two approaches: 1) SVMs trained on the ratings from the NRC list, and 2) the fine-tuned BERT model trained on

sentences for five different emotions. The objective is to assess the impact of the predicted ratings when utilising the predicted SVM/BERT assignments. The results of the SVM-based emotion predictions, represented as single-digit ratings, combined with the metaphor classification model, are presented in Tables 3.5, 3.7, and 3.9. On the other hand, Tables 3.6, 3.8, and 3.10 display the results of the classification experiments when using the predicted ratings in the form of probability distributions.

When observing each emotion individually with the classification model, it was found that *anticipation* and *fear* emotions showed the best overall results for the experiments on the VUAMC dataset, with *fear* achieving a slightly better F1 score. However, the recall for *fear* was better than the recall for *anticipation* and the baseline recall. An increase in recall indicates that the model correctly predicted the metaphorical sentences. On the other hand, when combining all eight emotions together, the overall performance was negatively affected, resulting in an F1 score of 43.12%. Although precision increased from 56.2% to 61.82%, it means that the model made more correct predictions for both literal and metaphorical sentences, rather than accurate metaphor predictions alone. The recall decreased to 33.106% from the initial 51.1%, indicating that the model did not identify many metaphors among the correctly predicted sentences. Table 3.5 provides the hyperparameters used to achieve these results.

Similarly to the single-digit predictions, the probability distribution had both positive and negative effects on metaphor classification testing. The recall and F1 scores for *polarity* slightly decreased, while the precision showed a slight increase to 57.35%. The recall for the *anger* emotion decreased significantly, dropping from the initial 51.1% to 26.46%, and the F1 score dropped to 37.22% from 53.5%. A similar pattern was observed for the *sadness*, *surprise*, and *trust* emotions. However, there was a slight increase in the F1 score for *disgust*, from 53.5% to 53.9%, while the recall decreased from 51.1% to 50.19%. On the other hand, the *fear* emotion showed improved results, with a slight decrease of 1.27% in precision and

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC Single-Digit											
Gao et al., 2018 Reported	53.4%	65.6%	53.4%								
Baseline	*56.28%*	*51.107%*	*53.57%*	128	0.01	0.3	0.0	0.2	64	1	20
Polarity	55.1%	58.8%	**56.9%**	128	0.001	0.3	0.0	0.2	32	1	20
Anger	57.5%	45.7%	50.9%	300	0.001	0.3	0.0	0.2	64	1	20
Anticipation	54.8%	55.7%	55.3%	300	0.001	0.3	0.0	0.2	64	1	20
Disgust	60.31%	42.9%	50.1%	300	0.001	0.3	0.0	0.2	64	1	20
Fear	53.9%	**59.2%**	**56.4%**	300	0.001	0.3	0.0	0.2	64	1	20
Joy	58.9%	40.7%	48.1%	300	0.001	0.3	0.0	0.2	64	1	20
Sadness	57.7%	49.4%	53.3%	300	0.001	0.3	0.0	0.2	64	1	20
Surprise	**62.02%**	33.67%	43.6%	300	0.001	0.3	0.0	0.2	64	1	20
Trust	57.75%	47.58%	52.17%	300	0.001	0.3	0.0	0.2	64	1	20
All eight emotions	61.82%	33.106%	43.12%	300	0.001	0.3	0.0	0.2	64	1	20

Table 3.5: A sample of trials showing some of the effect results of step 1 experiments on the VUAMC dataset with the single-digit SVM prediction.

an increase of 5.1% and 2.03% in recall and F1, respectively. All metrics for *joy* increased, with precision, recall, and F1 scores improving by 1.1%, 4.1%, and 2.7%, respectively. Table 3.6 presents the hyperparameters used in all experiments.

In the experiments with single-digit predictions on the MOH-X dataset, the *polarity* emotion showed improvements in recall, reaching 81.37% with a 5.13% increase, and the F1 score reached 77.76%. All NRC emotions exhibited some improvements in F1 scores, with the highest being for *disgust* at 77.74%, while the highest recall was observed for *joy* at 79.61%. The precision increased for all emotions, with *disgust* having the highest increase, but it decreased for *polarity*, *anger*, and *joy*. The hyperparameters used to achieve these results are presented in Table 3.7.

The probability distribution rating format with classification metaphor detection on the MOH-X dataset provided similar results to the single-digit predictions. There was an increase in F1 scores in all cases, with better improvements observed in recall and precision

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC Prob. Dist.											
Gao et al., 2018 Reported	53.4%	65.6%	53.4%								
Baseline	*56.28%*	*51.107%*	*53.57%*	128	0.01	0.3	0.0	0.2	64	1	20
Polarity	57.35%	50.02%	53.44%	300	0.001	0.3	0.0	0.2	32	1	20
Anger	**62.718%**	26.46%	37.22%	300	0.001	0.3	0.0	0.2	64	1	20
Anticipation	60.05%	40.03%	48.04%	300	0.001	0.3	0.0	0.2	64	1	20
Disgust	58.31%	50.19%	53.95%	300	0.001	0.3	0.0	0.2	64	1	20
Fear	55.483%	**53.71%**	54.587%	300	0.001	0.3	0.0	0.2	64	1	20
Joy	56.822%	53.2%	**54.95%**	300	0.001	0.3	0.0	0.2	64	1	20
Sadness	58.55%	44.12%	50.32%	300	0.001	0.3	0.0	0.2	64	1	20
Surprise	60.017%	39.46%	47.61%	300	0.001	0.3	0.0	0.2	64	1	20
Trust	62.28%	28.22%	38.843%	300	0.001	0.3	0.0	0.2	64	1	20

Table 3.6: A sample of trials showing some of the effect results of step 1 experiments on the VUAMC dataset with the probability distribution's SVM prediction.

in some instances. For the *polarity* emotion, precision and F1 increased from 75.44% to 78.19% and from 76% to 76.796%, respectively. The *trust* emotion exhibited the highest F1 score, reaching 77.38%, representing a 1.81% increase. The recall showed a slight increase of 0.77%, while precision improved by 2.24%. Table 3.8 presents the hyperparameters used to achieve these results.

The addition of sentiment to the TroFi dataset has a slightly different impact compared to the MOH-X dataset. For the *polarity* emotion, there was a slight increase of 1.9% in recall and 0.34% in F1 score. Except for *trust*, all other emotions showed a small increase in F1 scores, with *anger* having the highest at 71.779%. In terms of recall, all emotions generally showed an increase starting from 73.04%. The *fear* emotion had the highest precision, reaching 70.31% compared to the baseline precision of 69.661%. Table 3.9 provides the hyperparameters for each experiment.

The TroFi experiments using the probability distribution prediction format showed

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
MOH-X Single-Digit											
Gao et al., 2018 Reported	75.3%	84.3%	79.1%								
Baseline	*75.44%*	*77.393%*	*76%*	300	0.001	0.2	0.0	0.2	10	1	30
Polarity	74.88%	**81.37%**	**77.76%**	300	0.001	0.2	0.0	0.2	10	1	15
Anger	74.68%	79.38%	76.63%	128	0.001	0.2	0.0	0.2	10	1	15
Anticipation	77.961%	77.082%	77.222%	300	0.001	0.2	0.0	0.2	10	1	15
Disgust	**78.38%**	77.638%	**77.74%**	256	0.001	0.2	0.0	0.2	10	1	15
Fear	76.13%	77.89%	76.72%	256	0.001	0.2	0.0	0.2	10	1	15
Joy	73.766%	79.61%	76.26%	256	0.001	0.2	0.0	0.2	10	1	15
Sadness	75.84%	78.07%	76.7%	300	0.001	0.2	0.0	0.2	10	1	15
Surprise	75.2%	78.59%	76.6%	300	0.001	0.2	0.0	0.2	10	1	15
Trust	75.73%	78.93%	76.89%	128	0.001	0.2	0.0	0.2	10	1	15

Table 3.7: A sample of trials showing some of the effect results of step 1 experiments on the MOH-X dataset with the single-digit SVM prediction. The hyperparametres used to reach these result are recorded.

an overall increase in F1 scores. The emotion *sadness* had the highest F1 score, reaching 72.05%. Similarly, the recall increased for all emotions, with *trust* showing the highest increase, reaching 76.77% from the initial 73.04% baseline. However, the precision exhibited varying rates of decrease for all emotions, with *trust* having the most significant decrease, reaching 66.8% from the starting 69.661%. The hyperparameters used to produce these results are in Table 3.10.

The BERT model was trained to classify emotions in short messages using several datasets annotated for each one of five emotions. After training the model, it was utilised to predict the emotions of the VUAMC, MOH-X, and TroFi datasets, and these predictions were tested with the metaphor classification model. The results of the metaphor prediction using BERT sentiment are presented in Table 3.11, and it has shown the best results so far in Step 1.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
MOH-X Prob. Dist.											
Gao et al., 2018 Reported	75.3%	84.3%	79.1%								
Baseline	*75.44%*	*77.393%*	*76%*	300	0.001	0.2	0.0	0.2	10	1	30
Polarity	**78.19%**	75.72%	**76.796%**	256	0.001	0.2	0.0	0.2	10	1	15
Anger	74.61%	80.55%	77.13	300	0.001	0.2	0.0	0.2	10	1	15
Anticipation	74.96%	78.44%	76.28%	256	0.001	0.2	0.0	0.2	10	1	15
Disgust	73.75%	**81.73%**	77.31%	300	0.001	0.2	0.0	0.2	10	1	15
Fear	76.12%	76.86%	76.08%	256	0.001	0.2	0.0	0.2	10	1	15
Joy	77.41%	77.7%	77.209%	256	0.001	0.2	0.0	0.2	10	1	15
Sadness	75.764%	77.25%	76.065%	256	0.001	0.2	0.0	0.2	10	1	15
Surprise	74.91%	80.93%	77.37%	300	0.001	0.2	0.0	0.2	10	1	15
Trust	77.13%	77.989%	**77.38%**	300	0.001	0.2	0.0	0.2	10	1	15

Table 3.8: A sample of trials showing some of the effect results of step 1 experiments on the MOH-X dataset with the SVM's probability distribution predictions. The hyperparametres used to reach these result are recorded.

For the VUAMC dataset, the F1 score increased significantly to 59.297% from the initial baseline of 53.5%, representing a notable improvement of 10.83%. The recall also experienced a substantial increase of 33.229%, reaching 68.08%. However, the precision decreased to 52.51%, showing a decrease of 6.5%. The increase in F1 scores for the MOH-X and TroFi datasets was relatively slight compared to the increase observed in the VUAMC dataset. MOH-X's F1 score increased by 2.28%, the recall by 4.006%, and the precision by 0.106%. Similarly, for the TroFi dataset, the F1 score and recall saw slight increases of 0.36% and 2.98%, respectively, while the precision slightly decreased by 2.3%.

These findings demonstrate the efficacy of the BERT sentiment model in improving metaphor prediction, with the VUAMC dataset exhibiting the most significant improvements. Table 3.11 provides the hyperparameters used to produce these results.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
TroFi Single-Digit											
Gao et al., 2018 Reported	69.661%	73.04%	71.088%								
Baseline	*69.661%*	*73.04%*	*71.088%*	128	0.001	0.2	0.0	0.1	5	1	15
Polarity	68.71%	74.434%	71.336%	64	0.001	0.2	0.0	0.1	5	2	15
Anger	69.808%	74.21%	**71.779%**	256	0.001	0.2	0.0	0.1	5	1	15
Anticipation	68.228%	75.1409%	71.214%	300	0.001	0.2	0.0	0.1	5	1	15
Disgust	68.4333%	74.704%	71.303%	300	0.001	0.2	0.0	0.1	5	1	15
Fear	**70.31%**	73.829%	71.699%	128	0.001	0.2	0.0	0.1	5	1	15
Joy	67.127%	**76.122%**	71.261%	128	0.001	0.2	0.0	0.1	5	1	15
Sadness	68.761%	74.933%	71.586%	256	0.001	0.2	0.0	0.1	5	1	15
Surprise	68.006%	75.058%	71.211%	128	0.001	0.2	0.0	0.1	5	1	15
Trust	68.542%	73.723%	70.899%	256	0.001	0.2	0.0	0.1	5	1	15

Table 3.9: A sample of trials showing some of the effect results of step 1 experiments on the TroFi dataset with the single-digit predictions. The hyperparametres used to reach these result are recorded.

Step 2

In this step, the previous step was extended by combining the best aspects of both the NRC polarity and emotion. Step 1 tested each form of sentiment separately, Step 2 tested combining the best tests of polarity with best tested emotion. For this combined experiment, *polarity* with *anticipation* and *fear* was tested using VUAMC and MOH-X tested *polarity* and *disgust*, and TroFi tested *polarity* and *anger*. By incorporating the strengths of both sentiment models, the aim is to further enhance the performance of sentiment prediction in the metaphor detection task.

Emotions, *anticipation* and *fear*, and polarity were combined with the GloV+ELMo vector representation to test their impact on the metaphor detection task on the VUAMC dataset. Table 3.12 shows the summary of this stage. The results showed an overall decrease for all three measurements. The increase in precision indicates that the model did not

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
TroFi Prob. Dist.											
Gao et al., 2018 Reported	69.661%	73.04%	71.088%								
Baseline	*69.661%*	*73.04%*	*71.088%*	128	0.001	0.2	0.0	0.1	5	1	15
Polarity	69.38%	73.733%	**71.293%**	256	0.001	0.2	0.0	0.1	5	1	15
Anger	69.03%	74.314%	71.37	256	0.001	0.2	0.0	0.1	5	1	15
Anticipation	67.96%	75.9%	71.71%	300	0.001	0.2	0.0	0.1	5	1	15
Disgust	69.51%	73.9%	71.34%	256	0.001	0.2	0.0	0.1	5	1	15
Fear	68.7%	74.11%	71.21%	256	0.001	0.2	0.0	0.1	5	1	15
Joy	68.35%	75.72%	71.71%	256	0.001	0.2	0.0	0.1	5	1	15
Sadness	69.47%	75.08%	**72.05%**	128	0.001	0.2	0.0	0.1	5	1	15
Surprise	68.306%	76.31%	72.024%	128	0.001	0.2	0.0	0.1	5	1	15
Trust	66.8%	**76.77%**	71.22%	300	0.001	0.2	0.0	0.1	5	1	15

Table 3.10: A sample of trials showing some of the effect results of step 1 experiments on the TroFi dataset with the SVM's probability distribution predictions. The hyperparametres used to reach these result are recorded.

necessarily get the metaphor prediction right but increased its correct prediction in general. The decrease in recall shows that the model missed predicting more metaphors. Overall, the classification model performance was affected negaively by combining these two sentiment forms in once experiments. The hyperparameters are in Table 3.12 for both *polarity* with *anticipation* and *polarity* with *fear*.

The result of the experiments conducted on the MOH-X dataset, when combining *polarity* and *disgust* in Table 3.12, differs from VUAMC. While the recall slightly decreased to 76.207% by 1.53%, the precision and F1 improved to 78.941% and 77.192% by 4.64% and 1.5%, respectively. The increase in precision indicates that the model has become more precise in predicting the correct instances. On the other hand, the decrease in recall suggests that the model missed more correct instances of metaphors. In other words, it failed to identify some positive instances that should have been detected. The overall result of the balanced F1 score shows the combined effect of precision and recall. It indicates

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC											
Gao et al., 2018 Reported	53.4%	65.6%	53.4%								
Baseline	*56.2%*	*51.1%*	*53.5%*	128	0.01	0.3	0.0	0.2	64	1	20
BERT emotion	52.51%	68.08%	59.297%	128	0.001	0.3	0.2	0.2	64	1	25
MOH-X											
Gao et al., 2018 Reported	75.3%	84.3%	79.1%								
Baseline	*75.44%*	*77.39%*	*76%*	300	0.001	0.2	0.0	0.2	10	1	30
BERT emotion	75.52%	80.49%	77.733%	300	0.001	0.2	0.2	0.2	10	1	15
TroFi											
Gao et al., 2018 Reported	69.661%	73.04%	71.088%								
Baseline	*69.661%*	*73.04%*	*71.088%*	300	0.001	0.2	0.0	0.1	10	1	15
BERT emotion	68.05%	75.21%	71.34%	256	0.001	0.2	0.0	0.1	5	1	15

Table 3.11: The BERT emotion with the classification model on the three metaphorical datasets and their hyperperemeters.

the model benefited from combining the sentiment features, *polarity* and *disgust* in the detection experiment on the MOH-X dataset specifically. This improvement in performance is in contrast to the overall results for the VUAMC experiments. The hyperparameters that produced these results are in Table 3.12.

In TroFi's experiment in Table 3.12, the combination of *polarity* and *anger* resulted in the following changes in performance metrics: the recall improved from 73.04% to 76.75% by 5.07%. This suggests that the model became better at correctly identifying metaphors out of all the actual correct instances and fewer correct instances were missed. The precision reduced by 3.78%, from 69.661% to 67.026%. This means that the model became less precise in predicting metaphors correctly out of all the correct instances. The decrease in precision suggests that the model generated more false positives (a metaphor predeicted as non-metaphor). Lastly, the F1 score experienced a slight increase from 71.088% to 71.3% by

0.2%. The F1 score considers both precision and recall and provides a balanced measure of a model's performance. This small increase in the F1 score suggests that the combination of *polarity* and *anger* in TroFi's experiment overall improved the model's ability to detect metaphors. Table 3.12 provides the hyperparameters used to acquire these results.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC											
Gao et al., 2018 Reported	53.4%	65.6%	53.4%								
Baseline	*56.2%*	*51.1%*	*53.5%*	128	0.01	0.3	0.0	0.2	64	1	20
Pol+anticipation	58.61%	43.27%	49.78%	300	0.01	0.3	0.0	0.2	64	1	20
Pol+fear	62.11%	27.65%	38.27%	300	0.01	0.3	0.0	0.2	64	1	20
MOH-X											
Gao et al., 2018 Reported	75.3%	84.3%	79.1%								
Baseline	*75.44%*	*77.393%*	*76%*	300	0.001	0.2	0.0	0.2	10	1	30
Pol+disgust	78.941%	76.207%	**77.192%**	128	0.001	0.2	0.0	0.2	10	1	15
TroFi											
Gao et al., 2018 Reported	69.661%	73.04%	71.088%								
Baseline	*69.661%*	*73.04%*	*71.088%*	300	0.001	0.2	0.0	0.1	10	1	15
Pol+anger	67.026%	76.75%	**71.3%**	128	0.001	0.2	0.0	0.1	5	1	15

Table 3.12: A sample of trials showing some of the effect results for this step on the three dataset and their hyperparameters.

Step 3

The experiments in step 1 and 2 involved explicit testing of each sentiment form, both individually and in combination, using the classification model on the three datasets. In this step, a different approach was taken. The experiments focused on categorising words as either *sentiment* or *non-sentiment*, regardless of the specific sentiment value. The BERT fine-tuned model was utilised for classification, where each token assigned one of the four emotions (*joy*, *sadness*, *fear*, and *anger*) was replaced by the label 1 to indicate that the word

evokes *sentiment*. Tokens assigned as *neutral* were labeled 0, indicating *non-sentiment*. For all three datasets, these new labels (1 or 0) were concatenated with the GloVe+ELM0's 1324-dimension vector, which provides additional word/context information. The results of these experiments and their hyperparameters are presented in Table 3.13.

Among the three datasets, VUAMC exhibited the worst performance in this step. Both recall and F1 score significantly decreased to 35.83% and 45.54%, respectively. The increase in precision, reaching 62.47%, indicates that the model correctly predicted more instances out of all instances predicted as correct. However, the decrease in recall suggests that the model missed more metaphors in this dataset.

When compared to the BERT experiment, the performance of MOH-X was reduced. However, compared to the baseline, the model showed some improvements. The F1 score slightly increased by 0.9% over the baseline and by 1.31% over the BERT experiments. Precision achieved the highest value, surpassing the baseline by 4.01% and BERT's by 3.90%. On the other hand, the recall was lower compared to the baseline by 1.89% and by 5.66% compared to BERT's.

In the case of VUAMC and MOH-X datasets, utilising BERT's explicit emotion assignment benefited the classification model more. This contrasts with the results obtained from the TroFi experiments, where recall and F1 score showed more significant improvements compared to the baseline and BERT experiments. Recall increased to 77.121% from the baseline's 73.04% (by 5.58%) and from BERT's 75.21% (by 2.54%). The increase in F1 score was slight, reaching 71.614% from the baseline's 71.088% and from BERT's 71.34% (by 0.73% and 0.38%, respectively). However, precision experienced a slight decrease to 67.165% from the baseline's 69.661% and from BERT's 68.05%. The sentiment/non-sentiment approach proved to be more successful for the TroFi dataset.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC											
Gao et al., 2018 Reported	53.4%	65.6%	53.4%								
Baseline	*56.2%*	*51.1%*	*53.5%*	128	0.01	0.3	0.0	0.2	64	1	20
BERT emotion	52.51%	**68.08%**	**59.297%**	300	0.01	0.3	0.0	0.2	64	1	20
Sent, non-sent	**62.47%**	35.83%	45.54%	300	0.01	0.3	0.0	0.2	64	1	20
MOH-X											
Gao et al., 2018 Reported	75.3%	84.3%	79.1%								
Baseline	*75.44%*	*77.393%*	*76%*	300	0.001	0.2	0.0	0.2	10	1	30
BERT emotion	75.52%	**80.49%**	**77.733%**	256	0.001	0.2	0.0	0.2	10	1	15
Sent, non-sent	**78.47%**	75.93%	76.708%	256	0.001	0.2	0.0	0.2	10	1	15
TroFi											
Gao et al., 2018 Reported	69.661%	73.04%	71.088%								
Baseline	*69.661%*	*73.04%*	*71.088%*	128	0.001	0.2	0.0	0.2	10	1	15
BERT emotion	68.05%	75.21%	71.34%	128	0.001	0.2	0.0	0.2	5	1	15
Sent, non-sent	67.165%	**77.121%**	**71.614%**	128	0.001	0.2	0.0	0.2	5	1	15

Table 3.13: The BERT emotion with the classification model as sentiment or non-sentiment testing. The table provide the hyperparameters.

3.5.5 Sequence Labelling

The chosen baseline model was built based on the approach presented in Mao et al. (2019). After multiple code executions with hyperparameter adjustments (as described in Chapter 1), the baseline model was selected for achieving the best results across all three evaluation metrics: precision, recall, and F1 score. In the sequence labeling model, words in a given sentence are classified as either 1 for metaphor or 0 for non-metaphor based on the overall context. The VUAMC dataset was divided into three sections: training, validation and testing, while MOH-X and TroFi datasets experiments were preformed using the 10-fold cross validation technique.

Step 1

For the sequence labeling experiments, Step 1 was tested. Two types of tests were conducted: the first test involved the explicit assignment of sentiment (*polarity* and *emotion*) for the NRC list using SVM. The second test involved fine-tuning BERT for emotion classification. In both cases, each token from the metaphorical dataset was assigned a label for *polarity* or *emotion* (single-digit and probability distribution), which was then concatenated with the GloVe+ELMo 1324-dimension vector to provide additional word/context information. The BERT assignment was also added to this vector.

Table 3.14 presents the results of the experiments done on the VUMAC dataset. NRC polarity and emotions all exhibited noticeable improvements in F1 and recall. The best recall was for *anticipation* emotion that reached 77.33% from the baseline's 64.45%, huge jump by 19.98%/. Lowest recall improvement reached 72.56% by 12.58% for joy emotion. Highest F1 is for anger, reached 74.42% increased by 4.49%. Lowest F1 was for polarity the reached 73.26, an increase of 2.86%. Similarly for BERT emotion, the recall and F1 increased over the baseline by 16.72% and 3.45%, respectively. The precision decreased from 79.58% to 72.205% (i.e. by 9.2%). The increase in recall indicates that the sequence model benefited greatly from adding these emotion to the vector representation. The model correctly identified more metaphor out of the correctly predicted, the model recognised more metaphors. BERT's prediction had a positive effect on the sequence labeling model. Similar to the NRC sentiment, the model's recall and F1 score improved, while the precision decreased. The increase in recall was not as substantial as with the NRC sentiment analysis, but it was still considered significant, reaching 69.1% from 64.45% (an increase of 7.21%). The F1 score also showed a slight increase, reaching 72.9% from 71.22% (a rise of 2.35%). However, the precision experienced a decline to 77.1%,(a decrease of 3.11%). Table 3.14 shows the hyperparameters used to produce these results.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC Single SVM											
Mao et al., 2019 Reported	71.8%	76.3%	74.0%								
Baseline	*79.58%*	*64.45%*	*71.22%*	256	0.2	0.5	0.0	0.2	2	1	29
NRC Polarity	72.76%	73.77%	**73.26%**	300	0.2	0.5	0.0	0.1	2	1	29
NRC Anger	74.36%	74.48%	**74.42%**	300	0.2	0.5	0.0	0.1	2	1	29
NRC Anticipation	71.21%	77.33%	74.15%	300	0.2	0.5	0.0	0.1	2	1	29
NRC Disgust	73.68%	74.24 %	73.96%	300	0.2	0.5	0.0	0.1	2	1	29
NRC Fear	71.66%	75.39%	73.48%	300	0.2	0.5	0.0	0.1	2	1	29
NRC Joy	74.78%	72.56%	73.65%	300	0.2	0.5	0.0	0.1	2	1	29
NRC Sadness	72.23%	74.49%	73.34%	300	0.2	0.5	0.0	0.1	2	1	29
NRC Surprise	73.801%	74.49%	74.14%	300	0.2	0.5	0.0	0.1	2	1	29
NRC Trust	72.56%	74.37%	73.45%	300	0.2	0.5	0.0	0.1	2	1	29
BERT emotion	77.1%	69.1%	**72.9%**	300	0.2	0.5	0.0	0.1	2	1	29

Table 3.14: The list of the successful sequence labelling experiments, their results and their hyper-parameters for the VUAMC dataset.

The addition of the probability distribution of the NRC sentiment to the VUAMC dataset had a positive effect, although it fell slightly behind the single-digit prediction. The increase in F1 score ranged from 2.21% for *anticipation* (72.8%) to 3.2% for *polarity* and *disgust* (73.5%). On the other hand, the increase in recall was more substantial, with the lowest increase being 7.83% for *anticipation* (69.5%) and the highest being 11.7% for *anger* (72%). However, the precision decreased in all experiments, with the lowest precision of 71.7% occurring in the experiment for *joy*, which experienced a decrease of 9.9%. The complete results and the corresponding hyperparameters are shown in Table 3.15.

The use of the single-digit prediction of the NRC sentiment on the MOH-X dataset led to an increase in F1 scores for all cases in the sequence labelling task. The highest increase was observed for *anticipation*, which reached 80.311% from the starting baseline of 78.46% (an increase of 2.35%). The smallest increase was for *surprise*, with only a 0.2% improvement. Regarding recall, there was a slight increase for *anger* by 0.19% and *disgust*

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC Prob. SVM											
Mao et al., 2019 Reported	79.7%	79.8%	79.8%								
Baseline	*79.58%*	*64.45%*	*71.22%*	256	0.2	0.5	0.0	0.2	2	1	29
Polarity	76.9%	70.4%	**73.5%**	300	0.2	0.5	0.0	0.1	2	1	29
Anger	75%	**72%**	73.4%	300	0.2	0.5	0.0	0.1	2	1	29
Anticipation	76.4%	69.5%	72.8%	300	0.2	0.5	0.0	0.1	2	1	29
Disgust	76.8%	70.4%	**73.5%**	300	0.2	0.5	0.0	0.1	2	1	29
Fear	76.4%	70.3%	73.2%	300	0.2	0.5	0.0	0.1	2	1	29
Joy	71.7%	71.3%	73.3%	300	0.2	0.5	0.0	0.1	2	1	29
Sadness	76.9%	70%	73.3%	300	0.2	0.5	0.0	0.1	2	1	29
Surprise	77.4%	71.5%	73.3%	300	0.2	0.5	0.0	0.1	2	1	29
Trust	77.8%	68.8%	73%	300	0.2	0.5	0.0	0.1	2	1	29

Table 3.15: The list of the successful sequence labelling experiments, their results and hyperparameters for the VUAMC dataset using the probability distribution.

by 0.24%, while the other classes experienced a decrease, with polarity having the lowest recall decrease of 2.95%. In contrast, the precision generally showed an increase from the baseline value of 76.408%, with the highest precision achieved for *anger* at 80.66%. Overall, using the NRC sentiment as a single-digit prediction had a positive impact on the sequence labeling model's detection task. The increased recall for *anger* and *disgust* suggests that the model recognised more metaphorical words. For detailed results and their hyperparameters are in Table 3.16.

