

Neuro-Symbolic Architectures for Natural Language Comprehension

Gihan Sameera Wiriththamulla Gamage

-Bachelor of Science in Computer Science and Engineering (Honours),

University of Moratuwa 2020

A thesis submitted in total fulfilment of the requirements for the degree of
Doctor of Philosophy

La Trobe Business School
La Trobe University, Victoria, Australia

March 2025

For being my guiding light, for showing me the way and for being there everytime I fall,

*To my parents,
Amma & Thaththa*

For always standing by me during the hurdles of life by being my companion,

*To my loving wife,
Poornima*

Abstract

The introduction of Generative Artificial Intelligence (AI), particularly through connectionist large language models (LLMs), has transformed computer systems that process natural languages, revolutionising the way machines interact with and understand human communication. Traditional language processing methods have relied heavily on explicit rule-based techniques, whereas modern connectionist AI has evolved to perform more implicit and intelligent tasks such as Natural Language Generation (NLG) and Natural Language Understanding (NLU). These advancements have enabled significant progress in automated text processing, conversational agents, and knowledge retrieval. Despite these advancements, contemporary models still face critical limitations, particularly when compared to the depth and flexibility of human Natural Language Comprehension (NLC). In my PhD thesis, I investigate these limitations in the context of human language comprehension to determine novel AI algorithms and techniques that advance the state of the art of Generative AI for language comprehension. To address these challenges, I propose Neuro-Symbolic Architectures for NLC, an approach that integrates principles from both Symbolic AI and Connectionist AI. This hybrid framework is inspired by cognitive processes in human comprehension and is designed to overcome the constraints of purely connectionist or purely rule-based models. By leveraging the complementary strengths of these two paradigms, my research aims to build explainable, controllable, and robust AI systems. These architectures are empirically evaluated across various aspects of language comprehension, including emotion elicitation, natural language interactions, subtext comprehension, and the enhancement of word representations. Additionally, several practical applications and algorithms have been developed and deployed to support sustainable, safe, and trustworthy AI in complex domains such as healthcare, energy, and finance.

Keywords: Generative Artificial Intelligence, Large Language Models, Natural Language Comprehension (NLC), Connectionist AI, Symbolic AI, Neuro Symbolic Architectures

Statement of Authorship

Except where reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis accepted for the award of any other degree or diploma. No other person's work has been used without due acknowledgement in the main text of the thesis. This thesis has not been submitted for the award of any degree or diploma in any other tertiary institution. This work was supported by a Centre for Data Analytics and Cognition (CDAC) Stipend and a La Trobe University Full-Fee Research Scholarship.

Gihan Gamage
11th March 2025

Acknowledgements

With deep appreciation, I reflect on the invaluable guidance and support extended by my supervisors, Prof. Daswin De Silva and Dr. Harsha Moraliyage, in navigating this academic pursuit. Being my principal supervisor, Prof. Daswin De Silva guided me as I explored novel avenues in neuro-symbolic artificial intelligence for natural language comprehension. It was his knowledge, advice, and support that continually inspired me to think innovatively and advance my research in many directions, which paved the way for me to publish my research work in several high-impact journals and conferences. Thank you, this would not have been possible without your constant guidance and encouragement. My co-supervisor, Dr. Harsha Moraliyage, has supported me greatly by encouraging me to reach unprecedented heights in my research and academic career. I admire and appreciate your time and efforts advising me throughout my PhD journey, especially with your knowledge and guidance in generative artificial intelligence. During my four years as a research scholar, I met with many colleagues at the Research Centre for Data Analytics and Cognition at the La Trobe Business School. I cherish and appreciate all the fond memories of working together as a team. Thank you for being wonderful companions to me at La Trobe University. Furthermore, I would like to express my gratitude to La Trobe University for supporting my research by providing me with a Centre for Data Analytics and Cognition (CDAC) Stipend and a La Trobe University Full-Fee Research Scholarship, which helped me greatly in completing my research and covering travel costs for overseas research conferences.

Table of contents

List of figures	ix
List of tables	xii
Publications originated from the thesis	xiv
1 Introduction	1
1.1 Background	1
1.2 Motivation	2
1.3 Research Objectives	4
1.4 Research Questions	5
1.5 Research Contributions	6
1.6 Thesis Structure	8
2 Setting the Stage	9
2.1 Chapter Introduction	9
2.2 Natural Language Comprehension	10
2.3 Natural Language in Machines	11
2.4 Emergence of Large Language Models (LLMs)	12
2.5 Advantages of LLMs	15
2.6 Limitations of LLMs	16
2.6.1 Growing Complexity and Unsustainable Trends	17
2.6.2 Lack of Explainability in LLMs	19
2.6.3 Lack of Compositional Behaviours	20
2.7 Chapter Summary	23
3 Towards Neuro Symbolic Approach	25
3.1 Human Language Comprehension	25
3.1.1 Compositionality	26

3.1.2	Systematicity	28
3.1.3	Grounded Language Learning	29
3.1.4	Construction-Integration (CI) Theory	30
3.2	Leveraging Human Aspects in Computational Models	31
3.3	Neuro-symbolic Architectures for Natural Language Comprehension	32
3.3.1	Symbolic AI Methods for Natural Language Comprehension	33
3.3.2	Connectionist AI Methods for Natural Language Comprehension	35
3.3.3	Attempts in Building Neuro-Symbolic Architectures in the Literature	37
3.3.4	Enhancing the Effectiveness of LLMs	38
3.3.5	Proposed Neuro Symbolic Framework for Natural Language Com- prehension	40
3.4	Chapter Summary	42
4	Evaluating the Neuro-Symbolic Architectures in Natural Language Comprehension	43
4.1	Chapter Introduction	43
4.2	Enhancing Word Representations	45
4.2.1	Visually Grounding Word Embeddings	46
4.2.2	A Robust Word Embedding for Implicit and Explicit Expressions of Emotion	51
4.3	Disinformation Detection	57
4.3.1	Introduction	57
4.3.2	Methodology	58
4.3.3	Experiments	62
4.4	Detecting Idioms in Natural Languages	65
4.4.1	Introduction	65
4.4.2	Methodology	66
4.4.3	Experiments	69
4.5	Extremism Detection in Natural Languages	71
4.5.1	Introduction	71
4.5.2	Methodology	73
4.5.3	Experiments	75
4.6	Chapter Summary	81
5	Neuro-symbolic Architectures Emotion Elicitation	82
5.1	Chapter Introduction	82

5.2	Emotion AWARE: An Artificial Intelligence Framework for Adaptable, Robust, Explainable, and Multi-Granular Emotion Analysis	83
5.2.1	Background	84
5.2.2	Methodology	89
5.2.3	Experiments	96
5.3	Ablation Studies of Emotion Aware	109
5.3.1	Emotion Analysis of Cancer Council Victoria	109
5.3.2	Analysing the Role of Manager Emotions on Stock Returns	116
5.4	Chapter Summary	122
6	Neuro-symbolic AI for Natural Language Interactions	124
6.1	Chapter Introduction	124
6.2	La Trobe Energy Analytics Platform - LEAP	125
6.3	Interacting with Dynamic Structured Data Spaces Using Natural Language .	127
6.3.1	Introduction	127
6.3.2	Cooee-internal - Neuro-symbolic Architectures for Interacting with Structured Energy Spaces	128
6.4	Interacting with Multi-modal Unstructured Data Spaces using Natural Language	139
6.4.1	Introduction	139
6.4.2	Cooee-external - Neuro-symbolic Architectures for Interacting with Unstructured Energy Spaces	140
6.4.3	Advancing the Retrieval Process	142
6.4.4	Hierarchical Vector Spaces for Optimized Semantic Search	143
6.5	Leverage Neuro-symbolic Architectures to Implement Multi-agentic Behaviours in Energy Environments	146
6.5.1	Introduction	146
6.5.2	Multi-agentic Framework for Energy Task Automation	147
6.5.3	Experiments	150
6.6	Chapter Summary	153
7	Conclusion	155
7.1	Summary of the contributions	155
7.2	Addressing the research questions	157
7.3	Future directions	162
	Bibliography	164

List of figures

2.1	Evolution of Natural Language Processing Systems	11
2.2	From Conventional Techniques to Advanced Systems Powered by Generative AI	12
2.3	Generative AI Paradigm	13
2.4	Transformer Architecture (Vaswani <i>et al.</i> , 2017)	14
2.5	Different Types of LLM Architectures(Schoeninger, 2023)	15
2.6	Limitations of LLMs	16
2.7	Expanding Trend in LLM Size and Complexity (Asia, 2024)	17
2.8	LLMs Showing Fundamental Errors with Lack of Compositionality	22
2.9	The Reversal Curse of LLMs (Berglund <i>et al.</i> , 2024)	23
3.1	Elements of Human Language Comprehension	26
3.2	Examples of Compositionality	27
3.3	Systematicity in Human Language Comprehension	28
3.4	Grounded Language Learning	29
3.5	Construction-Integration (CI) Theory	31
3.6	Leveraging Human Aspects in Computational Models	32
3.7	Instance of Symbolic NLP on Sentiment Classification (Hardeniya and Borikar, 2016)	34
3.8	Connectionist Model Architecture - ANN (Asaad and Ali, 2019)	36
3.9	Enhancing LLM Outputs Through Prompt Engineering	38
3.10	Transfer Learning: Adapting Pretrained Models for Specialized Tasks	39
3.11	Guardrails on LLMs	39
3.12	Proposed Framework of Neuro-symbolic Architectures for NLC	41
4.1	Overview of the Applications Utilizing the Neuro-Symbolic Framework	44
4.2	Visually Grounding Word Embeddings	47
4.3	Image and Caption Pairs from COCO2017 Dataset	48

4.4	Emozen Model Architecture	52
4.5	Heatmaps of Cosine Similarity Scores Between Words with Paired Emotions	56
4.6	Multi-faceted Neuro-Symbolic Architecture for Fake News Detection . . .	59
4.7	Extended Pre-training for Token Classification with BERT	67
4.8	Term-level Annotation for Identifying Parts of Idiomatic Expressions	68
4.9	Common Facets in Extremism	71
4.10	Extremism Detection in Natural Languages	73
4.11	Distribution of Top Extremist Themes in COVID-19 Vaccination Tweets . .	77
4.12	Distribution of Top Extremist Themes in US Election 2020	79
5.1	PRISMA 2020 Flow Diagram for Updated Systematic Reviews Which In- cluded Searches of Databases, Registers and Other Sources	85
5.2	The Modular Composition of the Emotion AWARE Framework	89
5.3	The Emotion Embedding Space Generated by Module 3 of the Framework .	92
5.4	An Emotion Assemble Generated by the Emotion AWARE for Mixed Polarity Sample Text	95
5.5	Demonstrating Explainability of Emotion Assembles using Constituent In- tensity	107
5.6	Contributing Terms and Corresponding Intensity Scores for Emotion Ex- plainability	107
5.7	Utilising neuro-symbolic architectures to improve telehealth services in Cancer Council Victoria	111
5.8	Emotion trend calculation in the call transcripts	112
5.9	Call volumes by month from 2018-2021	114
5.10	Call volumes by positive and negative emotions for the years 2018-2021 . .	114
5.11	Emotion profiles by Call duration and four negative emotions	115
5.12	Average emotion trend variation for the patient and the nurse	115
5.13	Pipeline for analyzing the impact of manager emotions on stock returns . .	117
6.1	La Trobe energy analytics platform(LEAP) Data Space	126
6.2	High-level architecture of Cooee internal chatbot	129
6.3	Cooee-internal Q&A Dashboard Implementation	130
6.4	Cooee-internal Chat Pop-up Implementation	130
6.5	Instance of Entity Recognition in a Sample query	131
6.6	Resolving Contextualized questions using LLMs	136
6.7	Contextualized question answering examples	137
6.8	Cooee External Framework for interacting unstructured data spaces	140

6.9	Cohee-external Implementation (a)	141
6.10	Cohee-external Implementation (b)	142
6.11	Advanced retrieval augmented generation techniques for documented knowl- edge retrieval	142
6.12	Time complexity comparison of linear and hierarchical vector search with growing DB sizes	144
6.13	Multi-Agent Chatbot Architecture for Decision Support in Net-Zero Emis- sions Energy Systems	147
6.14	Prompt template for the intent recognizer component	148
6.15	Instances from event notifications and summary reports stores for RAG . . .	151
6.16	Chatbot workflow for decision support in net zero energy systems	152
6.17	End-to-End Workflow of the Visualizer Agent in the Chatbot System	153

List of tables

4.1	Comparison of Word Similarity Benchmarks and Models	50
4.2	Emotion-related keywords used to extract the training dataset	53
4.3	Comparison Against State-of-the-art Word Embedding Models	56
4.4	Description of Semantic Aspects and Corresponding Features	62
4.5	Performance Comparison of Classification Models Before and After Stan- dardization	63
4.6	Datasets Used to Cross-validate the Proposed Fake-News Detection Approach	64
4.7	Accuracies Obtained using Voting Classifiers with Different Base Models .	64
4.8	Idiom Dataset Statistics	68
4.9	Idiomatic token classification results for the experiments	69
4.10	Comparing the Proposed Method with Non-contextual Embedding-based Classifiers	70
4.11	HateXplain Dataset Statistics(Mathew <i>et al.</i> , 2020)	73
4.12	Target Groups Considered for the Annotation	74
4.13	Monthly Statistics of COVID-19 Vaccination Tweets	76
4.14	Example Tweets Showcasing Various Top Extremism Themes in COVID-19 Vaccinations	77
4.15	Hate/Offensive Text Analysis for US Election 2020 Tweets	79
4.16	Example Tweets Showcasing Various Top Extremism Themes in US Election 2020	80
5.1	Comparison of Emotion AWARE with related work in emotion detection . .	88
5.2	Alignment of the 2, 8 and 14 emotion classification schemes	91
5.3	Description of Datasets Used in the Experiments	97
5.4	Nine studies evaluating and demonstrating capabilities of the proposed Emo- tion AWARE framework	99
5.5	Comparison of Results for Two-emotion Assembles using ISEAR and Twitter100	
5.6	Comparison of results for four emotion assembles (anger, fear, sadness, joy)	101

5.7	Comparison of results for four emotion assembles (disgust, surprise, trust, anticipation)	101
5.8	Demonstrating the Elicitation of 2, 8 and 14 Emotion Assembles in Increasing Granularity	102
5.9	Comparison of results generated by Emotion AWARE adapted for the finance sector	103
5.10	Comparison of Results Generated by Emotion AWARE When Adapted for the Technology Sector	103
5.11	Emotion Profiles Based on Sentence Valence and Intensities	104
5.12	Demonstrating Robustness across Intensifiers and Inhibitors	105
5.13	Performance of Inhibitor and Intensifier Detection	105
5.14	Results for Robustness of Emotion AWARE in Negation Detection	106
5.15	Ten Most Frequent Keywords per Emotion Category in Fairy Tales Dataset	108
5.16	Sample distribution by quarters	118
5.17	Sample distribution by industry	119
5.18	Emotion Classification of CEO Statements during the Q&A	120
5.19	Emotion Statistics for Different Sections of the Transcripts	121
5.20	Dependent and Control variables of the analysis	121
5.21	Correlation matrix of selected variables with statistical significance levels	122
6.1	Entity Recognition Classes in NLU Module	131
6.2	Regex classes and their corresponding standard templates	132
6.3	Comparing Cooe QnA capability with TAPAS and Text-to-SQL	134
6.4	Comparison of the response quality in linear and hierarchical search	146
6.5	Expert Agents in the Multi-Agent System for Net-Zero Emissions	149

Publications originated from the thesis

The research contributions of this thesis have been published in the following peer-reviewed journals and conference proceedings.

1. **G. Gamage**, D. De Silva, N. Mills, D. Alahakoon, and M. Manic, “Emotion AWARE: an artificial intelligence framework for adaptable, robust, explainable, and multi-granular emotion analysis,” *Journal of Big Data*, vol. 11, no. 1. Springer Science and Business Media LLC, Jul. 10, 2024. doi: 10.1186/s40537-024-00953-2.
2. **G. Gamage**, D. De Silva, A. Adikari, and D. Alahakoon, “A BERT-based Idiom Detection Model,” 2022 15th International Conference on Human System Interaction (HSI), vol. 235. IEEE, pp. 1–5, Jul. 28, 2022. doi: 10.1109/hsi55341.2022.9869485.
3. **G. Gamage**, N. Mills, P. Rathnayaka, A. Jennings, and D. Alahakoon, “Cooee: An Artificial Intelligence Chatbot for Complex Energy Environments,” 2022 15th International Conference on Human System Interaction (HSI), vol. 21. IEEE, pp. 1–5, Jul. 28, 2022. doi: 10.1109/hsi55341.2022.9869464.
4. **G. Gamage** et al., “Augmenting Industrial Chatbots in Energy Systems using ChatGPT Generative AI,” 2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE). IEEE, Jun. 19, 2023. doi: 10.1109/isie51358.2023.10228101.
5. **G. Gamage** et al., “Multi-Agent RAG Chatbot Architecture for Decision Support in Net-Zero Emission Energy Systems,” 2024 IEEE International Conference on Industrial Technology (ICIT), vol. 201. IEEE, pp. 1–6, Mar. 25, 2024. doi: 10.1109/icit58233.2024.10540920.
6. A. Adikari, **G. Gamage**, D. de Silva, N. Mills, S.-M. J. Wong, and D. Alahakoon, “A self structuring artificial intelligence framework for deep emotions modeling and analysis on the social web,” *Future Generation Computer Systems*, vol. 116. Elsevier BV, pp. 302–315, Mar. 2021. doi: 10.1016/j.future.2020.10.028.

7. L. Langer, **G. Gamage**, D. Ranasinghe, D. De Silva, and P. R. Mather, “Do Emotions Matter? The Role of Manager Emotions on Stock Returns,” SSRN Electronic Journal. Elsevier BV, 2023. doi: 10.2139/ssrn.4526589.
8. S. Ranasinghe, **G. Gamage**, D. de Silva, N. Mills et al. “An Artificial Intelligence Framework for the Detection of Emotion Transitions in Telehealth Services,” 2022 15th International Conference on Human System Interaction (HSI), vol. 13. IEEE, pp. 1–5, Jul. 28, 2022. doi: 10.1109/hsi55341.2022.9869503.
9. P. Rathnayaka, **G. Gamage**, D. De Silva, D. Alahakoon, and M. Manic, “EmoZen: A Robust Word Embedding for Implicit and Explicit Expressions of Emotion,” IECON 2023- 49th Annual Conference of the IEEE Industrial Electronics Society, vol. 12. IEEE, pp. 1–4, Oct. 16, 2023. doi: 10.1109/iecon51785.2023.10312470.
10. K. Prabagar, K Srikandabala, **G. Gamage**, D. De Silva, “Investigating COVID-19 Vaccine Messaging in Online Social Networks using Artificial Intelligence,” 2022 15th International Conference on Human System Interaction (HSI), vol. 13. IEEE, pp. 1–6, Jul. 28, 2022. doi: 10.1109/hsi55341.2022.9869484.
11. A. Matharaarachchi, **G. Gamage**, D. De Silva., “Optimizing Generative AI Chatbots for Net-Zero Emissions Energy Internet-of-Things Infrastructure,” Energies, vol. 17, no. 8. MDPI AG, p. 1935, Apr. 18, 2024. doi: 10.3390/en17081935.

Chapter 1

Introduction

1.1 Background

In today's digital landscape, natural languages or text can be considered as one of the major forms of data we have. Especially with the arrival of the World Wide Web(WWW), social media, and online messaging platforms, natural language has become an increasingly popular mode of communication. As a result, Natural Language Processing (NLP) has emerged and well-established as an extensively researched field in Artificial Intelligence (AI). Today we have a wide spectrum of NLP systems which extends from basic word-matching methods to sophisticated and complex Large Language Models (LLM) such as Generative Pre-trained Transformers (GPT).

Out of diverse NLP methods which solves various natural language problems computationally, LLMs can be considered as the current state-of-the-art in most of the NLP downstream tasks. These models utilizes internet scale of unstructured natural language data in complex transformer neural architectures to provide Natural Language Understanding (NLU) and Natural Language Generation (NLG) capabilities that closely match the human creativity. These advanced NLU and NLG skills of LLMs unlock a wide range of abilities in NLP once believed to be impossible with traditional methods. LLMs aid in tasks like chatbot development, brainstorming, content creation, and coding assistance, among others. Furthermore, with specific training, these models can even surpass human performance in certain benchmarks. However, despite their impressive capabilities in various use cases, these models also have inherent challenges. Understanding these limitations is important to enhance these models as well as utilizing and customizing them for real-world applications beyond information retrieval and text generation.

One major limitation of LLM is its inherited complexity. Due to the nature of transformer architecture, the accuracy of these models scales up with both the size of the model(number

of parameters) and the volume of data used for the training process. As a result, language models are in a rat race to beat benchmark scores by expanding the model sizes. Furthermore, along with the model and training data, these models demand more sophisticated computing resources to carry out these extensive training processes. So this ultimately narrows down the development of language models to a selective cohort big companies that has access to internet-scale datasets as well as advanced computing resources.

Another main limitation of LLM is the lack of explainability. These models are often referred to as “opaque models” because their internal workings and decision-making processes are not transparent or easily interpretable. This is a primary concern when building responsible and safe AI applications in sensitive domains like healthcare, finance, or legal systems. In addition to that, explainability is also an essential factor in controllability. So lack of explainability also results in lack of controllability in these models. So without controllability, these models sometimes can end up with unintended behaviours such as hallucinations and biases. Additionally, these models show a lack of compositionality in specific contexts, especially in tasks related to commonsense reasoning, arithmetic, and world knowledge, where their performance often falls short. This presents a significant issue given the substantial financial investment in training these models, only to find they struggle with basic arithmetic tasks that a young child could manage.

When considering the recent development trends and ongoing challenges in LLMs, it underscores the need that these sophisticated NLP systems must be redirected to gain features such as affordability, explainability, and controllability, thereby enhancing their sustainability. In this premise, the goal of this thesis is to investigate how to build more sustainable solutions rather than complexifying the neural architectures that can comprehend natural languages more effectively. It further aims to propose innovative solutions to key challenges in the current literature on NLP with inspiration from human language comprehension. Through a combination of theoretical analysis, empirical experiments, and practical implementations of real-world problems, this research seeks to advance existing NLP approaches, thereby fostering a future where the benefits of AI can be harnessed.

1.2 Motivation

In artificial intelligence, the ultimate goal is to create systems that can perform tasks with the same efficiency and understanding as humans. This involves developing AI that can think, learn, and adapt in ways that closely mimic human intelligence and decision-making processes. To achieve this, it is not sufficient to focus solely on technical advancements; we must also incorporate strategies and techniques that humans use to learn and make decisions.

Aligning technological progress with an understanding of human cognition and behaviour is crucial for creating AI systems that truly resemble human performance.

According to neuroscience and psychology, language comprehension is a cognitive process involving many brain activities. These skills start to develop in the early childhood stages of humans like the age of three and then grow with education and experience (Gernsbacher and Kaschak, 2003). A combination of components, including vocabulary, understanding of language structures, memory, inference and reasoning, and background knowledge underpins this evolution (Beck *et al.*, 2013; Daneman and Merikle, 1996; Gernsbacher, 1999). Moreover, the effectiveness and efficiency of language comprehension are greatly enhanced by techniques such as compositionality, systematicity and grounded language learning.

Compositionality is a core principle in how humans understand language. It refers to our ability to grasp the meaning of a complicated phrase or sentence by piecing together the meanings of its components. This allows humans to understand new combinations of words they have never encountered or learned before (von Heusinger *et al.*, 2019). In addition to that, another underlying technique is the systematicity. It explains the systematic and rule-governed manner of language understanding. This technique differentiates the meanings of expressions, even though they are made from the same compositional elements (Fodor and Pylyshyn, 1988). Thus, rules of the language, such as grammar, word order and other syntactical rules, are also a significant part of language comprehension. In addition to that, grounded language learning enables humans to get a more comprehensive understanding of language by grounding the language symbols in multi-modal real-world references (Suglia *et al.*, 2024). Together, these techniques reduce the need for infinite memorization of language variations, which would otherwise impose an impractical cognitive load. Instead, humans rely on a limited set of language primitives and inferencing capabilities to interpret meaning, achieving both efficiency and scalability.

In contrast, modern computational methods of language comprehension, such as LLMs, rely heavily on connectionist, information-intensive methods driven by complex neural architectures. Though these methods provide higher accuracies by leveraging larger quantities of data, they often present many challenges, as discussed earlier, which question the sustainable computational methods in language comprehension (Buyl *et al.* (2024)). Additionally, they are inherently limited to the text modality, which restricts their ability to capture the full spectrum of human language understanding. Thus, current tendencies clearly widen the gap between contemporary computational approaches to human strategies of language comprehension. Traditional rule-driven language processing methods, on the other hand, are inherently more aligned with techniques such as compositionality and systematicity. These methods are often open systems based on explicit rules and structured reasoning, which offer

better explainability and controllability compared to modern connectionist methods(Dziri *et al.*, 2023). However, they often lack scalability and are typically suitable only for simpler NLP tasks, making them insufficient for handling complex, real-world problems.

A promising way forward is the adaptation of neuro-symbolic approaches that integrate the strengths of contemporary connectionist models with the structured reasoning and transparency of rule-based systems. In this way, it overcomes the drawbacks of modern systems by fusing the explainability, controllability, and multimodal grounding of symbolic methods with the scalability and learning power of neural architectures. It enables a more holistic alignment with human language comprehension processes, minimizing the need for unsustainable model expansion and allowing systems to incorporate rich, multimodal understanding matching to how humans learn and interpret language.

1.3 Research Objectives

Motivated by the challenges and limitations of contemporary state-of-the-art natural language comprehension methods, this research aims to explore and address the growing need for sustainable, explainable, and efficient approaches to computational language understanding. Drawing inspiration from human language comprehension, the primary goal of this thesis is to design and develop a neuro-symbolic framework for natural language comprehension that effectively integrates connectionist and symbolic methods.

The specific objectives of this research are:

1. Investigate the limitations and challenges of current computational approaches, particularly focusing on the unsustainable trends and gaps in controllability, explainability, and alignment with human language comprehension techniques.
2. Study the mechanisms of human language comprehension to understand how humans achieve efficient, effective, and systematic language understanding through principles such as compositionality, systematicity, and grounded language learning.
3. Explore the potential of neuro-symbolic AI to bridge the gap between human-inspired principles and computational systems, enabling natural language comprehension methods that better align with human cognition.
4. Design and develop a neuro-symbolic AI framework that encapsulates various aspects of natural language comprehension, including controllability, explainability, and learnability. This framework will serve as the primary contribution of this research.

5. Apply the proposed framework to build practical applications in sensitive and high-stakes domains such as healthcare, finance, and energy. These applications will prioritize safety, trustworthiness, and alignment with human language comprehension features.
6. Conduct empirical evaluations and demonstrations to validate the effectiveness of the proposed neuro-symbolic framework in addressing the identified challenges. This will include assessing the framework's capabilities in achieving sustainable, explainable, and efficient natural language comprehension.

1.4 Research Questions

This doctoral dissertation seeks to answer the following questions;

- RQ1: What is the role of current computational methods in enabling machines to intelligently process natural language?
 - (a) Has the evolution of computational methods over time led to significant advancements in language comprehension capabilities?
 - (b) Which underlying mechanisms and key advantages enable large language models (LLMs) to outperform previous approaches and become the current state-of-the-art in natural language processing?
 - (c) What current trends and limitations of large language models (LLMs) reveal significant research gaps that require further investigation?
- RQ2: What cognitive and linguistic principles from human language comprehension can inform the design of effective AI systems for natural language comprehension?
 - (a) To what extent do the principles of compositionality and systematicity enable humans to comprehend language more efficiently and effectively?
 - (b) How significantly do humans leverage other modalities to ground the language comprehension process?
 - (c) To what extent does the construction-integration theory explain systematic and hierarchical aspects of language understanding?
- RQ3: How can neuro-symbolic architectures be leveraged to bridge the gap between human and computational methods for natural language understanding?

- (a) To what extent do traditional symbolic methods provide advantages and face limitations in language processing?
 - (b) Which strategies can effectively enhance the performance and applicability of large language models (LLMs) within the connectionist paradigm?
 - (c) Could neuro-symbolic architectures, by merging symbolic reasoning with connectionist learning, create more robust systems that align with human cognitive strategies?
- RQ4: How can neuro-symbolic architectures be utilized to develop practical applications for sustainable development?
 - (a) How significantly can incorporating implicit elements like emotion and visual knowledge improve the robustness of word representations in natural language processing?
 - (b) To what extent can neuro-symbolic architectures effectively capture and analyze complex language comprehension aspects such as figurative, deceptive, and extremist language usage?
 - (c) How effectively do neuro-symbolic architectures facilitate robust, explainable, and controllable emotion elicitation in natural language processing systems?
 - (d) To what extent can neuro-symbolic architectures improve natural language interactions in critical energy infrastructures by enabling more effective, multi-modal, and automated communication systems?

1.5 Research Contributions

- Conducted a comprehensive review examining the natural language comprehension problem and the evolution of computational methods attempting to solve it. Then, it further explores the emergence of LLMs as state-of-the-art, highlighting the strengths of LLMs while critically analyzing three major limitations: increasing complexity, lack of explainability, and the absence of compositional behaviour. The review underscores the requirement for sustainable and human-aligned approaches. (**Chapter 2**)
- Performed an extensive examination of human language comprehension strategies, focusing on their efficiency and effectiveness. This analysis identified critical gaps between human cognitive processes and the contemporary methods employed by LLMs, providing insights for developing neuro-symbolic strategies. (**Chapter 3**)

- Investigated the integration of human dimensions of comprehension to computational techniques with the intersection of symbolic and connectionist AI in NLP. Based on this analysis, a novel neuro-symbolic framework was proposed, designed to enable safe and trustworthy deployments by combining the strengths of both paradigms. **(Chapter 3)**
- Proposed innovative techniques to improve word representation learning. These include the development of EmoZen: Emotion-fused word embeddings to incorporate emotional context into word representations and visually grounded word embeddings that integrate visual modality to ground language understanding in real-world contexts. These contributions enhance the overall effectiveness of word embeddings in neuro-symbolic systems. **(Chapter 4)**
- Adapted the proposed neuro-symbolic framework to address various natural language comprehension challenges, including deceptive language detection, extremism elicitation, and figurative language detection. Through empirical evaluations, the framework demonstrated its ability to provide features such as explainability, controllability, and sustainability while avoiding the unsustainable scaling associated with purely connectionist approaches. **(Chapter 4)**
- Introduced the Emotion AWARE framework, a robust, adaptable, and explainable AI system for multi-granular emotion analysis. The framework was validated across eight benchmark datasets and nine studies, demonstrating its effectiveness in emotion elicitation, domain adaptability, robustness, and explainability. Furthermore, it was deployed in two case studies: emotion analysis for Cancer Council Victoria and the impact of managerial emotions on stock returns. **(Chapter 5)**
- Leveraged the neuro-symbolic framework to enhance natural language interaction through the development of the Cooee Energy AI Chatbot. This system was deployed and evaluated within the La Trobe Energy Analytics Platform (LEAP), featuring three integrations: Cooee-Internal for structured data interactions, Cooee-External for multimodal unstructured data spaces, and a multi-agent implementation of Cooee for task automation in energy systems. These applications demonstrate the framework's scalability, flexibility, and real-world impact. **(Chapter 6)**

1.6 Thesis Structure

The rest of this thesis is organised as follows. Chapter 2 provides a comprehensive review of the literature on the current state of natural language comprehension, covering the evaluation of NLP systems, the rise of large language models, and their various capabilities. It also categorizes and explains the limitations of large language models. Chapter 3 introduces the Neuro-Symbolic AI Framework for Natural Language Comprehension, drawing inspiration from human language processing. It discusses key components and cognitive theories that support effective language understanding in humans, as well as how combining symbolic and connectionist paradigms can enhance neuro-symbolic architectures. Chapter 4 outlines practical applications of neuro-symbolic AI in natural language comprehension. It describes experiments aimed at improving word representations through visually grounded learning and emotion infusion and discusses approaches for detecting extremism, figurative language, and deceptive content using neuro-symbolic methods. Chapter 5 focuses on emotion elicitation using neuro-symbolic approaches, detailing the Emotion AWARE framework and presenting two ablation studies. Chapter 6 explores the use of neuro-symbolic AI for natural language interaction, with a focus on the Coeee Energy AI chatbots and their interactions in structured, unstructured, and multi-agent task automation. Chapters 3 to 6 represent the primary contributions of this thesis. Chapter 7 concludes the thesis by summarizing the contributions, aligning the findings with the research questions, and discussing future directions and potential areas for further research.

Chapter 2

Setting the Stage

2.1 Chapter Introduction

With the rise of social media and contemporary online systems, the popularity of natural language data has increased along with that of NLP systems. When considering the evolution of NLP systems, traditionally, these systems were limited to explicit language processing, primarily through rule-based methods. However, recent developments in NLP have expanded its capabilities beyond explicit processing to include implicit aspects like natural language understanding (NLU) and natural language generation (NLG). One of the driving forces behind this transformation is the introduction of Generative AI, particularly large language models (LLMs). These models have revolutionized the NLP landscape by enabling more advanced and human-like interactions with language and making many tasks that were once thought to be unachievable, possible. Currently, LLMs can be considered state-of-the-art in most benchmark NLP tasks with significant margins.

Although these models show advanced performance across many tasks in NLP, they still have several limitations. Understanding and resolving these limitations is also crucial for improving the current state of LLMs and adapting them for more sophisticated use cases beyond language generation. In this chapter, I analyze and summarize these limitations, categorizing them into three main areas: complexity, lack of explainability, and lack of compositionality. The word prediction accuracy of LLMs scales up with the expansion of training data and model sizes. As a result of this, there is the current trend of expanding the models to beat the benchmark scores, which ultimately makes language models even larger. This restricts the development of LLMs to a limited cohort that possesses access to vast datasets and substantial computing resources. In addition to that, the opaque nature of these complex neural architectures results in limited controllability and explainability, introducing challenges for safe and trustworthy deployment in sensitive domains. Additionally, LLMs'

lack of compositional behaviour frequently leads to basic mistakes that limit their usefulness and applicability even after lengthy training procedures. In this chapter, by comprehensively analysing these challenges of contemporary language comprehension systems, I set the stage for exploring more effective and sustainable approaches to natural language comprehension.

2.2 Natural Language Comprehension

Natural languages are the common languages that humans use naturally to communicate with each other, including spoken, written, and signed languages. Natural languages are rich in cultural, social, and contextual underpinnings and have developed naturally, in contrast to artificial languages (like computer languages). Although these languages follow grammar and syntax rules, they are flexible and frequently implicit or ambiguous, enabling speakers to express nuanced meanings, emotions, and intentions.(Jurafsky and Martin, 2024). Natural language comprehension (NLC) stands for the process by which an agent (a human or computational system) understands, interprets, and responds meaningfully to natural language inputs or commands(Trischler *et al.*, 2016). Computationally speaking, it involves building models that can comprehend and produce human language in ways that resemble human comprehension, frequently for tasks like summarising, discussion, language translation, and question answering.

Natural language comprehension is central to several important components that enable systems to process language meaningfully. One of the main components is Syntax and Grammar, which build the foundation by defining the structural rules for sentence formation(Jurafsky and Martin, 2024). To complete the comprehension, it needs to extend beyond simply processing words and syntax into capturing the meaning, context, and intent(Jurafsky and Martin, 2024). Semantics builds on this by interpreting word meanings based on context, managing ambiguities and polysemy. Pragmatics adds depth by incorporating social context, intent, and implied meanings, making it essential for nuanced understanding(Clark, 1996). Furthermore, the world and contextual awareness help to comprehend linguistic references and implications by grounding language in actual facts(Bisk *et al.*, 2020). Logical deduction and hypothesis testing are made possible by reasoning and inference skills, which deduce relationships in language that aren't explicitly stated. Finally, Sentiment Recognition captures the emotional undertones and attitudes embedded in language, enhancing interpretive accuracy in applications sensitive to user emotions(Poria *et al.*, 2017).

2.3 Natural Language in Machines

Rapid digitalization enables us to access, participate in and experience the world virtually through applications, devices and networks. In this digital era, text data is rapidly emerging as one of the fastest-growing types of big data. We come across textual data in numerous ways in everyday life. One of the main sources of textual data is social media. Various platforms and applications such as Facebook, Twitter, and Reddit produce thousands of terabytes of text data daily with increasing popularity. Messaging services such as WhatsApp and Messenger, as well as emails, also produce a considerable amount of text data. Apart from social media, eBooks and academic databases, websites, blogs and various news sites also contain large volumes of text data.

With the rapid enhancement of text data, language-based AI systems have also advanced by leaps and bounds, changing traditional notions of what this technology can do. Natural language comprehension systems span a wide spectrum, ranging from basic word-matching techniques to complex and expansive generative pre-trained architectures. These technologies enable a wide range of language processing tasks such as sentiment and emotion analysis, machine translation, information extraction, entity recognition, document clustering, keyword extraction, information retrieval (search), topic modelling, text classification, text summarization, natural language understanding and generation.

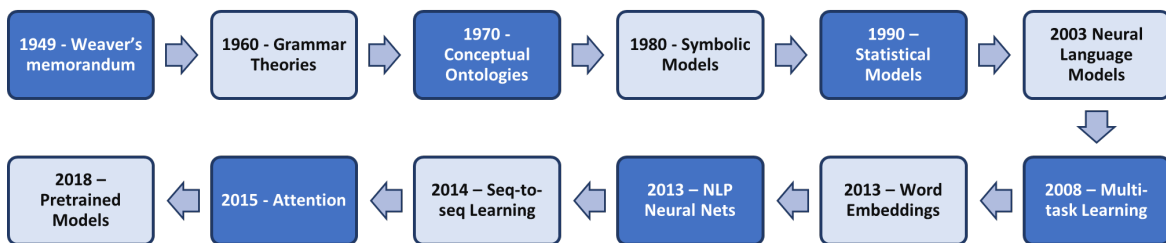


Fig. 2.1 Evolution of Natural Language Processing Systems

Figure 2.1 showcases the evolution of NLP systems highlighting major developments from Warren Weaver's initial proposition of machine translation(Chérargui, 2012) to the introduction of sophisticated pre-trained large language models(Vaswani *et al.*, 2017) which is the state-of-the-art today. After Warren's experiments, NLP was further motivated by foundational grammar theories(Bierwisch, 2015) and conceptual ontologies(Sowa, 1984) in the mid-20th century, advancing through symbolic and statistical models which laid the groundwork for the use of neural networks around 2000s(Lecun *et al.*, 1998). The progression includes key innovations such as multitask learning, word embeddings(Mikolov *et al.*, 2013a), and sequence-to-sequence models(Sutskever *et al.*, 2014a), culminating in

the introduction of attention mechanisms and the development of robust pre-trained models like BERT(Vaswani *et al.*, 2017) and GPT(Radford and Narasimhan, 2018) by 2018. This timeline further highlights a shift from rule-based to data-driven approaches in NLP, driven by advancements in computational power and machine learning techniques, which have significantly enhanced the field's capability to process and understand complex language data.

2.4 Emergence of Large Language Models (LLMs)

When summarizing the evolution of the NLP systems, it's clear that they are traditionally limited heavily to rule-based approaches, which process only the explicitly available information from the language. However, the invention of neural networks marked a significant advancement, leading these approaches to more data-intensive methods. This invention enabled us to reach more profound and more intelligent areas by leveraging large amounts of human language datasets in extensive training processes with learning patterns. The word embeddings especially helped explore more advanced aspects of natural languages, such as contextual understanding and figurative speech. This trend was further accelerated notably with the emergence of Generative AI technologies, which helped achieve natural language understanding and generation capabilities that closely matched human performance, reaching milestones once thought to be unattainable.

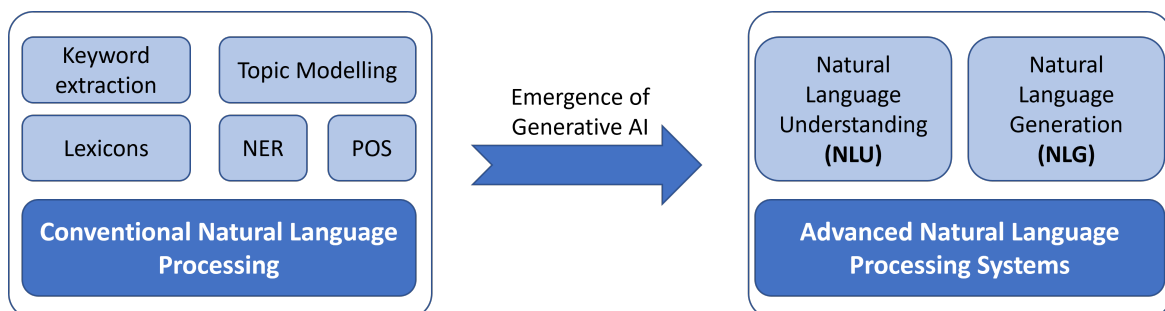


Fig. 2.2 From Conventional Techniques to Advanced Systems Powered by Generative AI

Generative AI is a branch of AI that aims to produce new media, including text, graphics, and other types, that closely mimic content produced by humans(Mondal *et al.*, 2023). These methods utilize deep learning technologies such as generative adversarial networks(GAN)(Goodfellow *et al.*, 2014), generative pre-trained transformers(GPT)(Brown *et al.*, 2020; OpenAI, 2023) and variational autoencoders (VAE)(Pinheiro Cinelli *et al.*, 2021) to imitate the patterns, styles and characteristics found in the training data to generate fresh content. Because they collect and analyse large quantities of data throughout their

training phases and incorporate meaningful human feedback, the final works frequently have a high degree of realism, making it difficult to tell them apart from content produced by real people (Lambert *et al.*, 2022; Ziegler *et al.*, 2019). The use of Gen AI technologies is expanding because of their growing sophistication and usefulness. Gen AI models, such as LLMs and diffusion models, have especially extended to various domains. These applications encompass general conversational services like ChatGPT (OpenAI, 2024), software engineering and programming assistance such as Copilot (GitHub, 2024), text-to-speech and voice cloning services like ElevenLabs (ElevenLabs, 2024) and creation of highly realistic artwork demonstrated by DALL-E 3 (Ramesh *et al.*, 2022), Stable Diffusion (AI, 2024), and Midjourney (MidJourney, 2024).

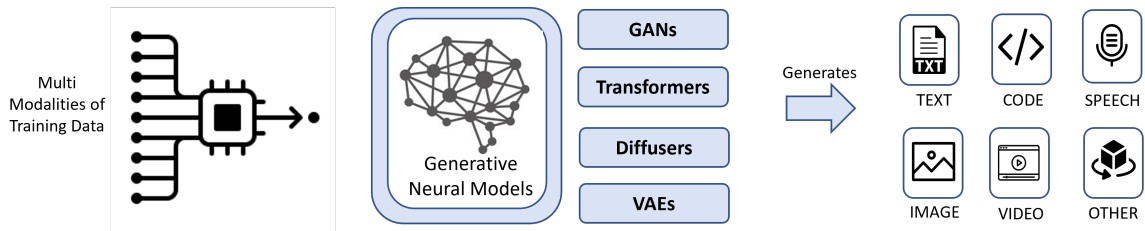
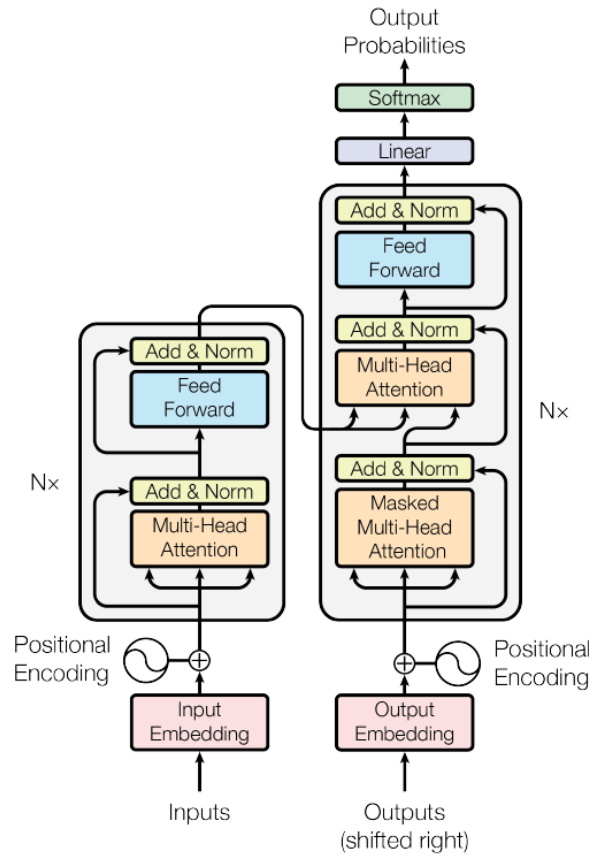


Fig. 2.3 Generative AI Paradigm

Most generative language models popular today are built upon the seminal innovation of the transformer architecture introduced by Google in 2017 with BERT (Bidirectional Encoder Representations from Transformers) (Vaswani *et al.*, 2017). Transformer architecture's ability to leverage the self-attention mechanism, support parallel processing, and scalability distinguished it from the existing best-performing methods such as RNNs (Rumelhart *et al.*, 1986) and LSTMs (Hochreiter and Schmidhuber, 1997), leading to significant advancements in the field of NLP. BERT demonstrated state-of-the-art performance across numerous benchmark tasks, including major NLP benchmarks like the Stanford Question Answering Dataset (SQuAD) (Rajpurkar *et al.*, 2016) and the General Language Understanding Evaluation (GLUE) (Wang *et al.*, 2018a) benchmark. This groundbreaking work laid the foundation for developing even more advanced models, such as ChatGPT (OpenAI, 2024) and other LLMs we have today.

In general, the transformer architecture consisted of both encoder and decoder stacks, which allow text inputs to be converted into text outputs. This process is powered by an internet-scale corpus of text data, books, and websites, allowing it to capture the nuances and complexities of human natural language. For instance, the transformer architecture of BERT allows one to read the entire sequence of words from left to right and right to left, understanding the context of words based on their surrounding words, resulting in more accurate language comprehension. When it comes to more recent LLMs like ChatGPT (OpenAI,

Fig. 2.4 Transformer Architecture (Vaswani *et al.*, 2017)

2024), GPT-4(OpenAI, 2023), and Claude Opus are also based on transformer architectures. These models, however, are typically designed using only the decoder part of the transformer. Unlike BERT, which focuses on understanding text, models like GPT are geared towards generating text. This difference in design intention leads to different capabilities and use cases.

Thus, LLMs are commonly classified into three main architectures. Encoder-based models (e.g., BERT, RoBERTa) focus on understanding and contextualizing input text, excelling at tasks like classification and generating word representations. Decoder-based models (e.g., GPT-3, GPT-4) are optimized for text generation, predicting the next word in a sequence for tasks like text completion. Encoder-Decoder Models (e.g., T5, BART) combine encoding and decoding, making them suitable for tasks that transform input to output, like translation and summarization(Schoeninger, 2023). Moreover, LLMs today utilize methods such as Human Feedback Reinforcement Learning (HFRL)(Lambert *et al.*, 2022) to enhance their performance by incorporating human feedback into the training process. This integration helps models adapt better to align with humans, with their feedback

and present more human-like behaviours. Another significant trend in LLM development is the tendency to create bigger models, both in terms of training corpora and the number of parameters. Larger models tend to be more accurate in modelling word sequences because they can capture finer details about word arrangements. Primarily, these pre-trained LLMs can be utilized in two main ways to cater to downstream tasks: first, to generate text outputs, and second, to create embeddings, which are numerical representations of the texts.

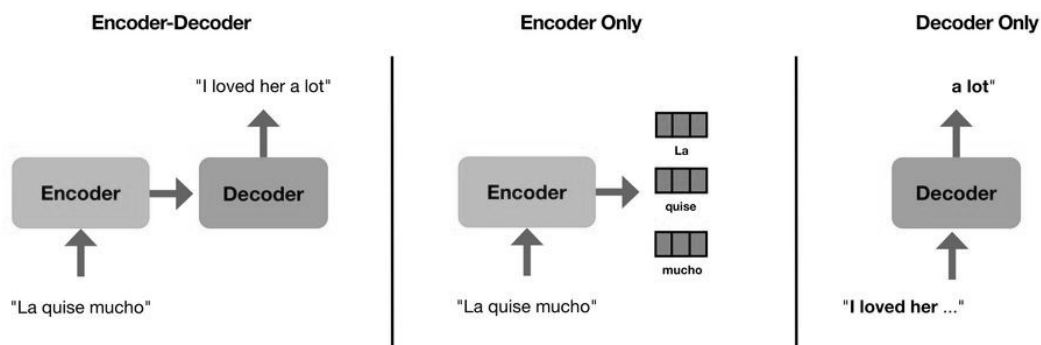


Fig. 2.5 Different Types of LLM Architectures(Schoeninger, 2023)

2.5 Advantages of LLMs

Today, we have a variety of sophisticated LLMs that are open-source as well as commercially available. Popular commercial models include Claude Opus from Anthropic, which is made for safe and interpretable AI applications(Anthropic, 2023), and OpenAI's GPT-4, which demonstrates state-of-the-art performance in numerous language tasks with human-like accuracy and fluency(OpenAI, 2023). On the open-source front, models such as EleutherAI's GPT-Neo and GPT-J offer easily accessible substitutes that allow broader experimentation and development within the AI community(Black *et al.*, 2021). Furthermore, new open-source models that offer great performance and versatility for a range of applications, like LLaMA 3 from Meta(Touvron *et al.*, 2023) and Mistral modals(Jiang *et al.*, 2023), push the limits of what is feasible in natural language processing.

Understanding context, generating human-like text, answering questions, and having meaningful conversations are just a few of the many abilities that LLMs like ChatGPT provide across a range of applications. These abilities allow LLMs to respond in a logical and

contextually appropriate manner, which makes them essential for activities like information retrieval, content creation, and text completion.(Radford and Narasimhan, 2018). The remarkable natural language understanding of LLMs is one of their biggest benefits. With extensive training in vast amounts of human languages, they can interpret and comprehend text input, mimicking human language understanding, which has numerous applications. Compared to chatbots based on conventional methods, chatbots powered by LLMs can comprehend user enquiries and deliver pertinent responses, greatly improving customer assistance and engagement.(Gamage *et al.*, 2022).

In addition to that, LLMs also show exceptional skills in content generation. This skill can be utilized in various fields, including marketing, where it can create coherent and contextually relevant articles, marketing copy, product descriptions, and reports for enhanced customer interaction. Moreover, question-answering and retrieving information from vast data sources make them invaluable tools for research and data analysis. This utilization can significantly enhance productivity by automating tasks, summarizing information, and aiding in research by saving time and human labour efforts.

2.6 Limitations of LLMs

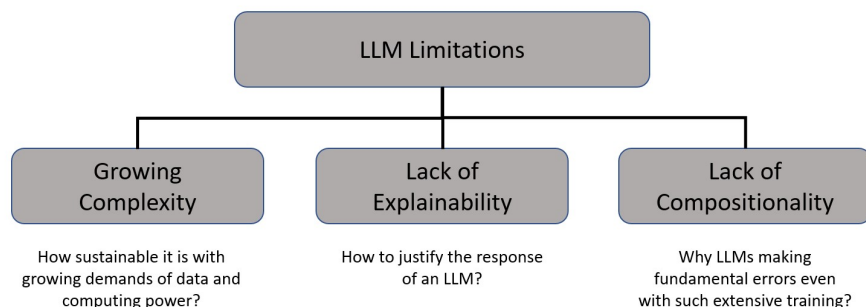


Fig. 2.6 Limitations of LLMs

Although LLMs showcase significant skills and performances across various aspects of language comprehension, they also have several limitations. Because of this, while these models might show great accuracies in benchmark tasks, their performance can vary significantly in certain other practical aspects. Especially in real-world situations, it requires consideration of many other matrices that benchmarks may not fully capture. Understanding these shortcomings is crucial for uplifting the current state of natural language comprehension systems beyond mere text generation or information extraction activities. In my thesis, I analysed and categorize these limitations into three main areas: Growing complexity, lack of

explainability and controllability, and lack of compositionality. The following sections will explain each of these categories in detail.

2.6.1 Growing Complexity and Unsustainable Trends

With the current trends, AI technologies are evolving and transitioning at a remarkable pace with various innovations and technologies. Firstly, the revolution started in November 2022, with the release of OpenAI's ChatGPT (OpenAI, 2024). It gained vast popularity, especially with the public web interface, which gave accessibility to the majority of technical as well as non-technical users. Today, most of the capabilities of ChatGPT, once considered surprising, have become common practices. Additionally, today we have all big tech giants, including Anthropic, Google, and Meta, have introduced their own language models, such as Claude (Anthropic, 2023), Gemini (Petrov *et al.*, 2024), and Llama (Touvron *et al.*, 2023), which improve upon ChatGPT in various ways. This trend for innovation has only accelerated. In March 2024, Anthropic launched Claude 3, surpassing the previous leading models from OpenAI and Google on various benchmarks. Without passing one month, In early April 2024, OpenAI reclaimed the top spot again by tweaking its existing model, and later on April 2024, Meta released Llama 3, which early results suggest is the most capable open model to date. OpenAI is expected to make another revolutionary impact at the end of 2024 with the anticipated release of GPT-5. This is expected to be far beyond those of any currently available LLMs in terms of capabilities as well as model sizes.

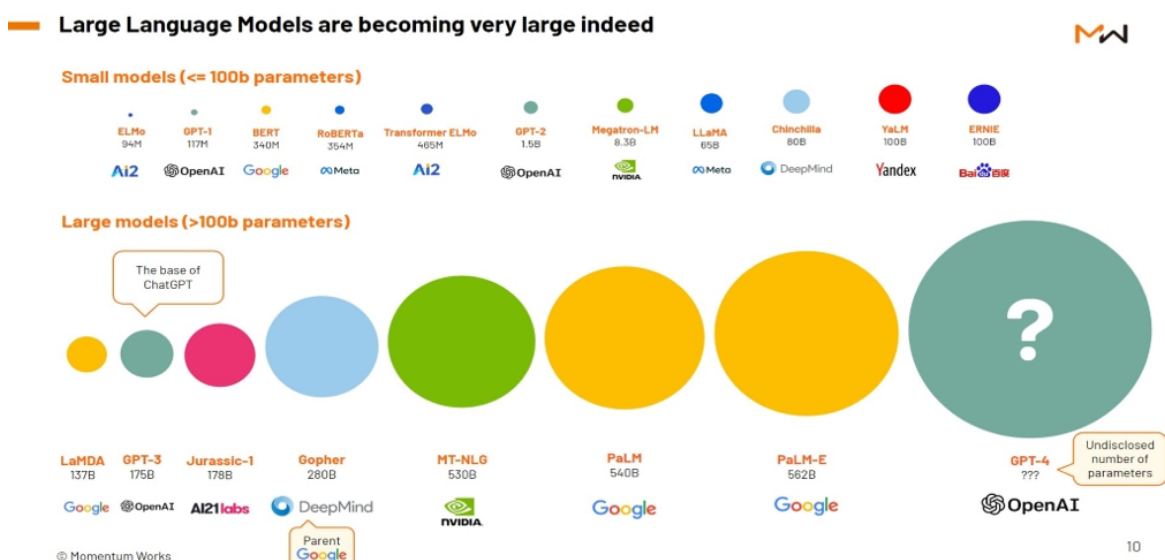


Fig. 2.7 Expanding Trend in LLM Size and Complexity (Asia, 2024)

Along with the model sizes and number of parameters, the complexities of these models are also growing rapidly. Figure 2.7 shows a good overview of the evolution of LLMs within the last few years (Asia, 2024). More current models, such as Google's PaLM and GPT-4, are on the right of the second row, whereas older models like ELMo, GPT-1, and BERT are on the left of the first row. When looking at the variation of the circle sizes, it's evident how large these models are becoming even in a short period of time. The primary reason for the expanding nature of LLMs lies in the transformer neural architecture's design. In a transformer neural model, the primary task is to predict the most probable word sequence. In other words, it tries to predict the next most suitable word given a sequence of words based on the probability patterns it has learned during the training. Therefore, the quantity of training data needed to learn various word arrangements and the model's capacity to remember the finer nuances of linguistic occurrences determine how accurate these predictions are. To achieve this, the neural model must be expanded to accommodate larger training sets and hold more information in its learnable parameters. Consequently, increasing the model's size becomes a necessary approach to enhance its performance.

For example, when considering Open AI's models such as GPT-3 and GPT-4, they are trained with the internet scale of the dataset with many language combinations. These datasets include book corpora, academic databases, vast collections of web pages, and more. GPT-3 was trained using 570GB of text data, requiring 1,024 GPUs over 34 days and costing approximately \$4.6 million (Narayanan *et al.*, 2021). When it comes to GPT-4, even though the technical details of the model have not been fully disclosed, it is estimated to have significantly more parameters and training data than GPT-3, indicating an even greater demand for computational resources and data.

This tendency ultimately makes the development of LLMs a highly resource-intensive and extremely complex task. Thus, the development of LLMs is only limited to large organizations with access to internet-scale unstructured data and sophisticated computational power. So that this further widens the gap between technology leaders and others in the field. Additionally, this behaviour also can result in various biases in LLM development where these models can often reflect the ideologies of their own creators. A study by Buyl *et al.* (2024) analyses this behaviour in their recent research.

In addition to that, when considering the substantial computational and energy consumption required during the model training, it also poses significant sustainability challenges as well. Popular models such as Claude and GPT, especially given their extensive usage statistics, demand sophisticated infrastructure, often necessitating large server farms. Thus, even during the inferencing, in order to upkeep efficiency while accommodating millions of requests, it requires substantial energy consumption. A study by Samsi *et al.* (2023) also

analyses and highlights the substantial energy costs related to LLM inferencing, mentioning the requirement for more efficient designs and hardware strategies to manage resource utilization. Beyond energy and carbon footprints, it was also estimated that 5.4 million litres of water were used to cool Microsoft's high-performance data servers when training GPT-3 and that 500 mL of water was consumed to respond to 10–50 queries in ChatGPT services(Jiang *et al.*, 2024). As these models continue to grow in size and complexity, the sustainability of such operations becomes increasingly questionable, raising concerns about the environmental impact and long-term viability of maintaining and expanding these extensive computational infrastructures(Miao *et al.*, 2023).

2.6.2 Lack of Explainability in LLMs

LLMs uses transformer architecture which is a complex neural network designed to interpret and generate human languages using vast amounts of data. Thus, along with the growing complexity, another limitation is the opaqueness or black-box nature of these models, which poses substantial difficulties in explaining the internal workings and decision making process of these models, making it challenging to ensure their reliability and fairness.

Explainability refers to the ability to explain or present the behaviour of models in human-understandable terms(Doshi-Velez and Kim, 2017; Du *et al.*, 2019). Explainability in LLMs is a crucial factor due to two main reasons. Firstly, for general end users, explainability fosters trust by clarifying the reasoning mechanisms behind model predictions in an easily comprehensible manner without necessitating technical expertise. This feature let the users to understand the model capabilities, limitations, and potential flaws. Secondly, for researchers and developers, explaining model behaviours offers useful insights that help identify unintended biases, risks, and areas needing performance improvements. In other words, explainability is a debugging tool that facilitates the rapid advancement of model performance in downstream tasks(Strobelt *et al.*, 2018; Bastings *et al.*, 2022). It aids in tracking model capabilities over time, comparing different models, and developing reliable, ethical, and safe models for real-world deployment(Zhao *et al.*, 2023).

LLMs make probabilistic decisions based on extensive training with unstructured natural language data. Essentially, they predict word sequences with the highest probability of occurrence matching the preceding set of words or the context. Although LLMs appear to understand and respond to queries, their process merely mimics human language comprehension through an information-intensive mechanism(Yao *et al.*, 2024). In addition to that, due to their stochastic nature, the predictions can vary even for the same input prompt. While adjusting model parameters like temperature and top_p can provide some control, the internal steps remain invisible. Consequently, customizing or amending the output typically

requires altering the input or prompt and relying on trial and error until the desired results are achieved. This inherent opacity ultimately makes LLMs less controllable and customizable for specific use cases.

Moreover, the output of an LLM is highly influenced by the data on which it is trained. Again, when considering its volume, it can be a challenge to ensure safety in every aspect. This dependency can result in unexpected responses, including hallucinations, biases, and other unsafe outputs. When LLMs generating responses they can generate content which reflect features such as prejudice, discrimination, or even extremism(Yao *et al.*, 2024). Even-though the biases are also common in other opaque neural architectures, LLMs have received high attention due to ethical and societal concerns with their vast popularity and adaptation to the existing systems(Dong *et al.*, 2023). The societal impact of LLMs has prompted discussions about the ethical responsibilities of organizations and researchers developing and deploying these models, leading to increased research on bias and fairness(Shaikh *et al.*, 2023).