The sequence labeling experiments on the MOH-X dataset using the probability distribution prediction showed a similar impact to the single-digit prediction. The overall F1 score increased from the starting value of 78.46%. *Fear* had the highest increase, reaching 80.1% with a rise of 2.09%, followed by *polarity* with 80.05% and an increase of 2.02%. The lowest improvement was observed for *trust*, with a 1.07% increase, reaching 79.3%. However, the recall decreased slightly for all NRC sentiment tests from the starting value of 81.63%. The smallest decline was for *disgust*, with only a 0.07% decrease (reached 81.57%), while

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
MOH-X Single SVM											
Mao et al., 2019 Reported	79.7%	79.8%	79.8%								
Baseline	*76.408%*	*81.63%*	*78.46%*	256	0.001	0.5	0.0	0.2	2	1	20
Polarity	79.315%	79.22%	79.84%	256	0.001	0.5	0.0	0.1	2	2	20
Anger	80.66%	81.79%	80.83%	300	0.001	0.5	0.0	0.2	2	3	20
Anticipation	78.102%	83.29%	**80.311%**	256	0.001	0.5	0.0	0.1	2	3	20
Disgust	78.62%	81.83%	80.016%	300	0.001	0.5	0.0	0.2	2	3	20
Fear	78.576%	80.92%	79.62%	256	0.001	0.5	0.0	0.1	2	1	20
Joy	78.56%	81.25%	79.69%	300	0.001	0.5	0.0	0.1	2	1	20
Sadness	77.798%	80.728%	78.813%	128	0.001	0.5	0.0	0.1	2	1	20
Surprise	78.015%	79.76%	78.65%	128	0.001	0.5	0.0	0.1	2	1	20
Trust	78.54%	80.33%	79.17%	300	0.001	0.5	0.0	0.1	2	1	20

Table 3.16: A sample of trials showing some of the effect results of step 1 experiments on the MOH-X dataset with the single-digit SVM prediction.

the highest decline was for *sadness* with a 3.05% decrease, reaching 79.14%. On the other hand, the precision increased overall, reaching 80.56% from the starting 76.408%, indicating an increase of 5.4% for *sadness*. The increase in precision suggests that the model recognised more positive instances (both metaphorical and non-metaphorical) out of all positive instances. However, it does not necessarily mean that it performed better at recognising metaphors specifically. For detailed results and the hyperparameters are in Table 3.17.

The results of experimenting with the NRC sentiment on the TroFi dataset using the sequence labeling model were promising. The increase in recall and F1 scores (except for joy) indicated that the model benefited from the additional information, enabling it to recognise more metaphors. The highest increase in F1 score was observed for *surprise*, with a rise of 1.46%, reaching 72.62% from the starting baseline of 71.575%. However, *joy*'s F1 score slightly decreased by 0.4%, though its recall had the highest increase of 8.33%, reaching 80.699% from the starting baseline of 74.489%. The lowest increase in recall was for *disgust*, with a 0.5% improvement. For precision, *joy*'s precision decreased to 65.433% from 68.98%,

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
MOH-X Prob. SVM											
Mao et al., 2019 Reported	79.7%	79.8%	79.8%								
Baseline	*76.408%*	*81.63%*	*78.46%*	256	0.001	0.5	0.0	0.2	2	1	20
Polarity	80.53%	80.5%	80.05%	128	0.001	0.5	0.0	0.1	2	1	20
Anger	78.57%	80.92%	79.62	256	0.001	0.5	0.0	0.1	2	1	20
Anticipation	79.76%	79.24%	79.34%	256	0.001	0.5	0.0	0.1	2	1	20
Disgust	78.54%	81.57%	79.79%	256	0.001	0.5	0.0	0.1	2	3	20
Fear	80.48%	79.97%	**80.1%**	256	0.001	0.5	0.0	0.1	2	1	20
Joy	78.86%	80.38%	79.47%	300	0.001	0.5	0.0	0.1	2	1	20
Sadness	**80.56%**	79.14%	79.56%	256	0.001	0.5	0.0	0.1	2	1	20
Surprise	78.146%	81.56%	79.42%	256	0.001	0.5	0.0	0.1	2	1	20
Trust	78.57%	80.71%	79.3%	256	0.001	0.5	0.0	0.1	2	1	20

Table 3.17: A sample of trials displaying some of the effect results for the NRC sentiment experiments on the MOH-X dataset with SVM's probability distribution predictions and their hyperparameters.

a decline of 5.14%. On the other hand, the highest increase in precision was observed for *anticipation*, with a rise of 2.03%, reaching 70.382% from 68.98%. The overall results and their hyperparameters are presented in Table 3.18.

Similar to the previous experiments, the F1 scores generally increased over the baseline of 71.575%. *Surprise* had the highest increase in F1, with a rise of 1.3%, reaching 72.515%, while the lowest increase was for *disgust*, which increased by 0.2% to reach 71.763%. The recall increased for all sentiment classes except *polarity* and *fear*, which experienced slight declines of 0.9% and 0.3%, respectively. The highest increase in recall was observed for *surprise*, with a rise of 4.49%, reaching 77.834% from the starting baseline of 74.489%. Regarding precision, the results increased over the baseline of 68.98%, with *polarity* showing the highest increase of 4.87% to reach 70.855%. However, precision decreased for *anticipation* (by 1.59% to 67.879%), *disgust* (by 0.54% to 68.605%), *sadness* (by 0.21% to 68.829%), and *trust* (by 0.62% to 68.551%). Overall, the impact of adding the probability distribution with

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
TroFi Single SVM											
Mao et al., 2019 Reported	67.4%	77.8%	72.2%								
Baseline	*68.98%*	*74.489%*	*71.575%*	256	0.001	0.5	0.0	0.1	2	1	20
Polarity	66.582%	79.914%	72.521%	128	0.0001	0.5	0.0	0.1	2	2	20
Anger	69.817%	75.408%	72.365%	128	0.001	0.5	0.0	0.1	2	1	20
Anticipation	**70.382%**	75.061%	72.611%	128	0.001	0.5	0.0	0.1	2	1	20
Disgust	70.342%	74.872%	72.433%	300	0.001	0.5	0.0	0.1	2	1	20
Fear	69.804%	75.537%	72.478%	300	0.001	0.5	0.0	0.1	2	1	20
Joy	65.433%	**80.699%**	71.261%	128	0.001	0.5	0.0	0.1	2	2	20
Sadness	70.231%	75.1404%	72.561%	128	0.001	0.5	0.0	0.1	2	1	20
Surprise	70.358%	75.12%	**72.62%**	128	0.001	0.5	0.0	0.1	2	1	20
Trust	70.27%	74.937%	72.492%	128	0.001	0.5	0.0	0.1	2	1	20

Table 3.18: A sample of trials showing some of the effect results from experiments with NRC sentiment on the TroFi dataset with the single-digit predictions and the hyperparameters.

the sequence labeling model when tested on the TroFi dataset is positive, as evidenced by the improvements in F1 scores and recall for most sentiment cases. Table 3.19 provide the complete set of results and Table **??** provides the hyperparameters.

The experiments involving the fine-tuned BERT model for emotion exhibited similar effects as the NRC sentiment model on the sequence labeling task. The recall and F1 scores for the VUAMC dataset increased over the baseline, reaching 69.1% (an increase of 7.2%) and 72.9% (an increase of 2.35%) respectively. However, the precision decreased to 77.1% by 3.1%. For the MOH-X dataset, not only did the recall and F1 scores improve, but the precision also increased. The recall showed an increase of 1.7%, the F1 score improved by 1.2%, and the precision increased by 0.8%. In contrast, the recall for TroFi decreased from 74.489% to 73.724%, showing a decline of 1.02%. The precision increased to 71.322% by 3.3%, and the F1 score also improved to 72.389% by 1.13%. Table 3.20 shows the reported results and their hyperparameters.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
TroFi Prob. SVM											
Mao et al., 2019 Reported	67.4%	77.8%	72.2%								
Baseline	*68.98%*	*74.489%*	*71.575%*	256	0.001	0.5	0.0	0.1	2	1	20
Polarity	**70.855%**	73.752%	72.163%	300	0.001	0.5	0.0	0.1	2	1	20
Anger	69.632%	75.811%	72.505%	128	0.001	0.5	0.0	0.1	2	1	20
Anticipation	67.879%	77.832%	72.496%	128	0.001	0.5	0.0	0.1	2	1	20
Disgust	68.605%	75.606%	71.763%	300	0.001	0.5	0.0	0.1	2	1	20
Fear	70.166%	74.263%	72.09%	300	0.001	0.5	0.0	0.1	2	1	20
Joy	69.146%	75.766%	72.204%	64	0.001	0.5	0.0	0.1	2	1	20
Sadness	68.829%	75.7865%	71.966%	128	0.001	0.5	0.0	0.1	2	1	20
Surprise	67.916%	**77.834%**	**72.515%**	128	0.001	0.5	0.0	0.1	2	1	20
Trust	68.551%	76.979%	72.478%	128	0.001	0.5	0.0	0.1	2	1	20

Table 3.19: A sample of trials showing some of the effect results from experiments with NRC sentiment on the TroFi dataset with the probability distribution predictions. The hyperparameters used to reach these results are recorded.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC											
Mao et al., 2019 Reported	79.7%	79.8%	79.8%								
Baseline	*79.58%*	*64.45%*	*71.22%*	256	0.2	0.5	0.0	0.2	2	1	29
BERT emotion	77.1%	69.1%	72.9%	128	0.2	0.5	0.0	0.2	2	1	29
MOH-X											
Mao et al., 2019 Reported	79.7%	79.8%	79.8%								
Baseline	*76.408%*	*81.63%*	*78.46%*	256	0.001	0.5	0.0	0.2	2	1	20
BERT emotion	77.079%	83.073%	79.477%	300	0.001	0.5	0.0	0.2	2	1	20
TroFi											
Mao et al., 2019 Reported	67.4%	77.8%	72.2%								
Baseline	*68.98%*	*74.489%*	*71.575%*	256	0.001	0.5	0.0	0.1	2	1	20
BERT emotion	71.322%	73.724%	72.389%	300	0.001	0.5	0.0	0.1	2	1	20

Table 3.20: The BERT emotion with the sequence labelling model on the three datasets with their hyperparameters.

Step 3

Building on the results of Step 1 for BERT emotion analysis, the experiments extended the concept discussed in Step 3 of the classification experiment. In this experiment, the emotion labels were replaced by 1 for sentiment words and 0 for non-sentiment words. Each prediction for the four emotions *joy*, *sadness*, *fear*, and *anger* was considered a sentiment-evoking word, while the prediction for the *neutral* emotion was considered a non-sentiment word. This approach allowed the model to be tested and evaluated on its ability to distinguish between sentiment and non-sentiment words. The tests were done on the three dataset. The VUAMC dataset was divided into three files (train, val, and test) while the other two, MOH-X and TroFi, were tested using a 10-fold cross validation. Table 3.21 show the result of this step.

Based on the results of the experiments, the effect observed on the three datasets when using sentiment words (1) and non-sentiment words (0) as labels was similar. The recall for the VUAMC slightly increased compared to the baseline, reaching 66.81%. However, this value was lower than the recall achieved by the BERT emotion model. The recall for MOH-X and TroFi recall was higher, reaching 83.091% and 75.42%, respectively. The F1 score for MOH-X increased over the baseline by 1.5% and over the BERT emotion model by 0.29%, reaching 79.71%. The VUAMC F1 score was higher than the baseline by 1.72% but lower than the BERT emotion model by 0.6%. Lastly, TroFi's F1 score decreased by 0.41% from the baseline and by 1.5% from the BERT emotion model.The precision for VUAMC and MOH-X increased compared to the baseline, while TroFi's decreased to 67.91% from 68.98%. It appears that the BERT emotion fine-tuned model's predictions using the sequence labelling approach produced better results overall, considering the various evaluation metrics and datasets used in the experiments. The hyperparameters are shown in Table 3.21.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC											
Mao et al., 2019 Reported	79.7%	79.8%	79.8%								
Baseline	*79.58%*	*64.45%*	*71.22%*	256	0.2	0.5	0.0	0.2	2	1	29
BERT emotion	77.1%	**69.1%**	**72.9%**	300	0.2	0.5	0.0	0.2	2	1	29
BERT sent, non-sent	79.13%	66.81%	72.45%	300	0.2	0.5	0.0	0.2	2	1	29
MOH-X											
Mao et al., 2019 Reported	79.7%	79.8%	79.8%								
Baseline	*76.408%*	*81.63%*	*78.46%*	256	0.001	0.5	0.0	0.2	2	1	20
BERT emotion	77.079%	83.073%	79.477%	300	0.001	0.5	0.0	0.2	2	1	20
BERT sent, non-sent	**77.086%**	**83.091%**	**79.71%**	300	0.001	0.5	0.0	0.2	2	1	20
TroFi											
Mao et al., 2019 Reported	67.4%	77.8%	72.2%								
Baseline	*68.98%*	*74.489%*	*71.575%*	256	0.001	0.5	0.0	0.1	2	1	20
BERT emotion	**71.322%**	73.724%	**72.389%**	300	0.001	0.5	0.0	0.1	2	1	20
BERT sent, non-sent	67.918%	**75.42%**	71.28%	300	0.001	0.5	0.0	0.1	2	1	20

Table 3.21: The BERT emotion with the sequence labelling model testing as sentiment and non-sentiment. The hyperparameters are recorded.

3.5.6 Discussion

The experiments in this chapter aimed to assess the effectiveness of incorporating sentiment information into Deep Learning models to enhance the automated metaphor detection. Sentiment was added as a main lexical feature to the vector representation of GloVe (capturing word meaning) and ELMo (capturing the context). These tests were conducted on three distinct metaphorical datasets: VUAMC, MOH-X, and TroFi (Gao et al., 2018; Mao et al., 2019).

The two models were built upon the proposed models of Gao et al. (2018) for classification and Mao et al. (2019) for sequence labeling (RNN_HG), respectively. In order to establish baselines for comparison, the provided code was run with further hyperparameter

adjustments. However, as mentioned in Chapter 1, Section 1.8.1, it was noted that the results obtained from Gao et al. (2018)'s model may vary from their reported results due to factors such as hyperparameter settings, early stopping, and other training adjustments, which were not explicitly specified. To address this uncertainty and ensure fair comparisons, the build-on models were selected, and the new tests were then compared against each other to evaluate the impact of adding sentiment information on the automated metaphor detection task. This approach allowed for a more reliable and controlled comparison between the different models and their performances.

The classification and sequence labelling models tested on the VUAMC dataset showed promising results across the three steps. In the tests where each NRC sentiment form was evaluated separately (*polarity* and *emotion*), the classification model demonstrated the most significant improvement when incorporating *anticipation* and *fear* sentiments. These sentiments yielded the best recall and F1 scores.

The VUAMC dataset for the classification model contained 23,113 sentences. After removing duplicate sentences, the final number was 10,489 sentences. The classification model encountered 634 SVM predicted words for *anticipation* across 7,566 non-metaphorical sentences and 334 *anticipation* predicted words across 2,923 metaphorical sentences. On the other hand, the model encountered 1735 *fear* predicted words across 7,566 non-metaphorical sentences and 987 predicted *fear* words across 2,923 metaphorical sentences.

For the sequence labelling model, the VUAMC dataset had a total of 945 instances of *anticipation* words out of 18,657 filtered tokens. Out of these, 28 tokens were considered metaphorical, while 114 tokens were considered literal. The model showed the highest increase in recall compared to all other NRC sentiments, while the *anger* emotion had the highest increase in F1. *Anger* predicted tokens were in total 1732, seen 81 times as metaphor and 228 times as non-metaphor.

In the probability distribution predictions with the classification model tested on the VUAMC dataset, it was observed that some sentiments had a negative effect on the metaphor detection task. The F1 results for *anger* and trust emotions were the worst, while *joy* emotion had the highest F1 score, with a slight increase over the baseline (by 2.7%). For the *joy* emotion, the number of predicted words was 571 tokens across 2,923 metaphorical sentences and 1,235 tokens across 7,566 non-metaphorical sentences. However, despite the increase in F1 score, the overall improvement for joy emotion was not substantial.

On the other hand, when using the sequence labelling model with probability distribution predictions, there was an increase in performance compared to the classification model. The F1 score for the *disgust* emotion was the best out of all the NRC sentiments. The model saw *disgust* 36 tokens as metaphors and 120 tokens as non-metaphors.

The slight improvements in F1 scores for the single-digit and probability distribution predictions on both the MOH-X and TroFi datasets, using both the classification and sequence labelling models, indicate that the models benefited from both predicted forms applied to these datasets, in contrast to the VUAMC dataset.

Disgust emotion provided the best F1 score for the MOH-X dataset when using the single-digit prediction. The model tested 33 disgust-predicted words in metaphor-labeled sentences and 25 times in non-metaphor sentences across the total 647 sentences. For the probability distribution prediction on the MOH-X dataset, *trust* provided the best result. The model tested 25 *trust*-predicted tokens in metaphorical sentences and 17 tokens in literal sentences.

In contrast, for the TroFi dataset, the single-digit prediction's best result was for *anger* emotion. The TroFi dataset had a larger number of total tokens compared to the MOH-X dataset. The classification model encountered 1,181 *anger*-predicted words in metaphorical sentences and 1,601 times in non-metaphor sentences. As for the probability distribution

prediction on the TroFi dataset, *sadness* achieved the best result. The model encountered 560 *sadness*-predicted tokens in metaphor-labeled sentences and 861 tokens in non-metaphor-labeled sentences.

The MOH-X dataset showed the best F1 score when testing the single-digit prediction for the sequence labelling model, with the *anticipation* emotion providing the highest performance. Out of the 647 sentences, the model encountered *anticipation* tokens 17 times in metaphorical sentences and 10 times in non-metaphor sentences. When using the probability distribution prediction on the MOH-X dataset, *fear* emotion yielded the best result. The model encountered 78 *fear*-predicted tokens in metaphor sentences and 55 tokens in non-metaphor labeled sentences.

For the TroFi dataset, both forms (single-digit and probability distribution) produced the best result with the *surprise* emotion. Out of the total 3,737 sentences, *surprise*-predicted words appeared 200 times in sentences labeled as metaphor and 252 times labeled as non-metaphor.

The difference in approaches between using SVM and fine-tuned BERT models to obtain emotion predictions is clearly evident in their respective results. While the SVM employs unigrams as the training input, represented by pre-trained BERT vectors, these vectors are isolated and do not consider any contextual information. Consequently, the meaning represented remains basic and uniform for all words, regardless of potential encounters with polysemy. In other words, regardless of a word having different sense in various contexts, the basic meaning in any context remains consistent. This uniformity leads to the same prediction value, which may not be ideal in situations where a word's true sentiment is closely connected to its context. For instance, the word *kiss* was predicted by the NRC's SVM as having *anticipation*, *joy*, and *surprise* emotions, even though in the phrase "*the last goodbye kiss*" it may be more appropriately considered as evoking a sentiment of *sadness*.

The fine-tuned BERT model, on the other hand, was trained on sentences where the emotion is assigned to the entire context. Despite producing results for unigrams, it demonstrated better performance by assigning interesting sentiments to certain words that have inherent sentiment in one's mind, such as the words *kiss* and *love*, which were both assigned a *neutral* emotion. Interestingly, when BERT tested the sentence "*the last goodbye kiss*", it assigned a *sadness* emotion to the whole sentence, whereas each individual word *the last*, *goodbye*, and *kiss* was assigned a *neutral* emotion. Similarly, BERT's emotion model assigned *fear* to the sentence "*tough love*", while each word was again assigned a *neutral* emotion.

This contrast illustrates that the SVM approach lacks contextual understanding, limiting its ability to capture the nuanced and contextual sentiment associated with words. Consequently, this highlights the advantages of using fine-tuned BERT models for more accurate emotion predictions. This observation was further supported by the experimenters in step 3, where they replaced the labels for binary testing (1 for *sentiment* and 0 for *non-sentiment*). Specifying the emotion generally yielded better results.

It is important to note that the rate of increase between the two models is different, with the classification model showing more prominent improvements. This difference in performance could be related to the detection structure: the classification model tests sentences as a whole, while the sequence labelling model targets individual words for prediction.

Furthermore, the gap in the number of non-metaphorical and metaphorical sentences/words in the datasets could pose a challenge for the models during training and testing. Having more instances where words are considered non-metaphorical than instances where they are considered metaphorical could bias the models to predict non-metaphorical words more frequently. This imbalance issue should be carefully addressed to ensure a balanced and accurate model performance in automated metaphor detection.

These findings illustrate how the different emotions impacted the models' performances on each dataset and emphasise the potential benefits of incorporating sentiment information into the classification and sequence labelling model for automated metaphor detection. In addition to the quality of the added information, the role of the machine learning model responsible for assigning predictions to words is crucial, as evident when comparing the SVM to the BERT fine-tune model for multi-classification of sentiment.

The SVM utilised BERT pre-trained vectors for individual words, in isolation from their context, resulting in the same vector representation for polysemous words. As a consequence, the SVM's sentiment predictions for these words were the same. As opose to the BERT fine-tune model, which was trained on sentences, allowing the model to learn the emotions of words based on the context of the sentence. Although the predictions were assigned to individual words, the BERT model was more likely to assign the most appropriate emotion, as demonstrated by the experimental results across all datasets. This contextual understanding of the BERT model contributed to its improved performance in sentiment prediction compared to the SVM method.

The analysis of the correct metaphor prediction and incorrect metaphor prediction files provided more insight into the effect of sentiment information on the deep learning models used for metaphor detection. As mentioned earlier, the emotion *anticipation* improved the model's performance on the VUAMC dataset. The incorrect prediction files revealed that the word *time* was among the most frequently misclassified as a metaphor, occurring 20 times. The word *time* also appeared 44 times in the correct prediction file. The number of non-metaphorical sentences containing the word *time* was 145, while the number of metaphorical sentences that contained the word was 85. This difference indicates that the model observed the word *time* more often in literal senses than in metaphorical contexts, potentially contributing to the incorrect predictions. Furthermore, the sentences themselves predominantly used *time* in a conventional manner, such as in the example, "*This* ***time***

the pause was almost eloquent, then he gave a richly contagious laugh.". The GloVe word representations provided the basic meaning of words, but it seems that ELMo's contextual representations did not offer sufficient information to distinguish the intended sense of the word in context.

The word *poison* appeared in the MOH-X incorrect predicted file when the emotion *disgust* was tested. The word *poison* was one of the predicted words for *disgust* by the SVM. It was incorrectly predicted as non-metaphor in the sentence "*poison someone's mind.*" Although the word *poison* could be understood in the context of the *disgust* emotion, the word/context information supplied by GloVe and ELMo was not sufficient. ELMo's contextual vectors could not distinctly indicate whether *poison* was metaphorical or literal due to the length of the sentence. In addition to GloVe's vectors providing the basic meaning, all may have contributed to the incorrect prediction. On the other hand, the word *abuse* was correctly predicted as a metaphor in the sentence "*Her husband often abuses alcohol.*" The word *abuses* was assigned the *disgust* label and is compatible with the overall meaning of the sentence, which in turn supports the metaphor prediction. Here, *abuses*' does not refer to *physical harm* but rather to *excessive use to the degree of being harmful.*

The word *pump* was among the correctly predicted words in the TroFi dataset when testing for the emotion *trust.* Although the word *pump* itself was not assigned with *trust* by the SVM, the word *experience* was predicted to evoke the *trust* emotion in the sentence "*The Australian experience might be of interest to the U.S., which just agreed to pump nearly $1 billion of development aid into the Philippines.*" In this context, the word *pump* is used metaphorically to mean *to allocate*, *disburse*, or *invest*, rather than its literal meaning of *to move or transfer fluids*, *gases*, or *substances.* The correct prediction indicates that the model was able to understand the metaphorical use of the word *pump* in this particular sentence, showing the model's capability to distinguish between different senses of words if sufficient contextual information were provided.

The word *drag* was among the misclassified words in TroFi's incorrect prediction file. In the sentence "*As a result, gold and silver were dragged down more by platinum than by crude oil.*" the phrase *dragged down* meant a *decrease in value* rather than *the physical movement of pulling something downwards with force*. The sentence's intended meaning is that the "*decline in gold and silver prices was more influenced by platinum prices than by crude oil prices*", and the phrasal verb *dragged down* was used to indicate *the severity of the price declines*. The metaphor annotations were assigned partially for the word *drag* only. Similarly, the word vector was assigned to this part only, disregarding the other sense of the complete phrase *drag down*. Interestingly, when tested in isolation, the BERT model for emotion assigned the phrase *drag down* as *neutral*, but when seen in the overall context of the sentence, the model assigned the emotion *anger* to the entire sentence. This demonstrates the importance of providing enough context for effective testing.

The experiments involving the semantic richness element, sentiment information, provided interesting results for both classification and sequence labelling models. Comparing the sentiment predictions provided by the SVM and fine-tuned BERT models offered further insight into the nature of sentiment and how deep learning models interacted with these predictions during the metaphor detection task. The experiments in this chapter, used two models, both utilised GloVe and ELMo for word/context information. The analysis of the results provided answers for the research questions.

RQ1 was evaluated using multiple sentiment resources, including *polarity* and *emotion* from human-based studies and deep learning models, such as classification, sequence labelling. Both SVM and BERT model predictions were examined. Notably, the predictions from the BERT model contributed to the best F1 score among all the experiments. These findings demonstrate that deep learning models can effectively recognise and utilise sentiment in metaphor detection, thereby addressing RQ2. Furthermore, the discussion section revealed that metaphorical sentences carry emotions (addressing both RQ1 and RQ3). The

use of metaphor goes beyond the words; it conveys a deeper message. Metaphors allow individuals to express their feelings and intentions beyond straightforward information. For instance, the metaphorical use of **JOURNEY** in the context of a cancer patient and healthcare provider conveys *struggle, sorrow*, and *hope*. Similarly, *WAR*-related phrases are used by healthcare providers to express their *support, offense*, and *defense* against the illness.

However, several observations were made regarding the hindrances to effectively utilising sentiment information and using the full potential of metaphorical neural network models for metaphor detection. These limitations and further analyses on how to address them are extensively discussed in the conclusion, Chapter 7.

3.6 Conclusion

This chapter aimed to enhance the automated metaphor detection task by utilising deep learning models. The focus was on testing the impact of incorporating one perceptual semantic richness element, the sentiment. Two metaphor detection models were employed for this purpose: the classification model, which the was the baseline for sentence-level experiments (Gao et al., 2018), and the sequence labeling model for word-level detection, which the chosen baseline was based on Mao et al. (2019). The tests involved incorporating sentiment predictions from two sources: SVMs trained on unigrams and a fine-tuned BERT model trained on sentences. These sentiment predictions were then added to the GloVe+ELMo vector for word and context representation. These were evaluated on three datasets: VUAMC, MOH-X, and TroFi (Gao et al., 2018; Mao et al., 2019).

In the next chapter, imageability and sensory experience will be experimented with as features for metaphor detection. These two elements are very similar in their definition, and as a result, they will be compared in a single chapter. By investigating these two features,

the study aims to gain insights into how perceptual and sensory aspects affect neural network models for metaphor identification.

Chapter Four

Sensory Experience and Imageability

4.1 Introduction

The previous chapter, Chapter 3, covered sentiment as a part of the perceptual aspect of semantic richness. The elements were tested with neural networks for the metaphor detection task. The objective is to evaluate the impact of enriching the contextual information by adding more lexical and linguistic features to the word/context representation used in the detection experiments. Thus, two forms of sentiment were tested: *polarity* and *emotions*. The model's results show improvement on the three metaphorical datasets.

This chapter addresses the third and fourth elements due to their similar definitions. The fundamental idea behind imageability and sensory experience is that certain words evoke an image scheme, along with sensory experiences drawn from one's own encounters. Consequently, the experiments conducted in this chapter assess the impact of imageability and sensory experiences on detecting metaphor utilising neural network models. These experiments are carried out on three metaphorical datasets, employing two types of deep learning models: sentence classification and word classification. The chapter encompasses a discussion of relevant literature and the underlying theories guiding the execution of these

metaphor tests. The findings are analysed, and limitations are identified. Detailed solutions are provided in the Conclusions chapter, Chapter 7.

4.2 What are Sensory Experience and Imageability?

Certain words possess the capacity to evoke images in the mind upon being read or heard (referred to as imageability). When these words evoke additional sensations, they are described as having a certain degree of sensory experience. Hence, sensory experience pertains to the extent to which a word can elicit sensory and/or perceptual encounters within the mind (Juhasz and Yap, 2013). For instance, upon encountering the word *incense*, the mind creates a mental image associated with the concept of *incense*; furthermore, there are instances when the *fragrance* of *incense* can even be *smelled*.

4.3 Imageability, Sensory Experience and Metaphor Theories

Metaphors are a form of figurative language that is integrated in the day-to-day written/spoken communication. The metaphor meaning is not a direct translation of the literal meaning of its parts, but requires an additional information of the context. Therefore, it is a challenging task for the field of NLP to automatically detects metaphor in natural text.

For instance, the phrase "*weigh my options*" refers to analysing the advantages and disadvantages of a decision that is being made. The **WEIGHING** is not a literal scaling, but meant as **CONSIDERATION**. The words from the **WEIGHING** is the source domain that its words is being used to describe the target domain, **CONSIDERATION**. This

mapping is called the *The Conceptual Metaphor Theory* (CMT) (Lakoff and Johnson, 1980) discussed in Section 1.2.