The above-mentioned issues can be critical considerations, especially when we leverage the LLMs to build solutions and deploy those in sensitive domains such as healthcare, education, and law. The reason is, in such areas, the potential for generating harmful impacts on society underscores the urgent need for robust mechanisms to ensure the reliability, fairness, and safety of LLMs. Because of these limitations, it is crucial to reconsider the utilization of LLMs with proper guardrails in such sensitive domains to maintain trust and reliability in their outputs. This is especially important in applications where accuracy and fairness are significant, as the consequences of errors or biases can be severe and far-reaching.

2.6.3 Lack of Compositional Behaviours

While LLMs show excellent performance in various tasks, they tend to behave very poorly in certain cases where it does not capture some fundamental guidelines despite their extensive training. Particularly, in scenarios such as arithmetic, common sense and multi-step reasoning, and tasks requiring world knowledge, LLMs often fail to demonstrate true language understanding (Bender and Koller, 2020). However, some of these limitations can still be mitigated by expanding LLMs to include more training data that simulates such scenarios. This approach is not ideal, as we explained with the complexity and explainability limitations of the expansive nature.

One of the primary reasons for this behaviour is lack of compositionality. Compositionality refers to the principle that the meaning of complex expressions is determined by the meanings of their constituent parts and the rules used to combine them(von Heusinger *et al.*, 2019). In other words, it can be understood by understanding how individual words

combine to form phrases and sentences, with the meaning of the whole being derived from the meanings of the parts. However, it reaches more profound levels, such as compositional generalization, where it abstracts the systematic behaviours and applies them to novel concepts without being specifically trained on them. This principle is crucial for tasks that involve multi-step reasoning, logical deductions, and handling novel combinations of known elements.

More recent LLMs appear to exhibit more natural and human-like language understanding and generation, which mimic compositional behaviours. For instance, these models can produce coherent and contextually appropriate conversation responses, solve complex question-answering tasks, and even create creative texts that seem to reflect a human-like understanding of language and the world (OpenAI, 2023; Brown *et al.*, 2020). However, this behaviour can be understood as utilization of the models' ability to learn and replicate compositional patterns present in the vast amounts of text data used in the training phase (Russin *et al.*, 2024). While LLMs can occasionally exhibit compositional behaviour, this does not guarantee consistent compositional understanding. One approach to improve this is the application of innovative prompting techniques, such as chain-of-thought prompting and self-ask strategies, which improve their ability to handle compositional tasks.

In humans, compositional language understanding involves rule-driven compositional mechanism where various cognitive components such as vocabulary, language structures, memory, inference and reasoning, and background and world knowledge connect together to build a coherent understanding of the language. Humans naturally integrate these elements when comprehending and generating language, allowing for more flexible and robust understanding (Marcus, 2020). While LLMs can simulate some aspects of this process, their approach based on probabilistic deduction remains fundamentally different from the deeply integrated and systematic nature of human cognition (Press *et al.*, 2023). In contrast, traditional NLP methods align more closely with compositional behaviour by explicitly consolidating components within rule-driven systems, ensuring predictable and reliable performance. However, these systems often lack the compositional generalizability required to handle novel or unseen contexts, a strength that LLMs inherently possess. Therefore, integrating traditional NLP methods with LLMs offers a promising pathway to enhance and stabilize the compositional behaviours of LLMs along with prompting strategies.

Figure 2.8 showcases two examples where LLMs exhibit fundamental errors because they lack compositionality. These experiments were conducted using the early 2023 GPT3.5-turbo model (February 13 release). In the first example shown in figure 2.8, it provides a simple arithmetic expression as the query to Chatgpt: $241 - (-241) + 1$. Ideally, according to the correct mathematical operator precedence, the correct evaluation should be $[241 - (-241)]$

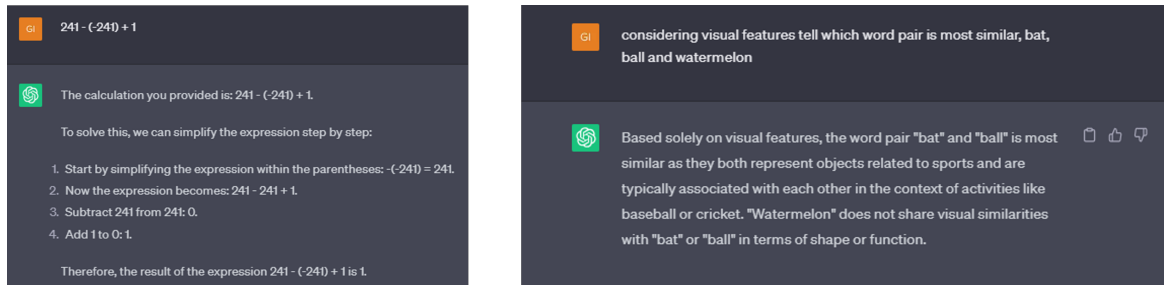


Fig. 2.8 LLMs Showing Fundamental Errors with Lack of Compositionality

$+ 1] = [241 + 241 + 1] = [482 + 1] = [483]$. However, the model produces an incorrect result by considering the arithmetic expression as a set of words and predicting the most probable next set of words as the answer. Critics may argue that LLMs are designed for language generation rather than arithmetic tasks, but this example highlights their lack of compositional and systematic behaviour, where a simple rule-based calculator logic could solve the problem without extensive training.

This behaviour of LLMs was further discussed in the paper (Brown *et al.*, 2020). This study analyses the arithmetic skills of LLMs with the varying number of digits. It shows that LLMs can perform arithmetic with two and three digits with high accuracy but struggle more with four-digit and five-digit calculations. This limitation arises because smaller numbers appear more frequently in the training data compared to larger numbers. Consequently, the models develop a stronger capability in handling 2-3 digit operations, which are more common in real-world texts as well as in its pre-training dataset. The paper also explains that by expanding model sizes, LLMs can achieve higher accuracies in these more complex arithmetic problems (Brown *et al.*, 2020). However, rather than just expanding models, it should leverage compositional rules and multi-step processes to solve these problems more effectively.

In the second example, three words/concepts are provided to the LLM: bat, ball, and watermelon. The model is asked to consider visual features to find the most similar word pair. It is expected to select watermelon and ball due to their obviously rounded shapes, the same as a human does. However, the model chooses bat and ball, as these words most commonly occur in similar contexts. This phenomenon exemplifies how LLMs consider word relatedness instead of similarity holistically, as humans do. For instance, cars and petrol are more related as they occur in similar contexts. However, they demonstrate lower similarity since similarity involves more grounded aspects, such as shared physical or functional properties, beyond mere co-occurrence (Hill *et al.*, 2015a). This also exemplifies that LLMs are not grounded during the training process, unlike humans when they learn the language. LLMs rely solely on language modality for training. In contrast, humans get

support from the other modalities in the corresponding real-world environment, integrating sensory experiences and contextual information, which provides a more comprehensive understanding of language and concepts. This lack of grounding is another major limitation of LLMs, leading to fundamental errors (Merkx *et al.*, 2022).

The reversal curse is another occasion that highlights the lack of compositionality of LLMs. This issue occurs when a model trained on a sentence of the form "A is B" does not automatically generalize to the reverse direction "B is A." (Berglund *et al.*, 2024). As shown in the 2.9, the model is first asked, "Who is Tom Cruise's mother?" and correctly responds with "Mary Lee Pfeiffer." However, when the reverse is prompted with "Who is Mary Lee Pfeiffer's son?" the model fails to retrieve "Tom Cruise" as the answer. This example highlights a fundamental limitation of LLMs, demonstrating their struggle with bidirectional compositional reasoning. In contrast, humans can quickly identify these connections as their understanding is built upon the compositional elements and relationships.

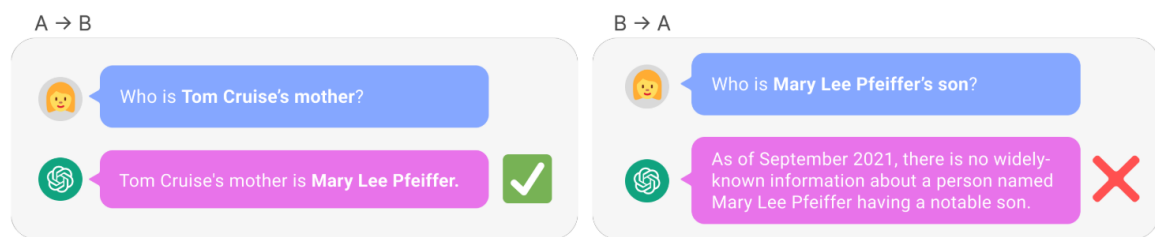


Fig. 2.9 The Reversal Curse of LLMs (Berglund *et al.*, 2024)

Despite extensive training and significant computational resources, LLMs still exhibit critical errors on several occasions, as discussed above. This ongoing challenge raises important questions about the current approaches in developing and refining these models, suggesting a need to rethink strategies for overcoming such limitations.

2.7 Chapter Summary

This chapter starts by providing a comprehensive overview of the significant role of natural language data in modern computer systems and discussing how various computational models attempt to process language intelligently. It further examines the evolution of traditional NLP systems, transitioning from explicit, rule-based methods to advanced neural architectures capable of handling complex language tasks.

A significant portion of the chapter is dedicated to discussing the emergence of large language models (LLMs), detailing their architectures, behaviours, and the numerous advantages they offer across various applications. The chapter also provides a comprehensive

analysis of the limitations inherent in these systems, including their increasing complexity and unsustainable growth trends, lack of explainability, and lack of compositional language understanding, which often lead to fundamental errors.

The chapter concludes by widening the viewpoints on the need for more effective approaches to address these limitations also maintaining the benefits of the models. It highlights the importance of bridging the gaps between computational systems and human language comprehension to achieve better alignment and sustainability in NLC advancements.

Chapter 3

Towards Neuro Symbolic Approach

As we discussed in detail in Chapter 02, current trends with natural language comprehension approaches are skewed towards data-intensive complex neural architectures to create more advanced systems, which presents various challenges in practical development. In addition to that, when we compare existing computational approaches to how humans comprehend language, it's evident that there's a significant gap. As humans, we exhibit a more efficient and effective language comprehension process, utilizing considerably less information than current state-of-the-art LLMs. Furthermore, humans keep a balance between heuristics and memory, which is often absent in current computational models.

To bridge this gap between human comprehension and machine comprehension and bring human aspects into computational approaches, one of the most promising ways is leveraging neuro-symbolic architectures. These methods focus on integrating both symbolic and connectionist paradigms of NLP rather than relying heavily on connectionist approaches such as LLMs to build solutions. This hybrid mechanism helps to enhance the strengths of both paradigms while mitigating the limitations. In this chapter, we explore how neuro-symbolic architectures can be leveraged to construct natural language comprehension systems with properties such as learnability, explainability and controllability which ensures sustainable and trustworthy deployments.

3.1 Human Language Comprehension

According to neuroscience and psychology, comprehension is a cognitive process involving several brain activities. Human language comprehension skills start to develop in early childhood and are enhanced with the help of education and experiences (Gernsbacher and Kaschak, 2003). In addition to that, neuroscientific studies have proven that language comprehension involves various regions of the brain. For instance, neuroimaging studies

have recognized the left inferior frontal gyrus (Broca's area) as being involved in language production and processing, while the left temporoparietal region (Wernicke's area) is crucial for comprehension of the language (Gernsbacher and Kaschak, 2003). These different regions work together to enable effective understanding and generation of language. According to Kendeou *et al.* (2009), the development of reading comprehension skills can be predicted by early oral language and decoding skills. Their study demonstrates that these foundational skills are crucial for later reading success, suggesting that comprehension is built on a combination of linguistic and cognitive abilities that form at a young age.

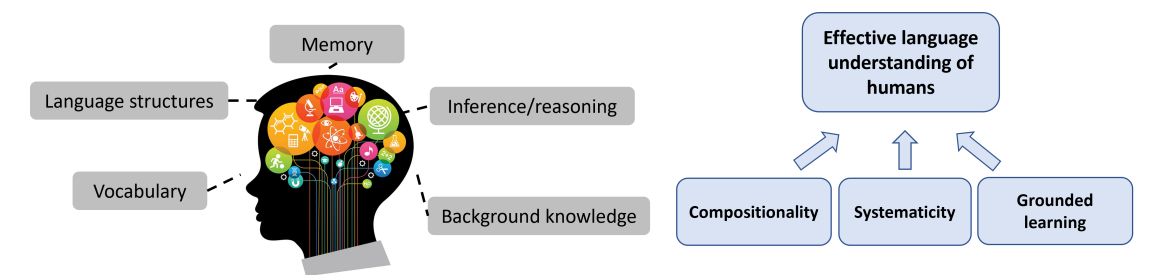


Fig. 3.1 Elements of Human Language Comprehension

In addition to that, human language comprehension is also centred on various components, such as vocabulary, language structures, memory, inference and reasoning, and background knowledge. Vocabulary knowledge is an essential component for decoding and interpreting sentences, allowing readers to infer meanings of unfamiliar words from the context (Beck *et al.*, 2013). Furthermore, knowledge of language structures, including grammar and syntax, helps in comprehending word relationships in sentences. Memory, particularly working memory, also plays an important role in language comprehension. It enables holding and manipulating required information while reading, facilitating coherence across sentences and paragraphs (Daneman and Merikle, 1996). Inference and reasoning skills help to fill in gaps beyond explicit information, which is essential for constructing coherent mental representations. Background knowledge helps readers make sense of new information, draw inferences, and understand references (Gernsbacher, 1999). Moreover, humans utilize various techniques to comprehend language efficiently and effectively, which will be discussed in detail in the next sections.

3.1.1 Compositionality

Compositionality is a central concept in human language comprehension. Compositionality refers to the ability to derive the meaning of a complex expression by consolidating the meanings of its components (von Heusinger *et al.*, 2019). This principle of language

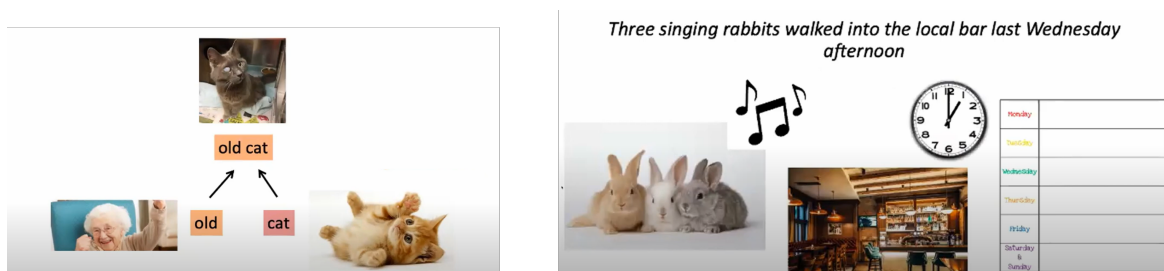


Fig. 3.2 Examples of Compositionality

comprehension enables humans to understand new phrases or sentences even if they have not heard or learned them before. Figure 3.2 showcases two examples that elaborate on the compositionality of human language understanding. For example, when interpreting the phrase "old cat," what humans are essentially doing is simply combining the meanings of the concepts "old" and "cat" to understand the concept of old cat. This process illustrates how compositionality enables us to make sense of various word arrangements based on our knowledge of individual words and their meanings (Partee, 1984).

This ability extends to more complex and imaginative expressions as well. Consider the second example of figure 3.2, "Three singing rabbits walked into the local bar last Wednesday afternoon." Although it is unlikely to encounter such an event in the real world or have heard or learned this exact sentence before, compositionality allows us to comprehend it without any problem. By knowing the meanings of the individual words and concepts such as "three," "singing," "rabbits," "walked," "bar," and "Wednesday", we can comprehend the entire sentence without any difficulty. This capability underscores the efficiency and effectiveness of compositionality in language comprehension, enabling us to interpret and make sense of intricate and previously unseen linguistic constructs (Dziri *et al.*, 2023).

The concept of compositionality can be further extended by going beyond mere word meanings as compositional elements. As discussed before, in human cognition, language understanding requires various integrals to be more holistic. While words and phrases play the main role as the building blocks of the meaning, other dimensions such as contextual knowledge, emotionality and various inputs from other modalities can also be significant compositional elements based on the situation. One occasion that highlights this behaviour is figurative language usage, where the collective meaning of the words is different from that of the phrase. For instance, the phrase "*It's cold in here*" could imply a mere observation of temperature or serve as a polite request to close a window, depending on the situation (LITKOWSKI and RENKEN, 2013). Moreover, a sarcastic remark like "*Oh, great job!*" might mean the opposite of what the words suggest, highlighting the role of tone and emotional context.

3.1.2 Systematicity

Systematicity is another important technique related to compositionality. Usually, these two techniques apply together during human language comprehension. Systematicity explains the structured and rule-governed nature of how we understand and produce language. It highlights that our language comprehension is based on applying a finite set of grammatical rules, allowing for flexible and adaptive use of language across numerous contexts (Fodor and Pylyshyn, 1988). When comparing systematicity and compositionality, compositionality focuses on various individual components that combine to generate the meaning of a complex expression. In contrast, systematicity focuses on how these different components combine during the comprehension process.



Fig. 3.3 Systematicity in Human Language Comprehension

Figure 3.3 illustrates two examples of how systematicity contributes to the process of language comprehension. In this scenario, there are two sentences: 1) "The dog bit the man" and 2) "The man bit the dog". Both of these sentences are composed of the same set of linguistic compositional elements. However, the arrangement of these elements is different in the two cases. Thus, even though these sentences are composed of the same elements, their meanings are completely different to each other. This highlights the significance of the systematic nature of language comprehension. In the first sentence, "The dog bit the man," the subject is "the dog", indicating that the dog is the one biting. In the second sentence, "The man bit the dog," the subject is "the man", indicating that the man is the one biting, though it is a less likely scenario in the real world. Thus, in addition to compositionality, various components such as word order, grammar, syntax, semantics, and morphological rules are equally important, all of which are covered by systematicity (Calvillo *et al.*, 2021).

Furthermore, systematicity also explains the cognitive ability to generalize and apply learned linguistic rules to new situations. This enables humans to use more flexible and adaptive language. For instance, if someone understands the expressions "brown triangle" and "black square," this implies that they can also understand the expressions "brown square" and "black triangle" (Spencer and Blutner, 2007). This is also supported by the inferencing capability of humans along with the systematicity. Apart from the explicit linguistic rules of combining words and phrases, systematicity also extends to the systematic consolidation of context, emotionality and other modality inputs, as discussed under compositionality.

Altogether, systematicity and composition help humans comprehend infinitely many language expressions by utilizing limited lexical and syntactic guidelines. This largely reduces the cognitive load in human language comprehension.

3.1.3 Grounded Language Learning

Grounded language learning is another important technique in human language comprehension. It explains that humans always ground language symbols in real-world reference when they are learning a language. In other words, humans learn the language by always interacting and experiencing the real world. Unlike the LLMs, tries to learn the concept only by leveraging the text modality and frequent surrounding terms, in grounded language learning, humans build the context from other modalities and real-world situations to have more overall grounded comprehension. This connection between language and the physical world enables a richer and more robust understanding of language, making it easier for individuals to grasp abstract concepts and meanings(Suglia *et al.*, 2024).

For instance, let's assume that we are trying to teach a young child the concept of an "elephant" for the first time in his life. One approach would be to describe the elephant using language alone, with a verbal explanation or a paragraph of text. However, relying solely on verbal descriptions may make it difficult for the child to fully understand the concept. However, if we combine the verbal explanation with a picture or a drawing of an elephant, it will largely enhance understanding, making the concept more concrete. By showing the child an image of an elephant, we add a visual component that solidifies the concept further. The child can now link the word "elephant" not only to the concepts in the verbal description but also to an image, making the idea of an elephant more vivid and accessible. It would be even better if we took the child to a zoo or safari and let him observe a live elephant in a real-world environment. In this immersive, multi-sensory setting, the child can see the elephant's size and shape, hear its sounds, and even feel the vibrations of its movements. This direct interaction provides a more holistic understanding, allowing the child to grasp the "concept of an elephant" in a way that words alone could never achieve.

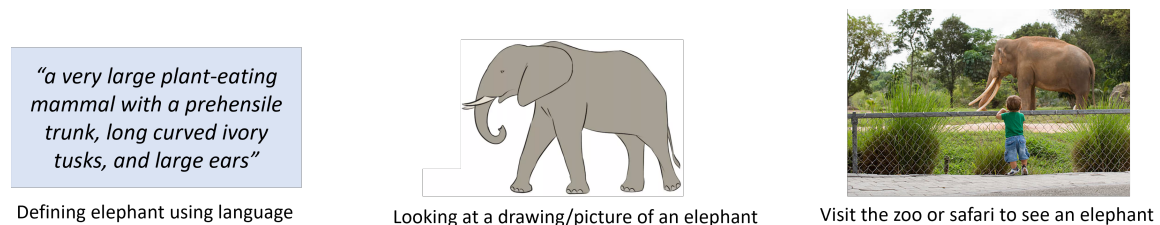


Fig. 3.4 Grounded Language Learning

3.1.4 Construction-Integration (CI) Theory

Construction Integration theory is another foundational theories which explains the multi-level human language comprehension process. This theory describes how individuals construct and integrate meaning from language by building personalized mental representations, often called mental models. Mental model construction is a main part of language comprehension according to this theory, as they combine information directly extracted from the text with prior knowledge and experiences, creating a complete and coherent understanding of the material(Kintsch, 1998). The CI Theory explains that comprehension is an active process where meaning is not simply received but constructed by the reader, drawing on both the immediate context of the text and the reader's background knowledge to fill in gaps, make inferences, and resolve ambiguities.

This theory describes the mental model construction occurring in three main stages: the surface level, the textbase level, and the mental model level. At the surface level, readers simply process the literal words and syntax of the language, performing a shallow, initial analysis. This is similar to the explicit processing of the languages in traditional NLP approaches. Moving to the text-base level, readers begin to extract concepts and analyze relationships between them, creating a foundational understanding of the text. Here, they enrich the extracted concepts with their prior knowledge and experiences. Finally, at the mental model level, readers achieve a deeper and complete comprehension by integrating the extracted and enriched concepts at surface and text-base levels. This stage involves further analysis of the relationships among concepts and incorporates a compositional understanding, which allows for a richer, more personalized model of the text's meaning(Graesser *et al.*, 1994). Thus, CI Theory illustrates that comprehension is a multi-step process where language is actively mapped to existing knowledge, refining and solidifying understanding in a way that simple textual processing alone cannot achieve.

When considering the LLMs, they operate probabilistically to predict the next most suitable term for a given sequence rather than constructing the meaning in a multi-step approach, as CI theory explains. The probabilistic nature of LLMs and other deep learning approaches avoids the controllability of these methods in identifying, enriching, and combining language concepts with building mental model representations, as in human language comprehension. However, these methods approximate these behaviours and mimic human understanding by focusing on language usage patterns with large quantities of datasets, which is a totally different approach. Comparing human language comprehension techniques with those of LLMs provides clear insights for resolving the current limitations in language processing systems. Although LLMs are not controllable or transparent enough to incorporate these

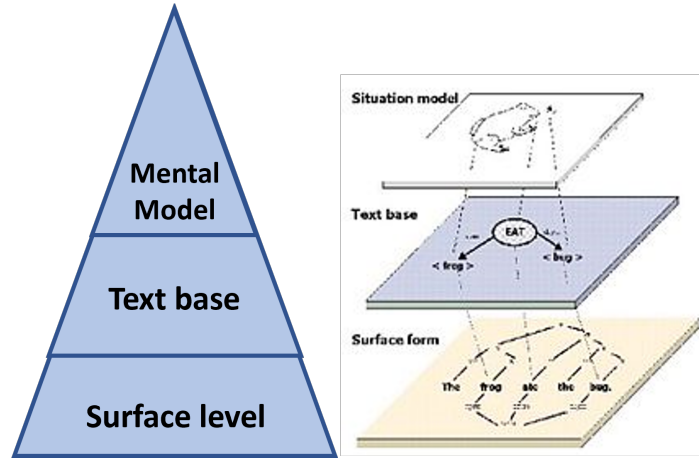


Fig. 3.5 Construction-Integration (CI) Theory

techniques fully, merging them with controllable symbolic NLP pipelines helps us to mitigate the limitations to a greater extent.

3.2 Leveraging Human Aspects in Computational Models

As discussed in the previous section, human language comprehension is empowered with the integration of strategies such as compositionality, systematicity, grounded language learning, and construction-integration theory. One common phenomenon in these mechanisms is that they enable individuals to process the language with a lesser cognitive load, effectively balancing memorization and inference. In contrast, LLMs primarily rely on probabilistic predictions based on extensive text data, highlighting a fundamental difference in language processing between humans and LLMs. This can also be interpreted as a human trying to utilize his memory to capture all language occurrences to mimic the true understanding of the language. This can be further referenced by John Searle's Chinese room argument, which explains that systems can generate contextually appropriate responses based on statistical patterns in data, yet they lack genuine understanding and comprehension strategies(Cole, 2024).

Basically, in LLMs, similar contexts within the training datasets are the foundations to behave in specific manners aligned with human comprehension. Thus, LLMs require vast and diverse datasets to capture the numerous contexts present in human language, leading to increased model sizes and computational demands. For instance, while compositionality and systematicity help humans utilize a finite set of rules and principles to understand and generate a wide range of language expressions, in LLMs, similar behaviours are archived by exposing these models to diverse language occurrences. Grounded language learning further

exemplifies this disparity. Humans naturally integrate multiple modalities—such as visual and auditory experiences—into their language comprehension, leading to a more holistic understanding. On the other hand, LLMs attempt to approximate grounded comprehension with language usage patterns that indirectly convey information from other modalities. Thus, ultimately, LLMs follow extreme connectionist trends with these behaviours. In addition to that, since these models operate as black boxes, it also becomes a challenging task to intervene or customize their processing strategies to incorporate such human-like mechanisms.

To address these limitations, it is essential to move beyond purely data-intensive methods and incorporate neuro-symbolic approaches that blend rule-based systems with data-driven learning. Symbolic AI offers systematic and interpretable frameworks that can complement the flexible learning capabilities of connectionist models. By integrating these approaches, we can develop more sustainable and effective language comprehension systems that mirror the balance found in human cognition. It holds promise for advancing natural language understanding, leading from expansive LLMs which cover all aspects of comprehension to multi-component pipelines that are capable of effective comprehension utilising hierarchical controllable strategies such as construction integration theory. This approach not only enhances efficiency but also makes advanced language processing capabilities more accessible and sustainable.

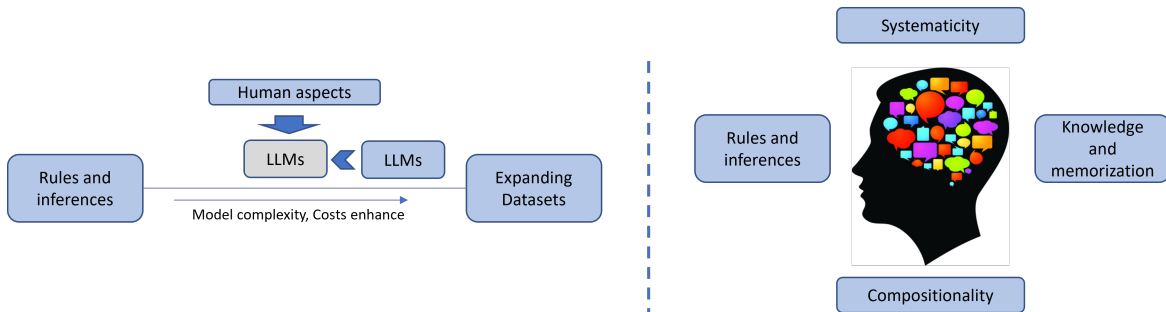


Fig. 3.6 Leveraging Human Aspects in Computational Models

3.3 Neuro-symbolic Architectures for Natural Language Comprehension

Neurosymbolic architectures combine symbolic and non-symbolic or connectionist approaches to build hybrid pipelines (Yu *et al.*, 2023). Here, the main intention is to use both paradigms rather than building a solution utilizing only one. In the NLP roadmap, the early stages of language processing were dominated by symbolic methods, which excel at

logical reasoning, rule-based decision-making, and handling structured knowledge. These techniques relied highly on specialists with domain knowledge to craft the rule sets carefully. However, these methods struggle with learning from raw data and reaching implicit areas of comprehension. These symbolic components capture techniques such as systematicity and compositionality inherent in human language comprehension.

In contrast, connectionist approaches, such as LLMs, are powerful in learning from large datasets, recognizing patterns, and making predictions. Especially with the introduction of neural networks, connectionist methods have gained vast popularity compared to the traditional symbolic methods. However, these approaches often lack the explainability, interpretability, and compositional behaviours that traditional methods have. These approaches are information-intensive and excel in retrieval and memorization, aspects of human language comprehension. Thus, rather than using these aspects isolatedly, modularizing and consolidating them in pipelines provides us with an occasion to mitigate the negatives and expand on advantages. Furthermore, merging symbolic and connectionist approaches into neuro-symbolic architectures aims to build systems that closely mimic human language comprehension, offering features such as explainability, robustness, controllability, and compositional behaviour, thus avoiding the need to expand neural architectures continually.

3.3.1 Symbolic AI Methods for Natural Language Comprehension

These approaches in AI are also referred to as Good Old-Fashioned AI (GOFAI). They use high-level, human-readable symbols to represent knowledge and processes. To construct these systems, it's required to have experts with domain knowledge of the problem that is trying to be solved. These systems rely on predefined and usually hand-crafted rules and logical operations to manipulate symbols and solve problems. One of the important features of these approaches is they are explainable and transparent. Because these systems operate using well-defined explicit rules and logic, it is straightforward to trace back the reasoning process and steps to understand how final outputs were generated.

Figure 3.7 illustrates an example symbolic approach for sentiment analysis task by (Hardeniya and Borikar, 2016). In this system, the main intention is to collect and iterate through Amazon product reviews and classify them into three different sentiment classes: positive, negative, or neutral. The approach depicted in the figure consists of several components corresponding to specific steps in the process that form a well-defined pipeline. The pipeline gets the reviews as text descriptions, and each review text in the database goes through three main steps: text preprocessing, sentiment scoring, and sentiment classification. In the first step of preprocessing, the review is normalized to prepare for the sentiment scoring step by using various natural language preprocessing methods such as stop word removal, stemming,

POS tagging and n-gram detection. Afterwards, the preprocessed texts will be forwarded to the sentiment scoring unit. During the sentiment scoring stage, these normalized texts are compared against a lexicon, a human-annotated list of words tagged with corresponding sentiment labels. During this step, based on the presence and frequency of sentiment terms in the reviews, a suitable sentiment score is assigned to each review in the dataset. In the final classification step, these scores will be compared against a set of predefined threshold bands to categorise the sentences into one of the three sentiment classes based on predefined sentiment threshold bands.

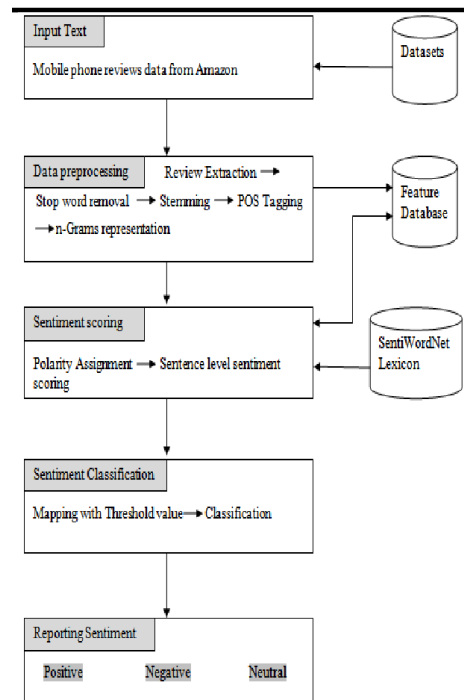


Fig. 3.7 Instance of Symbolic NLP on Sentiment Classification (Hardeniya and Borikar, 2016)

Since the end-to-end sentiment classification procedure is transparent, we can explain how a given review is categorized as positive, negative, or neutral using its sentiment score and the underlying words that match the annotated lexicon. This level of explainability enhances the system's trustworthiness because it operates according to the guidelines provided by human experts rather than relying on stochastic behaviour based on probabilities. In addition to that, the labels can only be one of these defined classes where it's unlikely to end up with unintended behaviours. This makes the system more suitable for deployment in sensitive areas such as medical and legal fields. For instance, if we need to convert this into a bipolar sentiment classifier that predicts only positive or negative sentiments, we can simply overwrite the classification rule. The modularized design of most symbolic approaches, with

well-defined communication routines between components, further enhances the system's debugging capabilities and customizability. This design improves the overall capability of the solution.

When considering the negatives, the main disadvantage of symbolic approaches is their lack of generalizability beyond their configurations. These systems rely heavily on predefined rule sets and knowledge sources. In the above-mentioned scenario of sentiment analysis, for instance, the system matches words in a review with those in a lexicon to detect sentiments. Although this lexicon covers a range of aspects based on expert human knowledge, there may still be instances where words in a review are absent in the sentiment lexicon, leading to incomplete or inaccurate sentiment detection. Additionally, in particular to this scenario, symbolic approaches struggle with implicit sentiments embedded in the text, such as sarcasm and idiomatic expressions, where the literal words do not reflect the true sentiment. The rapid evolution of language, particularly on social media, introduces new words and phrases that may not be captured by static lexicons. So, such use cases cannot be explicitly defined beforehand. In such dynamic contexts, learnable systems that can capture linguistic details and context are more beneficial than heuristic approaches. Extending or complexifying rule sets and knowledge sources to match different use cases is impractical. The manual effort required for these systems makes symbolic approaches well-suited for specific use cases with well-defined characteristics but not for capturing a wide range of scenarios.

3.3.2 Connectionist AI Methods for Natural Language Comprehension

Connectionist AI, as the name suggests, represents a collection of approaches based on network structures. Artificial Neural Network (ANN) is the foundational technique to most connectionist AI methods, which feature multiple layers of nodes or neurons. These neurons process input signals, combine them using weight coefficients, and then apply an activation function before passing the result to the next layer. These models randomly initiate their weight coefficients first and then adjust these weights and activation levels along with data points in the training process. This design is mainly inspired by the interconnected nature of neurons in the human brain. Therefore, connectionism aims to imitate how the human brain processes and stores information, where information is the core component of these approaches. Additionally, connectionist systems are known for their distributed and parallel processing capabilities. Each artificial neuron, or "node," processes information and transmits it to other neurons through connections, enabling the encoding and processing of complex patterns and data. So, the main strength of these approaches is the ability to process large quantities of data to better learn from those data points and come up with the most optimal weight distributions to generate more accurate predictions. This method has

broad applications in AI, allowing the creation of systems that can learn from experience, recognise patterns, and generalize knowledge to new situations. LLMs are the most recent and state-of-the-art connectionist approach for natural language processing systems, which we discussed in detail in Chapter 2.

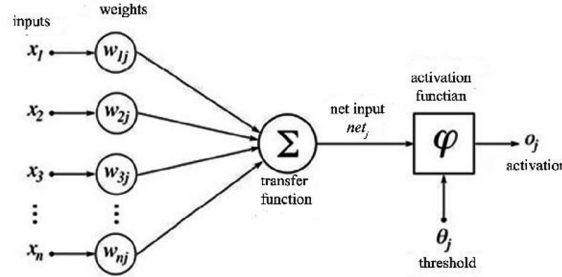


Fig. 3.8 Connectionist Model Architecture - ANN (Asaad and Ali, 2019)

In contrast to symbolic methods, the main difference of connectionist approaches is the capability to automatically derive the mappings between inputs and outputs rather than explicitly mentioning them to the systems. With this, a lot of manual efforts can be mitigated, and it can further capture the patterns that might overlooked even by human experts. These algorithms use statistical regression models to adjust the weight coefficients of their intermediate variables, aiming to minimize the cumulative error from all training data points through techniques such as gradient descent. This process enables connectionist models to focus on a variety of features and pay attention to even tiny details. This capability allows them to cover a large range of NLP problems, and we can utilize the datasets to generate the approaches themselves.

When talking about the downside of connectionist approaches, we come across the same limitations we discussed during Chapter 02 with LLMs. These approaches utilize probabilities based on training data to make decisions, which makes it challenging to fully interpret how a particular result is deducted. For instance, we can consider the same task of sentiment classification we went through with symbolic approaches. With the connectionist models, we can design sentiment classification using a neural network such as a random forest or XGBoost classifier. All required is a pre-annotated text-to-sentiment sample dataset to train the model to predict the final labels. Here, the mapping from input text expressions to corresponding sentiment classes will be automatically defined during the training, which is not transparent to us. All these mappings will rely heavily on the training data. It may internally consider statistical measures such as which words usually occur with a given annotation to support its predictions. By the other hand, if we use an LLM such as ChatGPT with few-shot or zero-shot setting, the only controllable part is the prompt instructions where

we ask it to classify the given review into one of the three groups: positive, negative or neutral.

In both of these cases, what we get is simply an output label of the sentiment. The process of how this is generated is not fully explainable. As these approaches are based on a probabilistic nature, sometimes they might not behave according to our expectations as well. In such cases, especially with supervised learning, we have to trial and test different parameters in the approach to get the expected behaviours as they are not fully interpretable. Similarly, with LLMs, we have to alter the prompts in strategic ways so that they provide optimal outcomes. This opaque nature of connectionist models reduces the confidence in deploying these in sensitive applications, given their stochastic behaviours.

3.3.3 Attempts in Building Neuro-Symbolic Architectures in the Literature

Building on the foundations of symbolic and connectionist AI methods, neuro-symbolic AI aims to integrate the explicit, rule-based reasoning capabilities of symbolic systems with the learning and pattern recognition strengths of neural networks. This hybrid approach seeks to create natural language processing systems that are both robust and interpretable, overcoming limitations inherent to either paradigm used in isolation. In the context of natural language comprehension, neuro-symbolic architectures attempt to unify the structured knowledge representation and logical inference of symbolic AI with the data-driven, flexible learning of connectionist models like large language models (LLMs). The goal is to build systems that can leverage both explicit domain knowledge and implicit contextual understanding, while ensuring transparency, controllability, and compositionality.

Recent developments in neuro-symbolic AI showcase diverse hybrid architectures across language, visual reasoning, and knowledge representation. The Neuro-Symbolic Concept Learner (NS-CL), for instance, excels in visual question answering by combining convolutional neural networks for object detection with a symbolic program executor for reasoning over visual scenes (Lu *et al.*, 2024). Another compelling recent neuro-symbolic system is Neuro-Conceptual Artificial Intelligence (NCAI), which integrates deep learning with Object-Process Methodology (OPM) for enhanced question answering. NCAI converts natural language text into rich OPM conceptual models—encompassing objects, processes, and states—via in-context learning, and demonstrates improved reasoning transparency and accuracy over traditional triplet-based knowledge representation (Kang *et al.*, 2025). In the domain of knowledge graph reasoning, Hitzler *et al.* (2024) present a neuro-symbolic system

that unifies deductive reasoning with sub-symbolic graph embeddings for improved link prediction, demonstrating that symbolic inferences can bolster embedding-based learning.

Although these hybrid systems successfully combine the strengths of symbolic and neural paradigms, they often fall short of delivering cognitively grounded, multi-level language comprehension, systematic compositionality, and fine-grained control over neural behavior that cognitive theories like construction–integration (CI) suggest crucial. In addition to that, the sustainability of these architectures is frequently compromised by computational complexity or reliance on monolithic large models, rather than modular, scalable designs. These challenges underscore the need for neuro-symbolic designs that embody compositional structures, controlled reasoning, transparent decision paths, layered cognitive modeling, and environmental feasibility—all while adapting and harnessing the latest capabilities of LLMs. Only such comprehensive hybrid architectures can bridge the gap between cognitive theory and practical, robust natural language understanding.

3.3.4 Enhancing the Effectiveness of LLMs

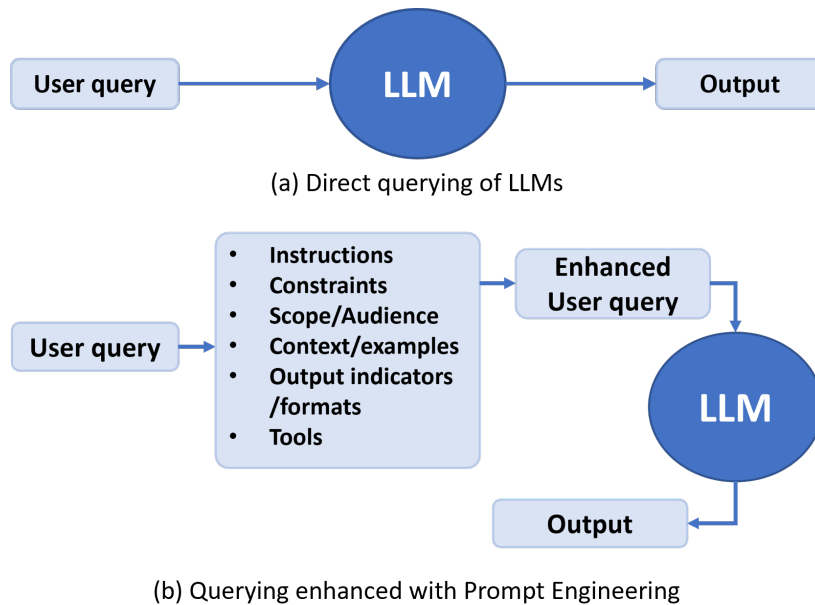


Fig. 3.9 Enhancing LLM Outputs Through Prompt Engineering

Several techniques can be used to minimize the limitations of LLMs we discussed earlier and fully utilize them in the NLP problems, as that's the best approach we have in the connectionist paradigm. Three main techniques are prompt engineering, transfer learning, and integrating rule-based components through prefixing and suffixing. These methods

enable LLMs to perform more accurately, reliably, and in a way that aligns with specific task requirements.

Prompt engineering is a collection of techniques that helps to carefully design prompts to guide the model in generating specific, desired responses. These techniques became popular with decoder-based language models such as GPT and Claude. It involves carefully designing prompts to guide the model in generating specific, desired responses. By strategically embedding instructions, constraints, contextual information, and formatting guidelines within the prompt, prompt engineering shapes the model's behaviour to produce outputs that meet user expectations and align with particular requirements. Additionally, crafting the prompts to solve the given tasks in certain templates and guidelines such as few-shot prompting (Brown *et al.*, 2020), Chain of thought(COT)(Wei *et al.*, 2022), and ReAct(reasoning and acting)(Yao *et al.*, 2023) enable more accurate responses. Altogether these techniques significantly enhances the relevance and usability of the model's responses, making them more suited to real-world applications.

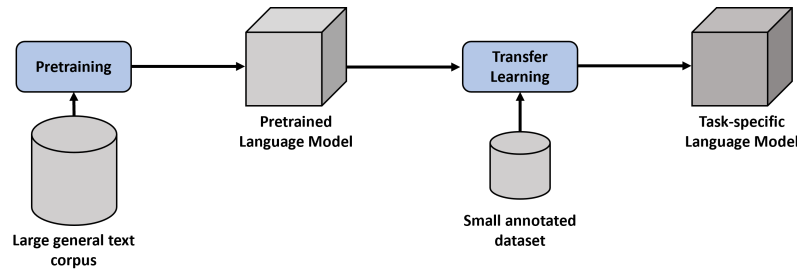


Fig. 3.10 Transfer Learning: Adapting Pretrained Models for Specialized Tasks

Transfer learning is another technique that lets the model adapt into more specialized or nuanced tasks from its inherited broad and general training. This approach is also referred to as fine-tuning, which enhances the skills of LLMs with task-specific datasets. This process improves the model's performance in areas with limited data, making it more effective and resource-efficient for specialized applications. Additionally, transfer learning can help fill gaps in knowledge that may have been missed during pretraining, such as incorporating visual knowledge through targeted transfer learning techniques.

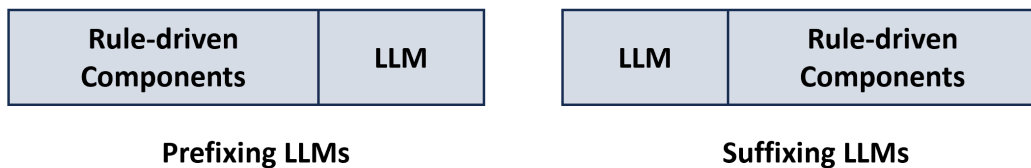


Fig. 3.11 Guardrails on LLMs

Adding guardrails to LLMs through prefixing and suffixing with rule-based components is another promising approach when leveraging these models. This hybrid method combines the flexibility of LLMs with the structured control of rule-based systems. In prefixing, rule-driven components are applied before the LLM generates a response, providing structured input or constraints that guide the model's behaviour. Conversely, suffixing involves post-processing the LLM's output with rule-based adjustments to ensure it meets specific criteria. This integration improves the controllability and explainability of LLM responses, making them more robust and adaptable for practical applications. Furthermore, the transparency of rule-based approaches adds an extra layer of explainability, enabling the use of smaller, resource-efficient language models rather than depending solely on large, resource-intensive LLMs.

3.3.5 Proposed Neuro Symbolic Framework for Natural Language Comprehension

Figure 3.12 showcases the proposed neuro-symbolic framework for natural language comprehension from this thesis. The core idea of the proposed framework is to modularize the components of NLP solutions, taking advantage of both symbolic and connectionist techniques. It leverages state-of-the-art LLMs from the connectionist paradigm with various techniques while utilizing traditional symbolic NLP methods according to the construction integration theory we discussed earlier in this Chapter. In this way, we can utilize both approaches within a unified system to build comprehensive NLP pipelines rather than relying solely on one approach, which can present limitations.

As depicted in the figure, on the symbolic side, the framework contains three main aspects to represent surface level, text-base and mental model construction steps in construction-integration theory. The surface level is represented by the explicit extraction of information from a natural language with techniques such as part-of-speech tagging, named entity recognition, regular expressions, topic modelling, keyword extraction, and tokenization. These methods facilitate the explicit parsing and understanding of language by identifying key elements and structures within natural languages. The text-base level represents the concept enrichment phase and is responsible for enriching the concepts with various sources aligned with world knowledge, such as lexicons, ontologies and other knowledge bases. The top layer of the construction integration theory, which is mental model construction represented by compositional semantics, involves using rules and heuristics, grammar and language structures, and logical reasoning to ensure the syntactic and semantic integrity of

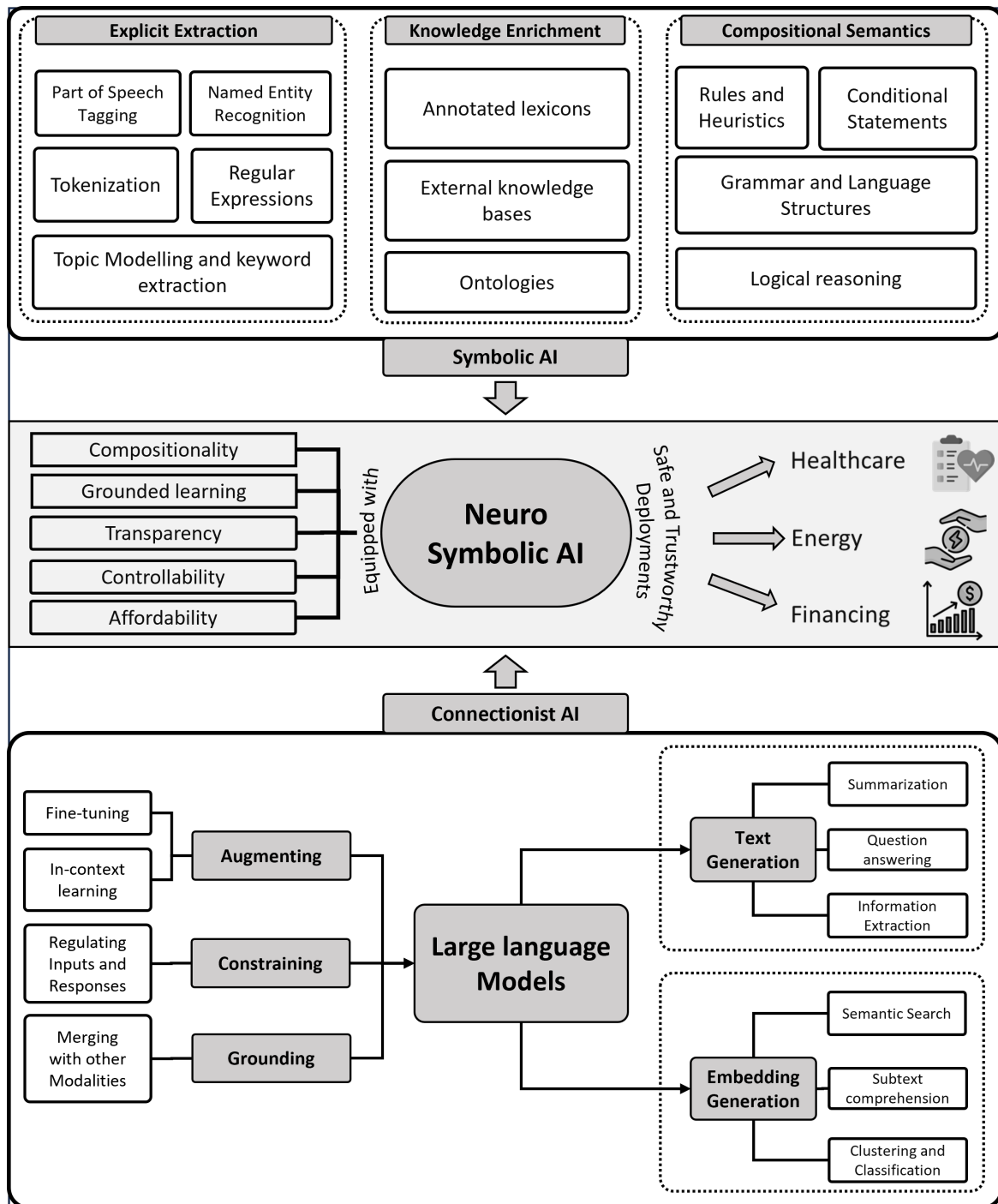


Fig. 3.12 Proposed Framework of Neuro-symbolic Architectures for NLC

language processing. This component ensures that the system can handle complex linguistic constructs systematically and modularly.

From the connections aspect, the framework is centred around the state-of-the-art large language models. LLMs will be augmented, constrained as well and grounded in order to avoid the limitations we discussed during the previous sections according to the corresponding use cases. Augmenting involves in-context learning and fine-tuning, allowing the models to incorporate specific additional information to ensure the expected behaviours. Constraining regulates inputs and responses to maintain the relevance and accuracy of LLMs by prefixing and suffixing them with rule-driven methods. Grounding specifically focuses on integrating the LLM's text-only comprehension capabilities with knowledge from other modalities, such as visual or auditory data, for a more comprehensive understanding that matches humans. LLMs can primarily generate two types of outputs: text and embeddings. Text generation capabilities can be utilized in scenarios such as summarization, question answering, and information extraction within neuro-symbolic pipelines. Embeddings, which are learned representations of natural languages, can be applied to tasks like semantic search, subtext comprehension, clustering, and classification. These capabilities enable the system to reach implicit areas that symbolic approaches alone cannot address.

In this way, it avoids the requirement of larger language models as, on many occasions, simpler encoder-based LLMs can be utilized with the guardrails to meet desired outcomes. This will make developing NLP solutions more sustainable and feasible for a wider cohort in society. Furthermore, often overlooked symbolic AI methods will add value to the NLP solutions by prefixing and suffixing the LLMs to better control and intended behaviours. Ultimately, the solutions build upon the proposed neuro-symbolic framework are expected to be equipped with several key features such as compositionality, grounded learning, transparency, controllability, and affordability with these architectures. Thus, it will enable the trustworthy and safe deployment of applications across domains such as financing, energy, and healthcare.

3.4 Chapter Summary

This chapter starts by comprehensively analysing several techniques that enable effective and efficient language comprehension in humans. Then, it compares these human comprehension strategies to the computational methods of language comprehension with a particular focus on neuro-symbolic NLP methods. The chapter explores symbolic and connectionist paradigms in detail, outlining the strengths and limitations of each approach, and underscores the necessity of their integration. Finally, it introduces a neuro-symbolic framework for natural language comprehension, offering a structured guide for developing NLP solutions that synergistically combine the advantages of both symbolic and connectionist methods.

Chapter 4

Evaluating the Neuro-Symbolic Architectures in Natural Language Comprehension

4.1 Chapter Introduction

Comprehensive studies and experiments that were conducted to evaluate the proposed neuro-symbolic framework for natural language comprehension are described in detail in Chapters 4, 5, and 6. An overview of these experiments is illustrated in figure 4.1. One of the primary focuses of the implementations centred around emotion elicitation in natural languages. Emotions play a vital role in language comprehension, enabling the perception of underlying sentiments and intentions in communication. Chapter 5 describes detailed experiments on emotion elicitation using the novel algorithm Emotion AWARE: an artificial intelligence framework for adaptable, robust, explainable, and multi-granular emotion analysis. This chapter also extends with two ablation studies based on the Emotion AWARE framework. In addition to emotion elicitation, another major focus of the applications covers improving natural language interactions. Chapter 6 explores the utilization of neuro-symbolic architectures for natural language interactions, discussing the implementation of the novel Energy AI Chatbot, Cooee. This section details how neuro-symbolic architectures are employed for interacting with both structured and unstructured data spaces and for facilitating multi-agent collaboration.

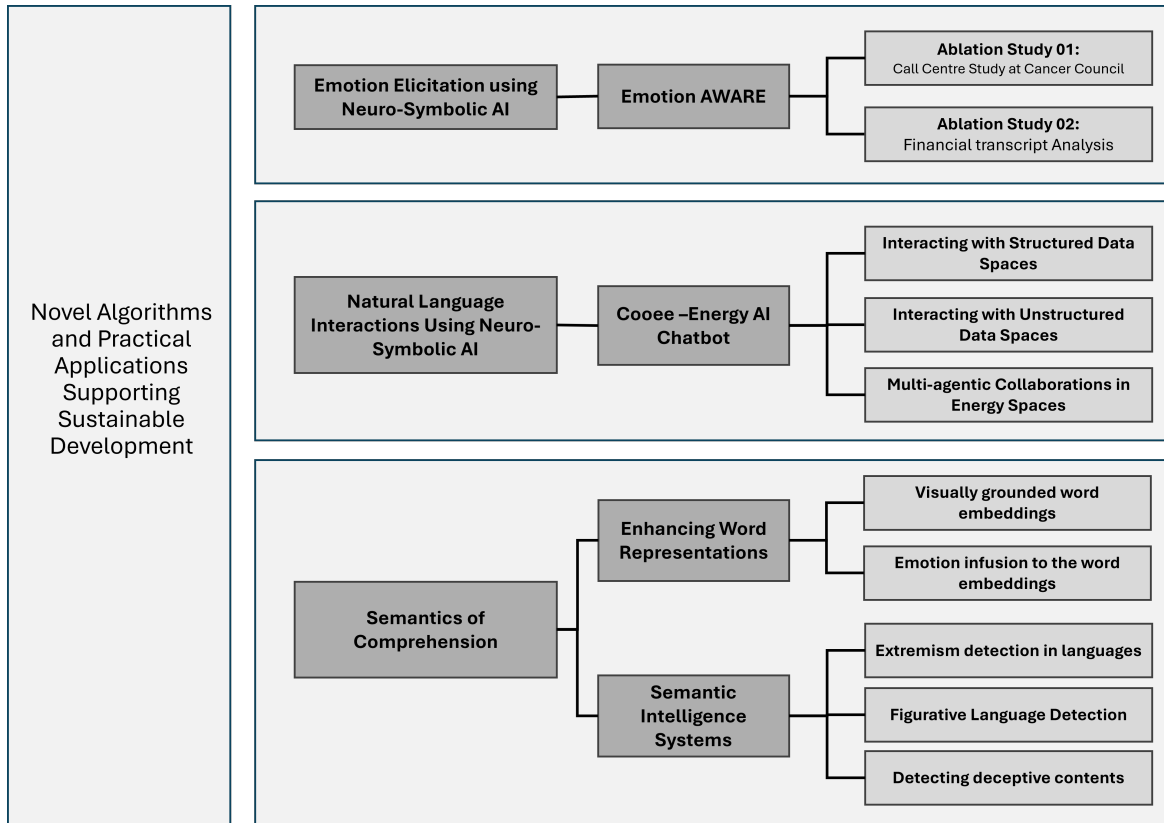


Fig. 4.1 Overview of the Applications Utilizing the Neuro-Symbolic Framework

Chapter 4 mainly focuses on discussing several other experiments and studies to evaluate various facets of natural language comprehension using neuro-symbolic architectures. This chapter is divided into two main sections: Enhancing Word Representations and Semantic Intelligence Systems. Word representation improvements consist of two main implementations, including Visually Grounded Word Embeddings, which integrate visual context to enhance the semantic understanding of words, providing a richer and more nuanced representation, and Emotion Infusion to Word Embeddings, which infuses emotional context into word embeddings, improving the system's ability to perform sentiment analysis by capturing the subtleties of human emotions. The semantic Intelligence systems section will cover several other facets of natural languages to provide overall coverage of neuro-symbolic architectures in the natural language comprehension spectrum. These implementations include Extremism Detection in Languages, identifying and analyzing extremist language patterns to prevent radicalization and promote safer communication environments; Figurative Language Detection, recognizing and interpreting idiomatic and figurative expressions; enhancing the system's ability to understand complex language usage and improving overall language comprehension; and Detecting Deceptive Content, developing systems to detect fake news

and other forms of deceptive content, helping to maintain the integrity of information and combat misinformation. Ultimately, these diverse applications and experiments will illustrate robust approaches to advancing natural language comprehension by validating the proposed innovative framework presented in the thesis.

4.2 Enhancing Word Representations

Word embeddings are numerical representations of the words, often represented as a sequence of numbers. Embeddings can capture and approximate word meanings, semantic relationships, and syntactic properties by positioning them in a continuous vector space. These representations are generated by distributional semantic models, which focus on various factors such as word appearance, frequency, and surrounding words. Building numerical representations enables many traditional machine learning algorithms to leverage natural language data as they do not support direct text inputs.(Merkx *et al.*, 2022).

Word embedding generation usually requires a large corpus of text to learn the statistical properties of the words and their contexts to create accurate numerical representations. There are several approaches for vectorizing words, including simple count-based methods to language models, which can capture the contextualize with higher accuracies. Count-based methods create word occurrence matrices that show the frequency of words. Techniques like Latent Semantic Analysis (LSA) (Deerwester *et al.*, 1990), and GloVe (Global Vectors for Word Representation) (Pennington *et al.*, 2014), refine these matrices into dense vector embeddings. In addition to that, there are several predictive methods that involve training neural networks to predict words from their context. Examples include Word2Vec (Mikolov *et al.*, 2013a), which uses CBOW and Skip-gram architectures, and FastText(Bojanowski *et al.*, 2017), which considers subword information. However, both of these approaches contain static word embeddings where the representations will not vary depending on usage, which is a major limitation of these representations. Contextual embeddings, such as BERT(Devlin *et al.*, 2018) and ELMo (Embeddings from Language Models)(Peters *et al.*, 2018a) and more recently GPT models(Radford and Narasimhan, 2018) address this limitation by generating dynamic embeddings based on word appearance within sentences, capturing polysemy and providing different representations for different contexts.

Word embeddings add great value in neuro-symbolic pipelines by creating numerical representations leveraging the learnability of connectionist approaches. These representations enable broader utilization in symbolic reasoning approaches rather than using raw text inputs, which ensures controllability. Furthermore, embeddings can enhance explainability by providing structured, quantifiable similarity measures. However, when considering even the

most sophisticated embedding techniques based on LLMs, one obvious behaviour is that their quality always improves with the model's complexities, which is not a sustainable trend. Thus, with these experiments, we are analysing the feasibility of improving the word representations by infusing them with usually overlooked implicit aspects such as emotionalities and visual knowledge to build more robust embeddings, which can be impactful in many use cases.

4.2.1 Visually Grounding Word Embeddings

Introduction

To improve the quality of word embeddings, currently, the most common approach is to expand the models to capture more word arrangements so that they can predict better probabilistic sequences, which leads to more accurate vectors. This behaviour is reflected by top-performing models with popular embedding benchmarks such as MTEB(Muennighoff *et al.*, 2023). For instance, when considering a much simpler model compared to many LLMs today, GloVe embeddings(Pennington *et al.*, 2014) were trained on 840 billion words, which would take a human about 80 years of continuous reading at 330 words per second to process(Chersoni *et al.*, 2022). In contrast, humans can comprehend language with much less exposure to support from other modalities from a richer sensory experience beyond just reading texts(Merkx *et al.*, 2022). This major gap between human and machine language comprehension motivates us to rethink the approaches to word representation learning in general.