This theory was further developed by Lakoff et al. (1999), proposing that the bodily interactions with the environment serve as the basis for projecting new conceptual notions in these metaphors. For example, consider the metaphor "*he is hungry for recognition*", in which **DESIRE** is metaphorically linked with **HUNGER** due to the association of *food* with *desire*. This physiological response is then extended to the abstract concept of *recognition*. The theory strives to explian how conceptual metaphors are comprehended, particularly in situations where there might be a lack of direct experiential connection between the *source* and *target* domains, resulting in some mappings being less effective.

Many metaphor theories attempt to outline the attributes of both the *source* and *target* domains (Lima, 2006; Velasco, 2001). Significantly, the *source* domain (i.e., the domain from which words are drawn) incorporates an image schema derived from bodily interactions. To illustrate, the **WEIGHING** domain encompasses activities such as *lifting/carrying objects* and *sensing the weightiness of an object*, among others. These tangible experiences serve as mental images that are employed to interpret and make sense of the *target* domain, **CONSIDERATION**.

From the discussion above, it could be inferred that some metaphors are constructed by building upon image-schema knowledge inherent to the sensory interaction system. In simpler terms, metaphors interpret abstract concepts by conjuring mental images of bodily interactions familiar to humans. As a result, the hypothesis of this chapter states that metaphors are inclined to employ words characterised by higher imageability scores, and these scores surpass the scores of their contextual surroundings. Furthermore, since comprehending metaphors entails a prior engagement with the projected concept, a certain level of human sensory input might be linked to this process. With this in mind, imageability and

sensory experience ratings are incorporated into the words within sentences to investigate whether these ratings could enhance the capability of neural networks to discern metaphors. These theories are tested on three different metaphorical datasets: the VUAMC, MOH-X and TroFi.

4.4 Related Work

Imageability has been a notable feature incorporated into numerous research studies related to metaphors and other NLP fields. For instance, Tsvetkov et al. (2014) conducted an experiment to assess whether metaphors are more conceptual than lexical. They trained a model in one language using lexical semantic features and effectively identified subject-verb-object (SVO) and adjective-noun (AN) metaphors in other languages. In a different approach, Rai et al. (2016) combined imageability with other conceptual, syntactic, and word embedding features to perform supervised classification for detecting metaphors in diverse contexts. Taking an even more comprehensive perspective, Shutova et al. (2016) constructed a metaphor identification system that employed meaning representations derived from a combination of linguistic and visual data. They employed a skip-gram model to learn linguistic meanings, while a convolutional network captured visual representations. Additionally, Parde and Nielsen (2018) utilised neural networks trained on imageability and other relevant attributes to identify novel metaphors within adjective-noun (AN) pairs. Overall, these studies underline the significance of imageability in enhancing the detection and understanding of metaphors, providing insights into their conceptual nature and cross-linguistic applicability.

Language provides a mutual platform for communication by utilising sensory inputs, such as *taste*, *sight*, *sound*, and others shared information derived from the environment.

This concept is valid for both non-figurative and figurative languages (Tekiroğlu et al., 2015). They attempt to measure the influence of these sensory elements on metaphor identification by analysing a dependency-parsed corpus of adjective-noun (AN) pairs. In a separate study, (Wan et al., 2020) explored the use of conceptual norms as a linguistic augmentation technique for detecting metaphors in VUAMC verbs. However, little have been done to explore the effect of the sensory experience with neural networks for automated metaphor detection. The paper Alnafesah et al. (2023) covers this part of the experiments.

While concreteness and imageability are closely related, there exist some distinctions between them. Certain concrete words may be challenging to visualise (e.g., the term *structure* with a concreteness score of 3 and an imageability score of 2). On the other hand, certain abstract words can evoke mental imagery, as seen with *frightfully*, which has an imageability score of 3 but a concreteness score of 1. This phenomenon is often attributed to the emotions or sensations these words evoke. In light of these instances and the hypothesis stated earlier, numerous experiments focused on metaphor detection have integrated imageability into their external features. This is interestingly aligned with the observations made regarding sensory experience scores as well. To take the word *nightmare* as an example: it possesses a high imageability score of 3, a low concreteness score of 2, and a relatively elevated sensory experience score of 5. These investigations highlight the significance of not only concreteness and imageability but also sensory experience in enhancing the understanding and identification of metaphors, further validating the premise that metaphors often rely on these cognitive dimensions.

Given the objective of enhancing automated metaphor detection, it becomes imperative to understand and grasp the complex and detailed aspect of language. These features, if considered, could potentially contribute to a more comprehensive understanding of metaphors and aid in refining the performance of automated detection models.

4.5 Metaphor Detection Experiment

This section offers in-depth explanations of experiments that utilise imageability and sensory experience as the main lexical linguistic features within deep learning models for the task of metaphor detection. The objective is to enhance these features by incorporating word/context representations, enriching their semantic information to improve the performance of the deep learning models. The models used in these tests are baselines build on the Gao et al. (2018) for sentence classification and on Mao et al. (2019) the RNN_HG sequence model for word classification. These experiments are performed on the three metaphorical datasets: the VUAMC, MOH-X and TroFi.

4.5.1 Dataset

There are two lexical datasets for words rated, one for imageability and the other for sensory experience because these are two semantic richness elements are being tested. The impact of testing these ratings, the predicted ratings from trained SVM are being tested.

Imageability

In their work, Ljubešić et al. (2018) produced lists of around 18,156 words rated for imageability on a scale of 4 across 77 languages. Each word's rating falls between 0 and 3 (for example, *factor* received a score of 1.84). These scores were then transformed into a unified numerical format, where words were assigned values from 0 to 3 based on their original decimal ratings (for instance, *factor*'s new score became 2). Lower scores near 0 denote low imageability (such as *necessity*), while those approaching 3 indicate high imageability (such as *debts*). In addition to test the existing ratings, the study examined two types of predicted ratings from Support Vector Machine (SVM): single-digit ratings and probability

distributions. These predicted ratings were incorporated into GloVe+ELMo embeddings for both word-level and sentence-level neural networks for metaphor detection.

Sensory Experience

In their work, Juhasz and Yap (2013) introduced a list of 5,000 English words, each rated for its sensory experience. They propose that words linked to sensory experience information can aid in making lexical decisions and influence response times for naming tasks. The words were rated on a scale of 7, with lower numbers indicating words with minimal sensory influence (for instance, *appoint* with a sensory experience score of 2), while words with higher sensory experience scores (such as *balloon* with a sensory experience score of 5) evoke more vivid sensory and image-based perceptions. The scoring system was converted into integers, resulting in ratings ranging from 1 to 6. Similar to previous experiments, in addition to testing the existing ratings, two types of SVM predicted rating formats were evaluated, and their impact on the metaphor detection task was observed and analyzed.

Metaphorical Datasets

In addition to the MOH-X and TroFi, the VUAMC annotated for both, sentence and word, are used in these experiments. These datasets are provided by Gao et al. (2018) and Mao et al. (2019). The VUAMC datasets are tested as training, validation and testing files. While the MOH-X and TroFi are experimented with the 10-fold cross-validation. These datasets were discussed in Chapter 1 Section 1.8.2.

4.5.2 Experiments Set-Up

The procedure described in this section are followed for both lexical lists. The procedure was described in Section 1.8.5 in Chapter 1. The experiments are divided into two steps: the first is training the SVM on the lexical lists' ratings to assign predicted ratings to the words in the metaphorical datasets. The ratings are tested in two formats: single-digit and probability distribution. The BERT pre-trained embeddings are utilised to obtain representations for every word in the lexical lists for SVM training. These BERT pre-trained embeddings are also applied to the metaphorical tokens to generate predicted ratings. These ratings are then combined with the GloVe+ELMo vectors that represent the word/context for each word. The second step involves the primary metaphor detection task, where neural network models are employed to test these incorporated ratings. The models and results has been published in Alnafesah et al. (2023).

Single-Rating Scores

As previously mentioned, the ratings within the lexical datasets are normalised, with words being assigned ratings ranging from 0 to 4 for imageability and 0 to 6 for sensory experience. These ratings are then transformed into integers within a range of 3 for imageability and 6 for sensory experience. This transformation is carried out to facilitate the training of SVMs and to ensure rating consistency. In this context, each word present in the lexical lists is associated with a BERT pre-trained vector that captures its semantic meaning, alongside its corresponding rating. These vectors and ratings collectively constitute the training input. Furthermore, the predicted rating is assigned to each token within the metaphorical datasets by extracting its vector representation from BERT's pre-trained embeddings. This process enables the establishment of associations between metaphorical tokens and their predicted ratings. Testing for this subsection involves two types of single-digit ratings:

1. The first single-score testing utilises the ratings that are already present within the lexical datasets, and it does not incorporate the predicted ratings from the SVM. It is important to note that the metaphorical words receiving ratings in this context must be existed in the lexical datasets. Essentially, this subsection examines the influence of human-produced ratings on the detection task. The single-digit rating is incorporated as an additional component within the vector, which is 300+1024 dimensions in length, resulting from the combination of GloVe and ELMo embeddings. In situations where words are out-of-vocabulary, a value of -1 is appended to the vector to indicate that the rating is *unavailable*. For reference, Table 4.1 presents the statistical data regarding the word counts within both lexical datasets and the extent to which they encompass tokens from the metaphorical datasets.

2. The second attempt of single-score testing involves employing the ratings that have been predicted by the trained SVM. One of the reasons for relying on BERT pre-trained vectors in the rating assignment process, as discussed earlier in Chapter 1 and Section 1.8.3, is its tokenisation capability. Consequently, all words present in the three metaphorical datasets are assigned a predicted rating. This particular stage aims to assess the effectiveness of the generated predictions in the metaphor detection task. Similar to the previous step, the predicted single-digit ratings are joined with the 1324-dimensional vector derived from the combination of GloVe and ELMo embeddings. Table 4.2 shows the F1 result of training the SVM on both dataset using the 10-fold cross-validation approach.

Probability Distribution Scores

The SVM is trained to assign a probability distribution as the rating for each word within the three metaphorical datasets. The length of these scores varies based on the rating scale.

Dataset	Total	IMG	SEN
IMG	26800	_	_
SEN	5856	_	_
VUAMC Cl.	17017	6928	3059
VUAMC Seq.	16979	7153	3156
MOH-X	1677	592	694
TroFi	13738	5501	2771

Table 4.1: The breakdown of the total tokens for each dataset and the number covered by the imageability and sensory experience lists.

SVM	F1 Mean	F1 MAX	F1 MIN
IMG-single	86.14%	88.73%	77.01%
IMG-prob.	86.01%	88.9%	72.7%
SEN-single	38.678%	43.247%	35.213%
SEN-prob.	40.026%	44.273%	37.606%

Table 4.2: The mean, maximum, and minimum F1 scores for each SVM trained on the imageability and sensory experience datasets.

These scores are then incorporated into the 1324-dimensional vectors originating from the combined GloVe and ELMo embeddings. Every digit within the probability distribution score signifies the likelihood of the word being associated with that particular rating. For the prediction of imageability, the score adopts a 3-dimensional vector structure, while for sensory experience, it takes the form of a 6-dimensional vector. The objective of this approach is to assess the efficacy of presenting more comprehensive rating information and to determine whether this format enhances metaphor detection capabilities. Table 4.2 shows the F1 result of training the SVM on both dataset using the 10-fold cross-validation approach.

4.5.3 Classification

The classification model, introduced by Gao et al. (2018), is employed to establish the baseline for comparative testing performance for sentence-level metaphors. In this model, sentences are labeled as 1 if they are predicted to be metaphors and as 0 if they are non-metaphorical. This assignment is made based on the context of the target verb within the sentence. Further details regarding this model can be found in Chapter 1 Section 1.8.1. The experiments involved the VUAMC datasets is divided into three files, the training, validation and testing files. While the 10-fold cross-validation approach is utilised on MOH-X and TroFi datasets.

Table 4.3 and Table 4.4 present the baselines used for each metaphorical datasets along with their hyperparatemters used to reach these results. In addition, the classification experiments for metaphor detection with the imageability and sensory experience are presented with their hyperparameters. The tables display the baseline results, along with the outcomes for the following scenarios: single-digit ratings that are not SVM-predicted (SD no SVM), single-digit ratings that are SVM-predicted (SD SVM), and probability distributions for all datasets.

In the imageability classification experiments, Table 4.3, improvements are observed across all three metaphorical datasets. The single-digit scores obtained without SVM involvement exhibit the most notable enhancement with the VUAMC. The F1 score raised to 55.5%. Additionally, the recall experiences a substantial increase, progressing to 56.55%. This heightened recall suggests that the model successfully predicts a greater number of metaphors within the correct predictions, implying that the incorporated ratings contribute valuable information for the VUAMC dataset. On the other hand, the probability distribution approach achieves the highest precision, but there are decreases in both recall and F1 score. This suggests that the utilisation of probability distribution might introduce more

noise to the predictions, leading to lower recall and overall performance.

In the MOH-X dataset experiments, the improvements are evident, with the highest increase observed in the F1 score for the SVM-predicted single-digit rating. The F1 score shows a slight enhancement from 76% to 77.054%. Notably, the single-digit score without SVM involvement demonstrates the highest precision, showing a 2.37% increase. In contrast, the probability distribution approach yields the most improved recall, experiencing an increase by 2.94%.

However, the results are different for the TroFi dataset. The probability distribution approach exhibits the most significant increase in recall and F1 score compared to the baseline. The recall rises to 75.933%, and the F1 score reaches 72.283%, a slight improvement from the initial 71.088%. Notably, the precision for all single-digit scores and the probability distribution slightly decreases from the initial 69.661%. In contrast to the results obtained from the VUAMC dataset, the utilisation of the probability distribution has demonstrated a positive enhancement in the metaphor detection task.

Table 4.4 presents the sensory experience with the classification model. Similar to the imageability tests conducted on the VUAMC dataset, the single-digit scores not predicted by SVM display the most remarkable improvements in terms of recall and F1 score. The recall shows a significant increase of 20.9%, rising from the baseline to 61.839%. The F1 score experiences a boost of 7.52%. However, all scoring formats result in a slight decrease in precision. When compared to the imageability scores on the VUAMC dataset, the sensory experience experiments exhibit a more enhancement in metaphor detection performance over the baseline.

In the context of the MOH-X dataset, the utilisation of single-digit predictions not generated by SVM yields the highest values for precision, recall, and F1 score. The precision sees a slight increases of 1.67%, the recall increases by 3.63%, and the F1 score increased by

2.3%. While both other scoring methods also exhibit improvements over the baseline, they fall short of the gains observed in the non-SVM single-digit prediction results. The same trend is observed for precision, but there is a less decrease in the results for SVM-predicted single-digit ratings. In this case, the score experiences a slight decrease, declining to 74.115%.

For the TroFi dataset, when considering sensory experience as a single-digit SVM-predicted score, the highest recall and F1 score are achieved. The recall notably improves by 3.9%, rising to 75.961%. Similarly, the F1 score sees a slight increase of 1.5%, moving to 72.159%. As seen in the MOH-X dataset, the precision for all scoring formats demonstrates a decrease from the baseline in the TroFi dataset as well.

4.5.4 Sequence Labelling

The sequence labelling baselines are build on the RNN_HG model proposed by Mao et al. (2019). Similar to the set up employed by the classification experiments, the experiments here follow the same approach. The sequence labelling model classify words in the sentence based on the surrounding context as 1, *metaphor* or 0, *non-metaphor*. More details regarding this model can be found in Chapter 1 Section 1.8.1. The experiments involved the VUAMC datasets is divided into three files, the training, validation and testing files. While the 10-fold cross-validation approach is utilised on MOH-X and TroFi datasets.

Table 4.5 and Table 4.6 present the baselines used for each metaphorical datasets along with their hyperparatemters used to reach these results. In addition, the sequence labelling experiments for metaphor detection with the imageability and sensory experience are presented with their hyperparameters. The tables displays the baseline results, along with the outcomes for the following scenarios: single-digit ratings that are not SVM-predicted (SD no SVM), single-digit ratings that are SVM-predicted (SD SVM), and probability distributions

for all datasets.

The imageability experiments are presented in Table 4.5 using the sequence model. The experiments with the VUAMC dataset show improvements over the baseline results. The single-digit SMV prediction has the best recall and F1 scores. The recall improved significantly by 14.9%, increasing to 74.115%. The F1 score also improved by 3.9%, rising to 74.055%. However, the precision decreased to 73.99%.

The results of the MOH-X dataset demonstrate the best outcomes with the experiment that incorporated the probability distribution. The MOH-X recall slightly decreased to 80.95%, while precision increased to 80.686%, and the F1 score improved to 80.539%.

For the TroFi dataset experiments, the best results were achieved when the non-SMV single-digit rating was incorporated. The precision declined to 65.68%. However, the recall and F1 score increased over the baseline. The recall increased by 8.64%, reaching 80.93%, while the F1 score increased by 1.12%, reaching 72.38%. These results indicate that imageability has a positive impact on the detection task.

The sensory experience experiments are presented in Table 4.6 using the sequence labelling model. Regarding the results concerning the VUAMC dataset, there was a decrease in precision across all sensory experience score formats. However, the recall demonstrated an increase when experimenting with the probability distribution, rising to 71.968% (an increase of 11.6%). The F1 score reached its best result with the not-SVM-single-digit rating, showing an increase of 3.6%.

For the MOH-X dataset, both precision and F1 score increased when the experiment utilised the not-SVM-single-digit rating. The precision increased by 3.156%, reaching 78.82%, while the F1 score improved to 79.257% (an increase of 1.01%). The recall saw a slight increase of 0.2% for the probability distribution format.

The improvements in TroFi dataset results are distributed across the three sensory experience score formats. The precision achieved its best result when the model was tested with the probability distribution format, reaching 69.963%, indicating an increase of 1.42%. The recall's best performance was observed when the model was tested with the not-SVM-single-digit rating, resulting in an increase of 2.66% and reaching 76.471%. The F1 score achieved its highest score of 72.212% through the experiment that tested the SVM's single-digit rating.

4.5.5 Discussion

The analysis of these results in the previous tables was done by examining the results of the prediction outputs. The output is divided into correctly and incorrectly predicted metaphors, for all metaphorical datasets.

The word *get* was one of the most incorrectly predicted words, with an imageability score of 3. The sentence "*They've improved the chimney recently, and it's much better than it was, but we still get the smells.*" was incorrectly classified as non-metaphor based on the verb *get* (all words around the verb were rated 3 for imageability). As discussed in the previous chapters, the challenge here is that the collective meaning of the phrasal verb deviates from the meanings of its individual parts. Furthermore, the metaphor annotation approach only annotates one part of the phrasal verb as the metaphor. Additionally, the lack of change in the ratings (all words have a 3 imageability rating) might have sent incorrect signals to the models, leading to their classification as literal, as in the example, "*He chuckled, produced two cardboard cups, and poured me a generous slug of the whiskey*", all words around the target verb *pure* has 3 for imageability.

The MOH-X incorrectly predicted files containing 101 words, including the word

push. In the sentence "*The guy hanging around the school is pushing drugs*", the consistent imageability rating of 3 for both the words *push* and *drugs*, along with the well-known frequent meaning of the word *push* as it typically refers to physical objects, contributed to the model's misclassification of the verb as literal. Additionally, the frequent meaning that appeared in a compatible context further added to this misclassification. The word *push* in this context refers to *selling* and *distributing*, as opposed to the *physical application of force or pressure to move objects*. This meaning was not captured by the models. Words that exhibit polysemy (words with multiple meanings) face greater challenges due to the inherent ambiguity in determining which specific meaning is being employed. The lack of rating shifts or shifts that do occur might not provide adequate assistance to the model in deciphering the intended meaning.

The TroFi dataset's incorrectly predicted files show 16 verbs. Among the most misclassified words is *step*. In the example, "*Last month, Judge Aguilar rejected a request by attorneys for Apple that he step aside because of his Hewlett-Packard connection*", the phrasal verb *step aside* is intended to convey *moving away* from the case and *resigning*. However, the word *step* typically means to *place one foot in front of another for movement*. The literal meaning of the word *step* differs from that of the phrasal verb. The context in which this phrase appears is compatible with the literal meaning of *step* and the *direction* indicated by *aside*. This situation is supported by the assigned ratings of 3 for both parts of the phrasal verb. Nonetheless, the annotation approach, where only the *step* part was marked as metaphor, could have contributed to the incorrect prediction.

The incorrectly predicted files analysis for the sensory experience classification model revealed that notable occurrence of prepositions, with the word *get* being featured in the list. In the example "*probably need to get Ken's permission!*", all words were assigned a sensory experience prediction score of 1, except for *Ken*, which received a score of 2. Despite the subtle change in ratings, particularly for *Ken* following the target verb *get*, the model still failed

to correctly predict it as metaphor. This misclassification might stem from the extremely low sensory experience ratings assigned to all words within the context, incorporated by the inherent metaphorical nature of the word *get*. Words like *get* and similar instances often constitute common words that have lost their metaphorical connotations and transitioned into literal usage. This shift often results in one of the polysemous meanings of the word. Moreover, the model's misprediction could be attributed to the absence of a noticeable rating shift, causing it to overlook the indicators of metaphoricity.

The word *produce* in the sensory rating was also misclassified in the sentence "*He chuckled, produced two cardboard cups, and poured me a generous slug of the whiskey.*" A similar situation was observed where the meaning of the target word matched the context, and there were no noticeable rating shifts. As a result, the model incorrectly classified the sentence as literal. In this case, the phrase "*produced two cardboard cups*" could be interpreted as literal since it refers to the actual action of creating the object *cup*. However, the intended meaning of the sentence was likely *bring out two cups* rather than *make two cups*. The phrase *to produce* can also mean *to bring out* or *to make apparent or present to the public*. This alternative meaning is what the sentence is likely trying to convey. The lack of noticeable sensory experience rating shifts (with sensory ratings of 2, 1, 3, and 3) might not have provided enough helpful information for the classification model to correctly understand the intended meaning of the sentence. As a result, it classified the sentence as literal based on the available information.

The TroFi dataset's incorrectly predicted files contain the word *stick*, which holds a sensory experience rating of 2. The verb *stuck* in the sentence "*Mississippi's penchant for sticking with a senator, once elected, magnifies the importance of this campaign*" is accompanied by ratings ranging from 1 to 2, without exceeding that range. This absence of a significant rating shift may have contributed to the inaccurate prediction. Furthermore, the words surrounding the word *stick* are contextually compatible. The term *sticking* could be

interpreted as the literal *act of holding* to Mississippi, rather than conveying the intended meaning of the *loyalty that elected senators maintain*, and for which they are praised.

When analysing the correctly predicted files, the word *theories* emerges as one of these words (with an imageability rating of 2). In the sentence "*As this is not a book on criminological theories, what follows will be a brief overview of a wide range of theories and explanations,*" the phrase *theories, what follows* exhibits a slight shift, with *theories* 2, *what* 3, and *follows* 3. This subtle shift could have contributed to the correct prediction.

About 119 verbs from the MOH-X dataset, the word *absorb* appeared multiple times. In the sentence "*The immigrants were quickly absorbed into society*", the word *absorbed* was predicted correctly. All words in the sentence have an imageability rating of 3. Although there is an absence of rating shifts, the model made the correct prediction in this case. This could be attributed to the combination of the word *absorb*, which refers to immigrants and goes against selectional preference norms.

The TroFi dataset's correctly predicted files include the word *step* as one of the accurately predicted terms. In the sentence "*My mission is to save the country, to establish a durable, irreversible democracy and to step down*", the word *step* is part of the phrasal verb *step down*. In this context, the phrase means *resigning* after achieving one's mission objectives. In a previous sentence from an incorrectly predicted file, the verb *step* was not predicted correctly due to several factors, including the unclear metaphorical meaning in that context. However, in this sentence, the discussion of a *mission* and *democracy* could contribute to the notion of *resigning* from a position.

The analysis of the correctly predicted files for sensory experience ratings revealed that the word *get* appeared more frequently in the VUAMC output. In the sentence "*You get sarcasm mainly*", the sensory experience ratings for the words are 1, 1, 4, and 1. The noticeable rating gap between *get* (rated 1) and *sarcasm* (rated 4) was clear, which could

have contributed to the correct classification. Despite *get* being a commonly used word that might have lost its metaphorical meaning, looking at this sentence through conceptual mapping could be insightful. The successful classification might be attributed to the fact that the **UNDERSTANDING** of the abstract concept *sarcasm* was the intended meaning, and the word *get* was metaphorically employed from the **ACQUISITION** source to represent **UNDERSTANDING** reinforced by the sensory experience ratings provided. In the sentence "*I felt it was so safe, they were so professional*", the word *feel* has a sensory experience rating of 4. It is surrounded by *I* (rated 1), *it* (rated 1), *was* (rated 1), *so* (rated 1), and *safe* (rated 2). This construction could represent the case of an emotion, which is *safety*, being perceived as a *physical item*. The comfort of safety as an attainable object is conveyed in the certainty that "*I feel it with my hand; it exists*".

The word *fill* has appeared multiple times in the correctly predicted TroFi files. In the sentence "*Last year the movies were filled with babies*", the word *fill* is used metaphorically. The rating for *fill* is 1, while *babies* has a rating of 5. This clear disparity in ratings, coupled with the compatibility of the two words, contributes to the correct classification. Typically, *fill* is used to describe the process of *pouring liquids*, but in this sentence, it's associated with *babies*. All these factors could contribute to the accurate classification. *Attack* is another word that appears in the files, with a sensory rating of 2. In the sentence "*Then two industry trade groups sued the state, attacking the lemon law's constitutionality*", the rating for *sued* is 3 and *law* is 1. Despite the subtle rating shifts among these elements, the model still correctly classified the word *attacking*. This could be attributed to the usage of the word *attacking*, which refers to a physical action combined with an abstract idea having low sensory/imageability characteristics.

The MOH-X correctly predicted files feature the word *swallow*, which appears multiple times. For instance, in the sentence "*She swallowed the last words of her speech*", the word *swallow* has a sensory experience rating of 3, while *words* has a rating of 4. Despite the subtle

difference in ratings, the utilisation of *swallow*, typically associated with food, in place of *words*, which represents an abstract idea, and the cumulative effect of this combined with the ratings, have likely strengthened this distinction. These factors could have played a role in the correct prediction.

Based on the analysis of the previously discussed results, it can be concluded that sensory experience ratings have a more positive impact than imageability ratings across the three datasets. In most cases where the model correctly predicted metaphors, the sensory ratings demonstrated shifts, even if they were slight. In contrast, imageability ratings often showed minimal or no shifts in most examples. It could be suggested that the accuracy of metaphor prediction based solely on imageability ratings might not solely result from rating shifts, but rather from a combination of factors including word compatibility, contextual usage, and ratings. To enhance the effectiveness of imageability ratings in both experiments, a more comprehensive rating scale might be beneficial. A 3-point scale might not provide sufficient differentiation to convey shifting signals effectively to the detection models. As observed in the sensory experience experiments, a 6-point scale facilitated better rating variations for words. Additionally, introducing specific senses (such as visual and others) evoked by words could enhance the ratings. For instance, if the specific sense evoked by the word *scream* was identified as *sound*, and *torture* evoked the senses of *sound* and *touch*, incorporating these sense labels into the ratings could potentially improve the performance of the models.

The neural network models have demonstrated noticeable improvements in metaphor detection with the addition of imageability and sensory experience ratings to the vector representation. The objective was to assess whether this lexical information could enhance the word/contextual representation, consequently leading to improved model performance. Both models demonstrated improvement with different rating formats on the metaphorical datasets. Following the analysis and discussion, the research questions are addressed.

These semantic richness elements utilise experience and construct an image schema when constructing a metaphor for a challenging, hard to understand abstract idea. This attempt is to provide the abstract concept with a clear meaning, consequently establishing the potential to correlate these notions with the nature of metaphor, answer for RQ1. Despite the observed improvements in the models, the rating scale could potentially cause a negative impact on performance. This is due to the limited range of the imageability scale, where a scale of 3 might not sufficiently capture rating shifts in the context that could more effectively indicate the presence of metaphor usage. Regarding sensory experience, both the sequence and classification models exhibit an increase in recall and F1 scores. The deep learning models effectively make use of this additional information to enhance performance, which address thr RS2. The results and discussion sections offer insights into RQ3, which investigates the correlations between these lexical resources and metaphors. Imageability is a widely recognised lexical resource that has been used in many application in numerous studies related to metaphor detection. This chapter further contributes by offering additional insights into its influence on classification and sequence labelling models.

The examination and discussion of the output files have highlighted several challenges that these deep learning models encounter in the detection task, some of which have been explored in earlier chapters. For instance, the annotation scheme employed for the phrasal verbs in the metaphorical dataset, where parts of the phrase is annotated as metaphor . Another challenge pertains to polysemy, which emerges as a crucial factor in enhancing metaphor detection, given that GloVe fails to offer precise intended meanings. ELMo, on the other hand, does provide contextualised representation; however, its quality falls short of the anticipated strength presented in BERT for example. Further discussion concerning the limitations and challenges is provided in Chapter 7, where these issues are also addressed.