As elaborated in Chapter 3, grounded language learning is one of the major processes that enhances the language learning process of humans by connecting language symbols to real-world environments through multiple modalities. This is further elaborated in the theory of embodied cognition, which asserts that our conceptual knowledge is rooted in the entirety of our sensory experiences(Barsalou, 2008; Foglia and Wilson, 2013). For example, reading "cat" evokes sensory experiences associated with cats, such as their sound and appearance. According to this theory, all sensory experiences contribute to our conceptual knowledge and processing, which should be evident in human behavior(Merkx *et al.*, 2022). Especially the contribution from the visual aspect plays a vital role in this sensory process of cognition(Schreuder *et al.*, 1985). So, this highlights the claim that if visual features are integral to our conceptual knowledge, then word embeddings that incorporate visual features should better explain human behavioural data than purely text-based methods. In other words, it explains that visual knowledge in compression cannot be fully conveyed by the text alone. As LLMs continue to grow in size and complexity, they face numerous challenges. By integrating other modalities into the language learning process, it is anticipated that

language representations will become more robust and effective compared to purely text-based approaches. Thus, in this experiment, we explore the effectiveness of incorporating visual features into smaller-scale LLMs to enhance word representations.

Methodology

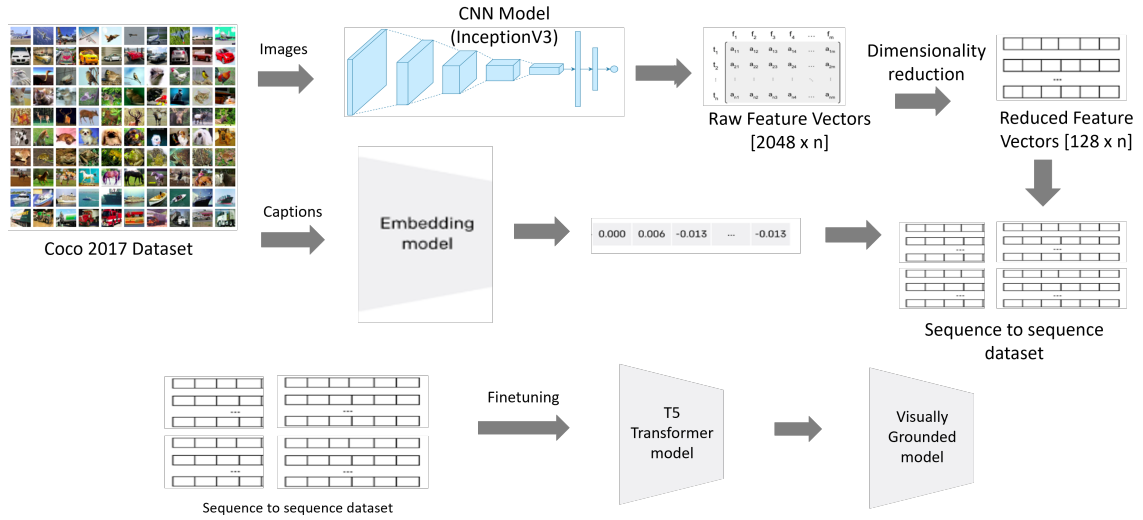


Fig. 4.2 Visually Grounding Word Embeddings

Figure 4.2 provides an overview of the methodology of visually grounded word embedding construction. The main intention of this work is to align language concepts with corresponding image references in order to extend the language learning capability. Thus, for this implementation, we leveraged the popular text and image dataset Microsoft COCO 2017 dataset (Lin *et al.*, 2014). Each sample in this dataset consists of an image accompanied by five distinct human-generated captions in text. Some examples from this dataset are illustrated in figure 4.3. Initially, we iterated through the captions for each image, creating a one-to-one mapping between captions and images. This ensures alternate descriptions for the same image reference, further generalising the visual grounding. To merge the images with the text captions, we employed a sequence-to-sequence learning approach. Therefore, it was necessary to convert images and captions into numerical sequences to prepare them for sequence-to-sequence learning (Sutskever *et al.*, 2014b).

In this method, a pre-trained convolutional neural network (CNN) model was used to extract and encode image features into numerical sequences. CNNs are a type of neural network designed specifically for processing grid-like data, such as images, by extracting local information and aggregating it through multiple layers of learnable parameters. These networks are typically trained on large datasets of human-annotated images to classify them



Fig. 4.3 Image and Caption Pairs from COCO2017 Dataset

into various categories (e.g., dog, horse, car). CNNs that were trained on such tasks can be further leveraged for encoding dense, meaningful numerical representations that align well with human intuitions (Bracci *et al.*, 2019; Lazaridou *et al.*, 2017). Here specifically, we used pre-trained Inception-V3 CNN model (Szegedy *et al.*, 2016) trained on the ImageNet dataset (Deng *et al.*, 2009), which is widely used in image processing downstream tasks. We set the dimension of the grounded embeddings to 1024, and a one-layer LSTM with 2048 units was applied. The image features, initially represented by numerical sequences of length 2048, were too lengthy for the sequence-to-sequence classification process. Therefore, we reduced the dimensionality of these vectors to 128 using Principal Component Analysis (PCA) (Jolliffe, 2011).

Encoder-based small-scale BERT language model (Devlin *et al.*, 2018) was used to embed the captions of the images. The BERT model has demonstrated groundbreaking performance in many natural language processing tasks, even using only text modality. During the pre-training phase, it has gone through extensive language learning with tasks such as mask language modelling and next-word prediction. In our sequence-to-sequence learning process, our main intention was not to replace this textual learning but to extend it with visual cues for better grounding. Thus, this mapping should ideally preserve the abstract knowledge derived from co-occurrence statistics in large textual corpora and align these textual embeddings

with their corresponding visual properties in images. By doing so, the grounded embeddings will effectively benefit from both concrete and abstract concepts.

Although it might appear logical to learn both modalities in a shared feature space, we contend that this approach is prone to causing the grounded embeddings to lose the abstract knowledge gained from textual co-occurrences. This often results in a bias towards concrete words, as observed by Park and Myaeng (2017) in their experiments. During the encoding, all the captions were processed using the default BERT base model to extract 128-length numerical vectors representing the image captions. After encoding both the images and captions, we constructed a sequence-to-sequence dataset for the transfer learning process. For transfer learning, we used the T5 base model(Raffel *et al.*, 2019) and tokenizer from the Hugging Face Transformers library(Wolf *et al.*, 2019), with a learning rate of 3e-4, weight decay set to 0.01, and a sentence length of 128. After training the T5 model, it was used to construct text embeddings utilizing the wrappers provided in the Hugging Face library.

Experiments

This section explains some experiments completed to evaluate the embeddings generated by the extended T5 model. These experiments are essential to carry out as it is crucial to evaluate the embedding quality before utilizing them in various neuro-symbolic approaches. As the evaluation task, we used word similarity and relatedness benchmarking, which highlights the value of visually groundedness of the word embeddings. To conduct this evaluation, as datasets, we used human-annotated lexical semantic similarity and relatedness datasets commonly employed for assessing embeddings(Mandera *et al.*, 2017; Rotaru and Vigliocco, 2020). The task involves estimating the similarity measures of a given pair of words, with the Pearson correlation coefficient serving as the evaluation metric. Relatedness is based on topical match, quantifying the degree to which two words are associated (e.g., child-play). Similarity, a subset of relatedness, is based on taxonomic closeness and quantifies how alike two words are (e.g., car-automobile). It is worth noting that some datasets do not distinguish between similarity and relatedness. We assessed the quality of our visually grounded word representations using the following datasets.

- **The MEN dataset** (Bruni *et al.*, 2014) is designed specifically for evaluating multi-modal models. It comprises 3,000 word pairs with semantic relatedness ratings collected through Amazon Mechanical Turk. Each relatedness score is given out of 50. For example, the pair (sun, sunlight) achieves a perfect MEN score of 50, while the pair (zebra, bakery) scores 0.

- **The WordSim353 dataset** (Finkelstein *et al.*, 2001) comprises 353 word pairs, each annotated by 13–16 human judgments. The judges did not differentiate between similarity and relatedness. Scores are given out of 10. For example, the pair (computer, keyboard) has a score of 7.62.
- **The SimLex999 dataset** (Hill *et al.*, 2015b) includes 999 word pairs annotated by 500 annotators via Amazon Mechanical Turk. This dataset has been utilized to explain human performance in psycholinguistic tasks. SimLex999 provides scores out of 10 for the word pairs.

To compare the embedding quality of the visually grounded model, we evaluated these benchmarks against several other existing approaches. Primarily, we compared the embeddings from the OpenAI text embeddings model, text-embedding-ada-002, which is widely utilized in various natural language processing tasks today. Additionally, we assessed the embeddings from the OpenAI’s CLIP model, popular text and image model, and the Cohere embedding model, embed-English-v2.0. Finally, we included the T5 base model in our comparisons to compare with the extended model to get insights on the generalizability trade-offs with extended pre-training. Table 4.1 outlines the results of these comparisons against the benchmark datasets, showing the Pearson correlation coefficients to human-annotated scores for each benchmark.

Benchmark	T5 base	CLIP (Open AI) – Text Image Model	Open AI(ada-002) Embedding model	Cohere	Visually grounded model
MEN	0.564	0.390	0.418	0.177	0.656
Simlex999	0.508	0.305	0.450	0.134	0.517
WordSim353	0.639	0.397	0.738	0.122	0.610

Table 4.1 Comparison of Word Similarity Benchmarks and Models

The results of the experiments reveal that for the MEN and SimLex benchmark datasets, the visually grounded model shows the highest correlation scores in comparison to the other approaches. In contrast, the highest correlation for the WordSim benchmark is achieved by the OpenAI ada-002 model, which is a text-only model. Although CLIP is a text-image model, it has a lower correlation score, likely because its primary objective is not text embedding extraction but to provide image reference to the text inputs and vice versa. Overall, this experiment demonstrates the potential of improving word similarity matrices with visual

grounding. However, the performance varies across different benchmarks, highlighting the complexities of measuring relatedness and similarity matrices. Text-only models sometimes show higher relatedness scores as their primary focus is to learn the context with surrounding words during the pretraining. Model sizes can also be a contributing factor to the score variations as larger models can approximate the visual clues using text in the training. Additionally, as seen in the WordSim benchmark, visually grounded models can lose some contextual abstractions, which the T5 base model retains, leading to reduced correlation scores. Nonetheless, this experiment emphasizes the potential of visually grounding word embeddings and the need for better benchmarks to evaluate embedding effectiveness accurately.

4.2.2 A Robust Word Embedding for Implicit and Explicit Expressions of Emotion

Introduction

Machine perception of emotions is an essential component in the human central AI systems. Understanding human emotions can improve user experiences, safety, and overall effectiveness of natural language comprehension. In industrial settings, emotion recognition of human operators has diverse benefits and applications, such as effective coordination of autonomous vehicles(Nallaperuma *et al.*, 2018), improved real-time face expression(Putro *et al.*, 2020), and autonomous video surveillance in smart city environments(Nawaratne *et al.*, 2017). However, capturing human emotions is a complex and challenging task as they are often conveyed in an implicit nature in languages rather than explicit expressions. Computational methods, specifically distributional semantics, seek to measure the semantic similarities between terms based on their distribution across large text datasets(Harris, 1954). Building on this concept, distributed word representations, or word embeddings, have become the most effective way to capture semantic meanings.

However, when considering traditional word embedding techniques such as Word2Vec(Mikolov *et al.*, 2013a), GloVe(Pennington *et al.*, 2014), ELMo(Peters *et al.*, 2018a) and embeddings based on recent LLMs, although they excel at capturing general semantics and contextual meaning they often overlook the subtle emotional cues that are essential for emotion-sensitive applications. Several existing studies attempt to address some of the challenges of word embeddings for human emotions(Wu and Jiang, 2019a; Agrawal *et al.*, 2018a). However, they are tightly coupled to specific expressions of emotion. This makes the embeddings less robust in situations where emotions are implied or vaguely described. To address this limitation, we propose a robust word embedding method for both implicit and explicit emo-

tion expressions. This method employs two techniques, Mask-EmoZen and Rogue-EmoZen, which are based on Gated Recurrent Unit (GRU) neural network architectures. Mast-Emozen learns the semantics of emotion based on the position of the term rather than the term itself, while Rogue-EmoZen learns semantics based on a rogue term instead of the actual term, which ultimately improves the word embedding with emotion infusion.

Methodology

In this section, we introduce our proposed approach, EmoZen, which comprises two techniques: Mask-EmoZen (EmoZen-M) and Rogue-EmoZen (EmoZen-R). Both techniques utilize neural network models based on Gated Recurrent Units (GRU) (figure 4.4). GRU(Chung *et al.*, 2015) is an improved version of a standard Recurrent Neural Network designed to overcome the vanishing gradient problem and exploding gradient problems(Bengio *et al.*, 1994).

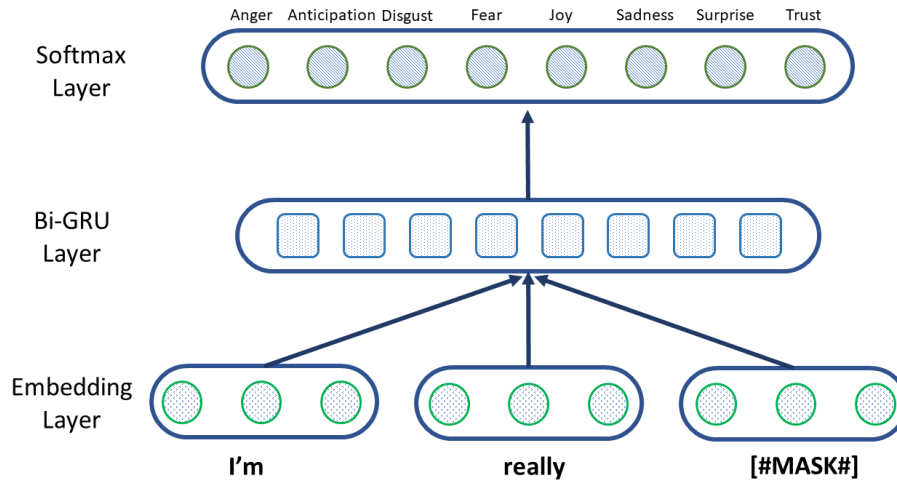


Fig. 4.4 Emozen Model Architecture

A single conversational element (usually a single sentence) expresses a single emotion in human communication. To capture this organic nature of communication, we annotated each sentence in our training dataset with one emotion label. Benchmark datasets such as ISEAR (Scherer and Wallbott, 1994a) and Crowdfower have adopted this approach. The task of learning distributed word representations requires a large corpus of textual data with a rich vocabulary. Finding emotion-rich textual data is a challenging task. Recent research indicates that Twitter, a social media microblogging platform, generates close to a billion emotion-rich textual data per day(Hasan *et al.*, 2019; Wang *et al.*, 2012). We collected seven million (7M) user-generated tweets containing specific keywords relating to eight

Emotion	Keywords	# of Tweets
Anger	angry, furious	743,978
Anticipation	looking forward, expect, confident	1,009,733
Disgust	disgusted, disgusting	981,960
Fear	afraid, frightened, scared, fearful	1,121,986
Joy	cheerful, happy, joyful	1,016,192
Sadness	sad, depressed, sorrowful	599,213
Surprise	surprising, surprised, astonished, shocked, startled, astounded, stunned	961,866
Trust	trust, faith, certainly, certain, belief, trusting	747,108
Total		7,182,036

Table 4.2 Emotion-related keywords used to extract the training dataset

emotions described in Plutchik’s wheel of emotion(Wallon, 1972). The emotion-related terms used to extract tweets are summarised in table 4.2. We employed a series of filters and normalization techniques to preprocess the acquired tweet dataset. First, we filtered tweets with non-English words and phrases and tweets with less than five words. Next, we normalized words such as usernames, dates, numbers, phone numbers, emails, and URLs by replacing them with unique tokens. Regarding the training corpus, we define a unique token vocabulary as v . Where $v = \{v_1, v_2, v_3, \dots, v_{|v|}\}$. Each word is represented by a d -dimensional vector $w_i \in \mathbb{R}^d$. The embedding matrix containing all unique tokens from the corpus is given as $W_e \in \mathbb{R}^{d \times |v|}$. We randomly initialize the word embedding matrix W_e to learn robust emotion word embeddings. In this case, we associate each sentence S with its corresponding target variable y , an emotion label. $d = \{(S_1, y_1), (S_2, y_2), \dots, (S_n, y_n)\}$ denotes the dataset with emotion label y_i for each sentence S_i . Given a sentence $S = [s_1, s_2, \dots, s_t, \dots, s_T]$ where s_t is the one hot vector representation of word at position t , the words are embedded into a continuous vector space using the embedding matrix W_e . X serves as the embedding of the input sequence and is fed to the bidirectional GRU Layer.

$$X = S * W_e \quad (4.1)$$

The output from the GRU layer (Say H) is fed to a dense layer that has a softmax activation. It outputs a vector of size 8 (number of classes). The values in the vector

components are probabilities for the presence of each of the eight emotions. The emotion with the highest probability is selected as the output.

$$Y = W_{dense}H \quad (4.2)$$

For all $y_i \in Y$, f_i is calculated as follows.

$$f_i = \frac{e^{-y_i}}{\sum_{y_j \in Y} e^{-y_j}} \quad (4.3)$$

Then, the class with the highest f_i is taken as the output.

$$\text{output} = \text{argmax}_i f_i \quad (4.4)$$

Looking at recent seminal developments in word embedding, it is evident that the learning objective is a significant consideration in determining the quality of the continuous word representation. For instance, C&W model (Collobert *et al.*, 2011) used corrupted n-gram, Word2Vec (Mikolov *et al.*, 2013c) model used skip-gram, and CBOW methods also some more recent models such as BERT (Devlin *et al.*, 2018) used masks. A shared task at EMNLP 2018 (Klinger *et al.*, 2018) focused on predicting emotion labels without having the emotion-related keywords in the input text. This task concluded that human performance is lower than that of automatic systems (which included neural network models). Building on this outcome, we posit that a neural network model designed with the appropriate learning objective can predict the emotion associated with a sentence based solely on the context words. We advance this position into the two techniques we propose, Mask-EmoZen (EmoZen_M) and Rogue-EmoZen (EmoZen_R).

Mask-EmoZen: Masked Robust Emotion Word Embedding

In Mask-EmoZen, keywords related to each emotion are hidden using a special masking token, but 10% of the dataset is unmasked in order to maintain the generalisability of the neural network. The learning objective is to predict emotion of the sentence using the context with masked keywords. The expectation here is that the neural network will learn the context of an emotion expression based on the position of the emotion-related keyword/s and not the actual word itself.

E.g. I'm really happy for you. $\rightarrow$ I'm really [#MASK#] for you.

Rogue-EmoZen: Rogue Robust Emotion Word Embedding

In Rogue-EmoZen, keywords related to each emotion are replaced with a rogue word from the same training vocabulary. The learning objective is to predict the emotion of the sentence using the context with rogue keywords. Just as in Mask-EmoZen, we retain the original emotion-related keywords in 10% of the dataset. In this case, it is more challenging than Mask-EmoZen as the neural network is required to generate word embeddings with neither emotion-related keywords nor the position of such words.

E.g. I'm really happy for you. $\rightarrow$ I'm really **computer** for you.

Experiments

Emotion Classification

We validated the effectiveness of our proposed robust emotion word embeddings by utilizing them for emotion classification on two benchmark datasets. Following the methodology of (Wu and Jiang, 2019b), we averaged the learned word embeddings across each dimension to create sentence representations. These representations were then used as input features to predict emotion labels at the sentence level and were fed into multi-class classifiers based on Logistic Regression (LR) and Support Vector Machine (SVM).

Datasets

- **ISEAR**: ISEAR dataset (Scherer and Wallbott, 1994b) was compiled by approximately 3,000 participants who have different cultural backgrounds and completed questionnaires about experiences and reactions to seven emotions. We use a subset of the dataset that only includes anger, disgust, fear, joy, and sadness emotions.
- **CrowdFlower** : The CrowdFlower(CrowdFlower, 2024) dataset consists of 39,740 tweets. It is annotated via crowdsourcing with one label per tweet. We have used a subset of the dataset with labels sadness, surprise, love, hate, happiness, and anger.

In order to compare our model with state-of-the-art (SOA) word embedding models, we trained each model with the acquired seven million (7M) tweet corpus. Classification performance of all seven models using 10-fold cross-validated macro F_1 score is tabulated in table 4.3. Mask-EmoZen and Rogue-EmoZen consistently outperform SOA word embeddings in both LR and SVM classification for both datasets. Between the two, Mask-EmoZen performs better, given that Rogue-EmoZen attempts the more challenging task of learning without the semantics of emotion-related words or the position of such words.

Methods	ISEAR		CrowdFlower	
	LR	SVM	LR	SVM
GloVe	0.498	0.506	0.522	0.514
Word2vec	0.472	0.477	0.504	0.511
SSWE	0.491	0.493	0.514	0.526
DeepMoji	0.578	0.583	0.537	0.533
EWE	0.583	0.581	0.542	0.547
EmoZen _R	0.613	0.621	0.603	0.604
EmoZen _M	0.628	0.633	0.611	0.619

Table 4.3 Comparison Against State-of-the-art Word Embedding Models

Emotion Similarity

NRC emotion lexicon (Mohammad and Turney, 2013) has word groups for eight emotions: Joy, Trust, Anticipation, Surprise, Fear, Anger, Disgust and Sadness and two polarities, negative and positive. We calculated the cosine similarity between emotion groups by comparing all individual words in each group, unlike (Agrawal *et al.*, 2018b) where comparison was on randomly selected samples. The standardised similarity scores are visualised in figure 4.5. This experiment further validates the robustness of emotionally infused word embeddings.

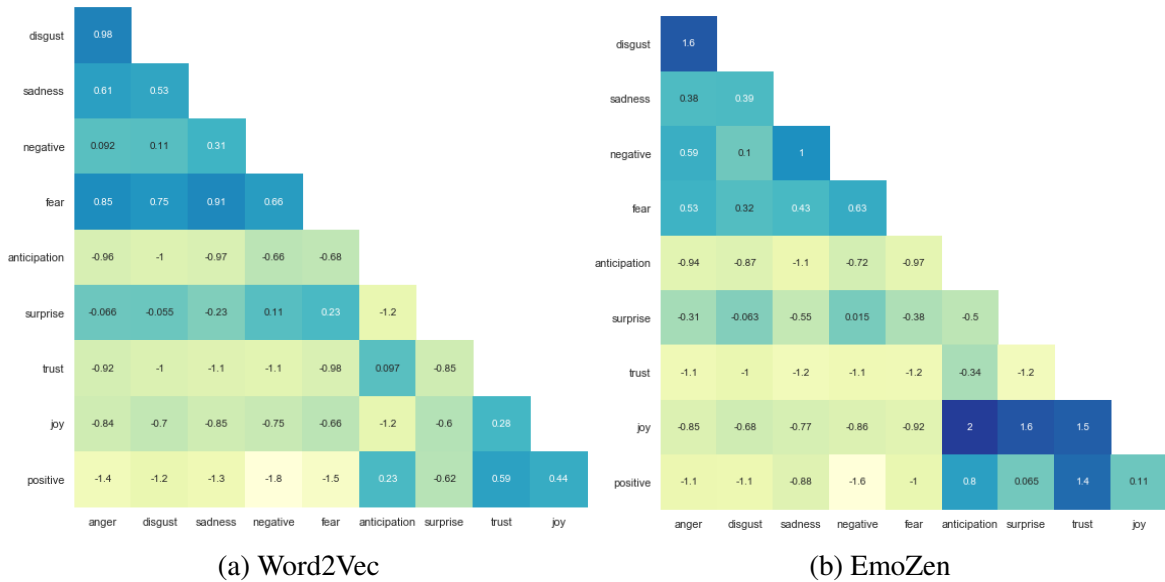


Fig. 4.5 Heatmaps of Cosine Similarity Scores Between Words with Paired Emotions

In conclusion, from this work, we focused on the robustness of word embeddings for emotion expressions. We proposed two techniques; Mask-EmoZen and Rogue-EmoZen. Mask-EmoZen learns the semantics of an emotion expression based on the position of an emotion term. Rogue-EmoZen learns to predict emotion using a context impacted by rogue keywords. We evaluated both techniques on emotion classification using two datasets, and compared performance with five SOA word embeddings. Both techniques outperform SOA, indicating the validity and robustness of EmoZen for implicit and explicit expressions of emotion. This methodology, in general, exemplifies a successful attempt to incorporate often overlooked implicit aspects of language comprehension in word representation learning.

4.3 Disinformation Detection

4.3.1 Introduction

The rapid spread of fake news or disinformation is a major challenge in the digital age, threatening societal harmony, democratic processes, and individual well-being. The growth of social media use has turned disinformation into a significant issue for governments, organizations, and communities. With the vast popularity of social media today, false information can be spread quickly and widely, reaching millions of users almost instantly. This quick spread makes it hard for accurate information to keep up, allowing disinformation to spread easily (Olan *et al.*, 2022). Fake news is often generated and spread through automated processes, resulting in large volumes, high speed, and diverse formats. However, fake news is also typically short-lived to avoid detection and counteraction by truthful responses.

Several studies from the literature highlight that fake news significantly affects public perception and behaviour. For instance, Lazer *et al.* (2018) discuss how spreading misinformation can influence election outcomes, minimise trust in public institutions, and polarize public opinion. These effects are amplified by the algorithms of social media platforms that prioritize engagement over accuracy, often leading to the expansion of misleading content (Vosoughi *et al.*, 2018). Allcott and Gentzkow (2017) highlight that the lack of strict content moderation with the popularity of these media allows fake news to flourish unchecked. The anonymity and rapid dissemination of information on platforms like Facebook, Twitter, and Instagram make it challenging to identify and control the spread of false information.

There are several approaches in the literature attempting to address this challenge of fake news detection and resolution. In summary, these approaches leverage various techniques, focusing on different aspects of the problem. One common approach is the usage

of LLMs for this task, similar to many downstream NLP tasks. Notable examples include SpotFake(Singhal *et al.*, 2019), exBAKE(Jwa *et al.*, 2019), and fakeBERT(Kaliyar *et al.*, 2021), which are based on small-scale encoder-decoder LLMs fine-tuned with fake news datasets. However, a major limitation of these approaches is their heavy reliance on training data, which can hinder generalization. Additionally, they often lack explainability and do not account for specific considerations of other aspects of fake news, focusing primarily on language as represented in the training data.

Fake news is intentionally written to mislead readers and is often crafted to trigger intense emotions, aiming to propagate widely on social networks. Some researchers have demonstrated that those disseminating fake news tend to use negative emotion words extensively in their writing style as a sign of guilt(Newman *et al.*, 2003). This nature of fake news construction can be used in reverse to detect fake news. In the literature, there are several attempts to leverage emotionalities in the fake news detection problem(Ajao *et al.*, 2019; Giachanou *et al.*, 2019; Guo *et al.*, 2019a). Furthermore, linguistic features with symbolic methods, such as n-gram analysis(Yang *et al.*, 2019), TF-IDF(Verma *et al.*, 2021), and lexicon-based approaches, are also employed in fake news detection. Although these methods may not always show the highest accuracies, their features, such as explainability and controllability, make them stand out in various use cases. While numerous symbolic and connectionist methods are used in isolated aspects of fake news detection, there is a lack of comprehensive studies that integrate multiple aspects using neuro-symbolic methods to ensure generalizability and explainability. This study experiments with ensemble neuro-symbolic pipelines to combine these various facets. The aim is to enhance the generalizability, explainability, and robustness of fake news detection.

4.3.2 Methodology

The multi-faceted neuro-symbolic framework for fake news detection is illustrated in figure 4.6. This architecture consolidates various symbolic and connectionist approaches in NLP to ensure a generalizable, robust, and comprehensive detection mechanism. There are three facets in this process: emotion, language, and semantics. In the framework, separate pipelines are dedicated to covering each facet and creating a feature vector representing each aspect. These feature vectors are then forwarded to their respective classifiers. The outputs from these classifiers are combined using a voting classifier to produce the final classification label of a news item as either true or fake.

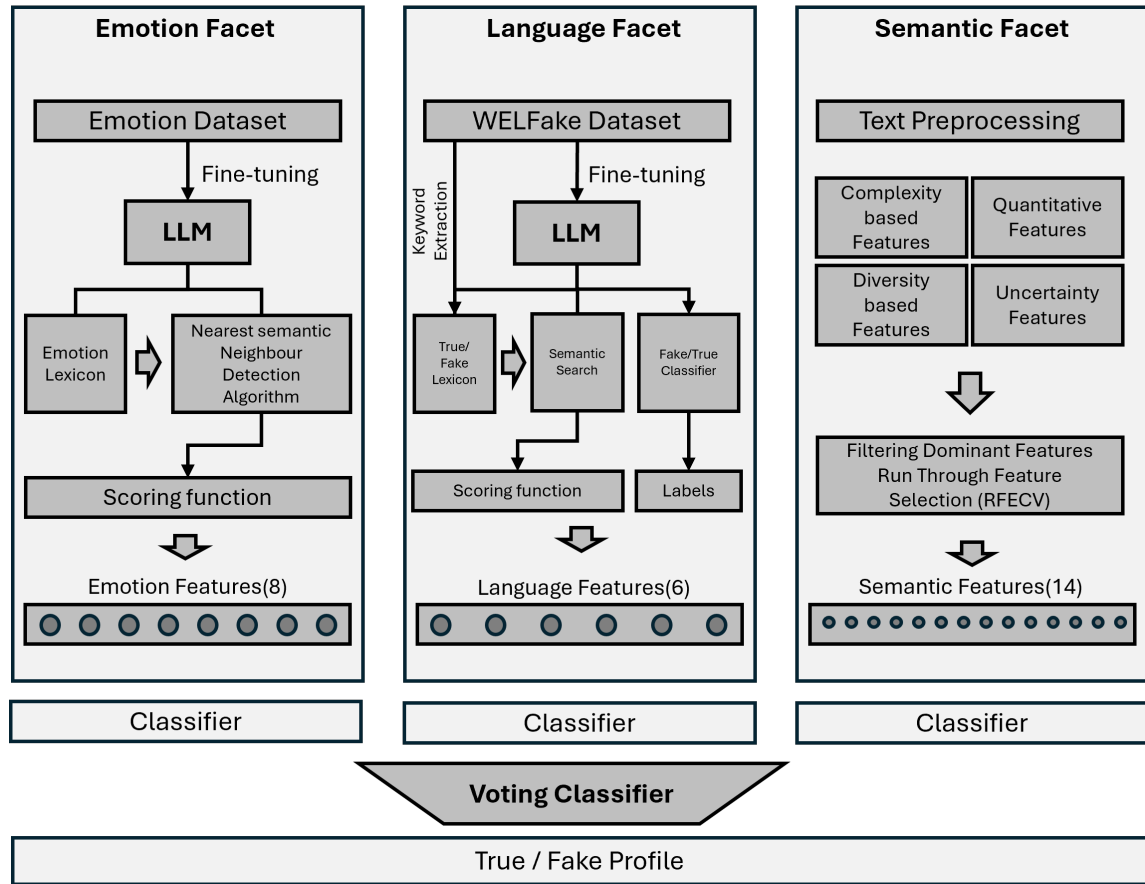


Fig. 4.6 Multi-faceted Neuro-Symbolic Architecture for Fake News Detection

Emotion Facet

In both the Emotion and Language facets, small-scale encoder-based LLMs along with lexicons are utilized. These LLMs are fine-tuned with corresponding datasets, enhancing the quality of the embeddings compared to baseline LLMs. Specifically, we use the DistilBERT model (Sanh *et al.*, 2019) through the Hugging Face Transformers library (Wolf *et al.*, 2019). DistilBERT is a transformer model that is smaller and faster than BERT, pre-trained on the same corpus in a self-supervised fashion using the BERT base model. For the Emotion facet, the approach aligns with the Emotion Aware algorithm (Gamage *et al.*, 2024), detailed in Chapter 5. It constructs eight emotion profiles based on Plutchik's eight basic emotions: joy, trust, anticipation, surprise, sadness, anger, fear and disgust (Wallon, 1972). Emotional profile construction is achieved using a nearest-neighbour detection mechanism combined with an embedded lexicon and scoring function. The final emotion profiles are represented as a vector with normalized scores for each of the eight emotions, ranging from 0 to 1.

Language Facet

In the language facet, the primary goal was to leverage connectionist learning capabilities to improve the separation between fake and true news. For this purpose, we used the WELFake dataset (Verma *et al.*, 2021), which contains over 72,134 news articles, including 35,028 real and 37,106 fake news pieces. This dataset was constructed by combining long texts from four different well-cited datasets to prevent classifier over-fitting (Verma *et al.*, 2021). Prior to the fine-tuning process, the WELFake dataset was subjected to several preprocessing steps. Firstly, the duplicates were removed, and then the title and text columns were concatenated together to create data points. Afterwards, the dataset was split into training, testing, and validation sets. Fine-tuning was performed using the Transformers library, achieving an accuracy of 0.99 on the test and validation datasets. This fine-tuned LLM serves two purposes: firstly, to predict label probabilities and then to extract the embeddings. To predict the true and fake probabilities of any given news item, LLM is used with the classification head based on its learned patterns, resulting in two scores—namely, Fake score and True score, as shown in equation 4.5. Here, the p and q correspond to the probability of being fake and true, respectively, for a given news item.

$$\text{Fake Score} = \ln \frac{p}{1-p}, \quad \text{True Score} = \ln \frac{q}{1-q} \quad (4.5)$$

In addition, following the Emotion AWARE approach, the WELFake dataset, along with LLM embeddings, was used to construct two embedded lexicons, namely the fake news lexicon (α) and the true news lexicon (β). Three steps follow this construction. The first step was to extract the most frequent unigrams, bigrams, and trigrams for both the fake and true labelled corpora. Secondly, tag each term with its corresponding embedding by passing it through the LLM. And finally, to attach four frequency-based measures to each term. These scores are FakeDocumentScore (Eq. 4.6(a)), TrueDocumentScore (Eq. 4.6(b)), FakeOccurrenceScore (Eq. 4.6(c)), and TrueOccurrenceScore (Eq. 4.6(d)). The TrueDocumentScore measures the proportion of true news items in the dataset that contain a given word, while the FakeDocumentScore does the same for fake news items. The TrueOccurrenceScore calculates the ratio between the total frequency of a given word in true news items and the total number of words in the true corpus. Similarly, the FakeOccurrenceScore does this for the fake news corpus. These scores enhance the lexicons' ability to distinguish between fake and true news.

$$\begin{aligned}
\text{FakeDocScore}(w) &= \frac{\sum_{i=1}^N f(w, \text{Fake News Item}_i)}{N} & \text{TrueDocScore}(w) &= \frac{\sum_{i=1}^N f(w, \text{True News Item}_i)}{N} \\
\text{(a)} & & \text{(b)} & \\
\text{FakeOccScore}(w) &= \frac{\sum_{i=1}^N g(w, \text{Fake News Item}_i)}{\text{Total Number of Words in } \alpha} & \text{TrueOccScore}(w) &= \frac{\sum_{i=1}^N g(w, \text{True News Item}_i)}{\text{Total Number of Words in } \beta} \\
\text{(c)} & & \text{(d)} & \\
f(\text{word}, \text{document}) &= \begin{cases} 1 & \text{if word in document} \\ 0 & \text{if word not in document} \end{cases} & g(\text{word}, \text{document}) &= \text{Word frequency in document}
\end{aligned} \tag{4.6}$$

During the nearest neighbour detection process, the corpora α and β are merged to create a single lexicon where each term is tagged with an embedding and the four aforementioned scores as metadata. For a given news item, the system performs a semantic search to identify the 100 nearest terms that best explain the article. The four scores of these neighbours are then averaged to create the corresponding scores for the news item. These scores, along with the True and Fake Scores, form a feature vector with six values that is forwarded to the classifier. This approach aims to incorporate language features in various ways, enhancing the robustness of the fake news detection method.

Semantic Facet

The duty of the semantic pipeline is to extract the language semantic features from incoming news items using symbolic NLP techniques. Initially, preprocessing steps are performed, including the removal of non-ASCII characters, URLs, and punctuation. Lemmatization is applied as part of the feature extraction process. Lexical semantic features are derived from a comprehensive survey by Zhou and Zafarani (2020). In this study, we considered features that cover diverse aspects of news items, including complexity, diversity, quantity and uncertainty. Using Recursive Feature Elimination with Cross-Validation (Awad and Fraihat, 2023), we selected the top 14 features that best match the semantic requirements of the proposed pipeline. These selected features ensure a comprehensive analysis of the semantic content in the news items, contributing to a robust and effective fake news detection system. Table 4.4 provides an overview of semantic aspects and corresponding selected features.

After processing the input news items from all three pipelines, a voting classifier is used to consolidate the features. Before passing the data into the final classifier, IQR (Inter-Quartile Range) filtering was applied to remove outliers from the emotion and semantic features. These two sets of features were then normalized using Z-score normalization. The language features were not standardized, as they performed well without it. The voting classifier

Semantic Aspect	Description	Feature
Complexity	Measures readability and comprehension through word and sentence structure.	Average # words per sentence(1), Average # characters per word(2), Average # punctuations per sentence(3)
Diversity	Assesses the variety and repetition of vocabulary used.	Type-to-token ratio(4), Redundancy(5), Noun ratio(6), Verb ratio(7), Adverb ratio(8), Adjective ratio(9)
Quantitative	Simple counts of text elements like words and sentences.	# Words(10), # Sentences(11)
Uncertainty	Evaluates the author's confidence and certainty in the text.	# Question marks(12), # Modal verbs (e.g., "shall") (13), # Certainty terms (e.g., "never" and "always") (14)

Table 4.4 Description of Semantic Aspects and Corresponding Features

employed two separate Gaussian Naive Bayes models for classifying emotion and semantic features, and a K Nearest Neighbour classifier for language features. These classifiers were selected based on their accuracy on test sets which will be elaborated in the experiments section. The final results were determined by a hard voting mechanism, considering the decisions from the three models.

4.3.3 Experiments

The proposed methodology was evaluated using several experiments to ensure the various aspects to ensure the accuracy and generalization. First, each modelling facet was evaluated individually to assess the performance and effectiveness in each aspect of the framework. Various algorithms, including tree algorithms, boosting algorithms, and other classification approaches, were employed to identify the most effective models for the classification tasks. Additionally, an examination was carried out to evaluate the impact of the standardization process. This phase of the study involved comparing and contrasting the average generalization (cross-validation) accuracies of these models. Results of these experiments are tabulated in table 4.5.

Out of these experiments, Gaussian Naive Bayes classifier performed exceptionally well in both emotion and semantic categories, achieving accuracies of 0.58 and 0.53, respectively. Additionally, the K Nearest Neighbours classifier excelled in linguistic features, with an accuracy of 0.6. The ablation trials also indicated that standardization positively

Table 4.5 Performance Comparison of Classification Models Before and After Standardization

Model	Emotion		Semantic		Language	
	Before	After	Before	After	Before	After
Extra Tree	0.54	0.55	0.47	0.54	0.55	0.54
Decision Tree	0.51	0.52	0.48	0.52	0.57	0.53
Gaussian Naive Bayes	0.56	0.58	0.53	0.58	0.58	0.55
Random Forest	0.53	0.54	0.48	0.53	0.57	0.54
K Nearest Neighbours	0.53	0.53	0.48	0.53	0.60	0.56
Ada Boost	0.54	0.55	0.48	0.55	0.54	0.54
XGBoost	0.55	0.54	0.48	0.54	0.57	0.54

impacted emotion and semantic characteristics, enhancing accuracies to 0.584 and 0.58, respectively. Conversely, language features maintained consistent performance regardless of standardization.

Next, the effect of combining the features was investigated in order to evaluate if the combination of the multiple facets had any significant impact on improving the generalizability of fake news detection. The voting classifier created using the Gaussian Naive Bayes classifier for normalized emotion and semantic features and the K Nearest Neighbours classifier for unnormalized language features, which performed hard voting (got the majority of the votes as the prediction label), was observed to make the most accurate prediction. The usual evaluation strategies and metrics are not sufficient to evaluate the generalizability of an approach. Therefore, a relatively new metric that has been used to evaluate model generalizability in previous research, the Cross Validation Average Accuracy (CVAA), was used to evaluate the experiments.

The CVAA of a classifier model (M) across five different datasets (D1, D2, D3, D4, and D5) is calculated according to Equation 4.7.

$$CVAA(M) = \frac{\sum_{i=1}^5 CVAA(M, D_i)}{5} \quad (4.7)$$

Where,

$$CVAA(M, D_i) = \frac{\sum_{j=1}^5 A(M, D_j)}{4} \quad (4.8)$$

$$A(M, D_j) = \begin{cases} \text{Accuracy of Model M on Dataset } D_j & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \quad (4.9)$$

The proposed fake news detection pipeline was tested on five major benchmark datasets commonly used in fake news detection research. Table 4.6 provides an overview of these datasets.

Table 4.6 Datasets Used to Cross-validate the Proposed Fake-News Detection Approach

Dataset	Description
CONSTRAINT(Kolluri <i>et al.</i> , 2022)	Composed of nearly 11,000 tweets related to COVID-19, with a near equal split of fake and true tweets.
Fake News – Kaggle(Chitic, 2020)	Composed of fake and true news from mainstream media. It comprises 3164 fake news articles and 3171 true news articles.
FakeNewsNet(Shu <i>et al.</i> , 2020)	Short statements collected from PolitiFact and Gossipcop, which are two popular fact-checking platforms. This dataset contains 5755 fake statements and 17441 true statements.
LAIR(Wang, 2017)	Short statements collected from PolitiFact. It has data belonging to six different categories: true, mostly true, half true, mostly false, false, and pants on fire. The labels were converted to binary labels by taking the first three as true and the latter three as fake. Accordingly, this dataset has 5657 fake statements and 7134 true statements.
ISOT(Ahmed <i>et al.</i> , 2017)	Long texts collected from websites verified as legit and fake news spreaders by popular fact-checking engines. It has 21417 fake news articles and 23481 true news articles.

Table 4.7 presents the accuracies achieved by the voting classifier when trained and tested on five benchmark datasets, along with the Cross-Validation Average Accuracy (CVAA) for each combination. The table also highlights the cumulative CVAA value of 0.62.

Table 4.7 Accuracies Obtained using Voting Classifiers with Different Base Models

Dataset trained on	Dataset tested on					
	CONSTRAINT	FNN	ISOT	Kaggle	LIAR	CVAA
CONSTRAINT	0.70	0.59	0.85	0.62	0.54	0.65
FNN	0.53	0.75	0.66	0.57	0.56	0.58
ISOT	0.50	0.49	0.93	0.78	0.52	0.57
Kaggle	0.49	0.51	0.87	0.83	0.55	0.61
LIAR	0.61	0.67	0.84	0.61	0.56	0.68
Cumulative CVAA						0.62

With the above experiments it reveals that the combination of multiple facets in detecting fake news significantly outperforms individual approaches alone. Another important insight

from these experiments was the contribution of the datasets to the experiment outcomes. The datasets themselves emerged as a critical factor in determining accuracy, as models may not generalize well to unseen data even if they show high performance on the same split of train and test data in supervised methods. Extensive comparisons of various classifiers and training datasets underscore the importance of careful selection of supervised models and training sets. Additionally, it showed that integrating symbolic and connectionist approaches across multiple facets consistently enhances the robustness and generalizability of the methodology across diverse datasets.

4.4 Detecting Idioms in Natural Languages

4.4.1 Introduction

Idioms are multi-word expressions derived and established by convention and cultural experience. Although a limited subset of language, idioms are frequently used in formal and informal communication for creative, subtle, simple, or stimulating effects. Idioms are distinguished from other figures of speech because their meanings cannot be derived from their individual terms. However, the non-literal meaning of an idiom is familiar to native speakers. For instance, “*to break a leg before an interview*” means wishing someone good luck, a highly figurative connotation beyond the literal meaning. Although idioms enhance the expressiveness of language, an idiomatic expression’s contradiction of compositionality means these phrases present a unique challenge in Natural Language Processing/Understanding (NLP/U) settings, from simple techniques such as keyword extraction, topic extraction, sentiment analysis, information retrieval, to complex techniques such as text summarization, machine comprehension, machine translation, question answering, and emotion analysis (Salton, 2017; Salton *et al.*, 2014; Avramidis *et al.*, 2020; Pelosi, 2020). For instance, the example “*The movie’s final scene left me with a lump in my throat*” contains strong negativity, which literal sentiment classifiers often fail to capture.

Recent literature reports several methods for idiom detection and classification. One common approach is the lexicon-based method, which maintains a large set of idioms and uses string matching techniques to identify the idioms in the text (Klebanov *et al.*, 2013; Cacciari, 2014). These methods consider idioms to be frozen phrases. However, idioms can exhibit syntactic changes such as verb and tense alteration that do not impact their figurative meaning. This allows the same idiom to appear in multiple forms (Yusifova, 2013). In such cases, simple string matching fails, and in response, regular expressions (Feldman and Peng, 2013) and local grammar-based methods have been proposed (Williams *et al.*,

2015; Priyanka and Sinha, 2014). These two types of methods are primarily affected by the time and effort required of human experts to produce idiom dictionaries and local grammar. Furthermore, detecting diverse forms of idioms and new idiom discovery is infeasible because the classification is limited to the local grammar or regular expressions generated by the human expert.

While symbolic approaches alone cannot sufficiently address the challenge of figurative language detection, the learnability of connectionist models can be merged with them to effectively improve these systems. There are some attempts in the literature which showcase the usefulness of supervised methods in this context. Muzny and Zettlemoyer (2013) proposed a binary classification model based on Wiktionary to predict whether a given phrase is idiomatic or literal. Wang *et al.* (2020) presented an attention-based mechanism to choose the best matching idiom to be picked for the Chinese idiom cloze test. More recently, language models based on transformer architectures have been effective in metaphor detection (Chen *et al.*, 2020). Škvorc *et al.* (2020) used the contextual embeddings of BERT and ELMo (Embeddings from Language Models) (Peters *et al.*, 2018b) on deep neural networks to outperform existing approaches for detecting Slovenian idioms. Although these methods attempt to resolve these challenges based on purely supervised approaches, they do not locate the idioms or metaphors in given unstructured text to the token level, which can be useful in many contexts. Rather than labelling the entire input text with a single label, we can clearly position and differentiate the regular text from the figurative speech using this approach within the text itself. From this experiment, we combine symbolic tokenization to reach the token levels of unstructured text and then leverage the BERT language model to detect token-wise idiomatic expressions in natural languages.

4.4.2 Methodology

High-level architecture of the proposed methodology is illustrated in figure 4.7. This method combines the language learnability from small-scale LLM, BERT (Bidirectional Encoder Representations from Transformers) (Devlin *et al.*, 2018) with symbolic tokenization to do model the idiom detection problem as rule-guided classification task. BERT’s innovation lies in its use of masked language modelling for training, as opposed to traditional right-to-left or left-to-right language modelling. This enables BERT to effectively encode information from both directions at each transformer layer, enhancing its performance (Wu and Dredze, 2019). Specially in the use cases such as figurative language detection this learnability can utilize to detect the compositional variations in participating segments of the text. By combining the symbolic method of tokenization with BERT’s neural architecture, we achieve a robust token classification system. We utilize DistilBERT (Sanh *et al.*, 2019), a computationally

efficient variant of the BERT model, and employ the HuggingFace Transformers library for accessibility (Wolf *et al.*, 2020). We preprocessed four datasets for token classification by splitting sentences into words and generating word-wise labels. During label creation, we marked idiomatic tokens as 1 and non-idiomatic tokens as 0.

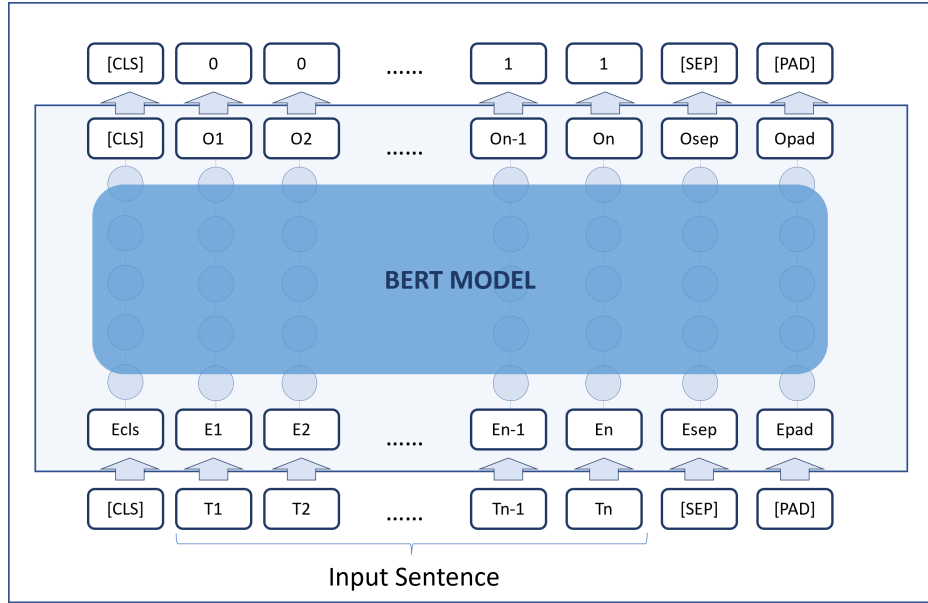


Fig. 4.7 Extended Pre-training for Token Classification with BERT

Idiomnet Dataset

Williams *et al.* (2015) have collected, annotated and verified the Idiomnet Dataset via crowdsourcing. This dataset consisted of 2525 sentences, of which 580 different idioms were included. Each idiom is presented across several sentences and in different forms of expression.

EPIE Dataset

The English Possible Idiomatic Expressions (EPIE) is a comparatively large dataset presented by Saxena and Paul (2020). This contains 25206 sentences along with lexical instances of 717 idiomatic expressions. Here, the extraction was done using several string processing steps using StringNet, other pattern-matching methods and some manual work. Similar to the Idiomnet dataset, here, each idiom is presented across several sentences. They have split the dataset into two classes. (1) Static idioms Dataset - idioms that do not undergo any lexical modifications, which contains 21891 sentences with 359 Static idioms. (2) Formal idioms

Dataset - idioms that occur in sentences with various lexical modifications, which contains 3135 sentences with 358 formal idioms.

Theidioms.com Dataset

This dataset was curated by us specifically for this study. We utilized the resources in theidioms.com (Idioms, 2024) which organizes idioms by topics. We collected 1106 unique idioms and corresponding example sentences. The main difference between this dataset and Idiomnet and EPIE is the one-to-one mapping between idioms and sentences. Table 4.8 summarises the four datasets.

Table 4.8 Idiom Dataset Statistics

Dataset Name	Sentences	Idioms	Sentences per Idiom
Idiomnet	2525	580	4.35
EPIE static	21891	359	60.67
EPIE formal	3135	358	8.75
theidioms.com	1106	1106	1.00

As the tokenizer is working as a word-piece tokenizer, a single word is further split into sub-tokens. Thus, token-label alignment was done by aligning the original token with the first-word piece and populating the rest with the non-idiomatic label 0. Sentence size was used as 40 where the padding and truncation were applied appropriately. Weight decay was set to 0.01, and warmup steps were set to 500. We selected the batch size as 16 for training and 5 for evaluation where used training epochs 3, 10 depending on the dataset.

0 0 0 0 0 0 0 0 0
 When we were returning from the picnic, it was
 1 1 1 1
 raining cats and dogs.

Fig. 4.8 Term-level Annotation for Identifying Parts of Idiomatic Expressions

Figure 4.8 depicts an instance of token classification. Here, non-idiomatic tokens are annotated with 0's, whereas idiomatic tokens are annotated with 1's. Generally, in a given text, there is a majority of non-idiomatic tokens. In an idiom, detection system, as well as correctly identifying the idiomatic tokens, it is also important that the model can recognize non-idiomatic tokens as not being a part of the idiom. Because they can have the same wording as idioms without idiomatic sense. Moreover, by modelling this as token classification, we can locate the idiom within the text rather than classify sentences for idiom existence.

4.4.3 Experiments

We conducted a series of experiments on all four datasets to evaluate the effectiveness of the proposed model. This includes four experiments focusing on the individual datasets, followed by three experiments that combine the datasets to investigate model robustness.

The experiment configurations are as follows:

1. Use the Idiomnet dataset. 80%, 20% train, and test split.
2. Use the EPIE Formal dataset. 80%, 20% train, and test split.
3. Use the EPIE static dataset. 80%, 20% train, and test split.
4. Merged EPIE Formal and Static datasets. 80%, 20% train, and test split.
5. Use the theidioms.com dataset. 80%, 20% train, and test split.
6. Using the Idiomnet dataset as the training set and theidioms.com dataset as the test set.
7. Using the EPIE Formal dataset as the training set and theidioms.com dataset as the test set.

The evaluation results for these experiments are tabulated in table 4.9. Here, several metrics were used, including accuracy, f1 Score, precision, and recall. Metrics other than accuracy are represented as pairs of values for non-idiomatic term prediction and idiomatic term prediction, respectively. The proposed idiom detection model consistently generates an accuracy of more than 94%. The highest accuracy is recorded for the EPIE - Static dataset, 99.4%.

Table 4.9 Idiomatic token classification results for the experiments

Experiment	Accuracy	F1 Score	Precision	Recall
1	0.980	[0.989 0.904]	[0.992 0.882]	[0.986 0.926]
2	0.972	[0.985 0.815]	[0.978 0.888]	[0.991 0.754]
3	0.997	[0.998 0.984]	[0.999 0.979]	[0.998 0.989]
4	0.994	[0.996 0.959]	[0.996 0.962]	[0.997 0.957]
5	0.960	[0.978 0.756]	[0.974 0.799]	[0.983 0.717]
6	0.948	[0.972 0.657]	[0.960 0.773]	[0.984 0.571]
7	0.944	[0.970 0.633]	[0.958 0.752]	[0.983 0.548]

Experiments 1 to 4 are shuffled and split into 80% training and 20% testing, respectively. Thus, the test set contains unseen idioms that were not included in the training set. Experiment 5 only uses *thedioms.com* dataset. When compared with other datasets, this dataset contains

a one-to-one mapping with the sentences and idioms. When we split the dataset into training and testing partitions, the test set contained entirely new idioms that were not included in the training set. In this experiment, the accuracy was 96% where all non-idiomatic matrices exceeded 97% and all idiomatic matrices exceeded 71%. This further validates that DistilBERT has learned the underlined figurative nature of the idioms and the ability to identify entirely new idioms unseen during training. In experiments 6 and 7, we used a combination of datasets, with nuances like different sentence lengths, and this, too, led to an accuracy of more than 94%.

Furthermore, this idiom detection model highlights the significance of contextual embeddings compared to first-generation neural embeddings such as Bag of Words (BOW)(Zhang *et al.*, 2010), Word2vec(Mikolov *et al.*, 2013b) in detecting figurative components in a text corpus. Contextual embeddings help to retain the phenomenon of polysemy, which captures different senses of a given word based on the context (e.g., paper as a material, as a newspaper, as a scientific work, and as an exam). Here, the scope of the context is typically a sentence and a word will get a unique vector for each figurative or literal meaning it takes(Škvorc *et al.*, 2020). This capability intensifies the uniqueness of the idiomatic segments in a given sentence, even if it is unseen by the model. As shown in table 4.10, we compared our language model-based token classification method against existing popular embedding generation methods for experiment 5. We used BOW and Word2vec as embedding generation methods. Naïve Bayes was used as the classifier. This further elaborates on the significance of contextual embedding in figurative language detection problems.

Table 4.10 Comparing the Proposed Method with Non-contextual Embedding-based Classifiers

Experiment	BOW embeddings with Naïve Bayes Classifier	Word2vec embeddings with Naïve Bayes Classifier	Our approach
Accuracy	0.439	0.549	0.960
F1 Score	[0.512 0.339]	[0.662 0.323]	[0.978 0.756]
Precision	[0.876 0.216]	[0.849 0.224]	[0.974 0.799]
Recall	[0.362 0.775]	[0.542 0.579]	[0.983 0.717]
Confusion Matrix	[[1608, 2829], [227, 784]]	[[2414, 2036], [427, 588]]	[[7951, 136], [213, 540]]

4.5 Extremism Detection in Natural Languages

4.5.1 Introduction

Extremism can be explained as phenomenon which is characterized by the advocacy of extreme measures or views. It often manifests in various spheres of society, including politics, religion, and social movements(Moghaddam, 2005). Similar to the fake news, the spread of extremism is another significant challenge in the social media age. Especially with the natural languages, extremism is harmful as it both influences and is influenced by the ways individuals and groups communicate. The linguistic aspect of extremism can spread over various directions such as usage of language to advocate radical ideologies, recruit followers, and incite violence or discrimination against perceived adversaries. Thus, detection and elimination of extremism in natural languages is a crucial requirement for the well-being of the society as well as digital social environments. Defining the extremism itself is a challenging task due to the diversity of this phenomenon. Thus, it is usually recognized through several distinct characteristics. These characteristics are crucial for recognizing and understanding extremist behaviors and ideologies in natural languages as well.(Bötticher, 2017; Neumann, 2013).

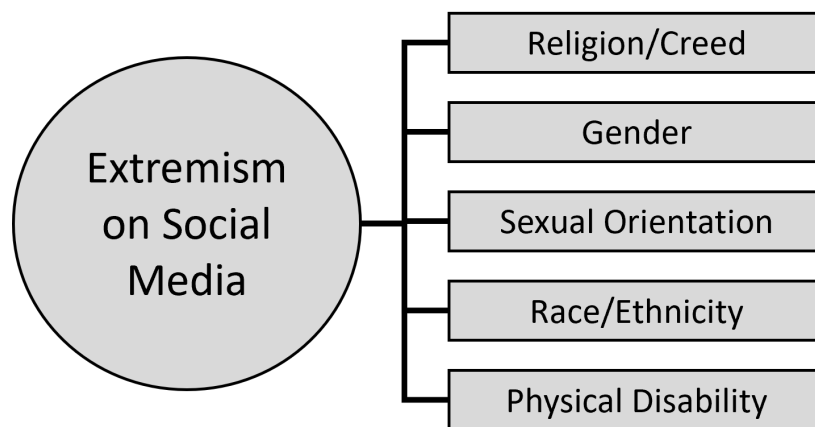


Fig. 4.9 Common Facets in Extremism

One of the distinguishable features of extremism is that these ideologies are usually centered around a specific set of themes such as religion, gender, sexual orientation, race, ethnicity, and physical disability. For instance, the majority of extremist movements are based on religious ideologies where they interpret religious texts or beliefs in an extremist manner that justifies their actions and goals(Ishaku *et al.*, 2021; Frazer, 2023). Gender and sexual orientation are two other areas of extremism. Gender extremists promote misogynistic views or advocating for extreme gender roles and identities(O’Hanlon *et al.*, 2024). Homophobic

and transphobic extremism targets individuals based on their sexual orientation or gender identity, often promoting discriminatory and violent actions. Furthermore, ideologies based on racial superiority or inferiority are another theme of extremism often leading to acts of racism, xenophobia, and ethnic cleansing (Matamoros-Fernández and Farkas, 2021). Apart from these themes, extremism can also target individuals with physical disabilities, promoting discriminatory practices against them (Faizan *et al.*, 2023). In addition to these themes, another characteristic of extremism is its nature of going against established social norms and practices. Extremists typically reject existing social norms, and promote the alternative systems that align with their views. As a result of that extremists usually express intense hatred towards those who hold opposing views or who belong to different social, religious, or cultural groups. (Gule, 2019). So thematic behavior and having strong negative emotions can be leveraged to identify the violent extremism in natural languages.

Extremists use social media as a powerful tool to spread their ideologies to wider as well as diverse audiences globally (Bright, 2018). This widespread reach facilitates the recruitment of new followers, particularly targeting vulnerable individuals who may be more susceptible to radicalization (Awan, 2017). Furthermore, social media gatherings such as pages and communities are being used for coordination and planning of extremist activities with the ease of communication. In addition, extremists use these platforms for fundraising, soliciting donations to support their activities through anonymous channels (Ganesh and Bright, 2020). Along with the ideologies, they also leverage modern tools and techniques in these platforms to share hatred against the opposition. With such activities extremists manipulate public opinion by spreading misinformation within societies to fulfil their necessities (Ganesh and Bright, 2020).

When considering the literature, there are only limited studies that have investigated the detection of extremism in natural languages or other modalities. These handful of studies are also focused on narrow instances of extremism, such as terrorism by groups like ISIS (Awan, 2017; Alfifi *et al.*, 2018), white supremacy (Gallaher, 2021; De Gibert *et al.*, 2018), and other similar cases. Additionally, these methodologies rely heavily on keywords and topic modeling techniques, which are specific to the underlying use cases, which limits the capability to generalize. Furthermore, a majority of these approaches are utilizing classification techniques, typically binary classification which labels a given text into positive or negative classes. This limits informativeness as well as the explainability of extremism detection. Thus, there is a high requirement for a generalized approach to extremism detection that considers various dimensions of extremism, integrates topics and scores, and incorporates explainability features. Moreover, there is a need for approaches that go beyond simple classification to construct explainable profiles of extremist behavior. However, this field faces

significant challenges, including a lack of prior studies and generalizable methods, as well as a scarcity of annotated datasets, making validation difficult. Addressing these challenges is crucial for developing robust and comprehensive extremism detection methods.