4.6 Conclusion

In this specific chapter, the focus was on assessing the impact of the third and fourth elements of semantic richness, the *imageability* and *sensory experience*. The concept of semantic richness suggests that cognitive processing tends to be faster for words and concepts that have richer associations. This idea is extended in this dissertation for testing the metaphor detection task. The suggestion is that enhancing the word representations, provided by the word/context embeddings, with additional information regarding imageability (how easily a word evokes imagery in the mind) and sensory experience (the sensory experience associated with a word) could potentially enhance the performance of deep learning models in metaphor detection. Therefore, this chapter investigates the impact of these elements on two models: the classification model built on Gao et al. (2018) for classifying sentences and the sequence labelling model built on Mao et al. (2019) for word labelling. These investigations were conducted on three metaphorical datasets: VUAMC, MOH-X, and TroFi. The token ratings, integrated with the vector representations of GloVe+ELMo, were obtained through trained SVMs. These ratings were presented in two formats: single-digit and probability distribution.

The results exhibited promising improvements across the two deep learning models (sensory experience results have been published in Alnafesah et al. (2023)). The analysis of the test outputs provided valuable insights into the detection process and highlighted the limitations that present challenges for the detection task. Further discussions on this matter will be presented in Chapter 7.

The next chapter introduces the final concept considered as part of the perceptual aspect of semantic richness, namely, body-object interaction. The experimental results from these investigations have been published in Alnafesah et al. (ibid.).

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC											
Gao et al., 2018 Reported	53.4%	65.6%	53.4%								
Baseline	*56.28%*	*51.107%*	*53.57%*	128	0.01	0.3	0.0	0.2	64	1	20
SD. no SVM	54.57%	**56.55%**	**55.5%**	300	0.01	0.5	0.0	0.2	64	1	20
SD. SVM	58.667%	46.507%	51.88%	300	0.01	0.5	0.0	0.2	64	1	20
Prob. Dist.	**57.74%**	45.08%	50.63%	300	0.01	0.5	0.0	0.2	64	1	20
MOH-X											
Gao et al., 2018 Reported	75.3%	84.3%	79.1%								
Baseline	*75.44%*	*77.393%*	*76%*	300	0.001	0.2	0.0	0.2	10	1	30
SD. no SVM	**77.229%**	76.39%	76.416%	300	0.001	0.2	0.0	0.2	10	1	15
SD. SVM	76.181%	78.444%	**77.054%**	128	0.001	0.2	0.1	0.2	10	2	15
Prob. Dist.	75.132%	**79.674%**	77.048%	256	0.001	0.2	0.0	0.2	10	1	15
TroFi											
Gao et al., 2018 Reported	69.661%	73.04%	71.088%								
Baseline	*69.661%*	*73.04%*	*71.088%*	300	0.001	0.2	0.0	0.1	10	1	15
SD. no SVM	69.461%	74.998%	71.644%	128	0.001	0.2	0.0	0.1	10	1	15
SD. SVM	69.21%	74.74%	71.763%	300	0.001	0.2	0.0	0.1	5	1	15
Prob. Dist.	69.22%	**75.933%**	**72.283%**	256	0.001	0.2	0.0	0.1	5	1	15

Table 4.3: The table displays the baseline and the outcomes generated by the classification model utilised in the imageability experiments. The results related to the single-digit ratings presented in two formats, along with the results derived from probability distributions. Additionally, the hyperparameters employed in these experiments are presented. Bold font highlights the best results when compared to the baseline.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC											
Gao et al., 2018 Reported	53.4%	65.6%	53.4%								
Baseline	*56.28%*	*51.107%*	*53.57%*	128	0.01	0.3	0.0	0.2	64	1	20
SD. no SVM	53.91%	**61.839%**	**57.603%**	128	0.01	0.3	0.0	0.2	64	1	20
SD. SVM	56.13%	50.65%	53.25%	300	0.01	0.3	0.0	0.2	64	1	20
Prob. Dist.	55.4%	56.7%	56.1%	300	0.01	0.3	0.0	0.2	64	1	20
MOH-X											
Gao et al., 2018 Reported	75.3%	84.3%	79.1%								
Baseline	*75.44%*	*77.393%*	*76%*	300	0.001	0.2	0.0	0.2	10	1	30
SD. no SVM	**76.7%**	**80.209%**	**77.807%**	128	0.001	0.2	0.0	0.2	10	1	15
SD. SVM	74.115%	78.943%	76.216%	128	0.001	0.2	0.1	0.2	10	1	15
Prob. Dist.	76.622%	77.328%	76.468%	256	0.001	0.2	0.0	0.2	10	1	15
TroFi											
Gao et al., 2018 Reported	69.661%	73.04%	71.088%								
Baseline	*69.661%*	*73.04%*	*71.088%*	300	0.001	0.2	0.0	0.1	10	1	15
SD. no SVM	68.937%	73.952%	71.312%	300	0.001	0.2	0.0	0.1	10	1	15
SD. SVM	69.068%	**75.961%**	**72.159%**	256	0.001	0.2	0.0	0.1	5	1	15
Prob. Dist.	68.852%	74.113%	71.227%	128	0.001	0.2	0.0	0.1	5	1	15

Table 4.4: The table displays the baseline and the results generated by the classification model utilised in the sensory experience experiments. The results encompass single-digit ratings in two distinct formats, alongside the results for probability distributions. Furthermore, the table presents details about the hyperparameters employed in these experiments. The bold font indicates the best results compared to the baseline.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC											
Mao et al., 2019 Reported	71.8%	76.3%	74.0%								
Baseline	*79.58%*	*64.45%*	*71.22%*	256	0.2	0.5	0.0	0.2	2	1	29
SD. no SVM	78.198%	66.87%	72.094%	300	0.2	0.5	0.0	0.2	2	1	29
SD. SVM	73.99%	**74.115%**	**74.055%**	300	0.2	0.5	0.0	0.2	2	1	29
Prob. Dist.	77.805%	68.95%	73.11%	300	0.2	0.5	0.0	0.2	2	1	29
MOH-X											
Mao et al., 2019 Reported	79.7%	79.8%	79.8%								
Baseline	*76.408%*	*81.63%*	*78.46%*	256	0.001	0.5	0.0	0.2	2	1	20
SD. no SVM	79.399%	80.179%	79.25%	300	0.001	0.5	0.0	0.1	2	1	20
SD. SVM	79.28%	81.003%	79.946%	300	0.001	0.5	0.2	0.1	2	2	20
Prob. Dist.	**80.686%**	80.95%	**80.539%**	256	0.001	0.5	0.2	0.1	2	2	20
TroFi											
Mao et al., 2019 Reported	67.4%	77.8%	72.2%								
Baseline	*68.98%*	*74.489%*	*71.575%*	256	0.001	0.5	0.0	0.1	2	1	20
SD. no SVM	65.68%	**80.93%**	**72.38%**	128	0.001	0.5	0.0	0.1	2	1	20
SD. SVM	68.505%	76.278%	72.042%	128	0.001	0.5	0.2	0.1	2	1	20
Prob. Dist.	67.772%	77.219%	72.007%	300	0.001	0.5	0.0	0.1	2	1	20

Table 4.5: The table shows the baseline and the outcomes produced by the sequence labelling model employed in the imageability experiments. The results are for single-digit ratings in two distinct formats, as well as the results for probability distributions. The table also provides details about the hyperparameters utilised in these experiments. The bold indicates the best results compared to the baseline.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC											
Mao et al., 2019 Reported	71.8%	76.3%	74.0%								
Baseline	*79.58%*	*64.45%*	*71.22%*	256	0.2	0.5	0.0	0.2	2	1	29
SD. no SVM	76.23%	71.199%	73.629%	128	0.2	0.5	0.0	0.2	2	1	29
SD. SVM	78.1%	70.046%	**73.85%**	300	0.2	0.5	0.0	0.2	2	1	29
Prob. Dist.	74.6%	**71.968%**	73.27%	300	0.2	0.5	0.0	0.2	2	1	29
MOH-X											
Mao et al., 2019 Reported	79.7%	79.8%	79.8%								
Baseline	*76.408%*	*81.63%*	*78.46%*	256	0.001	0.5	0.0	0.2	2	1	20
SD. no SVM	**78.82%**	80.053%	**79.257%**	256	0.001	0.5	0.0	0.1	2	1	20
SD. SVM	78.795%	79.809%	79.033%	300	0.001	0.5	0.2	0.1	2	1	20
Prob. Dist.	77.562%	**81.804%**	79.204%	128	0.001	0.5	0.0	0.1	2	2	20
TroFi											
Mao et al., 2019 Reported	67.4%	77.8%	72.2%								
Baseline	*68.98%*	*74.489%*	*71.575%*	256	0.001	0.5	0.0	0.1	2	1	20
SD. no SVM	68.598%	**76.471%**	72.172%	256	0.001	0.5	0.0	0.1	2	1	20
SD. SVM	69.358%	75.591%	**72.212%**	64	0.001	0.5	0.0	0.1	2	1	20
Prob. Dist.	**69.963%**	74.439%	72.001%	128	0.001	0.5	0.0	0.1	2	1	20

Table 4.6: The table shows the baseline and outcomes produced by the sequence labelling model employed in the sensory experience experiments. The results are for single-digit ratings in two distinct formats as well as for probability distributions. The table also provides details regarding the hyperparameters used in these experiments. The bold font indicates the best results compared to the baseline.

Chapter Five

Body-Object Interaction

5.1 Introduction

The preceding chapter addressed two attributes of semantic richness: imageability and sensory experience. This chapter will assess the influence of these elements on the metaphor detection task using deep learning models. These two elements are conceptually connected as they both evoke imagery in the mind. Consequently, they were presented together in one chapter for the purpose of facilitating comparison. Although they pertain to image and perceptual experiences, their outcomes differed when applied to the metaphor detection task using a neural network. In certain instances, sensory experience exhibited an advantage in identifying sentences with a more metaphorically accurate context, while in other cases, imageability marginally outperformed it.

This chapter introduces the final element of the perceptual aspect of semantic richness: body-object interaction. This element, body-object interaction, has been developed from the measurement of concrete imageability (as discussed in Chapters 2 and 4) along with other lexical variables. As a result, it is interesting to evaluate the influence of the body-object integration element using deep learning models for metaphor detection.

The approach employed to assess this element follows the same procedure as that of all the previously discussed elements, explained in Chapter 1 and in detail in Section 1.8. This consistency aids in facilitating a meaningful and relevant comparison for all elements. The body-object integration lexical resource ratings are subjected to testing by incorporating these ratings into the word/context vectors. This process serves to augment the contextual information and evaluate its impact on the metaphor detection task.

This chapter is organized as follows: Section 5.2 introduces this new concept that is novel in the realm of metaphor detection. The subsequent section, Section 5.3, discusses the theoretical connections established by this dissertation for metaphor detection. Following that, Section 5.4 describes the experimental setup for this chapter, following the procedure previously outlined in Section 1.8. This section also presents the experimental results of the two models used for the metaphor detection task. Section 5.4.5 discusses the outcomes of the experiments and evaluates the performance of the models across the three metaphorical datasets. In addition to addressing the research questions presented in Section 1.9.

5.2 What is Body-Object Interaction?

The body-object interaction rating was established as a means of reflecting the ease of physical interaction with a word's referent (Siakaluk et al., 2008). Words are rated along a scale of high-to-low body-object interaction. Words characterised by high body-object interaction, like *key*, tend to be more concrete and easily perceptible. Additionally, they exhibit strong associations with sensory, haptic, and visual experiences. On the other hand, words with low body-object interaction, such as *mountains*, might possess the previously mentioned characteristics to a lesser extent. The human physical behavior and interaction with such words differ. We can see *mountains*, but engaging with them through everyday

physical actions is not possible. In contrast, we can see, touch, and turn to unlock *keys*.

5.3 Body-Object Interaction and Metaphor

Body-object interaction is a growing subject of investigation within cognitive science. Researchers are examining the impact of this measure on diverse cognitive activities, including word recognition, acquisition, and decision-making tasks. Siakaluk et al. (2008) stated that conceptual knowledge is rooted in perceptual interactions with the environment, aligning with an embodied perspective on cognition. In essence, the acquisition and comprehension of novel concepts, even those that are abstract, stem from the prior knowledge acquired through interactions with the environment.

This notion could also be extended to the realm of metaphors. Lakoff et al. (1999) discussed the Conceptual Metaphors Theory (CMT), exploring how we take what we've learned and project it onto these conceptual constructs. Consider the metaphor "*this movie stinks*"; it's readily comprehended as a negative statement, implying the movie's poor quality. This understanding arises from the linking between one's experiential knowledge that *stink is bad*. In essence, such metaphors can be grasped due to the parallels between the tactile and visual experiences ones had and what the metaphor signifies. Consequently, measures of body-object interaction might aid in translating this acquired knowledge into a language suitable for explaining words and concepts. This is particularly relevant for enhancing the understanding of abstract ideas.

This concept draws from the notions of concreteness-abstractness and imageability, which were examined in the context of metaphor detection in Chapter 1 and Chapter 4 of this dissertation. Given that these investigations yielded promising outcomes, it is interesting to assess the applicability of the body-object interaction concept to the task of metaphor

detection.

To my knowledge, no investigations into body-object interaction and metaphor detection using a neural network have been conducted as of the date of this report. This chapter's results are published in the paper by Alnafesah et al. (2023).

5.4 Metaphor Detection Experiments

As has been done in the previous experiments involving semantic richness elements, the objective is to assist the neural network models in the metaphor detection task. This is accomplished by enhancing the word/contextual information provided in the GloVe+ELMo vectors for each word in the sentence. The performance of the neural networks is then evaluated based on their ability to detect metaphors in three metaphorical datasets: VUAMC, MOH-X, and TroFi.

The ratings are obtained from the body-object interaction dataset, which includes words rated on a scale determining their ease of physical manipulation. These acquired ratings, combined with the predicted ratings generated by the trained SVM, are concatenated with the GloVe and ELMo vectors. They are then used as input in both sentence classification (Gao et al., 2018) and sequence labeling (Mao et al., 2019) models, the RNN_HG. The performance of these models is evaluated under two rating formats: single-digit values and probability distributions.

5.4.1 Dataset

The Body-Object Interaction

Pexman et al. (2019) compiled a word list containing more than 9000 English words that were rated based on their *ease of body interaction*. The rating scale used for these words ranges from 1 to 7. Words that are challenging for the body to interact with are assigned lower values on the scale. In contrast, words with higher ratings indicate that they are easily interacted with by the body. For instance, the word *chair* is simple to engage with physically, as you can *touch*, *move*, and *sit on it*. Therefore, it received a higher rating on the scale. On the other hand, the word *ceiling* has a lower rating because it is less conducive to physical interaction (e.g., *it is hard to jump and touch the ceiling*).

To enhance the organisation of the training process, ratings within the range of 1 to 6 are converted from decimal to integer scaling. The first phase involves exclusively utilising words that have been rated for body-object interaction and exist in the three metaphorical datasets. This means that words present in the metaphorical datasets and also existing in the body-object interaction dataset are assigned ratings. The objective is to assess the human-provided ratings using metaphor detection models. Subsequently, the SVM is trained to predict a single-digit rating between 1 and 6 for each word found in the metaphorical datasets. In the final stage, an SVM is trained to predict a 6-digit probability distribution for every word within the metaphorical dataset. These predicted ratings are then combined with the GloVe+ELMo vectors employed in the detection models.

Metaphorical Datasets

The semantic elements are evaluated on three metaphorical datasets: VUAMC, MOH-X, and TroFi, using two neural network models. One of these models is a classification model

that classifies sentences based on Gao et al. (2018). The other model is the sequence labeling model, RNN_HG, based on Mao et al. (2019), which classifies words based on context.

The VUAMC datasets (the sentence-level and the word-level) are partitioned into training, validation, and testing files. In contrast, the MOH-X and TroFi datasets are evaluated using a 10-fold cross-validation approach. A thorough examination of these three datasets, along with comprehensive details about the models, is provided in Chapter 1, Sections 1.8.1 and 1.8.2.

5.4.2 Experiment Set-Up

The body-object interaction rating is established based on factors such as concreteness and imageability, along with other lexical variables. The concept behind this notion implies that words assigned higher values on the body-object interaction scale tend to possess greater concreteness, tactility, visual perceptibility, and susceptibility to physical interaction. In contrast, words with lower body-object interaction ratings exhibit the opposite characteristics. In connection to the hypothesis outlined in Chapter 2 regarding concreteness, changing in word ratings within context could offer insights into metaphoricity.

For instance, in the sentence "*he won her heart*", the word *won* holds a body-object interaction rating of 1, while *heart* possesses a rating of 5. The proposition is that when a verb like *won* is attributed such a low body-object interaction rating, but the object of the action, *heart*, is associated with a high body-object interaction rating, the likelihood of metaphorical usage increases.

The experiments conducted in this chapter follows the procedure detailed in Chapter 1, Section 1.8, and the subsequent chapters. This standardised procedure is employed using two metaphor detection models on three metaphorical datasets:

1. Only the body-object interaction rating is assigned to the metaphorical datasets tokens which are presented only in the body-object interaction list. These ratings are single-digit, ranging from 1 to 6, and they are added to the word/context vectors provided by the combination of GloVe and ELMo. Following that, these enhanced vectors are subjected to testing using two neural network models designed for metaphor detection. Table 5.1 presents the statistical breakdown of the metaphorical datasets and the body-object interaction dataset.

2. The BERT pre-trained vectors are extracted for each word in the body-object interaction datalist. These vectors, along with the associated ratings, are utilised as the training input for the SVM. Subsequently, the SVM utilises the BERT pre-trained vectors for each word within the metaphorical datasets to predict a single-digit rating, which falls in the range of 1 to 6. These predicted ratings are then combined with the GloVe+ELMo vectors and tested using the two neural network models. In Table 5.2, the minimum, maximum, and mean F1 results are displayed for training the SVM using 10-fold cross-validation for the single-digit rating.

3. Lastly, the SVM, which was trained on the body-object dataset, assigned a 6-digit probability distribution rating to each word within the metaphorical datasets. This process follows the previously established procedure, where both training and rating assignment employ BERT pre-trained vectors. These 6-digit ratings are then incorporated into the GloVe+ELMo vectors and subjected to testing using the two neural network models. In Table 5.2, the minimum, maximum, and mean F1 results are displayed for training the SVM using 10-fold cross-validation for the probability distribution rating.

Dataset	Total	BOI
BOI	9349	_
VUAMC Cl.	17017	3572
VUAMC Seq.	16979	3701
MOH-X	1677	811
TroFi	13738	3216

Table 5.1: The breakdown of total tokens for each dataset and the number that the body-object interaction list cover.

SVM	F1 Mean	F1 MAX	F1 MIN
BOI-single	38.881%	42.887%	36.791%
BOI-prob	39.03%	41.711%	35.508%

Table 5.2: The mean, maximum, and minimum F1 scores for each SVM trained on the body-object interaction datasets.

5.4.3 Classification

The classification model classifies sentences based on the verb and its context. The model assigns a value of 1 for *metaphors* and 0 for *non-metaphors*. Word and context representations are provided from GloVe's 300-dimensional vector and ELMo's 1024-dimensional vector, respectively. These vectors are augmented with the body-object interaction ratings and then tested using the model. Baseline results are obtained using the hyperparameters outlined in Table 5.3. These baseline results are utilised to assess performance in scenarios involving both single-digit and probability distribution ratings, across the three metaphorical datasets.

The results of the classification model, compared with the baseline, are presented in Table 5.3 for the three datasets. The first experiment tests the human-produced ratings using only the words that exist in the body-object interaction dataset, without the involvement of the SVM. The second experiment tests the SVM-predicted single-digit rating, while the third experiment tests the probability distribution.

The VUAMC dataset does not show any improvement over the baseline for the three rating formats. On the contrary, the performance decreased across all three metrics. The worst recall and F1 score decreased by 33.7% and 17.7%, respectively, for the first experiment's rating format. However, precision increased to 63.06%. Recall indicates the number of correct metaphor predictions out of all correct predictions. The decrease in recall suggests that the model's ability to detect metaphors has been worsened by the rating. The SVM-predicted single-rating format has the best F1 score, although it is slightly lower than the baseline, at 53.24%.

The MOH-X and TroFi datasets exhibit different results compared to the VUAMC dataset. The inclusion of body-object interaction has led to improved performance in the experiments related to the SVM-predicted single-digit ratings for the MOH-X dataset. The recall slightly improved to 78.39%, with the F1 score increasing by 1.4%. The increased recall suggests that the model has gained useful information from the body-object interactions. The results for the TroFi dataset with the inclusion of the body-object interaction show improvement for the experiments that did not involve SVM-predicted ratings. The recall increased by 4.8% compared to the baseline, and the F1 score increased by 1.2% to 71.98%. However, the precision for the three rating formats has slightly decreased from the initial 69.661%.

5.4.4 Sequence Labelling

The sequence labeling model utilised is derived from the RNN_HG model introduced by Mao et al. (2019). This model assigns classification labels to words within sentences, labelling them as 1 for *metaphor* or 0 for *non-metaphor* based on their context. Similar to the classification model's experiments, the sequence labelling model evaluates the three rating formats across the three metaphorical datasets. The body-object interaction ratings are con-

catenated to the vectors that represent each word/context. These vectors are obtained from GloVe's 300-dimensional and ELMo's 1024-dimensional embeddings. The baseline results for precision, recall, and F1 scores, along with the corresponding hyperparameters, have been presented in Table 5.4.

The incorporation of body-object interaction ratings using the sequence labeling model, RNN_HG, has shown promising results across the three metaphorical datasets. The VUAMC dataset demonstrates an increase in recall when utilising the body-object interaction rating as the probability distribution predicted by the SVM. This increase amounts to 9.8% over the baseline, raising to 70.79%. Moreover, there is a 2.6% increase in F1 score over the baseline for the single-digit rating format that does not involve SVM predictions.

However, it's worth noting that the precision results for the VUAMC experiments are lower than those of the baseline. Precision signifies the overall correct of predictions, whereas recall indicates the proportion of correct metaphor predictions among all correct predictions. This implies that the model's capability to detect metaphors has indeed improved. The results from the MOH-X dataset exhibit a similar pattern to the VUAMC dataset, showing improved metrics when incorporating body-object interaction ratings. Notably, the recall and F1 score demonstrate the most significant enhancements over the baseline when utilising the SVM-predicted single-digit rating. The recall increased by 1.9% and the F1 score by 1.6% compared to the baseline. The best precision is achieved with the probability distribution rating. Although its F1 score might not be the highest among all the experiments, it does reflect an improvement over the baseline by 1.03%. This highlights that the integration of body-object interaction ratings has consistently contributed to enhanced performance across these experiments.

In the case of the TroFi dataset experiments, all F1 scores exhibit improvements over the baseline. The highest F1 score is associated with the body-object interaction rating

without the SVM predictions. Conversely, the most notable increase in recall is observed with the SVM-predicted single-digit rating, amounting to a 3.7% improvement over the baseline.

Furthermore, precision displays enhancement when using the probability distribution rating. The initial result stands at 68.98%, and the increase amounts to 1.5%, resulting in a precision of 70.03%. These findings collectively highlight the positive impact of incorporating body-object interaction ratings on the TroFi dataset experiments, leading to improved model performance across various evaluation metrics.

5.4.5 Discussion

The discussion section dives into the analysis of the output files generated from the previous experiments. These experiments focused on evaluating the performance of neural network models when integrating body-object interaction ratings for the task of metaphor detection. The output files provide insights into correctly predicted and incorrectly predicted metaphors across the three datasets. Table 5.1 shows the metaphorical datasets and the body-object interaction dataset statistics.

In the VAUMC dataset, there were several instances of incorrectly predicted words, such as *got* (26 times), *go* (15 times), and *back* (14 times). These words had predicted body-object interaction ratings of 1, 2, and 5, respectively.

For example, the sentence "*having got some of the plumbing details wrong.*" was examined, (*having* 2, *got* 1, *some* 1, *of* 1, *the* 1, *plumbing* 3, and *details* 2). It was observed that there was a slight shift in ratings between the words, ranging from 1 to 2 to 3. Furthermore, the verb *got* with a rating of 1 and the words *plumbing* with a rating of 3, and *details* with a rating of 2, do not strongly suggest physical manipulation. This subtle shift in

ratings and the lack of strong indications for physical interaction might explain the decrease in model performance.

A similar reasoning can be applied to the sentence "*Well, hang on a minute*" where the ratings are 1, 1, 2, 1, and 1 for the words (*Well*, *hang*, *on*, *a*, *minute*). In this case, the verb *hang* is the target. The considerable shifts in ratings might not provide any significant additional information beyond what is already captured by the GloVe and ELMo embeddings. This could contribute to the model's reduced performance on these instances. However, the word *hang* is polysemous, with its primary meaning being to *suspend or be suspended from above*. The intended meaning to be used here relates to *remain static or wait*. Unfortunately, this intended meaning was not adequately provided by GloVe or ELMo embeddings. This discrepancy may have resulted in the model's inability to accurately predict this word as metaphorical.

In the case of a correctly predicted metaphorical sentence such as "*Dimples played in his cheeks*", the target verb *play*, along with its context ratings (*Dimples* 4, *played* 2, *in* 3, *his* 1, and *cheeks* 5), exhibited a shift from a rating of 4 to 2. This shift, coupled with the fact that *Dimples* suggests an action like *play*, could serve as an indication of metaphoricity.

This showcases how changes in ratings, particularly when aligned with the semantics of the words involved, can aid the model in recognising metaphors, as demonstrated by these instances.

The TroFi dataset contains instances where the word *step* has been incorrectly predicted. Specifically, it has been predicted incorrectly 87 times. This word *step* is polysemous, meaning it has multiple meanings. Many of the instances where it's misclassified involve phrasal verbs like *step down* or *step in*. For instance, in the sentence "*Deluxe balked and a competitor stepped in to accept the contract, startling investors*", the company Deluxe initially withdrew from the contract, and then a competitor took action (referred to as *stepped*

in) and agreed to the contract instead. This use of *stepped in* is metaphorical, as it signifies the competitor taking over the role rather than a physical movement. The body-object interaction rating for the word *competitor* is 3, while for *step* it is 5. Despite the relatively lower body-object interaction rating for *competitor* compared to the higher rating for *step*, they are still compatible in conveying the concept of someone taking a physical action, especially if the basic and frequently used meaning were considered. This compatibility could further explain the instances of misclassification.

These datasets includes a list of words known for their polysemy. In cases where the meaning provided by GloVe static and ELMo somewhat contextualised vectors did not aid in identifying the intended metaphorical meaning, coupled with the format used for annotating phrasal verbs (where the metaphorical annotation is partial), these factors collectively impose clear limitations that impede the effectiveness of metaphor detection models.

The list of words in the incorrectly predicted file for the MOH-X dataset includes words such as *reflect*, which appeared three times. In all the provided examples: ‘‘*This action reflects his true beliefs.*”, “*His lack of interest in the project reflects badly on him.*”, and “*The mess in his dorm room reflects on the student.*”, the word *reflect* is employed to establish a connection between two entities. In these instances, *reflect* is used to convey *an association or correlation* rather than its basic meaning of *light bouncing off a mirror's surface.*

The term *reflect* in these sentences signifies an indication or representation of one aspect through another. It communicates that there exists a meaningful relationship between the subject and the object being referred to. For instance, the action reflects the person's true beliefs, the lack of interest reflects negatively on him, and the mess in the dorm room reflects something about the student.

The body-object interaction rating for the word *reflect* is 1, while for *beliefs* and

project, it's 2, and for *room*, it's 4. This indicates that *reflect* is associated with a lower degree of physical interaction, whereas the other words have higher ratings, suggesting a stronger connection with physical presence or interaction.

However, the absence of a contextualised representation that signifies the meaning of *reflect* in each sentence could be the reason why the word was mispredicted. Both the GloVe and ELMo vectors fall short in providing the necessary quality of representations required for accurate prediction.

In their study, Pexman et al. (2019) noted that the provided ratings reflect the level of ease in terms of physical interaction associated with the examined words. Despite certain words having similar meanings, their body-object interaction ratings vary. For example, *he* holds a body-object interaction rating of 2.96, while *she* has a rating of 3.30, *boy* is rated at 4.9, and *girl* has a rating of 5.52. These discrepancies in the body-object interaction ratings for words that are conceptually related might have influenced the process of metaphor learning in connection with these ratings.

It is worth mentioning that Pexman et al. (ibid.) also mentioned that these ratings were derived from concreteness and imageability ratings, among other variables. However, the specific details of these variables were not provided. This lack of specificity raises questions about what factors might have contributed to the rating gap observed between *boy* and *he*, two words that might be expected to have similar ratings due to their semantic proximity. The findings and the subsequent discussion of this chapter have been documented and published in the paper Alnafesah et al. (2023).

Following the thorough analysis and subsequent discussion of the results and output files generated by the models, the research questions that were originally introduced in Chapter 1.1, specifically in Section 1.9, have been addressed and explored in depth.