4.5.2 Methodology

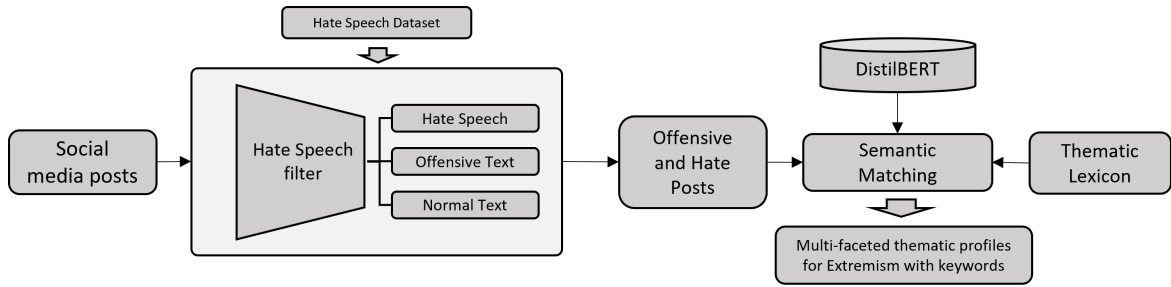


Fig. 4.10 Extremism Detection in Natural Languages

Figure 4.10 shows the high-level architecture of the proposed neuro-symbolic approach for extremism detection. This framework combines both connectionist and symbolic NLP techniques to construct multifaceted extremism profiles. As discussed in the introduction, a major challenge in extremism detection is the lack of resources in the literature. Especially, there are no general datasets for extremism detection to implement supervised approaches. Given these limitations, we leveraged hate speech detection resources that closely align with the emotionalities as well as the themes of the extremism detection problem. In this implementation, we utilized the HateXplain dataset by Mathew *et al.* (2020) as the foundation to build our approach. Dataset details are provided in the table 4.11. Here "Undecided" refers to the cases where all three annotators chose a different class.

	Twitter	Gab	Total
Hateful	708	5,227	5,935
Offensive	2,328	3,152	5,480
Normal	5,770	2,044	7,814
Undecided	249	670	919
Total	9,055	11,093	20,148

Table 4.11 HateXplain Dataset Statistics(Mathew *et al.*, 2020)

HateXplain is a comprehensive dataset that covers multiple thematic aspects of the hate speech detection problem. According to the authors, it is the first benchmark dataset that considers multi-facets of detecting hate speech in natural languages(Mathew *et al.*, 2020). The dataset construction was based on two steps, firstly to scrape social media platforms to

collect the data points and then use crowd-sourcing for the annotation process. This dataset has collected data-points from two popular social media platforms, namely Twitter and Gab, and leveraged Amazon Mechanical Turk (MTurk) to validate as well as annotate these data points with the help of crowd workers. In this dataset, each post is classified as hate, offensive, or normal speech. Additionally, annotators identify the target communities mentioned in the post and highlight parts of the text that justify their classification decisions. So ultimately, this comprehensive annotation includes the 3-class classification (hate, offensive, or normal), the community targeted by hate speech or offensive speech, and the rationales which are the specific portions of the post that justify the labeling decision(Mathew *et al.*, 2020).

As illustrated in figure 4.10 proposed framework employs a two stage pipeline. The stages are hate speech filtering and thematic profile construction. In the first stage, we utilized the HateXplain dataset along with the hate, offensive, or normal annotation labels to build a classifier for categorizing social media posts into these three categories. To achieve this, we used the DistilBERT(Sanh *et al.*, 2019) base model with a sequence classification head from the Transformers(Wolf *et al.*, 2019) library. For fine-tuning, we set the batch size to 64 and the learning rate to $2e-5$. Hate speech and offensive posts filtered during the first stage of the pipeline are then forwarded for the extraction of extremist themes. For this purpose, we employ a lexicon-based embedding matching approach, similar to the Emotion AWARE algorithm(Gamage *et al.*, 2024), which will be discussed in detail in Chapter 5. This thematic profile construction process utilizes a lexicon containing keywords corresponding to various extremism themes, each tagged with its respective embeddings. To construct this thematic lexicon, we leveraged the rationales from the posts along with the selected target victim category annotations from the HateXplain. Each annotation was replaced with target groups such as race, religion, gender, sexual orientation and Miscellaneous, as explained in the table 4.12. These keywords were then processed through the DistilBERT(Sanh *et al.*, 2019) model to generate embedding representations that reflect their meanings.

Target groups	Categories
Race	African, Arabs, Asians, Caucasian, Hispanic
Religion	Buddhism, Christian, Hindu, Islam, Jewish
Gender	Men, Women
Sexual Orientation	Heterosexual, Gay
Miscellaneous	Indigenous, Refugee/Immigrant, None, Others

Table 4.12 Target Groups Considered for the Annotation

Once the thematic lexicon was constructed, it was used for semantic search to build thematic profiles. When a post is presented, it is first converted into an embedding using the

same embedding model. This embedding is then compared with the thematic words in the lexicon, and the top 50 thematic expressions closest to the post embedding are identified. After separating these neighbours, their thematic labels are accumulated. The frequency of keywords per target group is then normalized by the total neighbourhood threshold size to generate the scores. For explainability purposes, keywords are also extracted by comparing the embeddings of the post with each of the substituting tokens' embeddings to identify the top keywords.

4.5.3 Experiments

Due to the lack of benchmark datasets for standard evaluations, we tested the proposed neuro-symbolic extremism detection approach on two case studies: the first involved analyzing extremism in COVID-19 vaccination-related tweets, and the second focused on extremism in tweets related to the 2020 US election.

Analyzing COVID vaccination-related tweets

The usage of social media platforms has massively increased during the COVID-19 pandemic resulting both positive and negative impacts. While it played a significant role in keeping people informed and connected, there were several negative impacts such as spread of misinformation and other harmful ideologies. Extremist ideologies and activities were also reflected in various occasions during the pandemic period. Particularly during the COVID-19 vaccination roll-out, social media became a hotbed. There were many extremist views highlighted in regard of vaccination with different themes such as sexual orientation, gender, religion/creed, and race/ethnicity.

LGBTQ+ individuals were one of the targets by extremist groups who leveraged misinformation and conspiracy theories, often falsely attributing the virus's spread or vaccine side effects to this community. Furthermore women and gender minorities faced increased online harassment and blame for vaccination policies(Adamson *et al.*, 2022). Additionally there were rise of religious extremism during the vaccination roll-out where some groups claimed the vaccine contradicted their beliefs or was part of a broader conspiracy against their faith, leading to the spread of false information.(Zarieczna *et al.*, 2023). Racial and ethnic minorities were often targeted during the entire period of pandemic subjected claims that certain races were more responsible for the spread of the virus or that the vaccine was a tool for racial control or experimentation.

Detecting extremism in social media in general is crucial for public health, as helps maintain social cohesion by mitigating divisions within society and protecting targeted groups

from harassment and discrimination. Social media platforms serve as a mirror to societal issues, amplifying both positive and negative sentiments (Lenadora *et al.*, 2020). During this experiment, we used the COVID Vaccine Tweets dataset from Kaggle. This dataset was built by extracting posts from Twitter (now known as X) based on the #CovidVaccine hashtag, starting from August 1, 2020, and updated daily. Here, we used version 35 of this dataset, which included tweets up to August 2021. Due to the higher volume and better focus on vaccinations, we used a subset of the data from January to August 2021, containing 261,081 tweets regarding COVID-19 vaccination. (Suresh, 2021).

Month	Total Tweets	Hate Speech/Offensive Tweets	Percentage (%)
January	61467	1166	1.897
February	56335	977	1.734
March	28946	512	1.769
April	23509	435	1.850
May	11887	197	1.657
June	10139	146	1.440
July	35466	907	2.557
August	33332	885	2.655
Total	261081	5225	2.001

Table 4.13 Monthly Statistics of COVID-19 Vaccination Tweets

Table 4.13 provides a summary of the statistics which shows monthly breakdown of the total tweets as well as the number of hate speech/offensive tweets identified in the first stage of the pipeline. The count of hate speech/offensive tweets ranges from 1.4% to 2.6% per month, with a cumulative total of 5,225 tweets flagged as hate speech/offensive compared to the total vaccination related tweets. These flagged tweets were then analyzed to identify extremist themes, with a focus on categories such as religion/creed, gender, sexual orientation, and race/ethnicity.

Thematic profile construction generates a multi-faceted extremism theme along with a corresponding score for each facet. Additionally, it emits extremist keywords for explainability. We have taken the extremism theme with the highest label to annotate the posts for ease of analysis. In our analysis, we observed that the majority of top extremism labels were directed towards race and ethnicity, which accounted for over 63% of the identified hate speech/offensive tweets. This was followed by extremist content with top themes in religion/creed, gender, and sexual orientation, respectively. Figure 4.11 illustrates the distribution of these predominant extremist themes, highlighting the significant focus on race and ethnicity. Additionally, Table 4.14 presents sample tweets where our algorithm extracts the dominant theme from each thematic category. This detailed analysis underscores the prevalence of targeted hate speech across different social dimensions during the COVID-19

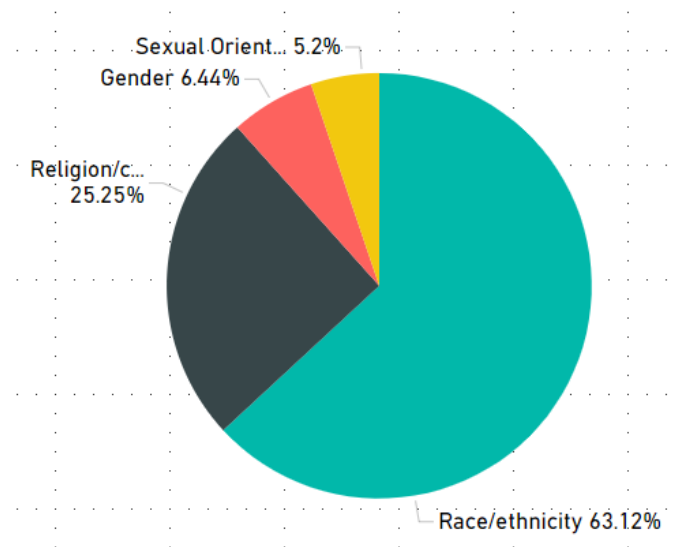


Fig. 4.11 Distribution of Top Extremist Themes in COVID-19 Vaccination Tweets

vaccination period, emphasizing the necessity for robust detection and mitigation strategies to combat online extremism and promote a safer social media environment. In the table, Offensive words are censored.

Table 4.14 Example Tweets Showcasing Various Top Extremism Themes in COVID-19 Vaccinations

Category	Tweets
Race/ Ethnicity	Not another fuXXing word about American citizens getting the #COVIDvaccine. JoeBiden and his administrations are #Traitors using #ILLEGALimmigrants in every way possible.
	If you're black you can't go out! Africans in China face racism...TALK ABOUT HYPOCRISY.. THE #Chinese UNLEASHED THE KUNG FLU #COVID19 #coronavirus ON THE WORLD BY EATING INSECTS BATS AND NOW R SCAPEGOATING THE AFRICANS
	Worse, antivaxxers are trying to make it about race by insinuating there's something about Blacks that make them suXX
	What a cover up again by these #Global murderers (#Chinese), Wow.#WuhanVirus #ChinaVirus
	CHINA IS A THREAT TO THE ENTIRE WORLD. #CCP china & Its Govt have Murdered Millions of people because of #COVID19 #CovidVaccine We will NEVER again trust this dangerous Chinese govt. #Wuhan
Religion/ Creed	FuXXing jews AGAIN! 7000 gather in Brooklyn last week, 400 here in London, do they think they are fuXXing immune!

Category	Tweets
	AXX why will #Pakistan wage war. They won't dare to do that unless some namak haram like ... decide to be traitor. What a stupid statement
	Boko Haram pending, covid scam is there, bandits, herdsman killings as well. Rubbish!
	People that refuses the vaccine because of "religious beliefs" are the absolute worst scum. And if they work with others, they should be fired for putting other in danger.
	Palestine burns flags of #USA, wishes death to America, danced and celebrated 9/11 attack almost 20 years ago and what America does? Donates 500,000 doses of #covidvaccine. It doesn't matter, they will continue to hate USA, Jews, Israel.
Gender	WHO The_Great_Wose #women just like #Children are full of shXX
	If you're gonna trust them to fix your wilXX you're dumb as fuXX.
	Unbelievable listening to that dolt jessphillips
	You gotta be some kind of CUXX to start charging for online vaccine appointments
	Listen here, you backwards non compliance loving cuXXs: You will be forcefully vaccinated.
Sexual Orientation	Dr. Fauci has doubts as to whether or not Russia's vaccine will work. Dr. Fauci also killed more fXggots than AIDS
	#ThisMorning Why don't you cover something more relevant than lesbian rabbits ?? Covid19 Vaccine reactions rise
	The unluckiest people in the world.... would be a black trans, lesbian who has disabilities who is doesn't believe in vaccines.
	If we say that a newly discovered COVID side effect is that it makes people gay, we will get herd immunity in no time.
	Don't go near those who have had the COVID vaccine. They have become homosexuals.

Analysing Extremism During US Election 2020

Similar to extremism analysis during COVID-19, we did another social media ablation study on US election 2020. The 2020 United States presidential election was a highly consequential and closely watched event, characterized by unprecedented voter turnout, significant political polarization, and also with the complications of the COVID-19 pandemic. The main candidates were the incumbent President Donald J. Trump from the Republican Party and former President Joseph R. Biden from the Democratic Party. During this period there was a significant surge of social media usage in US as well as considerable increase in extremist behaviors, both online and offline. X formerly Twitter as one of the most popular social media platform, played a big role in shaping public opinion, disseminating information,

and often misinformation. Deep political divisions, societal unrest, and the widespread dissemination of conspiracy theories were the campaign's defining features, which increased tensions across the country. Thus, considering these facts with this experiment, we analyzed the extremism spread across various themes on Twitter related to US 2020 election leveraging the proposed algorithm.

Tag	# Tweets	Hate/Offensive Text	Percentage (%)
Trump	970919	61964	6.382
Biden	776886	24759	3.187

Table 4.15 Hate/Offensive Text Analysis for US Election 2020 Tweets

As the dataset we used a Kaggle dataset containing 1.72 million tweets related to the US Election 2020(Manchunhui, 2020). These tweets were collected using combination of Twitter API statuses_lookup and snsrape, focusing on consisting keywords related to the election. This covers the period starting from October 15, 2020, up to November 4, 2020. Additionally, tweets were categorized based on the two main candidates, Donald J. Trump and Joseph R. Biden.

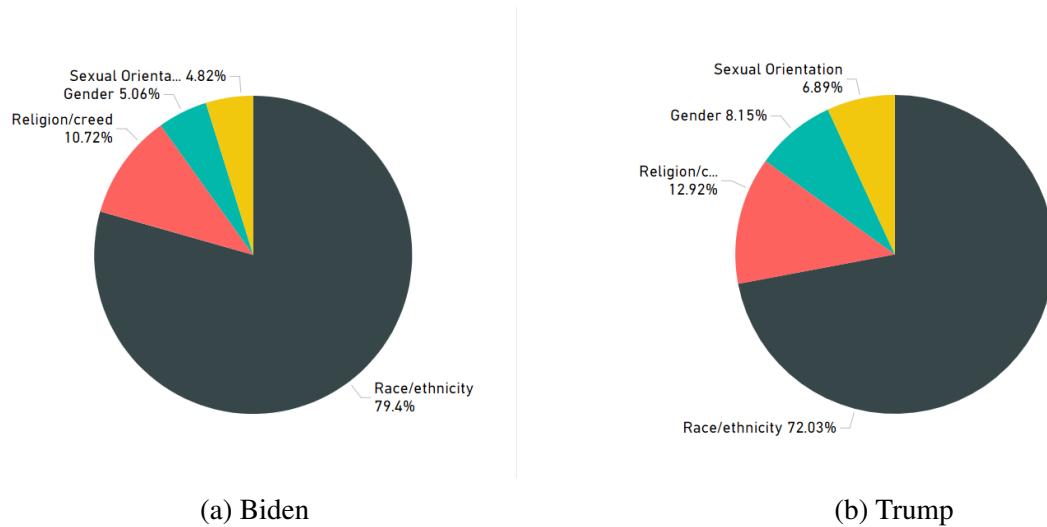


Fig. 4.12 Distribution of Top Extremist Themes in US Election 2020

Table 4.15 summarizes the statistics of the US election dataset. The total number of tweets mentioning Trump was approximately 200,000 higher than those mentioning Biden. The percentage of hate or offensive tweets was 6.38% for Trump and 3.18% for Biden, indicating that the proportion of hate/offensive tweets for Trump was more than twice that for Biden. When comparing this analysis to the study on COVID-19 vaccinations, these percentages are relatively higher. Figure 4.12 illustrates the distribution of dominant themes

in hate and offensive tweets of Trump and Biden classes. Consistent with the analysis of COVID-19 vaccination tweets, the majority of these tweets pertained to race and ethnicity, followed by themes related to religion, gender, and sexual orientation. A comparison between the two groups reveals that the Biden cluster had a relatively higher percentage (79.4%) of tweets themed around race and ethnicity compared to the Trump cluster (72.03%). Similar to the COVID-19 study, table 4.16 showcases some instances with top labels representing each of the thematic categories. Similarly offensive words are censored in the table.

Table 4.16 Example Tweets Showcasing Various Top Extremism Themes in US Election 2020

Extremist Theme	Trump	Biden
Race/Ethnicity	"DonaldTrump is THE worst racist, WhiteSupremacist EVER!!! Trump, the 'Racist,' Won More Minority Votes Than Any GOP Candidate Since 1960"	"BlackLivesMatter Listen how he talk about Blacks are poor,dumb & have not been socialized. Do you mean socialized like a dog? Biden is clearly a RACIST & says lock em up"
	JoeBiden Joe Biden is full of shXX. "You ain't Black" You can't go into 7-11 unless you have a Indian accent. C17Joe Biden is the racist and dividing the USA. Trump	"Least racist... Immigrants have low IQ, India is filthy. Debates2020 2020Election Biden BidenHarris2020"
Religion/Creed	"Europe is lost. IslamicState France stupid, stupid people letting your country be over run. Trump stopped it here and will keep us a Judeo-Christian nation!"	If you are Jewish, and you vote For Biden Democrats. You're stupid. They are screwing you. This is pure. HITLER-ESQUE like in NYC.
	No lie.. oldwhitemen in KatytX are voting trump because Obama turned America into an Islamic state. shitoldpeoplesay	"now were watching HILLBILLIES play GRAB-AXX w/ IMPERIAL WIZARD CARTOON CHARACTER the STUPID non-Christian, anti-life, TRUMP ZOMBIES R SAUCED on the CULT JUICE"
Gender	"trump at least paid high class whore Stormy 130,000.HepaidAmerica750."	See you cuXX. could put Kermit the fuXXing frog. in charge who gives a fuXXing shXX...stupid aXX country bitXX
	"First whoXX has no right to criticize anyone! Be worst first lady ever!"	Also, all you red X cuXXs can go back to your caves in dark times. your time is over, go get fuXXed GoodbyeTrump

Extremist Theme	Trump	Biden
Sexual Orientation	"Homosexuality is sin!" "Islam is a lie!"	if you vote trump, just realize that you are supporting a racist, homophobic, transphobic, underaged of girls predator piece of shXX who only cares about rich, white people. for a better future, vote biden! come on americans
	Biden2020 fuXX the faggot aXX trump supporters.	Does anyone actually have the leaked chats of people wanting to kill black and LGBT?

4.6 Chapter Summary

This chapter comprehensively discussed empirical evaluations in several natural language aspects utilising the proposed neuro-symbolic framework for natural language comprehension. It mainly covered two sections: firstly, the experiments enhancing word representations, and secondly, semantic intelligence systems constructed with hybrid architectures. These experiments showcase the potential of neuro-symbolic methods with features such as explainability, generalizability and controllability. In the following chapters, 5 and 6, I present the more sophisticated applications in the areas of emotion elicitation and natural language interactions.

Chapter 5

Neuro-symbolic Architectures Emotion Elicitation

5.1 Chapter Introduction

Emotions are fundamental to human behaviour. How we feel, individually and collectively, determines how humanity evolves and advances into our shared future. The rapid digitalisation of our personal, social and professional lives means we are frequently using digital media to express, understand and respond to emotions. Consequently, with the advancements in NLP technologies, it is crucial to integrate emotional understanding into language comprehension systems to create more effective and empathetic communication tools. Finding different emotional colours in natural languages helps as foundational components in applications of various domains. For example, in healthcare, analyzing the emotions in patient notes or telephone conversations can result in valuable, actionable insights. Similarly, in the business sector, emotions are crucial for interpreting feedback from customers and stakeholders for the improved quality of services. These examples highlight the broad and diverse applications of emotional analysis across different fields.

Although recent developments in Artificial Intelligence (AI) can analyse sentiment and detect emotions, they are ineffective at comprehending the complexity and ambiguity of affect and emotion expressed in natural human language. Most existing emotion detection systems rely on data-driven connectionist mechanisms, which pose challenges such as lack of explainability, limited adaptability, and increased complexity—issues discussed in Chapter 2 in the broader context of language comprehension. This chapter of the thesis introduces "Emotion AWARE," a neuro-symbolic approach to overcome these challenges. Emotion AWARE is an innovative AI framework that enables adaptable, robust, and explainable

detection of multi-granular assemblies of emotions. This framework integrates symbolic keyword-based mechanisms with small-scale encoder-only LLMs to generate multi-granular emotion assemblies covering two, eight, and fourteen emotions. It excels at handling ambiguous emotional expressions in conversation, adapts to domain-specific emotion semantics, and offers explainable outputs through constituent terms and intensity levels. Additionally, this chapter presents two real-world use cases of Emotion AWARE in the healthcare and financial sectors.

5.2 Emotion AWARE: An Artificial Intelligence Framework for Adaptable, Robust, Explainable, and Multi-Granular Emotion Analysis

The rapid digitization of society has empowered knowledge-focused human activities and communication to transpire on hyper-connected digital platforms. This spectrum of intrapersonal, interpersonal, and group activities has led to the generation and management of high volumes of big social data that represents patterns of behaviour of individuals and organizations and how they leverage insights drawn from that information for further engagement and collaborative activities (Olshannikova *et al.*, 2017). Expressions of emotion are encapsulated in these digital platforms, which is highly useful towards accurately modelling human behaviour (K. *et al.*, 2022). The persistence of this textual digital record enables computational approaches to process, analyse and synthesise emotional expressions.

Computational approaches for emotion detection have been classified using several schemes in existing literature. Acheampong *et al.* (2020) proposed three categories: rule-based, machine learning and hybrid methods. Alswaidan and Menai (2020) proposed a scheme of five categories: keyword-based, rule-based, classical learning, deep learning and hybrid. In reviewing these schemes, we have summarised them into three technical categories: 1) heuristics (which includes keywords, rule-based, probabilistic and statistical), 2) Artificial Intelligence (AI) (consisting of classical learning, machine reasoning and deep learning) and 3) hybrids of the two. In essence, the first category represents symbolic approaches, the second corresponds to connectionist methods, and the third combines both approaches. Despite the maturity of classification schemes and the prevalence of approaches across the three classes, the complexity and ambiguity of emotional expressions on digital platforms remain inadequately addressed. We have identified four key capabilities that challenge these issues: 1) output granularity of emotion detection, 2) domain specificity, 3) adaptability, and 4) explainability.

The Emotion AWARE framework addresses these challenges by integrating heuristics and connectionist techniques into a hetero-hierarchical structure. This framework takes natural language text expressions as input and generates a multi-granular assembly of emotions, complete with corresponding intensity values. These assemblies can be produced at three levels of granularity: two, eight, and fourteen emotions. The framework’s adaptability lies in its hetero-hierarchical structure, which can be revised and tailored to reflect specific domains or topics of interest. It is robust in detecting implied emotional expressions by analyzing the context of surrounding terms and scaling intensity values based on factors such as negations, intensifiers, and inhibitors. Additionally, the framework is explainable, as it identifies specific terms and phrases associated with each emotional expression, enabling the profiling and comparison of multiple emotion assemblies.

5.2.1 Background

We conducted a systematic literature review to analyze existing computational approaches of emotion detection, focusing on recent advancements. Using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, as illustrated in figure 5.1, we identified 83 articles that met our selection criteria. These studies were then evaluated based on four key capabilities: 1) the granularity of emotion detection output, 2) domain specificity, 3) adaptability, and 4) explainability. Our analysis highlights significant findings across the three main categories of emotion detection methods: Symbolic, connectionist, and hybrid approaches.

Symbolic approaches include keyword recognition, rule-based logical/grammatical affinities, and statistical and probabilistic methods. These methods are grounded in emotional lexicons, corpora and dictionaries that represent prior knowledge of how emotion is expressed in that domain or discipline. The emotion lexicon is typically a list of synonyms and related words used for each emotion category, where each word may also be assigned a fixed intensity value. Besides a list, the lexicon can also be organised hierarchically in a tree structure or interlinked as a graph or map structure. Several emotion lexicons reported in the literature are Plutchik’s emotional terms(Plutchik, 2001), the WordNet-Affect(Strapparava and Valitutti, 2004), EmoSenticNet(Poria *et al.*, 2014), DepecheMood(Staiano and Guerini, 2014), Senti-WordNet dictionaries(Esuli and Sebastiani, 2006). Keyword recognition methods(Adikari and Alahakoon, 2021; Adikari *et al.*, 2021a) rely on locating keywords representing emotions in a given text and assigning an emotion label based on these keyword counts and other statistics. These methods can be used for explicit emotion detection. For example, “*their arrival made me happy*” explicitly expresses the emotion happiness/joy with the keyword “happy”.

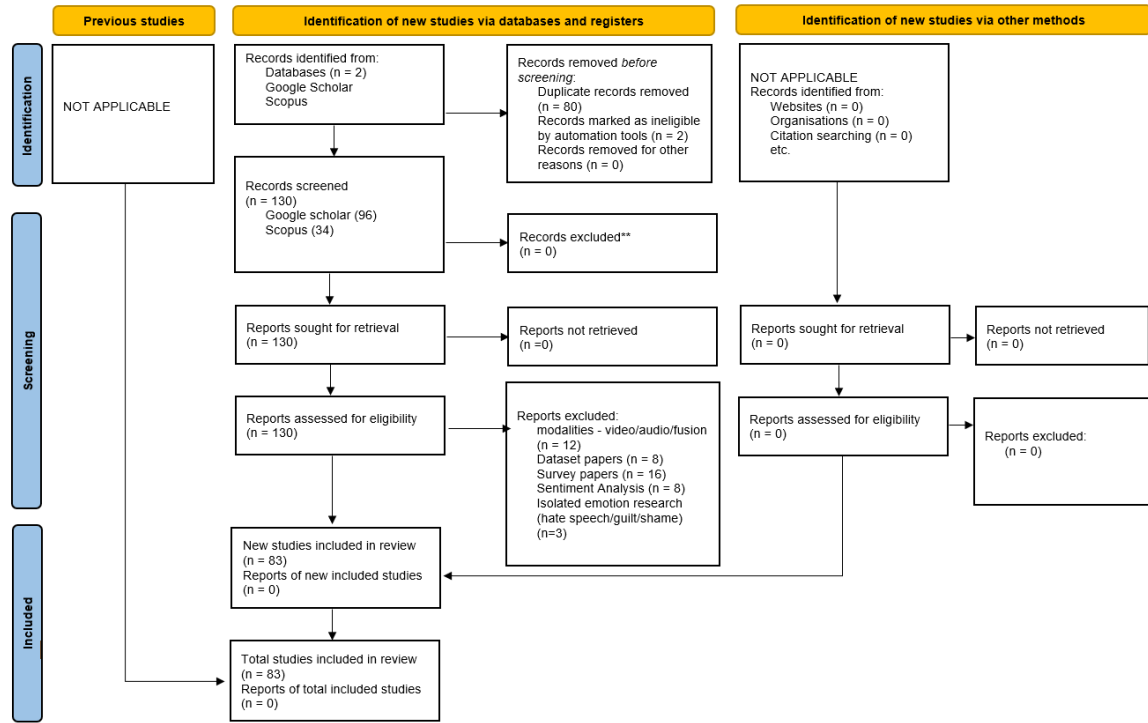


Fig. 5.1 PRISMA 2020 Flow Diagram for Updated Systematic Reviews Which Included Searches of Databases, Registers and Other Sources

However, emotions are often not explicitly mentioned and can be negated or modified to give different or opposing interpretations than a keyword search method would suggest. In such cases, more advanced heuristics are required. Rule-based approaches incorporate text processing methods such as tokenization, part-of-speech tagging, and dependency parsing along with corpora and lexicons to find the most effective rule sets for emotion detection (Udochukwu and He, 2015; Seal *et al.*, 2020). Several other approaches use lexical affinity with the support of lexicons to capture contextual and semantic relatedness to generate probabilistic values for each emotion category (Agrawal and An, 2012). Furthermore, some approaches utilize dimensionality reduction and categorical feature extraction methods such as Latent semantic analysis (LSA) (Landauer and Dumais, 1997), Probabilistic LSA (Hofmann, 2013) for improved emotion detection (Kim *et al.*, 2010). The use of lexicons enables domain adaptation in emotion detection, as lexicons can be easily extended or altered to suit the target domain. Furthermore, these methods can be extended for emotion intensity calculation, negation and modifier detection, as they can locate the keywords and evaluate the corresponding neighbourhood. However, a major drawback of all heuristic methods is that emotion expressions not specified in the lexicon and implied or ambiguous are not detected. Due to these reasons, methods purely based on symbolic approaches or heuristics are not

comparable to the benchmark performance of the connectionist methods(Acheampong *et al.*, 2020).