Due to the observed shifts in precision and recall, it can be suggested that the incorporation of the body-object interaction rating as a feature could indeed influence metaphor detection models, thus providing insights into addressing both RQ1 and RQ2. However, it's important to note that distinct disparities exist in the effects of each semantic richness resource on the detection task. For example, answering RQ2 entails also a comparison between the body-object interaction rating and concreteness ratings in terms of several aspects: dataset size (ranging from 40000 to 9000 instances), rating scale (4 classes versus 6 classes), rating methodology (concreteness gauges sensory perception while body-object interaction assesses physical interaction ease with the word referent), among others. These aspects, in many respects, mirror the performance outcomes of metaphor detection.

While concreteness ratings are defined based on the senses, body-object interaction incorporates alternative perspectives. For instance, *boy* and *girl* both possess a concreteness rating of 4, while *he* and *she* are rated at 3. This consistency potentially enhances the learning and detection of metaphors, as evidenced in the results of Chapter 2. However, it's plausible that the body-object interaction rating might lack the crucial information necessary to elucidate the nature of metaphors and their formation mechanisms (RQ3). Relying solely on the rating scale might fall short in fully comprehending the intricacies of metaphors. Introducing the parameters utilised to derive body-object interaction ratings into the deep learning model could offer a more comprehensive approach to understanding the essence of metaphors and why they manifest in their particular format.

A more thorough and comprehensive discussion regarding the observed limitations is elaborated upon in Chapter 7. The challenges identified in this section, and more, are ubiquitous across all the preceding semantic richness elements. It would prove advantageous to explore these challenges with a more focused approach, allowing for a comprehensive summary of all the elements. Furthermore, the recommendations for future work are also applicable to all these elements, given that the shared issues extend beyond individual com-

ponents. Addressing these challenges with could yield more effective solutions and strategies, ultimately enhancing the performance of the neural network models and refining the results of the metaphor detection task.

5.5 Conclusion

This chapter addresses the final element of the perceptual aspect of semantic richness: body-object interaction. Derived from concreteness and imageability, this concept describes the *ease of physical interaction with a word's referent.* A higher body-object interaction rating indicates greater ease of interaction with the word's referent by physical entities. Metaphors often combines physical interactions with abstract concepts to convey meaning. For instance, consider the metaphor "*his words cut the heart*". Since *words* are abstract and not easily physically interactive, combining them with the action *cut* conveys that these words are hurtful, much like a *knife cutting the heart.*

The results of this semantic richness feature in both models exhibited enhancements over the reproduced baselines across all three metaphorical datasets. However, the concreteness and imageability ratings yielded superior metaphor detection outcomes for both models. This could be attributed to concreteness offering more effective signals indicative of metaphoricity.

The following chapter summarises the undertaken work in this thesis concerning metaphor identification via neural networks and semantic richness elements. Additionally, the limitations section discusses common observations shared among all elements. Some pertain to the models' structure, the metaphorical and lexical datasets, and the embeddings responsible for providing word and context representations. The chapter also proposes solutions and recommendations for future work aimed at addressing these limitations.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC											
Gao et al., 2018 Reported	53.4%	65.6%	53.4%								
Baseline	*56.28%*	*51.107%*	*53.57%*	128	0.01	0.3	0.0	0.2	64	1	20
SD. no SVM	**63.06%**	33.84%	44.05%	300	0.01	0.3	0.2	0.2	64	1	20
SD. SVM	58.04%	49.17%	53.24%	128	0.01	0.3	0.2	0.2	32	1	20
Prob. Dist.	56.85%	48.2%	52.21%	300	0.01	0.3	0.2	0.2	32	1	20
MOH-X											
Gao et al., 2018 Reported	75.3%	84.3%	79.1%								
Baseline	*75.44%*	*77.393%*	*76%*	300	0.001	0.2	0.0	0.2	10	1	30
SD. no SVM	**77.566%**	77.655%	77.048%	300	0.001	0.2	0.0	0.2	10	1	15
SD. SVM	76.43%	**78.39%**	**77.097%**	128	0.001	0.2	0.0	0.2	10	1	15
Prob. Dist.	77.272%	77.358%	76.88%	256	0.001	0.2	0.0	0.2	10	1	15
TroFi											
Gao et al., 2018 Reported	69.661%	73.04%	71.088%								
Baseline	*69.661%*	*73.04%*	*71.088%*	300	0.001	0.2	0.0	0.1	10	1	15
SD. no SVM	68.18%	**76.613%**	**71.98%**	256	0.001	0.2	0.0	0.2	5	1	15
SD. SVM	68.439%	75.216%	71.643%	128	0.001	0.2	0.0	0.2	5	1	15
Prob. Dist.	68.416%	75.418%	71.664%	256	0.001	0.2	0.0	0.2	5	1	15

Table 5.3: The table shows the baseline and the outcomes produced by the classification model employed in the body-object interaction experiments. The results are for single-digit ratings in two distinct formats, as well as the results for probability distributions. The table also provides details about the hyperparameters utilised in these experiments. The bold indicates the best results compared to the baseline.

Exp.	P.	R.	F1	H_size	Lr	Drop1	Drop2	Drop3	B_size	Layer	Epch
VUAMC											
Mao et al., 2019 Reported	71.8%	76.3%	74.0%								
Baseline	*79.58%*	*64.45%*	*71.22%*	256	0.2	0.5	0.0	0.2	2	1	29
SD. no SVM	78.281%	68.58%	**73.115%**	300	0.2	0.5	0.0	0.2	2	1	29
SD. SVM	76.84%	69.46%	72.97%	300	0.2	0.5	0.0	0.2	2	1	29
Prob. Dist.	74.96%	**70.79%**	72.82%	256	0.2	0.5	0.0	0.2	2	1	29
MOH-X											
Mao et al., 2019 Reported	79.7%	79.8%	79.8%								
Baseline	*76.408%*	*81.63%*	*78.46%*	256	0.001	0.5	0.0	0.2	2	1	20
SD. no SVM	78.403%	80.292%	78.967%	256	0.001	0.5	0.0	0.1	2	1	20
SD. SVM	77.396%	**83.192%**	**79.767%**	256	0.001	0.5	0.2	0.1	2	2	20
Prob. Dist.	**79.439%**	79.841%	79.275%	128	0.001	0.5	0.2	0.1	2	2	20
TroFi											
Mao et al., 2019 Reported	67.4%	77.8%	72.2%								
Baseline	*68.98%*	*74.489%*	*71.575%*	256	0.001	0.5	0.0	0.1	2	1	20
SD. no SVM	68.98%	75.63%	**72.066%**	256	0.001	0.5	0.0	0.1	2	1	20
SD. SVM	67.57%	**77.31%**	71.99%	256	0.001	0.5	0.2	0.1	2	1	20
Prob. Dist.	**70.03%**	74.125%	71.925%	128	0.001	0.5	0.0	0.1	2	1	20

Table 5.4: The table shows the baseline and the outcomes produced by the sequence labelling model employed in the body-object interaction experiments. The results are for single-digit ratings in two distinct formats, as well as the results for probability distributions. The table also provides details about the hyperparameters utilised in these experiments. The bold indicates the best results compared to the baseline.

Chapter Six

Integrating the Different Semantic Elements

6.1 Introduction

The preceding chapters have demonstrated the effects of incorporating individual elements of the perceptual aspect of semantic richness in the detection of metaphors using deep learning models. These elements encompass the use of concrete concepts to explain abstract ideas, the inclusion of mental imagery and sensory experiences, and the connection between emotions and metaphorical language. Certain elements have exhibited performance improvements, suggesting their potential in shedding light on the intricacies of metaphors. In this chapter, a novel approach is taken by integrating all these elements into a unified experiment for both the classification and sequence labeling models.

6.2 Lexical Resources and Deep Learning in NLP

Deep learning models have proven to be highly effective for various natural language processing (NLP) tasks, especially when relying on distributional word representations, such as word embeddings. These embeddings capture and compute encoded lexical information based on data usage, providing valuable insights into word properties within corpora. However, the complexity of language cannot be fully addressed by this resource alone. It requires additional linguistic, world, and common sense knowledge for proper understanding and interpretation.

By incorporating these additional details, the gap can be bridged, as different lexical resources serve different purposes and act as implicit features. Together, they enhance the comprehension of language by capturing diverse aspects and mitigating individual limitations. In a study by Kuznetsov and Gurevych (2018), it was demonstrated that combining lemmatisation and part-of-speech (POS) tagging with word embeddings leads to improved performance on various related tasks. Additionally, incorporating human-readable annotations and metadata from lexical resources can enhance the interpretability and explainability of NLP models, shedding light on how these models make predictions.

In their work, Strohmaier and Tyen (2022) propose a notation that represents mappings and word sense repositories as a category theory. This formalization helps to articulate assumptions about these mapped repositories, facilitating a deeper understanding of their underlying structures and relationships.

In the previous chapters, each chapter focused on one or two features, aiming to test theories about metaphors and improve the automatic detection using neural networks. In this chapter, these features will be combined into a single test. The models will undergo three tests, with two scenarios involving BERT emotions, resulting in a total of six tests, as

outlined in the next section.

6.3 Experiments and Results

The classification model is responsible for categorising sentences into either metaphors or non-metaphors based on the context of the target verb. On the other hand, the sequence labeling model assigns labels of metaphorical or non-metaphorical to individual words based on their context.

For the experiments using the VUAMC dataset, the dataset is divided into three sections: train/val/test. However, due to the limited size of the other two metaphorical datasets, MOH-X and TroFi, a 10-fold cross-validation procedure is used for testing.

In these models, the impact of adding semantic richness elements to the word embeddings (GloVe and ELMo) is examined. The objective is to observe the effects of incorporating these concepts together in metaphor detection models.

Each lexical resource list varies in size, includes different words, and not all words in the metaphorical datasets are covered by these lists. Table 6.1 provides a breakdown of the tokens in each metaphorical dataset, the total number of words in each lexical list, and the number of words covered by these lists for each metaphorical dataset.

The experiments are divided into sets based on the score's source (the lists scores or the predicted SVM/BERT) and the type of the score (single or probability distribution) for both the classification and sequence labelling models:

- **Set_1 & Set_2**: Words from the metaphorical dataset that appear in the lexical lists for each semantic richness are assigned scores from these lists without involving SVM or

Lexical Feature	VUAMC Cl.	VUAMC Seq.	MOH-X	TroFi	Total Lexicon Tokens
Conc_Abst.	8781	8781	1202	7091	39954
Pos_Neg	2420	2420	298	1842	5474
Anger	509	509	64	407	1247
Anticipation	452	452	371	407	839
Disgust	384	384	47	278	1058
Fear	608	608	85	535	1476
Joy	357	357	58	292	689
Sadness	515	515	66	412	1191
Surprise	246	246	38	205	534
Trust	656	656	88	553	1231
IMG	6928	6928	592	5501	26800
SEN	3059	3059	694	2771	5856
BOI	3572	3572	811	3216	9349
Total	17017	16979	1677	13738	_

Table 6.1: The total number of token in each metaphorical dataset and how many got covered by each lexical list, with these list total number of tokens.

BERT. Out-of-list words are given a different indicator (0 or -2). Since these concepts are predominantly subjective, the aim is to compare the original human annotations with the results obtained from the trained SVM and BERT models. Additionally, the possibility and effectiveness of automating the generalization of these lexical resources across the entire text will be tested, thereby saving time and effort while covering a large amount of data.

In Set 1's experiments, 13 additional digits are added to the 1324-GloVe+ELMo vectors. These digits represent concrete-abstract distinctions, NRC sentiment (polarity and emotion), imageability, sensory experience, and body-object interaction. In Set 2's experiments, NRC sentiment is replaced with the BERT emotion score, resulting in the addition of 5 new digits. The objective is to examine the impact of these different sentiment resources on metaphor detection performance. Unlike the NRC sentiment, which includes a limited number of words, the BERT model assigns a predicted emotion

score to each word in the metaphorical dataset.

- **Set_3 & Set_4**: Follow the procedure outlined in Set1 and Set2. Initially, the models are tested on concrete-abstract, NRC sentiment (polarity and emotion), imageability, sensory experience, and body-object interaction. Subsequently, for the next step, NRC sentiment is replaced with BERT's predicted emotion scores. The key difference is that all words in the VUAMC, MOH-X, and TroFi datasets are assigned predicted scores from the trained SVMs. These scores are incorporated as individual predictions, which are then added to the 1324-GloVe+ELMo vectors.

- **Set_5 & Set_5**: Follow the previous experiment format, but the difference is that the predictions generated by the SVMs are probability distributions for concrete-abstract, NRC sentiment (polarity and emotion), imageability, sensory experience, and body-object interaction. The result is 37 additional digits incorporated into the 1324-GloVe+ELMo vectors. However, when the NRC sentiment is replaced with predictions of BERT emotions (in a single-digit format) in the last step, the number of added digits is reduced to 20.

The subsections that follow present comprehensive results for each step outlined above, pertaining to both models and each dataset. Evaluation of these results are depicted in Tables 6.2 and 6.4. Furthermore, Tables 6.3 and 6.5 cover essential details regarding the hyperparameters utilised in obtaining these results.

6.3.1 Set 1

Set 1 experiments focus on examining the impact of incorporating scores from the lexical lists for each semantic richness element on both the classification and the sequence labelling

models. These scores are derived from the original human-annotated lists, and only the words covered by these lists are assigned scores.

Table 6.2 provides the results of the classification experiments compared to the baselines. It is observed that the recall and F1 scores increased for the MOH-X and TroFi datasets, while they decreased for the VUAMC dataset, dropping by 8.46%. The precision scores slightly improved for VUAMC and MOH-X, while TroFi's precision slightly decreased. The improvements in recall indicate that the models correctly classified more metaphors among the total correct classifications, suggesting that these elements have contributed additional information that benefited the neural network.

As for the results in Table 6.4, are related to the sequence labelling model performance on the three datasets. In contrast to the classification results, the VUAMC dataset showed significant improvements in F1, reaching 74.28%, and in recall, increasing to 76.59%. The F1 scores for MOH-X and TroFi were slightly better than the baseline, with TroFi experiencing an increase in recall from to 78.401%, while the recall for MOH-X decreased to 79.825%. Overall, the VUAMC dataset exhibited greater benefits from the added lexical information compared to the other two datasets.

6.3.2 Set 2

This set follows Set 1's procedure. The scores for concrete-abstract, imageability, sensory experience, and body-object interaction are sourced from human-annotated lexical lists. The difference here is that the sentiment scores are obtained from the BERT model for emotion instead of the NRC list.

Table 6.2 presents the results of the classification experiments. It showed an increase in recall and F1 for all three datasets, specifically in the VUAMC dataset. The recall increased

by 15.273%, indicating that the classification model correctly predicted more metaphors, resulting in an improved F1 score of 57.814%. The precision in the MOH-X and TroFi datasets showed a slight decrease, while the VUAMC's precision decrease was more pronounced.

Similar patterns were observed for the sequence labelling model, as shown in Table 6.4. The F1 and recall scores increased across all three datasets, accompanied by a decrease in Precision. Notably, the VUAMC dataset demonstrated an improvement of recall, reaching 77.078% from the baseline of 64.45%, resulting in an overall enhancement in F1.

6.3.3 Set 3

The scores in this set are derived from SVMs trained on each lexical list for concrete-abstract, imageability, sentiment, sensory experience, and body-object interaction. These scores are single predictions added to the embedding. Table 6.2 presents the results for the classification model, where the MOH-X dataset shows slight increases in precision, recall, and F1 over the baseline. On the other hand, the precision decreases for the VUAMC and TroFi datasets, while their recall and F1 scores increase. Notably, the VUAMC dataset exhibits greater improvements over the baseline, with a recall of 60.59% and an F1 score of 56.395%. The sequence labelling results in Table 6.4 showcase improvements in almost all metrics for all three datasets. While precision decreases for the VUAMC, recall experiences a significant jump from the baseline's 64.45% to 70.51%.

6.3.4 Set 4

Here are the results when the BERT model for emotion replaced the NRC for sentiment. The classification model in Table 6.2 shows that the recall and F1 scores for all three datasets increased, while their precision scores decreased. The improvements in recall are greater,

with VUAMC reaching 55.423%, MOH-X reaching 80.814%, and TroFi reaching 76.348%. However, the increases in F1 scores are minimal due to the decrease in precision, since the F1 score balances precision and recall.

Table 6.4 for sequence labelling shows that MOH-X's scores improved over the baseline, indicating that the model provided useful information for this dataset. While the recall and F1 scores for the other two datasets also show increases, their precision scores decrease slightly. The increases in recall for VUAMC and TroFi are greater than the increases in the MOH-X results. The VUAMC's baseline recall of 64.45% improved to reach 72.481%, while TroFi's recall increased from to 77.92%.

6.3.5 Set 5

This section deals with the SVMs' scores in the probability distribution format, where each set of scores reflects the scale of the corresponding lexical resource list. For example, the concreteness-abstractness score scale ranges from 0 to 4, resulting in a 4-digit probability distribution. This format is implemented for all elements tested, adding 37 digits to the embedding.

In Table 6.2 for classification, the three datasets show improvements in all three metrics, except for precision in the VUAMC and TroFi. The VUAMC shows a significant increase in recall, reaching 61.896%. Similarly, TroFi shows an increase in recall, reaching 77.029%. The F1 scores for all three datasets also increased, with VUAMC showing a more pronounced improvement, reaching 56.084%. These patterns of precision, recall, and F1 increases are also observed in Table 6.4 for the sequence labelling model. While the increase in F1 scores for all three datasets is minimal, the increase in recall for VUAMC is more significant, reaching 71.856%.

6.3.6 Set 6

The BERT emotion predictions replace the NRC SVM predictions in this set, resulting in a reduction of the added digits to 20. BERT's prediction represents a single digit for one of the five emotions discussed in Chapter 3. In Table 6.2, the precision of the classification model and the F1 scores for the VUAMC slightly decreased (to 55.09% for precision and to 53.298% for F1), while the recall slightly increased to 51.618%. For MOH-X, there are slight improvements in all three metrics, while TroFi shows increases in recall and F1. When comparing the results in this set with those in set 5, overall improvements are observed. In the sequence labelling experiment for the VUAMC in Table 6.4, the results show improvement over baseline in terms of recall and F1, while the precision score decreased. The recall reached 75.748%. Similar improvements are seen in TroFi's results, where the recall increased to 79.914%. Although the recall for MOH-X decreased to 80.608%, its precision and F1 scores increased slightly compared to baseline.

The information recorded in Tables 6.3 and 6.5 is the hyperparameters used to obtain the results of each experiment presented in Tables 6.2 and 6.4. The bold in the experiment tabels represent the highest performance achieved for each dataset across all sections. The bold in the hyperparameters tables denotes the configuration that used to achieve that highest performance.

6.4 Discussion and Analysis

Set 1 in Table 6.2 for the classification experiments of the VUAMC dataset demonstrated the poorest performance among the six sets. However, the best result was obtained when BERT emotion prediction replaced NRC's list scores in set 2. Followed closely when utilising the SVMs' single predictions in set 3 and employing the SVMs' probability distribution prediction

Metrics	Gao et al. (2018)	Baseline	Set_1	Set_2	Set_3	Set_4	Set_5	Set_6
VUAMC								
Precision	53.4%	*56.28%*	56.788%	51.204%	52.743%	54.373%	51.269%	55.09%
Recall	65.6%	*51.107%*	46.791%	**66.382%**	60.59%	55.423%	61.896%	51.618%
F1	53.4%	*53.57%*	51.307%	**57.814%**	56.395%	54.89%	56.084%	53.298%
MOH-X								
Precision	75.3%	*75.44%*	76.061%	75.005%	76.099%	73.429%	76.178%	77.295%
Recall	84.3%	*77.393%*	79.111%	**81.195%**	78.832%	80.814%	79.14%	77.913%
F1	79.1%	*76%*	77.152%	**77.289%**	77.29%	76.778%	76.937%	**77.3%**
TroFi								
Precision	69.661%	*69.661%*	68.319%	68.63%	69.081%	68.05%	67.709%	69.233%
Recall	73.04%	*73.04%*	75.32%	75.4%	74.67%	76.348%	**77.029%**	74.26%
F1	71.088%	*71.088%*	71.49%	71.723%	71.603%	71.731%	**71.894%**	71.579%

Table 6.2: The results are compared to the reported results and the baselines built on Gao et al. (2018) classification model for each dataset. These experiments test the VUAMC, MOH-X, and TroFi with all the semantic richness elements based on the score's source: the lexical list or the SVM/BERT prediction. The 10-fold cross-validation procedure was applied to obtain the MOH-X and TroFi results.

scores in set 5. These observations suggest that the involvement of SVMs rather than the integration of BERT emotion prediction has benefited the VUAMC dataset more than the other test variations. However, the opposite was shown for set 2 for the sequence labelling experiment in Table 6.4. Without the use of SVMs or BERT, the lexical lists of human-annotated scores produced the dataset's highest F1 score (set 1). Notably, starting from the baseline's recall of 64.45%, all six sets exhibited substantial recall improvements. All in all, the integration of the semantic richness elements with the sequence labelling model on the VUAMC dataset demonstrated enhancement of the metaphor detection task. The VUAMC dataset findings show the effectiveness of adding SVMs and BERT emotion prediction with classification and sequence labelling models varies.

The results in Table 6.2 in the classification experiments of the MOH-X dataset in-

dicate that, in set 2, the highest recall was achieved (81.195%) and the best F1 score was 77.289%. The model in set 2 utilised semantic richness lists for score assignments but replaced the NRC sentiment with BERT emotion prediction. Overall, the classification model's performance on the MOH-X dataset showcases improvements for metaphor detection. Comparable results are shown in Table 6.4 for the sequence labelling experiments. The best recall was for set 2 and set 3 attained the highest F1 score. The F1 score in set 3 reached 81.007% over the baseline's 78.46% when the model utilised the single-score predictions of the SVMs. These findings showed that incorporating elements of semantic richness with the sequence labelling model for metaphor detection on the MOH-X dataset has enhanced.

Compared to the baseline's 71.088%, the TroFi experiments' F1 scores with the classification model in Table 6.2 demonstrated slight improvements in all sets. The best F1 score of 71.894% was achieved in Set 5. The recall metric, on the other hand, has shown more substantial improvement in set 5, increasing from 73.04% to 77.029%. Over the course of the six sets, this trend of improvement in recall was observed. However, precision scores decreased compared to the baseline. Table 6.4 shows similar patterns for the sequence labelling model. The highest F1 score was achieved when the model used the SVMs' probability distribution as the score format (set 4). Nevertheless, across all sets, the most significant improvement was observed in recall. The best recall reached 79.914% in set 6, compared to the baseline of 74.489% (the NRC sentiment analysis was replaced with BERT emotion prediction). In general, varying improvements in F1, recall, and precision were observed across different sets for the classification and sequence labelling models. The most notable improvement in recall for the sequence labelling model was when the BERT emotion was substituted for the NRC sentiment.

Samples of the output predictions, correctly predicted metaphors and incorrect predictions, were examined for the analysis phase. The total number of words in each file and the sentences in which they appeared were assessed. The prediction files for the VUAMC

showed that 309 tokens were accurately predicted as metaphors, while 389 tokens were incorrectly predicted. In the case of MOH-X files, it found that 139 tokens were correctly predicted as metaphors and 85 were incorrectly predicted, while TroFi's files showed that 49 tokens were correctly predicted as metaphors and 8 were incorrectly predicted.

Again, among the metaphoric words that are frequently incorrectly predicted are *get* and *go*. These words are often referred to as dead metaphors where their metaphorical meaning has become one of their literal meanings, or, in some instances, form parts of phrasal verbs. The phrase *gone wrong* in the sentence "*What has gone wrong?*" does not indicate a literal movement but rather signifies the end result of something, such as a situation or an experiment, depending on the context. Without contextual cues as additional information, the sentence "*What has gone wrong?*" does not raise any flags for the deep learning models to indicate a metaphorical meaning. This may be further supported by the additional information provided by the semantic richness features. For example, the word *gone* has a low concreteness score of 1, a neutral emotion, a high imageability score of 3, a low sensory experience score of 1, and a low body-object interaction score of 1. These scores align with those assigned to the word *wrong* [1, neutral, 3, 2, 1]. However, it is possible that the slight shifts in concreteness and sensory experience scores were not captured during the detection process.

Words like *push*, *revive*, and *rotate* were among the most incorrectly predicted metaphorical words in MOH-X files. The context in which the word *push* is used in the sentence "*The salesman is aggressively pushing the new computer model.*" is metaphorical, not in the literal sense of *the physical exertion of force on someone or something*. It signifies that the salesman is *promoting or advertising* the new computer model in an *assertive or compelling manner*. The misprediction could be a consequence of using the adverb *aggressively* which is compatible with the literal sense of the word *push*, in that both point at something physical. Additionally, the word *push* is commonly associated with tangible objects, which is the

computer mentioned in the sentence. Moreover, *model* here does not necessarily refer only to software but can also encompass an *upgraded physical design*. Additionally, the assigned scores for the words are also consistent, with minimal shifts between the three words: *aggressive* [2, anger, 3, 3, 1], *pushing* [3, neutral, 3, 2, 2], and *computer* [4, neutral, 3, 4, 6]. While there is a slightly larger shift in sensory experience and body-object interaction scores between the words *pushing* and *computer* compared to the previous examples, both words still convey a sense of physicality. In short, the agreement of the scores and with common usage of the contextual words resulted in the wrong prediction.

The most commonly incorrectly predicted metaphorical words in the TroFi dataset is *step*, that has [4, neutral, 3, 2, 5] as its scores, and the word *stick*, has [3, neutral, 3, 2, 1] as its scores. In addition to *step*'s multiple senses, it is frequently used in forming phrasal verbs. In the sentence "*To help make room, two members representing Newmont would step down.*" the phrasal verb *step down* indicates the act of *withdrawing* or *resigning from a position*. However, typically, the word *step* when considered alone notes the *physical act or movement of placing one foot in front of another*. The inclusion of the word *down* (scores are [3, neutral, 3, 2, 1]) adds a *direction* or *position* sense, aligning with the overall interpretation that the *movement is towards a lower or different position*. Moreover, their scores are compatible. The shift is minimal in the concreteness scores, while the difference in body-object interaction is bigger, the score by itself does not provide a clear differentiation between what is literal and what is not, it indicating the ease of manipulation.

Stick is another word that also presents challenges for these models for metaphor detection. Like *step*, it has multiple meanings and can be part of phrasal verbs. For example, in the sentence "*To me there is absolutely no fun being stuck within the same four walls every day.*", the word *stuck* means to *be fixed in a particular position or unable to move or be moved*. The sentence could mean literal sticking as the walls (that has [4, neutral, 3, 3, 5] as its scores) are objects used for blocking. Another sense for the word *stick* is in

the sentence "*He preferred to stick to instant photography.*" means to *remain in a specified place or situation.* As for the sense in the sentence "*If the charges stick, it could suggest that some concepts of freedom are a bit too broad.*" means *become convincing, established, or regarded as valid.* This is also supported by the scores assigned to the word *charges* ([2, anger, 3, 2, 1]), where the shift in concreteness is minimal. However, the change in emotion, despite the similarity in other scores, may have been neglected and resulted in an incorrect prediction. In all these examples, particularly when words have multiple meanings or are frequently used in metaphorical expressions, proper context and additional information play a significant role in distinguishing the intended metaphorical senses.

The correctly predicted metaphor files for the VUAMC dataset indicate that the word *make* is among the words that were accurately predicted. Consider the sentence "*Local people mutter and march, make speeches and throw things.*" In this sentence, the phrase *make speeches* refers to the act of uttering words using one's voice. The word *make* itself can be associated with activities like putting parts together or causing something to exist through manufacturing. While *make speeches* is a commonly used phrase, the detection models correctly identified this sentence as a metaphor. The accurate prediction could be attributed to the observed relation, enforced by explicitly added to the vectors, between the word *make* with scores of [2, neutral, 3, 2, 1] (related to tangible, manipulated, and imaginable concepts) and the word *speech* with scores of [3, neutral, 3, 2, 2] (related to abstract, audible, and invisible concepts) within the context.

The correct prediction files for the MOH-X dataset reveal that the word *swallowed* is one of the top correctly predicted words. In the sentence "*I swallowed my anger and kept quiet.*", the usage of the word *swallowed* here follows a similar logic to the word *make* in the VUAMC dataset. While *swallow* typically refers to *the physical action of something passing through the throat*, in this context, it is metaphorically connected to the intangible concept of *anger*. This connection is explicitly captured in the vector representation by incorporating

the semantic richness element, allowing the model to recognise the metaphorical usage of the word accurately.

TroFi's most correctly predicted metaphorical word is *roll*, which has various meanings depending on the context, but the primary meaning is related to *movements of turning over and over*. However, when it is a part of a phrasal verb, the known meanings of the word *roll* by itself are very different. In this example, "*As the first issue was rolling off the press, Mr. McGovern threw a party in Moscow and eagerly leafed through the magazine.*" The phrase *rolling off* the press refers to the process of an issue of a magazine or a newspaper being completed and printed. Similarly, in the sentence "*Still, the buses roll on.*" the phrasal verb roll on indicates that the buses *continue to operate or move forward despite any challenges*. The example "*When the TV cameramen motion for him to run closer to the water, Sen. Gore readily complies, the waves rolling up to his feet.*" describes the action of the waves *approaching and reaching* Sen. Gore's feet in a *rhythmic manner*. The *roll over* phrase suggests a *smooth, continuous movement*, which conveys the image of the waves *gradually and gently moving* towards Sen. Gore. The word *roll* is related to tangible objects that can be sensed and manipulated by the senses, and that can draw an image to what is being rolled; however, most of these examples refer to something different that the rolling is metaphorical, which is the end result of the prediction.

The previous examples discussed above reveal a consistent pattern observed throughout the previous chapters that examined these elements individually. It becomes evident that the accurate detection of metaphors heavily relies on selecting the appropriate sense that conveys the intended meaning when using a specific word in a given context, as in "*They held up the bright new diesel buses and, even worse, blocked the progress of private motorists*" and "*She had listened to Arlene's own voice and set about imitating it for she held her mentor in the highest esteem.*". This pattern is observed across all three datasets, with the VUAMC dataset containing longer sentences (the longest being 116 words) compared to

MOH-X (with the longest being 18 words) and TroFi (with the longest being 95 words).

GloVe provides a static representation of words regardless of the intended sense because it is derived from word co-occurrence statistics, which makes it non-contextual. Although ELMo considers the context, it still falls short in providing accurate senses, especially for metaphorical sentences where words agree, to some extent, with each other, such as in the phrase *step down* where it implies *movement in a particular direction.* Even in the example mentioned above, holding an object *bus* is a plausible interpretation, which might be why the model incorrectly predicted it as literal. Additionally, the added semantic elements are consistent. The word *hold* has a concreteness rating of 2, a neutral sentiment, a rating of 3 for imageabilty, 2 for sensory experience, and 2 for body-object interaction. The word *up* has ratings of [3, neutral, 3, 2, 1].

One solution could be to use longer sentences to train the SVMs. While it is true that BERT embeddings are better at capturing context, the training was still conducted on single words without context. Therefore, the full capabilities of BERT were not fully utilised. Moreover, the SVMs' predictions remain the same for the same word across different contexts. Adaptable predictions that are context-dependent could solve this limitation. For instance, *hold* as *carry* or *stop* would have higher concreteness, imageability, sensory experience, and body-object interaction, while *hold* in the second sentence, which implies *respect*, would have different scores for these five elements that reflect the intended sense.

6.5 Conclusion

This chapter is an integration of all previously discussed elements. Chapter 2 discussed the concreteness/abstractness element that focuses on the tangibility of concepts. One theory for metaphor is that in order to convey the meaning of a metaphorical sentence, an abstract

concept is described with a concrete one. Therefore, a shift in concreteness/abstractness score within the context of a sentence could indicate metaphorical usage. Chapter 3 covered a more subjective concept, which is sentiment. Theories on metaphor suggest that metaphorical sentences carry a degree of emotion. Thus, this chapter tested the effect of adding sentiment to the metaphor detection models. Following that is Chapter 4, which explains the concept of mental imagery and image-schemes, where certain concepts are associated with sensory involvement and can be imageable. The theory suggests that some difficult-to-understand concepts may be connected to imageable words and can also be understood through sensory experiences. Lastly, Chapter 5 covers body-object interaction, which has not been tested before in metaphor detection. This concept examines the ease with which concepts can be manipulated and interacted with. It is derived from concreteness and imageability. Therefore, the argument states that a concept with a high score, indicating ease of manipulation, describing a concept with a low score of manipulation, could hint at a metaphor.

This chapter combined all these elements into one experiment, testing metaphors at the sentence level using a classification model and at the word level using a sequence labeling model. These tests were done on three metaphorical datasets: VUAMC, MOH-X, and TroFi. The results show overall performance improvements across the three datasets. Similar results were observed for the individual tests. One problem encountered was assigning the prediction with the correct word sense (polysomy) intended in the sentence. GloVe is a context-independent word representation, while ELMo has some limitations in capturing the correct senses, especially in shorter or rare sentences. BERT representation could be used since it considers the context both to the left and right. The next chapter summarizes this dissertation, discusses the limitations found, and suggests possible future work.

Step	lr	hidden size	Batch size	# epochs	# lyres	drop 1	drop 2	drop 3
VUAMC								
Step_1	0.01	64	64	20	1	0.3	0.2	0.2
Step_2	**0.01**	**64**	**64**	**20**	**2**	**0.3**	**0.2**	**0.2**
Step_3	0.01	128	64	20	1	0.3	0.2	0.2
Step_4	0.01	256	64	20	1	0.3	0.2	0.2
Step_5	0.01	64	64	20	1	0.3	0.2	0.2
Step_6	0.01	256	64	20	1	0.3	0.2	0.2
MOH-X								
Step_1	0.001	300	10	15	1	0.2	0	0.2
Step_2	**0.001**	**256**	**10**	**15**	**1**	**0.2**	**0**	**0.2**
Step_3	0.001	128	10	15	1	0.2	0	0.2
Step_4	0.001	64	10	15	2	0.2	0.2	0.2
Step_5	0.001	256	10	15	3	0.2	0	0.2
Step_6	0.001	64	10	15	1	0.2	0	0.2
TroFi								
Step_1	0.001	300	5	15	2	0.2	0	0.1
Step_2	0.001	300	5	15	3	0.2	0.2	0.1
Step_3	0.001	300	5	15	2	0.2	0	0.1
Step_4	0.001	300	5	15	1	0.2	0	0.1
Step_5	**0.001**	**300**	**5**	**15**	**2**	**0.2**	**0.2**	**0.1**
Step_6	0.001	300	5	15	2	0.2	0.2	0.1

Table 6.3: These hyperparameters were the configuration used to generate the results of the classification experiments presented in Table 6.2 across the three datasets. The step indicated in bold highlights the parameters employed in that particular step that led to the best performance on each respective dataset.

Metrics	Mao et al. (2019)	Baseline	Set_1	Set_2	Set_3	Set_4	Set_5	Set_6
VUAMC								
Precision	71.8%	*79.58%*	72.104%	71.426%	77.69%	74.11%	75.662%	70.592%
Recall	76.3%	*64.45%*	76.59%	**77.078%**	70.51%	72.481%	71.856%	75.748%
F1	74.0%	*71.22%*	**74.28%**	74.144%	73.92%	73.291%	73.71%	73.079%
MOH-X								
Precision	79.7%	**76.408**	78.43%	77.01%	78.953%	79.085%	77.181%	77.944%
Recall	79.8%	*81.63%*	79.825%	**83.731%**	83.46%	82.614%	82.401%	80.608%
F1	79.8%	**78.46%**	78.794%	79.869%	**81.007%**	80.521%	79.077%	78.967%
TroFi								
Precision	67.4%	*68.98%*	67.615%	67.176%	68.446%	68.536%	68.323%	66.582%
Recall	77.8%	*74.489%*	78.401%	78.506%	77.219%	77.92%	76.893%	**79.914%**
F1	72.2%	*71.575%*	72.515%	72.35%	72.436%	**72.62%**	72.163%	72.521%

Table 6.4: The results are compared to the reported results and baselines built on Mao et al. (2019) sequence labelling model for each dataset. These experiments test the VUAMC, MOH-X, and TroFi with all the semantic richness elements based on the score's source: the lexical list or the SVM/BERT prediction. The 10-fold cross-validation procedure was applied to get the MOH-X and TroFi results.

Step	lr	hidden size	Batch size	# epochs	# lyres	drop 1	drop 2	drop 3
VUAMC								
Step_1	**0.1**	**64**	**2**	**29**	**1**	**0.5**	**0**	**0.2**
Step_2	0.2	256	2	29	1	0.5	0	0.2
Step_3	0.2	128	2	29	1	0.5	0	0.2
Step_4	0.2	256	2	29	1	0.5	0	0.2
Step_5	0.2	256	2	29	1	0.5	0	0.2
Step_6	0.2	64	2	29	2	0.5	0.2	0.2
MOH-X								
Step_1	0.001	128	2	20	1	0.5	0	0.1
Step_2	0.001	256	2	20	2	0.5	0	0.1
Step_3	**0.001**	**256**	**2**	**20**	**2**	**0.5**	**0.2**	**0.1**
Step_4	0.001	256	2	20	2	0.5	0	0.1
Step_5	0.001	256	2	20	2	0.5	0.2	0.1
Step_6	0.001	300	2	20	1	0.5	0	0.1
TroFi								
Step_1	0.001	128	2	20	2	0.5	0	0.1
Step_2	0.001	128	2	20	2	0.5	0.2	0.1
Step_3	0.001	300	2	20	1	0.5	0	0.1
Step_4	**0.001**	**300**	**2**	**20**	**1**	**0.5**	**0**	**0.1**
Step_5	0.001	128	2	20	2	0.5	0.2	0.1
Step_6	0.0001	128	2	20	2	0.5	0	0.1

Table 6.5: These hyperparameters were the configuration used to generate the results of the sequence labelling experiments presented in Table 6.4 across the three datasets. The step indicated in bold highlights the parameters employed in that particular step that led to the best performance on each respective dataset.

Chapter Seven

Conclusion

7.1 Introduction

Automatic metaphor detection poses a significant challenge due to the nature of this type of figurative language. Numerous researchers have endeavored to define metaphor, establishing the stages of creating, detecting, and understanding metaphors, while also distinguishing between basic, literal, and metaphorical meanings. Some of these theories emphasise the grammatical aspects, while others address the semantic perspective. Metaphors find their way into everyday communication, not exclusively for artistic reasons in literature, but also to express opinions, feelings, or ideas. This, coupled with the three states of metaphors in language (dead, conventional, and novel), amplifies the difficulty in detecting and interpreting them. All of these theories highlight the challenges faced by the field of Natural Language Processing (NLP) and its applications.

One theory discusses the creation of metaphors. It states that it involves employing words from one concept within the context of another concept. For instance, the metaphor "*he exploded at me*" uses the term *explode* to convey his state of *shouting angrily*. A connection exists between the state of *anger* and *explosion* because when one is angry, blood

pressure and temperature increase, the skin turns red, and other signs resemble, to some extent, a detonating bomb.

This thesis aimed to explore a semantic concept known as *semantic richness* and its elements within the context of the metaphor detection task when integrated with neural networks as external resources. These elements are: *concreteness/abstractness measure*, *sentiment*, *imageability*, *sensory experience* and *body-object interaction* (Findlay and Carrol, 2018). The investigation encompassed the examination of each element as an additional feature for word-level and sentence-level neural networks, enhancing metaphor detection.

7.2 Semantic Richness

Semantic richness pertains to the interlinked information within a concept, shaping a semantic network that represents the concept's meaning within the semantic space (Pexman et al., 2007; Findlay and Carrol, 2018). This network encompasses features that contribute to the concept's meaning, including two components tied to the network's strength measurement (such as the number of features, senses, first associates, and neighborhood density) and the network's perceptual aspect (including imageability, body-object interaction, sensory experience, and sentiment) (Findlay and Carrol, 2018).

Previous research in NLP primarily delved into the first component, where these features rely on numerical conversions of word counts, frequency, and other statistical formulas. Within the domain of metaphor detection, considerable attention has been given to the number of senses (Mason, 2004; Wilks et al., 2013; Mohler et al., 2013; Klebanov et al., 2016). Studies on semantic neighborhoods have centered around the proximity of words in the semantic space—words with akin meanings tend to be closely located. Numerous researchers have leveraged meaning theories in the context of metaphor detection (Goldberg, 2017; Pen-

nington et al., 2014; Do Dinh and Gurevych, 2016; Stemle and Onysko, 2018; Bizzoni et al., 2017; Haagsma and Bjerva, 2016; Tsvetkov et al., 2013; Köper and Walde, 2017a). The third element in Table 1.1 in Chapter 1 pertains to word associations, indicating that words often exhibit contextual associations with other words (Duñabeitia et al., 2008). Lastly, the number of features focuses on polysemous words and the count of synonyms (Pexman et al., 2002). A comprehensive discourse concerning these facets of semantic richness is covered in Chapter 1.

This thesis examined the effect of incorporating the elements of the perceptual part of the semantic richness on the neural network models for metaphor detection. This was achieved by including these elements as ratings within the word/context vectors provided by GloVe and ELMo, serving as additional features. The goal was to assess whether augmenting the contextual information with additional details would enhance the semantics of the analysed sentences and improve the models' ability to determine their metaphoricity.

The ratings are primarily evaluated in two formats: single-digit values and probability distributions. These ratings are generated as predictions from a Support Vector Machine (SVM) that has been trained on resources related to these elements. These resources consist of a list of words rated according to their corresponding element. Following the training phase, the SVM assigns predictions to each word within the metaphorical datasets that are used for these experiments.

7.3 Research Questions

One theory of metaphor relies on the broader context of a sentence to comprehend the intended meaning. Given that words from one domain are employed in different domains for specific purposes, grasping this implicit intention necessitates focusing our comprehension

of the context in which these words appear with our knowledge of the words themselves and their respective contexts. Neural networks have showcased their capacity to unearth concealed information devoid of the need for external, manually crafted attributes. At the very least, neural networks require the word meanings and the context to reach a decision. The hypothesis was constructed upon this foundation, as any alterations induced by the addition of features to this architecture could be observed. This study endeavored to address these inquiries:

RQ1 What are the connections between metaphor models, theories, and human-derived perceptual semantic richness, and how do they relate to the development of deep learning-based models of metaphors?

RQ2 Is it possible to utilise these perceptual semantic richness resources to create and test deep-learning metaphor models and correctly detect metaphors in the natural text?

RQ3 Finally, What can these models teach us about the function, distribution, and use of metaphor in natural language?

The baselines of the models used in this study, both word-level (based on Mao et al. (2019) named RNN_HG) and sentence-level (based on Gao et al. (2018)), rely on the vector representation of a word's literal meaning provided by GloVe and the word's context provided by ELMo to detect metaphors. These models have demonstrated their ability to extract the necessary contextual information to make accurate determinations. Each semantic richness element is incorporated as an additional feature to these vectors, enhancing the decision-making process of the metaphor models. Additionally, these elements are combined in experiments to assess their collective impact on the detection task. These elements may contain crucial information related to metaphors and non-metaphors, potentially assisting neural networks in their detection task.

7.4 Models and Datasets

It is reasonable to choose models that employ word meanings as the basis for detection as a starting point to ensure that the additional features are responsible for any changes in performance. Additionally, it was important to investigate the impact of these features on both word-level and sentence-level metaphors. Therefore, each experiment was divided into two parts: one related to sentence classification, using a baseline based on the (Gao et al., 2018) model, and the other focused on labeling word sequences, using the baseline based on (Gao et al., 2018; Mao et al., 2019) as the baselines. In both cases, the models classify targets (sentences or words) as *metaphors* by assigning a prediction of 1, and *non-metaphors* are assigned a prediction of 0. The results were calculated for metaphor detection performance, including precision, recall, and the F1 score. For training and testing, the VU Amsterdam Metaphorical Corpus (VUAMC) dataset (Tighe, 2010), annotated for both sentence and word-level metaphors, was used. Additionally, two other datasets, MOH-X (S. Mohammad et al., 2016) and TroFi (Birke and Sarkar, 2006), were employed to ensure generalization and mitigate overfitting and bias issues. The models rely on GloVe (Pennington et al., 2014) for word basic meaning and, in some cases, ELMo (Peters et al., 2018) for context meaning as the vector representations for words, which are provided as input to the model.

7.5 The Experiments

This section summarises the experiments on semantic richness features, including datasets and results. It provides an overview of the feature's definition and hypothesis and analyses key points.

7.5.1 Concreteness

Concreteness measurement is a very popular concept used concept in NLP experiments. It evaluates words based on their detectability and perceptibility. The higher the rating a word receives, the more concrete it is considered. This measurement is rooted in one theory of metaphors known as the Conceptual Metaphor Theory (CMT) (Lakoff and Johnson, 1980), which states that a sentence is a metaphor when its concept is understood by using words from another concept. This becomes apparent when an abstract concept is referred to using concrete terms from another concept. For instance, in the phrase "*you are wasting my time*", the concept of **TIME** (target domain) is mapped to **MONEY** (source domain) when the word *wasting* is used to convey the idea that ***TIME*** *is precious and should not be wasted like* ***MONEY***.

In this context, the connection between the abstract and concrete was the basis for conducting experiments with metaphors using this feature. It was hypothesised that a shift in concreteness ratings within the context of words is a sign of a metaphor. Concreteness ratings could provide additional insights into potential metaphors, assisting the detection task. Two forms of predicted ratings, provided by trained SVMs on the dataset from Brysbaert et al. (2014), were investigated.

For each experiment, every token in the three metaphorical datasets received both a single-digit rating and a 4-digit probability distribution predicted by trained SVMs. These assignments were then concatenated with the tokens' GloVe+ELMo representations and used as input for the neural networks.

One of the experiments in this chapter is slightly different as it involves a SharedTask competition for metaphor detection organised by Leong et al. (2020). In this competition, the VUAMC dataset provided differs from the VUAMC dataset used in all other semantic

richness tests, as described in Chapter 2, Section 2.4.1. The results from this experiment have been published in the paper by Alnafesah et al. (2020).

In Chapter 2, Section 2.4.2, the tables display the results of each model when the ratings were tested. The classification model with the VUAMC_18 dataset achieved its best performance when the rating took the form of a probability distribution. While precision decreased by 3.67%, recall and F1 increased to 62.4% and 58.04%, respectively. Similarly, for the MOH-X dataset, the F1 score increased to 77.175% from the initial 76% when the model was tested using the probability distribution. This improvement resulted from an increase in precision by 1.53% and recall by 1.46% compared to the baseline. In contrast, TroFi showed the best recall and F1 scores in experiments that tested the single-digit predicted concreteness rating. While there was a slight decrease in precision by 0.8%, recall increased by 2.65%, resulting in a balanced F1 score of 71.823%.

The probability distribution prediction case also benefited the sequence labeling model (SEQ_18 (Gao et al., 2018)) on the VUAMC_2020 dataset provided by the SharedTask competition. Precision increased by 0.6%, recall by 12.06%, resulting in a significant improvement of F1 by 6.8%. Similarly, VUAMC_18 showed the best results with SEQ_19 (Mao et al., 2019), the *RNN_HG* model, when the model tested the probability distribution rating. The 3.58% decrease in precision was offset by a significant increase in recall by 10.76%, resulting in an F1 score of 73.96%, up from the initial 71.22%. Furthermore, the probability distribution rating with the MOH-X dataset provided the best results. Precision slightly increased by 0.7%, and recall increased by 0.8%, resulting in a 1.2% increase in F1. However, the TroFi dataset benefited more when the model was tested with the single-digit rating. Precision increased by 0.17%, recall by 0.65%, resulting in an F1 that increased by 0.34% over the baseline.

These numbers suggest that the concept of concreteness plays a role in metaphors,

addressing RQ1. The neural classification for metaphor sentences and sequence labeling for metaphor words demonstrate the ability to extract valuable contextual information to make decisions across the three metaphorical datasets, providing an answer to RQ2. The concreteness experiment answered RQ3 by showing that each word's literal meaning dictates its concreteness rating. While this makes sense in some cases, it falls short with words that come in pairs or more, such as phrasal verbs, whose meaning differs from their constituent parts. Further observations are discussed in Section 7.6.1, and recommendations are explained in Section 7.6.2.

7.5.2 Sentiment

According to many research studies, metaphors convey emotions and attitudes. For example, the metaphorical sentence "*this movie stinks*" communicates the person's *dissatisfaction* with the movie. *Stinks* is a negative known feeling; if something *stinks*, then it is *terrible*. Detecting the sentiment of metaphors is an ongoing study. However, there has not been much research about using sentiment to detect metaphors. Therefore, sentiment was investigated as additional linguistic information for the neural metaphor models. The detection with two forms of sentiment was tested: *polarity*, where words were labeled as either negative or positive, and *emotions*, where words were labeled with one of eight emotions.

The experiments followed the same procedure and the steps applied in the concreteness study. Each type of sentiment was trained an SVM on the NRC list (S. Mohammad and Turney, 2010) for polarity and eight emotions. Each sentiment type was tested separately on the three metaphorical datasets. The effect of each emotion individually was also observed. The impact of the predictions' forms (single and probability distribution) on the detection models was examined. Additionally, a fine-tuned BERT sentiment model for sentence classification assigned a new emotion prediction (Garbas, 2019). This model assigned one of

five emotions after being trained on Ghazi et al. (2015), Y. Li et al. (2017), and Scherer and Wallbott (1994)

In Chapter 3, the steps under Section 3.5 explain the steps that the experiments followed to test different types of sentiment forms (i.e., polarity and various emotions) using the NRC dataset and the BERT fine-tuned model for sentiment classification. The information presented in this section reveals that different sentiment types, as well as the source of predictions (the lists, SVM, or BERT), have distinct effects on the metaphorical datasets.

For instance, the NRC list's *polarity*, *fear*, and *anticipation* performed better on the VUAMC dataset when using the classification model. *Polarity* yielded the best results, with an increase of 6.21% in F1 over the baseline, accompanied by a 15.05% increase in recall and a slight decrease in precision by 2.09%. *Polarity* also had a positive impact on the MOH-X dataset. Similar to the VUAMC dataset, there was a slight decrease in precision from 75.44% to 74.88%, but recall increased by 5.1% over the baseline, resulting in a 2.3% increase in F1 score. The TroFi dataset demonstrated the best results with the *anger* emotion. While the precision decreased slightly by 0.2%, the recall increased by 1.6%, resulting in a 0.9% increase in F1 score.

The comparison of the results from the BERT fine-tuned model for sentiment classification to those from the NRC list across the three metaphorical datasets is intriguing. The BERT model is trained on sentences that are annotated for one of the five emotions, and the assigned predicted emotion is done at the word level. The testing revealed that all three metaphorical datasets exhibited improvements over the baselines. In the case of the VUAMC dataset, precision showed a decrease of 6.5%, but this was offset by a significant increase in recall, marking a 33.2% improvement. This balanced the F1 score, which reached 59.297%, a boost of 10.8%. For the MOH-X dataset, precision, recall, and F1 score all increased compared to the baseline. Precision saw a very slight increase of 0.1%, recall increased by 4%,

and the F1 score improved by 2.2%. In the case of the TroFi dataset, precision decreased by 2.3%, and recall decreased by 2.9%, resulting in a minor 0.35% decrease in the F1 score.

The sequence labeling model, referred to as *RNN_HG*, exhibited a similar pattern. The use of the NRC list for sentiment prediction improved the results across all three metaphorical datasets. For the VUAMC dataset, both *polarity* and the *anger* emotion yielded the best results. The F1 score improved to 74.42%, with recall increasing to 74.48%, and precision decreased slightly to 72.76%. The MOH-X dataset benefited from the *anticipation* emotion, resulting in a 2.35% increase in F1 score, driven by a 2.03% increase in recall and a 2.21% increase in precision. On the other hand, the *surprise* emotion had a positive impact on the results for the TroFi dataset. Precision increased by 1.99%, recall increased by 0.84%, leading to a 1.46% increase in the F1 score. The results of the BERT model with the sequence labelling model showed the similar results of the classification model. In all three metaphorical datasets, there was a notable improvement in results compared to the baseline.

Experiments with *polarity* and *emotions* demonstrated that these features are inherent in metaphors, answering both RQ1 and RQ3. Both models, the classification and sequence labelling neural networks benefited greatly from the addition of sentiment to the detection task. In contrast to concreteness ratings, where rating shifts were more prominent, sentiment played a crucial role in labeling, highlighting the connection between semantic richness and metaphors, in addition to advancements in model development (answering RQ2). Furthermore, it's important to note that the NRC dataset is significantly smaller compared to the size of the concreteness list. Additionally, almost three-quarters of the words in the NRC and VUAMC datasets were labeled as *neutral*. Some sentences contained only one word with a sentiment rating other than *neutral*. This observation also addresses RQ3, indicating a link between sentiment and metaphor. Some of the misclassified sentences likely lacked enough labeled words with sentiment other than *neutral*. This made it challenging for the models to

capture the metaphorical information provided by sentiment. Further analysis is discussed in the next section, Section 7.6.1. Recommendations for future work are presented in the following section, Section 7.6.2.

7.5.3 Sensory Experience and Imageability

Imageability refers to the ability of words to potentially evoke mental images. Additionally, these terms are associated with having a certain degree of sensory experience, where additional sensations are elicited. Consequently, sensory experience refers to the extent to which a word can evoke a sensory and/or perceptual response in the mind (Juhasz and Yap, 2013).

For example, the word *incense* could form a mental image of what is commonly known as *incense*, which can also be *smelt*. Lakoff and Johnson (1980) further discussed how we grasp these conceptual metaphors. As previously demonstrated, based on the Conceptual Metaphor Theory (CMT), a metaphor maps a concrete concept to an abstract one during its formation. From this mapping, some image-scheme knowledge gained from sensory interactions can be drawn (Lima, 2006; Velasco, 2001).

For instance, consider the metaphorical sentence "*I don't know which way to go in life*", which connects the image of **MOVING_IN_A_DIRECTION** with **DECIDING**. The source domain, from which the words are drawn, is mapped to the target **DECIDING** to convey that when *choices* are made, *paths* in life open, and
textitwalks walking is a must on the chosen road. Therefore, it is reasonable to utilise image-related linguistic resources to assess whether metaphors indeed evoke mental images and are accompanied by perceptual experiences. Additionally, it is worth exploring whether neural networks can leverage this information to enhance detection capabilities. The sensory experience, to the date of conducting these experiments, have not been tested as the main

feature for metaphor detection with deep learning models. The experiments are published in the paper Alnafesah et al. (2023).

The experiments involving these features followed the procedure for preparing the datasets (imageability (Ljubešić et al., 2018) and sensory experience (Juhasz and Yap, 2013)) to train the SVMs. Each model was tested separately with one feature. The testing involved assigning predictions to the VUAMC, MOH-X, and TroFi tokens. The predictions were added to the GloVe and ELMo vectors as single-digit and multi-digit probability distributions. The tables in Chapter 4, Section 4.5.2, show the imageability and sensory experience results with the classification and sequence labeling models.

The VUAMC dataset showed better results with the single-digit imageability ratings from the original dataset when using the classification model. Precision decreased by 3.03%, offset by a 10.65% increase, resulting in a 3.6% increase in F1 score. For the MOH-X dataset, the results were better with the SVM's single-digit predicted rating, showing a 0.9% increase in precision and a 1.3% increase in recall, resulting in a 1.38% increase in F1 score. The TroFi dataset benefited the most from the probability distribution rating. Although precision decreased slightly by 0.63%, recall increased by 3.9%, resulting in a 1.68% increase in F1 score. Regarding sensory experience, the VUAMC dataset showed better results when the tested rating was the list-extracted rating, as was the case for the MOH-X dataset. The metrics increased by 1.6% for precision, 3.6% for recall, and 2.37% for F1 score. On the other hand, TroFi benefited from the SVM's single-digit predicted rating. Precision slightly decreased by 0.8%, but recall increased by 3.9%, resulting in a 1.5% increase in F1 score.

The sequence labeling model, when tested with imageability and sensory experience, also demonstrated performance improvements. For the VUAMC dataset, the single-digit predicted imageability rating yielded the best results. Although precision decreased by 7.02%, recall increased by 14.9%, resulting in a balanced F1 increase of 3.9%. Similarly,

the VUAMC dataset benefited from sensory experience. Precision decreased by 1.8%, but recall increased by 8.6%, resulting in a 3.6% increase in F1 score. In the case of the MOH-X dataset, imageability ratings showed the best results with the probability distribution. Precision increased to 76.408%, and recall slightly decreased to 80.95%, resulting in an F1 score of 80.539%. However, sensory experience benefited the most from the single-digit rating extracted from the list. Precision increased to 78.82%, while recall decreased to 80.053%. This resulted in a 1.01% increase in F1 score. For the TroFi dataset, imageability as a single-digit rating without the involvement of the SVM showed better results. Precision decreased by 4.78%, recall increased by 8.64%, and F1 increased by 1.12%. On the other hand, sensory experience showed better results with the single-digit rating from the SVM. Precision slightly increased by 0.54%, recall increased by 1.47%, and F1 increased by 0.88%.

Overall, the performance results showed that these two elements had a positive impact on the detection task over the three metaphorical datasets. It is true also that some form of raring performed worse. This could be attributed to the varying rating scale, where words were rated from 1 to 3 for the imageability, compared to the 6-range scale for the sensory experience, The small range may not capture the nuanced differences between words, which are crucial for understanding the connection between the semantic richness feature and metaphor, addressing both RQ1 and RQ2. The presence of such shifts in semantic richness was evident in these models' results, which addressed RQ1 and RQ2. These findings indicate that imageability and sensory experience ratings are inherent concepts in metaphors, providing an answer to RQ3. Further analysis details on this are covered in Chapter 4. More analysis about the observed limitations are presented in the following section, Section 7.6.1 with the future work recommendations that addressee these observations in Section 7.6.2.

7.5.4 Body-Object Interaction

The final element of the semantic richness perceptual aspect is the body-object interaction, which was created to reflect how easy it is to physically interact with a word's referent (Siakaluk et al., 2008). Body-object interaction ratings were derived from concreteness and imageability information. Words with high body-object interaction ratings are easier to handle physically, such as *books*. In contrast, lower ratings indicate the difficulty of physical interaction with words, even if they are perceivable and detectable by the senses, as in the case of a *mountain*. This feature supports the idea that prior physical experiences and knowledge can facilitate the understanding of conceptual metaphors. For example, in the sentence "*this movie stinks*", the knowledge gained from *rotten food stinks* is projected onto the movie's quality. This rating can potentially enhance metaphor detection using neural networks.

The experiments that utilised the body-object interaction features followed a specific procedure discussed in Chapter 5. Words in the dataset (Pexman et al., 2019) were rated on a scale from 1 to 6 for body-object interaction. A higher rating indicates that the word is easier to physically interact with, while a lower rating suggests greater difficulty in physical interaction. Two rating forms were tested using both the classification and sequence labeling models, the RNN_HG. Initially, the models tested metaphor detection by incorporating a single-digit prediction generated by an SVM. Subsequently, the SVM produced a probability distribution prediction with six digits. These additional ratings were concatenated with the GloVe+ELMo representations of the VUAMC, MOH-X and TroFi words. The models leveraged the information provided by both the literal meaning, contextual information and the ratings to make classification decisions. The results are published in the paper Alnafesah et al. (2023).

The results in Chapter 5, Section 5.4.2, showed that the effect of the body-object

interaction element on the VUAMC dataset, using the classification model, was negative. The three types of ratings (single without SVM, single SVM-predicted, and the probability distribution) all reduced the overall three metrics. Although the precision increased for the single without SVM rating, the recall significantly decreased, affecting the F1 score. However, the sequence labeling model showed a positive effect. The single without SVM rating had the best F1 score, which increased to 73.115%. The recall increased by 6.4%, and the precision slightly decreased by 1.6%.

In contrast, the MOH-X dataset showed a positive effect using the two models. The single-digit SVM predicted rating yielded the best results with the classification model, with precision increasing by 1.3%, recall by 1.2%, and the F1 score by 1.4%. This was also observed in the case of the sequence labeling model, where the single-digit SVM predicted rating improved all three metrics: precision increased by 1.2%, recall by 1.9%, and the F1 score by 1.6%.

Similarly, the single-digit, non-SVM predicted rating provided the most significant improvements for the TroFi dataset using the two models. In the classification model, precision slightly decreased by 2.1%, which was countered by an increase in recall by 4.8%, resulting in an F1 increase of 1.25%. On the other hand, the sequence labeling model showed that precision was not affected, remaining at 68.98%, while recall slightly increased by 1.5%, resulting in an F1 increase of 0.68%.

These results demonstrate that body-object interaction, as a feature, has a role in enhancing the detection model (RQ2). In the classification model portion of the test results on the VUAMC dataset, there was a performance decrease in all three metrics, which is in contrast to the other two datasets where the opposite trend was observed. Conversely, the results of the sequence labeling model were generally positive across all three metaphorical datasets. These observations support the claim that this feature contributes to and assists

neural metaphor detection models to some extent (answering RQ3). More detailed analyses can be found in Section 5.4.5. Additionally, it addressed RQ1 by showing that the nature of body-object interaction is compatible to some degree with the nature of metaphor. The model results were consistent with suggested metaphor theories, particularly those involving semantic richness in abstract concepts, sentiment, and imagery/sensory nature. Further discussions about the limitations faced in these experiments can be found in the next section, Section 7.6.1. Recommended solutions for future work are considered in Section 7.6.2.

7.5.5 Integrating the Different Semantic Elements

Deep learning models have excelled in various natural language processing (NLP) tasks. Word embeddings are one tool used to capture word information based on how they are used in data, providing valuable insights into their meanings. However, a full understanding of language requires more than just embeddings; it demands additional linguistic, real-world, and common-sense knowledge. By incorporating these additional details, such as linguistic and semantic information, the gap can be reduced. Different language resources serve distinct purposes, enhancing language comprehension by addressing various aspects and overcoming individual limitations. For example, Parts of Speech (POS) provide grammatical information and reveal word relationships in text. In previous chapters, each one focused on specific features, ranging from semantic richness to perceptual part, to test metaphor theories and enhance automatic detection with neural networks. In this chapter, Chapter 6, these features are combined into a single test: one for classification and another for sequence labeling. The models underwent three tests, including scenarios involving BERT fine-tuned for classifying emotions, resulting in a total of six tests, as explained in Section 6.3.

In Set 1, using the classification model provided positive results for the MOH-X and TroFi datasets, improving their performance. The F1 score for MOH-X increased by 1.5%,

attributed to an increase in precision by 0.82% and recall by 2.21%. The TroFi dataset also showed a slight enhancement in F1, to 71.49%, an increase of 0.5%. Its recall increased by 3.12% but with a slight decrease in precision of 1.9%. The sequence labeling model demonstrated performance improvements across all three metaphorical datasets. In the case of VUAMC, results increased for recall by 18.8% and F1 by 4.29%, although precision decreased by 9.39%.

In Set 2, the classification model exhibited the best performance on the VUAMC dataset. Despite a 9.01% decrease in precision, there was a significant 29.8% increase in recall, resulting in an F1 increase of 7.92%. The sequence labeling model also showed improvement on the VUAMC dataset, achieving the highest recall with a 19.5% increase over the baseline. The improvement patterns observed for MOH-X and TroFi datasets were consistent for both models. In the case of the classification model with the MOH-X dataset, it achieved the best recall, with a 4.91% increase over the baseline. Furthermore, it had the highest F1 score, with a 1.69% increase over the baseline score of 76%.

In Set 3, the classification model demonstrated improvements across all three datasets, with increased recall and F1 scores for VUAMC, MOH-X, and TroFi. However, a precision increase was observed only for the MOH-X dataset, with a minimal increase of 0.8%. The sequence labeling model also exhibited enhanced performance. Notably, the MOH-X dataset achieved the highest F1 score in all sets involving this dataset, with a 3.24% increase. This improvement was accompanied by a 3.3% increase in precision and a 2.2% increase in recall.

In Set 4, all three datasets exhibited performance improvements with the classification model. The VUAMC dataset showed a 2.46% increase in F1, while the MOH-X dataset experienced a 1.02% F1 increase. Finally, the TroFi dataset demonstrated a 0.9% F1 increase. The sequence labeling model also displayed performance improvements across the three datasets, with the TroFi dataset achieving its best performance in this set. Despite

a slight 0.6% decrease in precision, there was a 4.6% increase in recall, resulting in an F1 increase of 1.46%.

In Set 5, all three datasets demonstrated positive results with both models. Notably, the TroFi dataset delivered its best recall and F1 performance among all sets that included this dataset. Despite a 2.8% decrease in precision, recall saw a significant 5.4% increase, resulting in a 1.13% F1 increase. The sequence labeling model showed similar positive results across the three metaphorical datasets, with all of them experiencing improvements.

In Set 6, the final set, the VUAMC dataset with the classification model exhibited a slight decrease in performance. Precision decreased by 2.11%, but with a 0.99% increase in recall, the F1 score only decreased by 0.5%. On the other hand, the MOH-X and TroFi datasets provided positive results for the classification experiments. MOH-X showed a comparably high F1 score, similar to Set 2, with an increase of 1.7% compared to the 1.69% increase in Set 2. The TroFi dataset also showed improvement across all three metrics. The sequence labeling model yielded better results across the three datasets. Notably, the TroFi dataset achieved its best recall among all sets, with a 7.2% increase resulting in a 1.32% increase in F1.

7.5.6 Summary

According to the definition of semantic richness, the meaning of a concept can be conveyed through various elements, some of which pertain to the perceptual aspect of the semantic richness network. In this thesis, these elements were utilised to enhance neural networks for the metaphor identification task. Two models were employed for two types of detection: sentence-level experiments were managed by the classification model based on the work presented by Gao et al. (2018), while word-level experiments were handled by the sequence

labeling model introduced by Mao et al. (2019), known as the RNN_HG model.

Each model employed Glove and ELMo embeddings as their representations for the VUAMC, MOH-X, and TroFi datasets. SVMs were used to provide prediction ratings for various semantic richness elements, including concreteness, sentiment, imagery, sensory experience, and body-object interaction. Overall, all experiments yielded positive results across the three metaphorical datasets.

The collective testing of these elements in a single experiment using the two deep learning models consistently showed improvement over their individual contributions. While some elements were marginally affected negatively, the overall results remains positive. These elements share positive observations, as discussed in their respective analysis sections.

However, for a more more in-depth and thorough discussion, the limitations and observations identified during the analysis are consolidated in the following section. Recommendations for future work aimed at addressing these issues are introduced in the subsequent section.

7.6 Observations, Limitations and Future Work

Metaphors are inherently complex, as they evoke an array of concealed ideas, emotions, and past experiences within the structure of a sentence. The unique structure of metaphors often strays from the conventional usage of language, thus granting them the power to elicit multitudes of interpretations and a wealth of meaning. These complexities present a challenge for Natural Language Processing NLP.

The experiments in this dissertation were build mainly for attempting to enhance the automated metaphor detecting using deep learning model by adding semantic richness as

additional lexical information for words. The theory is build on the assumption that these semantic richness hold information about the word, adding additional layer of knowledge, which could support the detection task. The results for individual elements testing as well as combined showed promising results. Some of these elements improved the F1 measure of the detection significantly such as concreteness and sentiment, even with others slight increase of F1, the recall that is the an indication of the correct metaphor prediction out of all correct prediction has improved. These overall improvement in both models is a sign that the models have picked up some clues aiding the detection of metaphors.

Additionally, poor results or the absence of expected outcomes have led to other perspectives on examining the low performance factors. The analysis explored different factors or variables that may have contributed to the situation, which facilitated the planning and the recommendations for future work. These factors revolved around various aspects: the metaphor models employed in the experiments, the metaphorical datasets utilised by these models, the lexical resources of the semantic richness elements used to train the SVM, the method by which it assigned predicted scores to tokens, the handling of embedding and contextualised information and the inherent challenges posed by polysemy.

7.6.1 Observations and Limitations

The field of NLP previously approached the task of metaphor detection through modeling the task as classification of sentences problem, or as a classification of words problem (i.e., sequence labelling), and, in some instances, experiments with phrase-level classification (Zayed et al., 2020b). In the task of sentence classification, the objective is to label the whole sentence as metaphor or non-metaphor, while the phrase level models classify phrases of syntactical relation such as verb-noun as metaphor or non-metaphor. The majority of metaphor-related tasks adopt a word classification approach, where the goal is to assign the

metaphorical or non-metaphorical label to individual words within the sentence. It worth noting that most state-of-the-art models employed in the field tend to focus on the sentence or word classification approaches. Consequently, the chosen models used to examine the influence of the semantic richness on the detection task reflects the earlier observation. The experiments that used the the sentence-level classification was build on Gao et al. (2018) and the tests that used the word-level classification was build on Mao et al. (2019), known as RNN_HG.

The adopting of specific model architectures directly influences the structuring of the dataset utilised in these models where the dataset has to be aligned with the input requirement and be formatted accordingly. The issue of metaphor dataset has been a key focus of discussions of metaphor-related research (Veale et al., 2016). The inherent ambiguity in precisely defining what constitutes a metaphor is at the root of the issue. On the one hand, the words of the frequently used metaphors (i.e., conventional metaphors), over time, are being gradually integrated into their literal meaning as an additional literal word sense. At that point, would these words still be categorised as metaphors? On the other hand, language is dynamic and constantly evolving, which introduces new words into everyday discourse. Consequently, the discovery of novel, imaginative metaphors presents a challenge in providing datasets that capture these creative linguistic expressions. Sourcing an up-to-date dataset for metaphors that are both new and creative enough proves to be hard and time-consuming. One proposed solution was to follow a protocol of annotating metaphors called the Metaphor Identification Procedure (MIP) by the Pragglejaz-Group (2007). The flexibility of this protocol renders it applicable to all lexical items in natural discourse. However, this very flexibility also becomes a source of disagreement among experts. The procedure relies on the annotators' intuitions and their expertise in determining what constitutes basic meaning, contextual meaning, and the definition of a metaphor. A quality assessment of the VUAMC dataset conducted by Zayed et al. (2020b) revealed consistency issues when

annotating certain words in similar contexts. For example, the verb *commit* was labeled as a metaphor when paired with the nouns *acts*, *bag*, *government*, and *offence(s)* in some contexts, but as literal when paired with the nouns *rape* and *offence(s)* in very similar contexts. The authors concluded that approximately 5% of the annotated data raised doubts or exhibited inconsistencies regarding some verbs like *receive, form, create, use, make, recognise, feel, enjoy*, and *reduce*.

In addition to the issues raised above, these datasets structures disregard important information in favour of another. Take this mapping as an example for "*His letter has shattered my emotions*" is **EMOTION** as **FRAGILE OBJECT**. The sentence-level approaches disregards the explicit annotation of the source and target domain words and focus on the metaphoricty of the whole sentence. In the word-level annotation, the focus is on the metaphoricity of the source domain **FRAGILE OBJECT** (the vehicle) disregarding the target domain (tenor) **EMOTION**. In the relation level, while the relation between the target and the source is accounted for, it tend to lack the context it appear in because the focus is on certain relation as adjective-noun and subject-verb-object pairs (Tsvetkov et al., 2014) adjective-noun and verb-noun metaphorical expression (Shutova et al., 2016; Rei et al., 2017). Another annotation controversy is regarding the decision to label parts of phrasal verbs as metaphor and not the whole phrase such as the annotation of *pin him down* in word-level as 0, 0, 1.

The three models also encounters another challenge associated with the linguistic aspect. The analysis sections of the semantic richness experiments encountered a common challenge regarding the phenomenon of polysemy when dealing with words that possess multiple senses. The problem accrues when attempting to accurately assign vector embeddings that represent the intended meaning of the words used in each respective sentence. GloVe is static (i.e., the word has one vector representing its meaning). In other words, regardless of which is the word sense intended in a particular context, the vector representation is the

same. This is also seen when assigning for phrasal verbs where each part is considered in isolation form the other. The study conducted by Stowe et al. (2019) showed that verbs such as *hold, bring, put, get,* and *show* were among the verbs exhibiting high levels of ambiguity and misclassification. These verbs are commonly used to form verb phrases, adding complexity to their interpretation. It is to important to note that these same words also emerged as some of the most misclassified terms in the analysis conducted in this dissertation.

The next observation is closely connected to the words meaning previously discussed. The word meaning and context meaning in these metaphor detection experiments are provided by GloVe and ELMo embeddings, each possessing distinct characteristics and playing a specific role in the detection task. GloVe embeddings are static, assigning a single meaning to a word regardless of its context. This presents a challenge when facing polysemous words, as different sense of these words are relevant to specific contexts, which could affect the meaning of the whole sentence. Figuring which sense is the one which was intended to be used in the metaphorical sentence is key to improve the detection results. ELMo embedding is a better choice as these vectors are contextualised meaning they take the word's context into consideration. However, ELMo still has some limitations that could hindrance the full solution of polysemy and the context-specificity criterion. The study conducted by Ethayarajh (2019) measures to what extent the contextualise word representation are contextual. They compare three types of contextualised word representations: ELMo, BERT, and GPT-2 (Radford et al., 2019) over three contextual criteria:

1. The self-similarity: measures how contextual the word w in a particular layer l over number of different context n by measuring the averaged cosine similarity between its contextualised representations. Self-similarity of 1 means that the word w in that layer l did not contextualised the word, the lower the self-similarity of the word w, the more contextualised its representation.

2. The intra-similarity: it is the average cosine similarity between the word representations of a word in the sentence and this sentence's vector. It captures "*how context-specificity manifest in the vector space*". If both self-similarity and intra-similarity are low, the model contextualise words in that layer by giving each one a context-specific representation that is still distinct from all other word representation in the sentence. Meaning, each word in a specific sentence have a context related representation and that representation is distinct from other words in that sentence. If the intra-similarity score is high and self-similarity is low means less nuanced contextualisation, words in a sentence are contextualised by aligning their representations within the vector space.

3. The maximum explainable variance: tests how much in the word's contextualised representation in a given layer l can be explained by their first principal component. Meaning, how well a static embedding like GloVe could replace a word's representation. A value close to 0 suggests a poor replacement, while a value of 1 indicates a perfect replacement, indicating that a static embedding can adequately fulfill the task.

The study found that the 0th layer which is the input layer has self-similarity equals to 1, meaning it is not contextualised. However, the self-similarity for higher layers gets lower (closer to 0) indicating that contextualised representation is more context-specific in the higher the layers. This is supported by the findings of N. F. Liu et al. (2019) where the upper layer in LSTMs trained on different NLP tasks, such as knowledge of semantic relations, learn more task-specific representation. For ELMo's 2-layer biLSTM, the first is the input layer which is non-contextualised, the second is more context specific but words in the same sentence are more similar to one another. In contrast, the BERT's context specific contextualised representation is more dissimilar to one another in the upper layers suggesting that BERT posses the capability of capturing more detailed understanding of context. In other words, even though BERT considers the context of surrounding words, assuming that two words in the same sentence have similar meanings solely based on the

shared context is false. What is even more interesting is that stopwords such as *of, the, 's,* and *to* have been found to possess the most context-specific representations with the lowest self-similarity, despite not being polysemous. This discovery leads to the conclusion that contextualised representations reflect the contextual variations in which a word appears, rather than solely capturing its inherent polysemy or the selection of specific senses. Furthermore, BERT's non-contextualised embeddings of the input layer outperform on the Word Embeddings Benchmarks package (Jastrzebski et al., 2017) where it replaced the static embedding, which suggests the assignment is not from a finite number of word-senses representations. Undeniably, these tests showed how superior BERT to ELMo is, its capabilities of capturing the useful contextualised information and how context-specific it is due to its architecture of bi-directional 12-layer transformer-based language model.

The significance of context for the metaphorical sentence has been shown throughout the chapters. Similarly, depending on the context, the suitable semantic richness element scores are assigned accordingly. However, the semantic richness elements lists used for SVM training are consist solely of unigram words, lacking contextual information that could potentially provide valuable insights for more accurate score assignment, regardless of the BERT representation as the input. During the training of the SVM on the semantic richness elements lists, as well as the token-level score assignment step for metaphorical datasets, BERT is used as the input. However, BERT's full potential is not fully utilised in these two steps, as the contextual information is not taken into account.

7.6.2 Recommendations for Future Work

Moving forward, several recommendation are suggested for future work that address the discussed limitations. Regarding the pre-process step, where the SVM model is trained and the semantic richness scores are assigned, and the main metaphor detection step, improvements

can be made in two areas: enhancing the lexical resources used and refining the models employed.

As mentioned previously, the amounts and quality of the data used in the machine leaning and deep learning tasks are important. Comprehensive lists that are formatted as multiword, phrases or sentences would be beneficial compared to single-word setups. The lexical resources for the semantic richness elements used to train the SVM focus on single-word entries, where only one sense has been rated for each word. Thus, the limitation could accrue when another sense is encounter, leading to inaccurate assignment, which in turn minimises the impact of adding these information to the detection task.

Muraki et al. (2022) argue for the importance of multiword expressions, how they effect the word acquisition and their usage in everyday language. They extended the concreteness/abstractness rating for 62 thousand English multiword expressions. The list covers four type of expressions: Compound nouns are a type of word combination where multiple words are used to express a single concept (e.g., *washing machine*). Phrasal verbs is the second type covered such as *give in* and *give up*. The third is the idiomatic expressions that its overall meaning has nothing to do with the meaning of its components such as *spill the beans*. Lastly, fixed expression of familiar strings where the components deliver the single idea such as *thank you* and *go to bed*. Using this list for future metaphor detection alongside concreteness will be interesting to test. Furthermore, this list could serves as a building block for rating other elements of semantic richness. It is comprehensive, with quality and quantity items that converse more realistic, everyday language expressions.

Standardised and uniformity of the rating scales is another possibility that could positively improve the metaphor detection task as each of the semantic richness element's scale currently differ. For example, in Chapter 4, the imageability score scale was from 1 to 3, for 18156 words, while the sensory experience was from 1 to 7, for 5,000 words. This

difference in size could result in insufficient training due to a lack of training examples that could covers as much everyday language, leading to the loss of some useful hints during the detection tests that could have yielded better predictions. This is also the case for the scale difference, imageability's scale is ranged from 1 to 3 (also seen in Chapter 3, for the polarity experiments on sentiment) where the slight score shifts between the words in the sentences were not enough to signal to the neural network the existence of metaphorical usage. In contrast, the 4 range score for concreteness, Chapter 2 or the 6 range score for sensory experience, Chapter 4, provided better results in metaphor identification in some instances when the shift between words is 2 or more.

Enhancing the metaphor datasets is another recommendation. Problems such as: annotation disagreement and discrepancy (Zayed et al., 2020b; Ma et al., 2020), annotation schemes limitation (Zayed et al., 2020a), simple and easy sentences that fail to capture the complexity of realistic discourse and do not fully leverage the capabilities of the newer model (Ma et al., 2020), and focusing on context and ignoring the grammatical relations that is argued to be the more direct clue for metaphorical usage (Song et al., 2021; Stowe et al., 2019) all are limitations related to the metaphorical dataset.

Additionally, the structure of the detection task experiment's format requires refinement. As previously discussed, metaphor detection experiments tend to adopt one of three structures: sentence-level, word-level or phrase-level (i.e., relation-level). However, recent state-of-the-art models show a growing inclination towards shaping the tests at the relation-level rather than the sentence-level or the more widely adopted word-level approach. Even when the cases where the tests follows either the word-level or sentence-level approach, the relation-level bit is integrated into the process (Song et al., 2021). This is because the contextual information of metaphors is encapsulated within this relational component, which plays a crucial role in capturing the nuanced meaning and understanding the metaphorical expressions. Furthermore, this relational-level component could also help overcoming the

limitations caused by polysemy, where the sense of a word is more clearly distinguished when it is considered in combination with its surrounding words. By explicitly incorporating the relational context, the model can better disambiguate the intended meaning of a word in the given context, leading to more accurate metaphor detection and interpretation.

As previously highlighted about how word-level metaphor annotations favours source domain (the vehicle) and disregards the target domain (tenor), which is incomplete information. Zayed et al. (2020b) suggest that adding the missing information will improve the performance of the metaphor detection models. Thus, they adapted VUAMC and TroFi word-level benchmark dataset to be better suited for relation-level (phrase-level) metaphor identification. They also provided context to the adjective-noun relation-level in the TSV dataset. Stowe et al. (2019) extended the syntax in metaphor training dataset: VerbNet was utilised to enhance the identification of data related to both metaphoric and literal senses, and dependency parses was utilised to identify relevant sentences by leveraging syntactic characteristics of particular lexemes (basic unit of vocabulary).

It can be inferred from these previous findings that the research community is increasingly inclined to integrate syntax into the detection and expand the annotation and detection of metaphor to the phrase-level. Zayed et al. (2020a) adopted Feature-wise transformation techniques (FiLM) where the image part for visual reasoning is replaced by the sentence that has a metaphoric candidate (called the *contextual modulation* pipeline), and the linguistic interaction occurring between the different components within the candidate takes the place of the auxiliary input that employs an LSTM (called the *metaphor identification linguistic* pipeline). The model is trained on verb-noun and adjective-noun grammatical relations. The model show performance improvement on the TSV and MOH datasets compared to the state-of-the-art, as well as, showed a promising results when identifying conventional and uncommon metaphorical expressing.

Based on all the previous discussions regarding the observations and limitations, the following solution is proposed. The SVM that assigns semantic richness scores used BERT pre-trained embeddings as the word representations. This, as mentioned previously, does not solve the problem of polysemy because the sense of the predicted word, by default, is the basic meaning, which prompt the SVM to assign the same semantic richness score regardless of the specific sense in which the word is used. Therefore, in addition to building a multiword, phrase or sentences lists, models such as BERT should be used for training and assigning the predictions because they are more suited for language handing. Previous research experiments have shown that BERT can produce contextual representations that capture the context and are distinct from the surrounding words. This is a significant step forward in improving the handling of polysemy.

In addition, future research could adopt the approach introduced by Song et al. (2021) to test the impact of the semantic richness on the metaphor detection. Their methodology demonstrates an intriguing utilisation of context that has the potential to overcome multiple limitations. The model, Metaphor-relation BERT (Mr-BERT), view the metaphor detection as a relation between the verb and its grammatical, sentential and semantic contexts. This relationship is explicitly modeled and applied on the VUAMC, MOH-X and TroFi token-level datasets. Mr-BERT follows a three steps division:

1. Extract different types of verb's contextual components, represented by:
    - Global context represents the whole sentence.
    - Local context represents the grammatical relation to the target verb in form of verb-subject and verb-direct object.
    - Distant context represents the difference between the target verb's contextual meaning and its basic meaning, this step follows MIP role.

2. Represent the extracted contexts using BERT transformer model by following these steps:

   (a) Six component marker tokens, [subj], [/subj], [verb], [/verb], [obj] and [/obj], labelled the boundaries of the target verb's component after using the dependency parser to extract verb-subject and verb-direct object. This contextual information could be used to obtain the semantic richness elements' score.

   (b) Label the whole sentence by using [CLS] and [SEP].

   (c) BERT tokeniser takes the whole sentence and then to moved to the transformer layers.

   (d) The final transformer layer's hidden states are obtained to represenat the local context by calculating the average embedding between the start and end markers (i.e., [verb] and [/verb]) for each marked component.

   (e) the final hidden state of [CLS] is obtained as the global context representation.

   (f) The BERT tokeniser is independently used for the verb as the basic meaning contextual representation.

3. Integrate the three different contextual relation for metaphor prediction by either: concatenation, averaging or maxing-out.

The model showed that using the three contextual components is better and surpasses the use of one or two of them. Also averaging and maxing-out context strategies performs better than the concatenation. Concatenations needs more parameters which increase the difficulty of training. Additional effective strategy to improve the performance was tested by dividing long sentences by the number of the their existing clauses and utilising their grammatical relation. It suggested that the grammatical relation helps to balance the information within the sentences, as they are different, which helps to filter out noise.

7.7 Future Work Summery

The previous section discussed the observations and limitations faced by metaphor detection neural network models when incorporating semantic richness elements to test the detection performance. The main issues analysed can be examined from six main perspectives:

1. The lexical resource angle: that the lists of semantic richness elements used in the experiments are composed of unigrams with varying lengths, which can be considered relatively short lists of English words. Furthermore, it should be noted that these lists of semantic richness elements do not adequately represent the rich and complex everyday language that is commonly used. They lack popular and well-known phrases or expressions. Additionally, the scoring scales assigned to these elements vary, with some using a scale of 2 while others use a scale of 6. This variation in scoring scales could potentially hinder the models' ability to capture important hints for improving metaphor detection.

2. The model used to train the lexical resource angle: the model used for assigning the predicted semantic richness elements' scores is SVM. This model takes the unigram's representation in a form of a pre-trained BERT embedding, and then assign a predicted score to another unigram. This model is not suitable for large dataset.

3. The polysemy angle: correctly figuring the intended sense of the word used is a big problem when the process involved unigrams, for the semantic elements lexical list and for the metaphorical datasets. As the model assigned a semantic richness score by default for the basic meaning disregarding the intended sense, as well as the detection process handle words during the sequence or the sentence classification because of how the datasets are annotated.

4. The embeddings angle: the polysemy observation is reinforced by the utilization of

GloVe and ELMo representations in the experiments. GloVe is employed for word-level embeddings, while ELMo is used for capturing contextual information. However, as discussed earlier, GloVe embeddings are static, meaning that regardless of the sense or context, the embedding remains the same. On the other hand, ELMo falls short compared to other contextualized representations such as BERT, which have demonstrated superior performance in capturing contextual information.

5. The metaphor datasets angle: the metaphor/non-metaphor annotation strategies used are focus on words rather than phrase/multiword expressions, which results in incomplete information necessary for detecting metaphors. Additionally, the annotation process disregards grammatical relations and the inclusion of target domain words, which are important factors in understanding metaphors. Moreover, inconsistent annotations were observed, further adding to the challenges faced in accurately detecting metaphors.

6. The metaphor detection models angle: the structure of the two models used in the experiments has its own advantages and disadvantages. However, considering the previously discussed angles, adjustments are needed for both the word-level and sentence-level structures to align with the recommended changes.

Based on the previous points, several adjustments are necessary to facilitate the incorporation of semantic richness elements into the metaphor detection task. These adjustments are also aimed at handling multiword expressions and capturing contextual information, ultimately leading to more effective metaphor detection. Here are the recommended adjustments:

1. The lexical resource lists: adopt Muraki et al. (2022)'s comprehensive list of 62,000 English multiword expressions for future work on semantic richness element ratings. It

is a significant step towards improving the incorporation of such elements in metaphor detection. This extensive list covers the complex and everyday language used in discourse, making it a valuable resource for rating other words and expressions that are not initially included in the list. Moreover, identify the most efficient scoring range and standardise that scale for all semantic richness elements ratings.

2. The annotations of the metaphor datasets should be examined and revised to address any inconsistencies. Additionally, as suggested by Zayed et al. (2020b) it is recommended to adapt these datasets to suit relation-level experiments. It is beneficial to identify the target domain word, as demonstrated, and including the grammatical relation is encouraged. This step provides comprehensive information about the metaphors, as both the target and source domain words are explicitly annotated. Furthermore, by overcoming the limitations of single token annotation, the exact sense of the word can be easily captured, helping to mitigate the limitations introduced by polysemy.

3. After making adjustments to the lexical resource and metaphorical datasets, it is crucial to use a model that can handle complex and large data while effectively capturing contextual information. Using a form of BERT model is strongly recommended for the pre-processing step and for the detection step due to its ability to assign contextual representations that capture the overall context while maintaining distinctiveness for each word in the sentence. Additionally, for the detection task, it is recommended to utilise neural network models that incorporate BERT in their process, such as the Mr-BERT model introduced by Song et al. (2021). These models, unlike GloVe and ELMo, leverage the strengths of BERT to enhance metaphor detection by leveraging its contextual representation capabilities. Their usage of three different context information is a strong advantages as it handles multiple limitation faced in this dissertation experiments.

7.8 Large Language Models (LLM) and this Work

The research on metaphor detection in the field of Natural Language Processing (NLP) has become less aligned with the current AI paradigm, which leverages Large Language Models (LLMs) such as ChatGPT. Traditional deep learning models, including those like BERT and the models used in this dissertation, are fine-tuned on specific datasets and rely on word's basic meaning, contextual and semantic features to predict the metaphoricity of target words/sentences. On the other hand, LLMs utilise huge datasets trained on various architectures using different training strategies (Naveed et al., 2023), LLMs model size, pre-training data size, fine-tuning data size, and new fine-tuning parameter size (B. Zhang et al., 2024) to generate human-like conversations and produce creative content, demonstrating a significant advance in language understanding and generation capabilities.

The word's semantic richness and the context are used here to detect metaphors. The basic meaning is compared to the sentence's context to identify metaphorical expressions. However, the training dataset limits the scale of the model's understanding and its ability to generalise across different linguistic constructs and contexts, often resulting in less accurate metaphor detection compared to the capabilities of modern LLMs, which are pre-trained on very large-diverse corpora and massive parameter sizes. This difference allows them to develop a deep understanding of language, improving their ability to process and generate text with human-like accuracy.

For the metaphor detection research to remain relevant, the integration of LLMs into metaphor detection research is essential. Leveraging the vast pre-trained knowledge and contextual analysis capabilities of LLMs will ensure the development of more sophisticated metaphor detection models (Kim et al., 2023).

7.9 Conclusions

The presence of metaphors presents a challenge to the field of Natural Language Processing (NLP) due to their inherent tendency to convey meanings that is different from the literal interpretations of their constituent words. Many theories were discussed in an attempt to explain this form of language.

Semantic information is critical in the formation and comprehension of metaphors. Metaphors are subject to the influence of human knowledge, experiential background, and other contextual variables. These facets fall under the semantic richness, which serves as the central focus of this study.

The elements of semantic richness under investigation encompass several elements: the concreteness/abstractness measure, sentiment (*polarity* and *emotions*), imageability, sensory experience, and body-object interaction. This thesis extensively examines the feasibility and applicability of these semantic richness elements within the context of metaphor detection through neural networks. Additionally, it assesses the impact of these features on both word-level and sentence-level problems.

Two distinct models were employed for testing purposes. The first is the classification model, where the baseline was established based on the work proposed by Gao et al. (2018). This model classifies sentences as either *metaphorical* or *non-metaphorical* based on the target verb and the contextual information provided by the surrounding words. The second model is founded on the RNN_HG model, as introduced by Mao et al. (2019). This model classifies words based on their contextual surroundings. In both models, GloVe embeddings serve as the foundation for conveying the basic word meanings, while ELMo embeddings offer contextual information. The resulting word vectors are 1324-dimensional. Each semantic richness element's rating is concatenated with this vector either as a single-digit rating

or as a probability distribution rating. These ratings are derived either form the lists directly, through SVM predictions or as predictions from a BERT fine-tuned model, as in the sentiment experiments.

Upon a comprehensive analysis of the outcomes, which encompass both positive and negative results as examined in the respective correct-prediction and wrong-prediction files for each experiment, several important observations have surfaced. These findings are extensively discussed in the relevant chapters, as well as in this Chapter's Section 7.6.1. In light of the identified challenges this study also presents recommendations for future research that aim to address these points and further advance the field in Section 7.6.2.

References

Alkhatib, Manar and Khaled Shaalan (2018). "Paraphrasing Arabic metaphor with neural machine translation". In: *Procedia Computer Science* 142, pp. 308–314.

Alnafesah, Ghadi, Harish Tayyar Madabushi, and Mark Lee (2020). "Augmenting neural metaphor detection with concreteness". In: *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 204–210.

Alnafesah, Ghadi, Phillip Smith, and Mark Lee (2023). *Are you not moved? Incorporating Sensorimotor Knowledge to Improve Metaphor Detection*. URL: https://acl-bg.org/proceedings/2023/RANLP%202023/RANLP%202023%20Proceedings.pdf.

Aubusson, Peter J, Allan G Harrison, and Stephen M Ritchie (2006). "Metaphor and analogy". In: *Metaphor and analogy in science education*. Springer, pp. 1–9.

Barnden, John A (2007). "Metaphor, semantic preferences and context-sensitivity". In: *Words and intelligence II: Essays in honor of Yorick Wilks*. Springer, pp. 39–62.

Bernhardt, Mélanie et al. (2021). "Active label cleaning for improved dataset quality under resource constraints". In: *Nature Communications* 13.

Biddle, Rhys, Aditya Joshi, Shaowu Liu, Cecile Paris, and Guandong Xu (2020). "Leveraging sentiment distributions to distinguish figurative from literal health reports on Twitter". In: *Proceedings of The Web Conference 2020*, pp. 1217–1227.

Birke, Julia and Anoop Sarkar (2006). "A clustering approach for nearly unsupervised recognition of nonliteral language". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 329–336.

Bizzoni, Yuri, Stergios Chatzikyriakidis, and Mehdi Ghanimifard (2017). ""Deep" Learning: Detecting Metaphoricity in Adjective-Noun Pairs". In: *Proceedings of the Workshop on Stylistic Variation*, pp. 43–52.

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman (2014). "Concreteness ratings for 40 thousand generally known English word lemmas". In: *Behavior research methods* 46.3, pp. 904–911.

Cacciari, Cristina and Patrizia Tabossi (2014). *Idioms: Processing, structure, and interpretation*. Psychology Press.

Chakrabarty, Tuhin, Smaranda Muresan, and Nanyun Peng (2020). "Generating similes effortlessly like a pro: A style transfer approach for simile generation". In: *arXiv preprint arXiv:2009.08942*.

Chakrabarty, Tuhin, Xurui Zhang, Smaranda Muresan, and Nanyun Peng (2021). "MERMAID: Metaphor generation with symbolism and discriminative decoding". In: *arXiv preprint arXiv:2103.06779*.

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D Manning (2019). "What does bert look at? an analysis of bert's attention". In: *arXiv preprint arXiv:1906.04341*.

Coltheart, Max (1981). "The MRC psycholinguistic database". In: *The Quarterly Journal of Experimental Psychology Section A* 33.4, pp. 497–505.

Danguecan, Ashley N and Lori Buchanan (2016). "Semantic neighborhood effects for abstract versus concrete words". In: *Frontiers in psychology* 7, p. 1034.

Dankers, Verna, Marek Rei, Martha Lewis, and Ekaterina Shutova (Nov. 2019). "Modelling the interplay of metaphor and emotion through multitask learning". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-*

IJCNLP). Hong Kong, China: Association for Computational Linguistics, pp. 2218–2229. DOI: 10.18653/v1/D19-1227. URL: https://aclanthology.org/D19-1227.

Davidson, Donald (1978). "What metaphors mean". In: *Critical inquiry* 5.1, pp. 31–47.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Do Dinh, Erik-Lân, Steffen Eger, and Iryna Gurevych (Aug. 2018). "Killing Four Birds with Two Stones: Multi-Task Learning for Non-Literal Language Detection". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1558–1569. URL: https://aclanthology.org/C18-1132.

Do Dinh, Erik-Lân and Iryna Gurevych (2016). "Token-level metaphor detection using neural networks". In: *Proceedings of the Fourth Workshop on Metaphor in NLP*, pp. 28–33.

Duñabeitia, Jon Andoni, Alberto Avilés, and Manuel Carreiras (2008). "NoA's ark: Influence of the number of associates in visual word recognition". In: *Psychonomic Bulletin & Review* 15.6, pp. 1072–1077.

Dunn, Jonathan (2015). "Modeling abstractness and metaphoricity". In: *Metaphor and Symbol* 30.4, pp. 259–289.

Ethayarajh, Kawin (2019). "How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings". In: *arXiv preprint arXiv:1909.00512*.

Farias, Delia Irazu Hernandez, Emilio Sulis, Viviana Patti, Giancarlo Ruffo, and Cristina Bosco (2015). "Valento: Sentiment analysis of figurative language tweets with irony and sarcasm". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 694–698.

Ferreira, Roberto A, Silke M Göbel, Mark Hymers, and Andrew W Ellis (2015). "The neural correlates of semantic richness: Evidence from an fMRI study of word learning". In: *Brain and Language* 143, pp. 69–80.

Findlay, Holly and Gareth Carrol (2018). "Contributions of semantic richness to the processing of idioms". In: *The Mental Lexicon* 13.3, pp. 311–332.

Gao, Ge, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer (2018). "Neural metaphor detection in context". In: *arXiv preprint arXiv:1808.09653*.

Garbas, Lukas (Dec. 2019). *GitHub - Multi-class sentiment analysis lstm, finetuned bert*. GitHub. URL: https://github.com/lukasgarbas/nlp-text-emotion (visited on 02/04/2021).

Ghazi, Diman, Diana Inkpen, and Stan Szpakowicz (2015). "Detecting emotion stimuli in emotion-bearing sentences". In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 152–165.

Ghosh, Aniruddha and Tony Veale (2016). "Fracking sarcasm using neural network". In: *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 161–169.

Gibbs Jr, Raymond W (1992). "When is metaphor? The idea of understanding in theories of metaphor". In: *Poetics today*, pp. 575–606.

Goatly, Andrew (1997). *The language of metaphors*. Routledge.

Goldberg, Yoav (2017). "Neural network methods for natural language processing". In: *Synthesis lectures on human language technologies* 10.1, pp. 1–309.

Gould, Layla, Jacqueline Cummine, and Ron Borowsky (2012). "The cognitive chronometric architecture of reading aloud: semantic and lexical effects on naming onset and duration". In: *Frontiers in human neuroscience* 6, p. 287.

Haagsma, Hessel and Johannes Bjerva (2016). "Detecting novel metaphor using selectional preference information". In: *Proceedings of the Fourth Workshop on Metaphor in NLP*, pp. 10–17.

Hargreaves, Ian S and Penny M Pexman (2012). "Does richness lose its luster? Effects of extensive practice on semantic richness in visual word recognition". In: *Frontiers in human neuroscience* 6, p. 234.

Harrington, Kristine J (2012). "The use of metaphor in discourse about cancer: a review of the literature." In: *Clinical journal of oncology nursing* 16.4.

Hilton, Kelsey, Akbar Siami Namin, and Keith S Jones (2022). "Metaphor identification in cybersecurity texts: a lightweight linguistic approach". In: *SN Applied Sciences* 4.2, p. 60.

Jamrozik, Anja, Marguerite McQuire, Eileen R Cardillo, and Anjan Chatterjee (2016). "Metaphor: Bridging embodiment to abstraction". In: *Psychonomic bulletin & review* 23.4, pp. 1080–1089.

Jang, Hyeju, Yohan Jo, Qinlan Shen, Michael Miller, Seungwhan Moon, and Carolyn Rose (2016). "Metaphor detection with topic transition, emotion and cognition in context". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 216–225.

Jang, Hyeju, Seungwhan Moon, Yohan Jo, and Carolyn Rose (2015). "Metaphor detection in discourse". In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 384–392.

Jastrzebski, Stanisław, Damian Leśniak, and Wojciech Marian Czarnecki (2017). "How to evaluate word embeddings? on importance of data efficiency and simple supervised tasks". In: *arXiv preprint arXiv:1702.02170*.

Juhasz, Barbara J and Melvin J Yap (2013). "Sensory experience ratings for over 5,000 mono-and disyllabic words". In: *Behavior Research Methods* 45.1, pp. 160–168.

Karanasou, Maria, Christos Doulkeridis, and Maria Halkidi (2015). "DsUniPi: An SVM-based approach for sentiment analysis of figurative language on Twitter". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 709–713.

Karthiga, M, S Sountharrajan, E Suganya, and S Sankarananth (2021). "Sentence Semantic Similarity Model Using Convolutional Neural Networks". In: *EAI Endorsed Transactions on Energy Web* 8.35.

Keysar, Boaz (1989). "On the functional equivalence of literal and metaphorical interpretations in discourse". In: *Journal of memory and language* 28.4, pp. 375–385.

Kim, Jeongyeon, Sangho Suh, Lydia B Chilton, and Haijun Xia (2023). "Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing". In: DIS '23. Pittsburgh, PA, USA: Association for Computing Machinery, pp. 115–135. ISBN: 9781450398930. DOI: 10.1145/3563657.3595996. URL: https://doi.org/10.1145/3563657.3595996.

Kintsch, Walter and Anita R Bowles (2002). "Metaphor comprehension: What makes a metaphor difficult to understand?" In: *Metaphor and symbol* 17.4, pp. 249–262.

Klebanov, Beata Beigman, Ben Leong, E Dario Gutierrez, Ekaterina Shutova, and Michael Flor (2016). "Semantic classifications for detection of verb metaphors". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 101–106.

Klebanov, Beata Beigman, Ben Leong, Michael Heilman, and Michael Flor (2014). "Different texts, same metaphors: Unigrams and beyond". In: *Proceedings of the Second Workshop on Metaphor in NLP*, pp. 11–17.

Koch, Bernard, Emily Denton, Alex Hanna, and Jacob G Foster (2021). "Reduced, reused and recycled: The life of a dataset in machine learning research". In: *arXiv preprint arXiv:2112.01716*.

Köper, Maximilian and Sabine Schulte im Walde (2017a). "Applying multi-sense embeddings for german verbs to determine semantic relatedness and to detect non-literal language". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 535–542.

Köper, Maximilian and Sabine Schulte im Walde (2017b). "Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses". In: *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pp. 24–30.

Kuznetsov, Ilia and Iryna Gurevych (2018). "From text to lexicon: Bridging the gap between word embeddings and lexical resources". In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 233–244.

Lakoff, G and M Johnson (1980). *Metaphors we Live By*. University of Chicago press.

Lakoff, G, M Johnson, and John F Sowa (1999). "Review of Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought". In: *Computational Linguistics* 25.4, pp. 631–634.

Landauer, Thomas K, Peter W Foltz, and Darrell Laham (1998). "An introduction to latent semantic analysis". In: *Discourse processes* 25.2-3, pp. 259–284.

Le, Duong, My Thai, and Thien Nguyen (2020). "Multi-task learning for metaphor detection with graph convolutional neural networks and word sense disambiguation". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 05, pp. 8139–8146.

Leong, Ben, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen (July 2020). "A Report on the 2020 VUA and TOEFL Metaphor Detection Shared Task". In: *Proceedings of the Second Workshop on Figurative Language Processing*. Online: Association for Computational Linguistics, pp. 18–29. DOI: 10.18653/v1/2020.figlang-1.3. URL: https://aclanthology.org/2020.figlang-1.3.

Li, Yanran, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu (2017). "Dailydialog: A manually labelled multi-turn dialogue dataset". In: *arXiv preprint arXiv:1710.03957*.

Lima, Paula Lenz Costa (2006). "About primary metaphors". In: *DELTA: Documentação de Estudos em Lingüıstica Teórica e Aplicada* 22.spe, pp. 109–122.

Linjordet, Trond and Krisztian Balog (2019). "Impact of Training Dataset Size on Neural Answer Selection Models". In: *Advances in Information Retrieval*. Ed. by Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra. Cham: Springer International Publishing, pp. 828–835. ISBN: 978-3-030-15712-8.

Liu, Jerry, Nathan O'Hara, Alexander Rubin, Rachel Draelos, and Cynthia Rudin (2020). "Metaphor detection using contextual word embeddings from transformers". In: *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 250–255.

Liu, Nelson F., Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith (June 2019). "Linguistic Knowledge and Transferability of Contextual Representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1073–1094. DOI: 10.18653/v1/N19-1112. URL: https://aclanthology.org/N19-1112.

Ljubešić, Nikola, Darja Fišer, and Anita Peti-Stantić (2018). "Predicting Concreteness and Imageability of Words Within and Across Languages via Word Embeddings". In: *Proceedings of the 3nd Workshop on Representation Learning for NLP, RepL4NLP@ACL 2018, Melbourne, Australia, July 20, 2018*.

Ma, Weicheng, Ruibo Liu, Lili Wang, and Soroush Vosoughi (2020). "Improvements and Extensions on Metaphor Detection". In: *arXiv preprint arXiv:2012.04540*.

Mao, Rui, Chenghua Lin, and Frank Guerin (2018). "Word embedding and wordnet based metaphor identification and interpretation". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1222–1231.

Mao, Rui, Chenghua Lin, and Frank Guerin (2019). "End-to-end sequential metaphor identification inspired by linguistic theories". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3888–3898.

Mason, Zachary J (2004). "CorMet: a computational, corpus-based conventional metaphor extraction system". In: *Computational Linguistics* 30.1, pp. 23–44.

Maudslay, Rowan Hall and Simone Teufel (2022). "Metaphorical Polysemy Detection: Conventional Metaphor Meets Word Sense Disambiguation". In: *arXiv preprint arXiv:2212.08395*.

Merriam-Webster (2022). *Abulia*. In: *Merriam-Webster.com dictionary*. URL: https://www.merriam-webster.com/dictionary/bring%5C%20out (visited on 03/21/2022).

Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig (2013). "Linguistic regularities in continuous space word representations". In: *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751.

Mohammad, Saif, Ekaterina Shutova, and Peter Turney (2016). "Metaphor as a medium for emotion: An empirical study". In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pp. 23–33.

Mohammad, Saif and Peter Turney (2010). "Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon". In: *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pp. 26–34.

Mohammad, Saif M (2016). "Sentiment analysis: Detecting valence, emotions, and other affectual states from text". In: *Emotion measurement*. Elsevier, pp. 201–237.

Mohammad, Saif M (2021). "Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text". In: *Emotion Measurement*. Elsevier, pp. 323–379.

Mohler, Michael, David Bracewell, Marc Tomlinson, and David Hinote (2013). "Semantic signatures for example-based linguistic metaphor detection". In: *Proceedings of the First Workshop on Metaphor in NLP*, pp. 27–35.

Muñoz Martín, Ricardo (Dec. 2012). "Cognitive and psycholinguistic approaches". In: pp. 241–256. ISBN: 9780415559676.

Muraki, Emiko J, Summer Abdalla, Marc Brysbaert, and Penny M Pexman (2022). "Concreteness ratings for 62,000 English multiword expressions". In: *Behavior Research Methods*, pp. 1–10.

Naveed, Humza, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian (2023). "A comprehensive overview of large language models". In: *arXiv preprint arXiv:2307.06435*.

Ortony, Andrew (1993). *Metaphor and thought*. Cambridge University Press.

Parde, Natalie and Rodney Nielsen (2018). "Exploring the terrain of metaphor novelty: A regression-based approach for automatically scoring metaphors". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1.

Parvin, S Ashika, M Sumathi, and C Mohan (2021). "Challenges of Sentiment Analysis-A Survey". In: *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, pp. 781–786.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep contextualized word representations". In: *Proc. of NAACL*.

Pexman, Penny M, Ian S Hargreaves, Jodi D. Edwards, Luke C. Henry, and Bradley G. Goodyear (2007). "The Neural Consequences of Semantic Richness". In: *Psychological Science* 18.5. PMID: 17576279, pp. 401–406. DOI: 10.1111/j.1467-9280.2007.01913.x. eprint: https://doi.org/10.1111/j.1467-9280.2007.01913.x. URL: https://doi.org/10.1111/j.1467-9280.2007.01913.x.

Pexman, Penny M, Ian S Hargreaves, Paul D Siakaluk, Glen E Bodner, and Jamie Pope (2008). "There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition". In: *Psychonomic Bulletin & Review* 15.1, pp. 161–167.

Pexman, Penny M, Stephen J Lupker, and Yasushi Hino (2002). "The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks". In: *Psychonomic bulletin & review* 9.3, pp. 542–549.

Pexman, Penny M, Emiko J Muraki, David M Sidhu, Paul D Siakaluk, and Melvin J Yap (2019). "Quantifying sensorimotor experience: Body–object interaction ratings for more than 9,000 English words". In: *Behavior research methods* 51.2, pp. 453–466.

Pexman, Penny M, Paul D Siakaluk, and Melvin J Yap (2013). "Introduction to the research topic meaning in mind: semantic richness effects in language processing". In: *Frontiers in human neuroscience* 7, p. 723.

Prabhakaran, Vinodkumar, Marek Rei, and Ekaterina Shutova (2021). "How metaphors impact political discourse: A large-scale topic-agnostic study using neural metaphor detection". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15, pp. 503–512.

Pragglejaz-Group (2007). "MIP: A method for identifying metaphorically used words in discourse". In: *Metaphor and symbol* 22.1, pp. 1–39.

Pramanick, Malay, Ashim Gupta, and Pabitra Mitra (2018). "An lstm-crf based approach to token-level metaphor detection". In: *Proceedings of the Workshop on Figurative Language Processing*, pp. 67–75.

Rabinovich, Ella, Benjamin Sznajder, Artem Spector, Ilya Shnayderman, Ranit Aharonov, David Konopnicki, and Noam Slonim (2018). "Learning concept abstractness using weak supervision". In: *arXiv preprint arXiv:1809.01285*.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019). "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8, p. 9.

Rai, Sunny, Shampa Chakraverty, and Devendra K Tayal (2016). "Supervised metaphor detection using conditional random fields". In: *Proceedings of the Fourth Workshop on Metaphor in NLP*, pp. 18–27.

Ravin, Yael and Claudia Leacock (2000). "Polysemy: an overview". In: *Polysemy: Theoretical and computational approaches*, pp. 1–29.

Recchia, Gabriel and Michael Jones (2012). "The semantic richness of abstract concepts". In: *Frontiers in human neuroscience* 6, p. 315.

Rei, Marek, Luana Bulat, Douwe Kiela, and Ekaterina Shutova (2017). "Grasping the finer point: A supervised similarity network for metaphor detection". In: *arXiv preprint arXiv:1709.00575*.

Reisfield, Gary M. and George R. Wilson (2004). "Use of Metaphor in the Discourse on Cancer". In: *Journal of Clinical Oncology* 22.19. PMID: 15459229, pp. 4024–4027. DOI: 10.1200/JCO.2004.03.136. eprint: https://doi.org/10.1200/JCO.2004.03.136. URL: https://doi.org/10.1200/JCO.2004.03.136.

Rentoumi, Vassiliki, George A Vouros, Vangelis Karkaletsis, and Amalia Moser (2012). "Investigating metaphorical language in sentiment analysis: A sense-to-sentiment perspective". In: *ACM Transactions on Speech and Language Processing (TSLP)* 9.3, pp. 1–31.

Sahlgren, Magnus (2008). "The distributional hypothesis". In: *Italian Journal of Disability Studies* 20, pp. 33–53.

Scherer, Klaus R and Harald G Wallbott (1994). "" Evidence for universality and cultural variation of differential emotion response patterning": Correction." In.

Searle, John R, Ferenc Kiefer, Manfred Bierwisch, et al. (1980). *Speech act theory and pragmatics*. Vol. 10. Springer.

Shi, Nan, Dongyu Zhang, Lulu Li, Shengjun Xu, et al. (2021). "Predicting Mental Health Problems with Automatic Identification of Metaphors". In: *Journal of Healthcare Engineering* 2021.

Shutova, Ekaterina (2010). "Models of metaphor in NLP". In: *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 688–697.

Shutova, Ekaterina, Douwe Kiela, and Jean Maillard (2016). "Black holes and white rabbits: Metaphor identification with visual features". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 160–170.

Siakaluk, Paul D, Penny M Pexman, Laura Aguilera, William J Owen, and Christopher R Sears (2008). "Evidence for the activation of sensorimotor information during visual word recognition: The body–object interaction effect". In: *Cognition* 106.1, pp. 433–443.

Song, Wei, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu (2021). "Verb metaphor detection via contextual relation learning". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4240–4251.

Stemle, Egon and Alexander Onysko (2018). "Using language learner data for metaphor detection". In: *Proceedings of the Workshop on Figurative Language Processing*, pp. 133–138.

Stowe, Kevin (2019). "Syntactic and semantic improvements to computational metaphor processing". PhD thesis. University of Colorado at Boulder.

Stowe, Kevin, Sarah Moeller, Laura Michaelis, and Martha Palmer (2019). "Linguistic analysis improves neural metaphor detection". In: *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pp. 362–371.

Strohmaier, David and Gladys Tyen (2022). "A Category Theory Framework for Sense Systems". In: *Proceedings of Globalex Workshop on Linked Lexicography within the 13th Language Resources and Evaluation Conference*, pp. 42–51.

Swarnkar, Krishnkant and Anil Kumar Singh (2018). "Di-LSTM contrast: A deep neural network for metaphor detection". In: *Proceedings of the Workshop on Figurative Language Processing*, pp. 115–120.

Taler, Vanessa, Shanna Kousaie, and Rocio Lopez Zunini (2013). "ERP measures of semantic richness: the case of multiple senses". In: *Frontiers in human neuroscience* 7, p. 5.

Tekiroğlu, Serra Sinem, Gözde Özbal, and Carlo Strapparava (June 2015). "Exploring Sensorial Features for Metaphor Identification". In: DOI: 10.3115/v1/W15-1404.

Tighe, Brian (2010). *VU Amsterdam Metaphor Corpus Online*. URL: http://www.vismet.org/metcor/manual/tutorial.php#SEC2.

Tong, Xiaoyu, Ekaterina Shutova, and Martha Lewis (2021). "Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4673–4686.

Tsvetkov, Yulia, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer (2014). "Metaphor detection with cross-lingual model transfer". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 248–258.

Tsvetkov, Yulia, Elena Mukomel, and Anatole Gershman (2013). "Cross-lingual metaphor detection using common semantic features". In: *Proceedings of the First Workshop on Metaphor in NLP*, pp. 45–51.

Turney, Peter, Yair Neuman, Dan Assaf, and Yohai Cohen (2011). "Literal and metaphorical sense identification through concrete and abstract context". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 680–690.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Veale, Tony, Ekaterina Shutova, and Beata Klebanov (2016). "Metaphor: A computational perspective. Synthesis lectures on human language technologies". In: *San Raphael, CA: Morgan and Claypool Publishers* 10, S00694ED1V01Y201601HLT031.

Velasco, Olga Isabel Dıez (2001). "Metaphor, metonymy and image-schemas: an analysis of conceptual interaction patterns". In: *Journal of English studies* 3, pp. 47–64.

Wan, Mingyu, Baixi Xing, Qi Su, Pengyuan Liu, and Chu-Ren Huang (2020). "Sensorimotor enhanced neural network for metaphor detection". In: *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pp. 312–317.

Wetterstrand, K A (2016). *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. URL: www.genome.gov/sequencingcosts.

Wilks, Yorick (1968). *COMPUTABLE SEMANTIC DERIVATIONS (AND THEIR APPLICATION TO NATURAL LANGUAGE TEXTS)*. Tech. rep. INSTITUTE FOR FORMAL STUDIES LOS ANGELES CALIF.

Wilks, Yorick (1978). "Making preferences more active". In: *Artificial Intelligence* 11.3, pp. 197–223.

Wilks, Yorick, Adam Dalton, James Allen, and Lucian Galescu (2013). "Automatic metaphor detection using large-scale lexical resources and conventional metaphor extraction". In: *Proceedings of the First Workshop on Metaphor in NLP*, pp. 36–44.

Wu, Chuhan, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang (2018). "Neural metaphor detecting with CNN-LSTM model". In: *Proceedings of the Workshop on Figurative Language Processing*, pp. 110–114.

Zayed, Omnia, John P McCrae, and Paul Buitelaar (2020a). "Contextual modulation for relation-level metaphor identification". In: *arXiv preprint arXiv:2010.05633*.

Zayed, Omnia, John Philip McCrae, and Paul Buitelaar (2020b). "Adaptation of word-level benchmark datasets for relation-level metaphor identification". In: *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 154–164.

Zhang, Biao, Zhongtao Liu, Colin Cherry, and Orhan Firat (2024). *When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method*. arXiv: 2402.17193 [cs.CL]. URL: https://arxiv.org/abs/2402.17193.

Zheng, Wenfeng, Xiangjun Liu, and Lirong Yin (2021). "Sentence representation method based on multi-layer semantic network". In: *Applied sciences* 11.3, p. 1316.