AI-based connectionist methods can be subdivided into two: conventional supervised learning methods situated in annotated datasets and contemporary transfer learning methods that leverage pre-training contextual language models. Conventional methods require large, labelled datasets where a human expert typically pre-assigns each sentence, paragraph, or segment in the corpus to an emotion category (or label). This annotated dataset is used to train a multiclass classification model using supervised learning algorithms. Emotion classification and intensity calculation using XGBoost(Aziz and Dimililer, 2021; Winata *et al.*, 2019), Support Vector Machines (SVM)(Nida *et al.*, 2019; Tzacheva *et al.*, 2020), Naïve Bayes (NB)(Suhasini and Srinivasu, 2020; Jain *et al.*, 2017), k-Nearest Neighbor (kNN)(Jain *et al.*, 2017) and Decision Trees(Ghanbari-Adivi and Mosleh, 2019; Hasan *et al.*, 2019) are some prominent techniques reported in related literature. More recently, deep learning algorithms such as Long Short Term Memory (LSTM) networks(Chatterjee *et al.*, 2019; Vijayvergia and Kumar, 2021), Gated Recurrent Units (GRU)(Du and Nie, 2018; Rozental and Fleischer, 2018) and Deep Neural Networks (DNN)(Wang *et al.*, 2016; Park *et al.*, 2018) have also been used in the same supervised learning context but with increased performance. When evaluating the classification accuracies of supervised methods, these approaches typically achieve accuracies in the range of 65-80% on benchmark datasets (Acheampong *et al.*, 2020). However, they are hindered by two significant limitations: the scarcity of large, domain-independent labelled datasets and the difficulty of handling ambiguous and implicit emotional expressions.

More recent AI methods address these limitations by leveraging the semantic context of emotion expressions embedded in pre-trained language models. Unlike supervised methods, these methods can be fine-tuned with smaller labelled datasets using transfer learning. Emotion extraction using variations of BERT(Cambria *et al.*, 2020; Zeng *et al.*, 2021; Li *et al.*, 2021), GPT(Xiao, 2019; Fadel *et al.*, 2021), XLNet(Shen *et al.*, 2020; Alduailej and Alothaim, 2022) are such methods that leverage the contextual knowledge embedded in language models. These approaches report state-of-the-art accuracies for emotion detection from benchmark datasets in the range of 75-99%(Acheampong *et al.*, 2021). However, this strength is also a weakness due to the limited generalisability across new, unforeseen emotion expressions, as well as intensifiers, inhibitors, and negations of emotion expressions, lack of explainability and constrained domain adaptation. Collectively, these limitations question the practical value of the high accuracies reported in empirical evaluation(Dunietz *et al.*, 2020).

Several hybrid methods have also been proposed in the recent literature, combining symbolic with connectionist methods to improve accuracy and refine the emotion categories.

Tzacheva *et al.* (2020) proposed lexicon-based emotion annotation to train SVM classifiers for emotion extraction in tweets. Wu *et al.* (2006) utilized a rule-based approach to extract semantics related to emotions and combined it with lexicon ontology to extract emotions. In Sazzed and Jayarathna (2021), authors presented a self-supervised hybrid methodology for sentiment classification from unlabelled data that combines a machine learning classifier with a lexicon-based strategy. Li *et al.* (2012) proposed a hybrid emotion detection system combining hand-crafted rules and lexicon with a machine learning-based classifier to extract emotional levels in online blogs.

Collectively, across all three categories, the practical value of these methods in the management of information and extraction of patterns of behaviour of individuals and organizations is vast. Large-scale analyses of social media during elections (Ali *et al.*, 2022; Budiharto and Meiliana, 2018), patient-centred care for chronic illnesses such as Alzheimer's disease, cancer, and diabetes (Silva *et al.*, 2018; Saffar *et al.*, 2023; Ranasinghe *et al.*, 2022), real-time depression detection on social networks (Nijhawan *et al.*, 2022; Angskun *et al.*, 2022), expressions of emotion and sentiment during the COVID-19 global pandemic (Corti *et al.*, 2022; Adikari *et al.*, 2021b; Prabagar *et al.*, 2022), highlight the practical value in social and individual settings. In organisational settings, financial sentiment analysis (Sohangir *et al.*, 2018), understanding consumer satisfactions (Kumar and Zymbler, 2019), the role of social media in stock price moments (Smith and O'Hare, 2022), and the influence of review credibility and review usefulness (Fileri *et al.*, 2021) are pivotal studies that signify the continuing and incremental value of emotion analysis in digitalised content for all stakeholders.

When considering the various approaches, along with their limitations, strengths, and practical applications, we can elaborate on the four key capabilities and their effectiveness in addressing the challenges posed by the complexity and ambiguity of digital emotion expressions in knowledge-focused activities. The first capability is the output of the emotion detection approach. In most cases, this is limited to an emotion label without an intensity score for that emotion. This emotion label is also limited to a single granularity which cannot be further analysed in terms of its constituents. Most approaches assign a single emotion per atomic unit of text (sentence, paragraph or document) and overlook the presence of multiple emotions. The second capability, domain specificity, relates to the generalisability of the approach across diverse domains. Most approaches are highly specific to the syntax or semantics of a given domain, such as emotions in short text like tweets (Abdullah *et al.*, 2020; Sintsova and Pu, 2016), emotions in poetry (Khattak *et al.*, 2022; Ahmad *et al.*, 2020), emotions in code-switched text (Ilyas *et al.*, 2022; Lee *et al.*, 2015), and consumer reviews (Martin and Pu, 2014; Cao *et al.*, 2022). These are developed using supervised

learning and then evaluated using labelled custom datasets, which further limit generalisability and their application in diverse domains.

Approach	Output	Explainability	Adaptability	Modifier resolution and negation
Author: Mukherjee <i>et al.</i> (2021) Method: Machine learning algorithms with TD-IDF Dataset: Amazon reviews	Sentiment labels (positive, negative)	No	No	Negation prediction
Author: Adikari <i>et al.</i> (2021a) Method: Lexicon based Dataset: Online mental health forums	Emotion profiles (8 classes)	Limited to emotion keywords in the lexicon	No	Resolution limited to variations in the lexicon
Author: Kumar <i>et al.</i> (2022) Method: Deep learning and SVM Dataset: Twitter	Top sentiment label (positive, negative and neutral)	No	No	No
Author: Adoma <i>et al.</i> (2020) Method: Pretrained language models (BERT, RoBERTa, DistilBERT, and XLNet) Dataset: ISEAR	Top emotion label (7 classes)	No	No	No
Author: Shrivastava <i>et al.</i> (2019) Method: CNN (convolutional neural networks) and Word2Vec Dataset: TV shows transcripts	Top emotion label (7 classes)	No	No	No
Author: Tzacheva <i>et al.</i> (2019) Method: Hybrid (lexicon with SVM) Dataset: Tweets	Top emotion label (8 classes)	No	Challenging as large lexicon extensions are required	No
Author: Huang <i>et al.</i> (2019) Method: Pretrained language models (BERT) Dataset: Friends, EmotionPush	Top emotion label (8 classes)	No	No	No
Author: Rabeya <i>et al.</i> (2017) Method: Lexicon based Dataset: Social media and news	Sentiment (happiness and sadness)	Emotion keywords (limited lexicon)	No	No
Author: Emotion AWARE Method: Lexicon and finetuned language models Dataset: ISEAR, Twitter SemEval, Senti4SD, Goemotions, Fairy Tales, Affective Text, PhraseBank	Multifaceted emotion profiles with intensities; granularity levels of 2, 8, 14 classes	Emotion keyword extraction (captures a wide range of related words with embeddings)	Can adapt lexicons with small extensions of the lexicon; Lexicons and embeddings are customizable	Negation, intensifier, and inhibitor resolution followed by intensity rectification

Table 5.1 Comparison of Emotion AWARE with related work in emotion detection

Despite the custom datasets, some methods can be adapted (or retrained) for a new application, which is the third capability of adaptability. In recent work based on language models and annotated datasets, this capability is limited due to the large number of parameters and the opacity of transformer-based learning. They cannot be adapted without a significant volume of work on configuration and fine-tuning, which is equivalent to developing an entirely new approach. The fourth capability is the explainability of the detected emotion, which is becoming more important given our increasing dependence on AI and automation. Explainability has been overlooked in most approaches, mainly due to design limitations that have focused on producing emotion labels of singular granularity.

We do not consider accuracy as a core capability as it can be configured (or tweaked) in the design phase as an offset between the availability of annotated datasets for supervised learning and the need for generalisability across multiple domains. For instance, a high-quality human-annotated dataset can be leveraged by a supervised learning approach to produce highly accurate emotion classifications. In summary, the granularity of emotion detection output, domain specificity, adaptability and explainability are the formative capabilities of the proposed method for addressing the complexity and ambiguity of emotion expressions. Table ?? shows a capability comparison of Emotion AWARE against the most effective and relevant studies from the literature.

5.2.2 Methodology

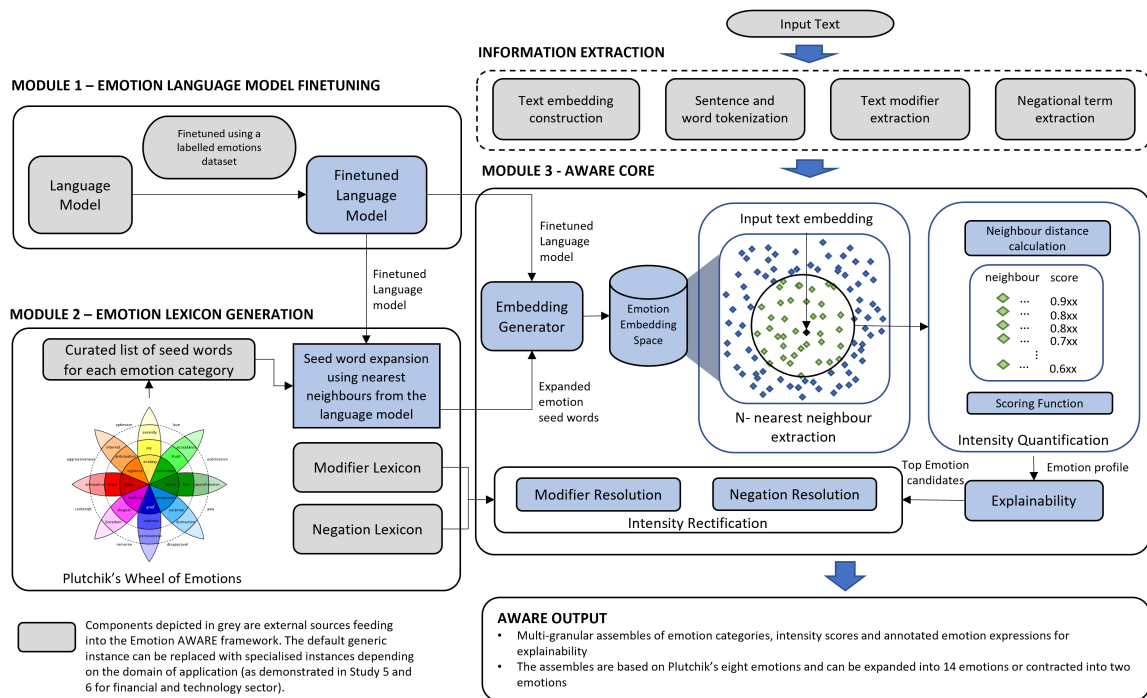


Fig. 5.2 The Modular Composition of the Emotion AWARE Framework

As illustrated in figure 5.2, the emotion AWARE framework consists of three modules, Module 1 - Emotion Language Model Finetuning, Module 2 - Emotion Lexicon Generation and Module 3 – AWARE Core. The components depicted in grey are external sources feeding into the Emotion AWARE framework, where the general instances we have used in this study can be replaced with specialised instances depending on the domain of application (this is demonstrated in Studies 5 and 6 for the financial and technology sector).

Module 1 begins with a state-of-the-art encoder language model, such as BERT(Devlin *et al.*, 2018), which has been effectively applied to diverse NLP tasks such as Reading Comprehension(Ramnath *et al.*, 2020; Joshi *et al.*, 2020) and Natural Language Inference(Lan *et al.*, 2019; Liu *et al.*, 2019). State-of-the-art language models are pre-trained on large volumes of unlabelled data to generate deep contextualised word representations by considering syntaxes and semantics(Qiu *et al.*, 2020). In application, these pre-trained models are finetuned using labelled datasets through transfer learning techniques. For this framework, we selected the DistilBERT(Sanh *et al.*, 2019) base-case model with Huggingface(Wolf *et al.*, 2019) PyTorch implementation for the finetuning. As the finetuning dataset, we selected the Emotion dataset(Saravia *et al.*, 2018a) due to its substantial size, the granularity of emotions, and widespread acceptance in the research community. It contained 20,000 tweets based on six emotions joy (33.5%), sadness (29.2%), anger (13.5%), fear (12.1%), love (8.2%), and surprise (3.6%). For the finetuning, we combined train and validation sets, randomized, and selected a subset of 5653 points, with 1000 samples per emotion except for surprise, which was 653 points. Finetuning settings were a default token length of 128 enabled by both padding and truncation and a batch size of 64 with eight epochs. At a learning rate of 0.00002 and weight decay of 0.01, the finetuning was completed with an F1 score of 0.9394 for the test segment of the dataset. The finetuned language model is utilised in Module 2 to expand a curated list of emotion seed words and in Module 3 for emotion embedding space generation. As noted earlier, DistilBERT can be replaced with any other language model closely aligned with the domain of interest.

Module 2 initiates with an emotion seed word list constructed and curated using a combination of automated and manual methods. In developing our emotion lexicon, we referenced Plutchik’s model(Plutchik, 1991), which identifies eight primary emotion classes, each further divided into three subcategories, resulting in a comprehensive 24-class system. Initially, seed keywords for each of these classes were manually curated from an online thesaurus(Merriam-Webster, 2024). However, we encountered a scarcity of unique terms for certain emotions, which necessitated the merging of closely related categories joy and ecstasy, amazement and surprise, disgust and loathing, interest and vigilance, anger (rage, anger, annoyance), fear (terror, fear, apprehension). As a result, we consolidated the model into 14 broader emotion classes, each supported by 15-20 thesaurus-derived terms.

While manually curating seed terms yielded high-quality initial seeds, the number of words was insufficient for comprehensive lexicon construction. Therefore, we utilized the vocabulary of the finetuned DistilBERT model itself and extracted embeddings for each of our seed words and compared them with the raw embeddings from the model’s vocabulary terms to find contextually and emotionally similar words. However, due to the ambiguity of

individual term embeddings, the relevance of these expanded terms was not highly consistent. To address this, we first clustered seed words into 4 subgroups using the constrained k-means algorithm (A. Demiriz P. S. Bradley, 2000) and then used the average embedding of each subgroup for the expansion. Here, the k value is extracted from the Elbow method (Syakur *et al.*, 2018). This process extended each subgroup to include the highest similar 25 terms from the model's vocabulary, aiming for a total of 100 terms per each of the 14 emotion classes. Subsequent refinement involved removing duplicates and terms conflicting with Plutchik's polar opposites to improve the lexicon coherence.

8 emotions	14 emotions	2 emotions
Joy	Serenity	Positive
	Joy/Ecstasy	
Trust	Acceptance	
	Trust	
	Admire	
Anticipation	Interest / Vigilance	
	Anticipation	
Surprise	Amazement/Surprise	Negative
	Distraction	
Sad	Sad	
Disgust	Disgust/loathing	
	Boredom	
Anger	Anger	
Fear	Fear	

Table 5.2 Alignment of the 2, 8 and 14 emotion classification schemes

The resulting vocabulary size for each emotion class contained between 80 to 100 terms. To standardize the lexicon, we pruned it by considering the centrality of term embeddings, where we compared each term's embedding to the average category embedding and retained the 80 most pertinent terms per class. The final emotion lexicon comprised 1,120 terms across the 14 classes. Table 5.2 depicts the alignment of the 2, 8 and 14 emotion classification schemes. The eight classes of emotion contained 80 words per class, with a total of 640 terms. The version with two classes contained 480 words per category with 960 terms. Module 2 also contains externally sourced lexicons for modifiers (inhibitors and intensifiers) and negations, which was based on the valence detection work described in VADER (Hutto and Gilbert, 2014). VADER employs an advanced process that integrates human annotations, heuristic rules, and statistical modelling to determine the valence and polarity of the modifiers. Module 2 provides these two lexicons and the expanded emotion terms as output into Module 3.



Fig. 5.3 The Emotion Embedding Space Generated by Module 3 of the Framework

Module 3 received the expanded emotion word terms and their corresponding embeddings to generate an emotion embedding space. If the lexicon is constructed from scratch, this step will be skipped, as words will already be tagged with embeddings during the expansion. For external lexicons, each word will go through an embedding extractor and be tagged with the corresponding embedding. The high dimensional vectors of this emotion embedding space can be visualised using the t-SNE algorithm on a 2-D grid, as shown in figure 5.3. Each point in the figure corresponds to an emotion term, with a clear separation between green and red, where green is for positive emotions, and red is for negative, in 14 emotion categorizations.

Next, Module 3 receives the sample input text or an entire text corpus. This text undergoes several information extraction steps, after which it is processed through the embedding generator and mapped onto the emotional embedding space. The n nearest neighbour extraction process identifies the closest emotion terms based on this projection. This process is depicted in Module 3, where the nearest neighbours are green dots, and the blue are all other emotion embeddings. Based on these nearest neighbours, the Intensity Quantification component calculates the intensities of each of the relevant emotion classes. Here, each neighbour will receive a score based on the proximity to the sample input. The terms are sorted and ranked based on similarity, then the terms are grouped based on emotion category, and the summed scores for each category are normalised to create an emotional assembly of

two, eight, and fourteen emotions per input text. See Equation 5.1. In this equation, θ_e is the intensity of emotion e , n is the number of nearest neighbors, A is the subset of nearest neighbors with emotion e , s_x is the distance score of the neighbor x .

$$\theta_e = \frac{\sum_{x \in A} s_x}{\frac{n \times (n+1)}{2}} \quad (5.1)$$

The next phase in Module 3 is the Explainability component. Explainability in AI aims to understand and interpret the model's output. In the context of Emotion AWARE, this is achieved by identifying and extracting the words that have contributed significantly to forming the emotion profile. Here, term embeddings extracted from the input text vector representation are compared with the mean embedding of the entire text. These terms are ranked based on similarity, and the top N terms are recorded for explainability and sent across to the intensity rectification component.

$$\theta_{e^k}^* = \theta_{e^k} + b \cdot a \quad (5.2)$$

$$\theta_e^{\text{normalized}} = \frac{\theta_e}{\sum_{i \in E} \theta_i} \quad (5.3)$$

The intensity rectification component consisted of two resolution processors for modifiers (intensifiers and inhibitors) and negations. The adjacent terms of the top N terms are passed through the corresponding lexicons to check for negative, intensive or inhibitive terms. Modifier resolution is completed prior to negations to detect intensified or inhibited negations. For detected intensifiers and inhibitors, the score of the top emotion in the profile is revised depending on the intensity of the modifier. Then, the emotion profile will be normalized so that the increment/decrement of the top emotion will affect the other emotions in the profile. In the case of negations, the emotion categories are revised based on Plutchik's polar opposites. See Equation 5.2 and Equation 5.3. In this context, the variables are defined as follows: $\theta_{e^k}^*$ represents the updated intensity of the top keyword's emotion, while θ_{e^k} denotes the current intensity of the top keyword's emotion. The variable b stands for the modifier polarity, acting as an intensifier (+1) or inhibitor (−1), and a reflects the modifier valence. The term $\theta_e^{\text{normalized}}$ is the normalized intensity of emotion e , and θ_e is the raw intensity of emotion e . Additionally, E encapsulates the set of all intensities in the emotion profile. Both the modifier and negation lexicons, as well as polarity and valences, are based on the prior work of VADER (Hutto and Gilbert, 2014).

Algorithm 1 THE EXPLAINABILITY AND INSENSITIVITY RECTIFICATION PROCESS OF EMOTION AWARE

Input: input text, nearest neighbours list, current emotion profile

Output: rectified emotion profile, emotional keywords

1 Initialize variables

2 $\text{model_output} \leftarrow$ Output after passing text through finetuned model

3 Initialize $\text{emotion_keyword_list}$, $\text{top_emotional_keywords}$

4 Initialize $\text{final_emotion_profile} \leftarrow \text{current_emotion_profile}$

5 $\text{input_terms_list} \leftarrow$ split input text to words

6 $\text{input_text_mean_emb} \leftarrow \text{extract_embedding}(\text{model_output}, \text{input_text})$

7 Calculate distances for significance

8 **for** *each* $w \in \text{input_terms_list}$ **do**

9 $w_emb \leftarrow \text{extract_embedding}(\text{model_output}, w)$

10 $w_dis \leftarrow \text{cosine_distance}(w_emb, \text{input_text_mean_emb})$

11 append $[w, w_emb, w_dis]$ to $\text{emotion_keyword_list}$

12 Finding top keywords and corresponding emotions

13 $\text{top_emotional_keywords} \leftarrow$ sort $\text{emotion_keyword_list}$ on w_dis and pick top n

14 $\text{best_emotion_keyword} \leftarrow$ select best candidate for intensity resolution

15 $\text{best_kwd_emo} \leftarrow \text{map_keyword_to_emotion}(\text{best_emotion_keyword}, \text{nearest neighbors list})$

16 Checking top keyword neighbourhood for modifier and negation resolution

17 $\text{best_kwd_neighborhood} \leftarrow \text{extract_kwd_neighborhood}(\text{input_text}, \text{best_emotion_keyword})$

18 Bool $\text{modifier} \leftarrow \text{check_modifier}(\text{best_kwd_neighborhood})$

19 **if** *modifier* **then**

20 $\text{rectified_emotion_profile} \leftarrow \text{modify_emotion}(\text{current_emotion_profile}, \text{best_kwd_emo})$

21 $\text{rectified_emotion_profile} \leftarrow \text{normalize}(\text{rectified_emotion_profile})$

22 Bool $\text{negation} \leftarrow \text{check_negation}(\text{best_kwd_neighborhood})$

23 **if** *negation* **then**

24 $\text{rectified_emotion_profile} \leftarrow \text{map_opposite_emotion}(\text{current_emotion_profile})$

25 **return** $\text{rectified_emotion_profile}, \text{top_emotional_keywords}$

5.2.3 Experiments

We designed a comprehensive set of nine studies to thoroughly evaluate and demonstrate the capabilities of the proposed Emotion AWARE framework in eliciting multi-granular, adaptable, robust, and explainable emotion assemblies. As detailed in table 5.4, each study involves a series of carefully structured experiments utilizing datasets from a state-of-the-art collection that captures realistic conversations and content across various digital media platforms (table 5.3). These datasets were selected to provide a diverse and representative sample of emotional expressions, ensuring that the framework’s performance could be rigorously tested across different contexts and scenarios.

The combination of these nine studies, spanning across eight distinct datasets, provides robust evidence of the framework’s effectiveness in detecting and analyzing the nuanced emotions conveyed in digital media. The consistent results across these varied datasets validate the framework’s ability to generalize well to different types of emotional content, reinforcing its potential for real-world applications.

Throughout all experiments, we maintained uniform configurations to ensure the reliability and comparability of the results. This included the use of the same finetuned language model, modifier and negation lexicons, and scoring and explainability modules. The emotion lexicons and embedding spaces employed were systematically based on corresponding 2, 8, and 14-class emotion categorizations, enabling a thorough exploration of the framework’s capabilities across varying levels of emotional granularity.

Dataset name	Description
ISEAR(Scherer and Wallbott, 1994a)	7665 sentences by 3000 respondents in 37 countries on all 5 continents describing situations where they experienced seven major emotions classified emotions are joy (14.3%), sadness (14.3%), fear (14.2%), anger (14.3%), guilt (14.2%), disgust (14.3%), and shame (14.3%) emotions
Twitter Senti-ment(Shahane, 2024)	3534 tweets extracted from figure eight’s data for everyone platform classified emotions are positive (31.2%),negative (28.3%) and neutral (40.5%). Test set is used.
SemEval 2007 (“Affective Text”)(Strapparava and Mihalcea, 2007)	1250 sentences from tweets, news headlines, Google News, and other major newspapers classified emotions are Ekman’s six basic emotions, fear (15.5%), anger (7.3%), joy (35.3%), sadness (21.2%), disgust (3.4%), and surprise (17.4%).
Continued on next page	

Table 5.3 – continued from previous page

Dataset name	Description
Fairy Tales(Alm, 2009)	All sentences with affective high agreement from 176 fairy tales(1207 datapoints) by B. Potter, H.C. Andersen, and Grimm’s classified emotions are angry-disgusted (18.1%), fearful (13.7%), happy (36.9%), sad (21.9%), and surprised (9.4%).
GoEmotions (Demszky <i>et al.</i> , 2020)	58,000 comments extracted from popular English-language subreddits on the reddit social media platform classified into admiration (9.4%), amusement (5.3%), anger (3.6%), annoyance (5.7%), approval (6.8%), caring (2.5%), confusion (3.1%), curiosity (5.1%), desire (1.5%), disappointment (2.9%), disapproval (4.7%), disgust (1.9%), embarrassment (0.7%), excitement (1.9%), fear (1.4%), gratitude (6.2%), grief (0.2%), joy (3.3%), love (4.8%), nervousness (0.4%), optimism (3.6%), pride (0.3%), realization (2.5%), relief (0.3%), remorse (1.2%), sadness (3.0%), surprise (2.4%), neutral (32.7%), some comments contained more than 1 emotion.
SemEval-2018(Park <i>et al.</i> , 2018)	10,983 manually curated English tweets from Affect in Tweets which included five subtasks on inferring the affectual state of a person from their tweet classified into 11 different emotions anger (36.1%), anticipation (13.9%), disgust (36.6%), fear (16.8%), joy (39.3%), love (12.3%), optimism (31.3%), pessimism (11.6%), sadness (29.4%), surprise (5.2%), trust (5.0%). Some comments contained more than 1 emotion.
Financial Phrase-Bank(Malo <i>et al.</i> , 2014)	4840 sentences from financial news on the LexisNexis database, classified emotions are positive (28.2%), negative (12.5%), and neutral (59.4%).
Senti4SD Dataset(Calefato <i>et al.</i> , 2018)	4424 posts extracted from stack overflow (www. stack overflow. com), classified emotions are positive (34.6%), negative (27.2%), and neutral (38.2%).

Table 5.3 Description of Datasets Used in the Experiments

Emotion AWARE Capability	Study ID	Description
Elicitation of emotion assemblies	Study 1	Objective: Elicitation of two-emotion assemblies – positive and negative
		Datasets: ISEAR(Scherer and Wallbott, 1994a) and Twitter Sentiment(Shahane, 2024) datasets
	Study 2	Study 2: Elicitation of four emotion assemblies - anger, fear, sadness, joy
		Datasets: SemEval 2007 (“Affective Text”)(Strapparava and Mihalcea, 2007), ISEAR and Fairy Tales(Alm, 2009)
	Study 3	Study 3: Elicitation of four emotion assemblies - disgust, surprise, trust, anticipation
		Datasets: GoEmotions(Demszky <i>et al.</i> , 2020) and SemEval-2018(Park <i>et al.</i> , 2018)
	Study 4	Study 4: Elicitation of 2, 8 and 14 emotion assemblies in increasing granularity
		Datasets: Demonstrated on sample sentences of GoEmotions
Adaptability	Study 5	Study 5: Emotion AWARE adapted for the finance sector
		Dataset: PhraseBank(Malo <i>et al.</i> , 2014)
	Study 6	Study 6: Emotion AWARE adapted for the technology sector
		Dataset: Senti4SD(Calefato <i>et al.</i> , 2018)
Robustness	Study 7	Study 7: Variation of emotion intensity score based on intensifiers and inhibitors
		Dataset: Demonstrated for sample sentence from SemEval-2018
	Study 8	Study 8: Robustness of Emotion AWARE across intensifiers and inhibitors
		Dataset: Manually curated dataset based on Fairy Tales
Continued on next page		

Table 5.4 – continued from previous page

Emotion AWARE Capability	Study ID	Description
Explainability	Study 9	Study 9: Explaining emotion assembles using constituent intensity scores and terms of emotional significance
		Dataset: Demonstrated for samples from Fairy Tales

Table 5.4 Nine studies evaluating and demonstrating capabilities of the proposed Emotion AWARE framework

Study 1: Elicitation of Two-emotion Assembles (positive and negative) Using ISEAR and Twitter Sentiment Datasets

This study demonstrates the generation of two-emotion assembles of positive and negative emotions, the accuracy of which is then validated with existing methods for the same binary classification. We used two datasets, Twitter Sentiment and ISEAR, in which we aggregated sadness, anger, fear, disgust, and joy as negative. The two-emotion assembles were evaluated with three other methods reported in the literature. They are 1) linear keyword matching using Plutchik’s emotion terms list (Udochukwu and He, 2015), 2) stemmed keyword matching (Adikari *et al.*, 2021a) with negation, inhibitor, intensifier detection components and 3) SentiWordNet 3.0 (Baccianella *et al.*, 2010). The evaluation was conducted across four metrics: accuracy, precision, recall and F1-score. Table 5.5 presents the results where Emotion AWARE surpasses all three methods.

Dataset	Approach	Positive			Negative			Accuracy
		Prec.	Rec.	F1	Prec.	Rec.	F1	
ISEAR aggregated	LKM	0.402	0.938	0.562	0.350	0.383	0.365	0.391
	SKM	0.422	0.954	0.585	0.339	0.427	0.378	0.405
	SentiWord Net	0.657	0.918	0.766	0.739	0.360	0.485	0.673
	Emotion AWARE	0.836	0.956	0.892	0.845	0.562	0.675	0.837
Twitter sentiment	LKM	0.335	0.854	0.482	0.407	0.873	0.555	0.373
	SKM	0.326	0.853	0.471	0.451	0.851	0.589	0.391
	SentiWord Net	0.568	0.774	0.605	0.639	0.729	0.681	0.605
	Emotion AWARE	0.715	0.905	0.797	0.928	0.782	0.849	0.827

Table 5.5 Comparison of Results for Two-emotion Assemblies using ISEAR and Twitter

Study 2: Elicitation of Four Emotion Assemblies (anger, fear, sadness, joy) Using SemEval 2007 (“Affective Text”), ISEAR and Fairy Tales

As noted previously, the proposed framework is capable of detecting all emotions in Plutchik’s wheel of emotions(Plutchik, 1991). However, only a handful of related works have proposed techniques to detect all eight emotions. Therefore, we split the eight emotions into two subsets (common and rare) in order to ensure that Emotion AWARE can be evaluated with state-of-the-art approaches in extant literature. Study 2 evaluates the common subset of anger, fear, sadness, and joy, while Study 3 evaluates the rare subset of disgust, surprise, anticipation, and trust. In study 2, we compared AWARE with rule-based, hybrid as well as machine-learning techniques. Rule-based methods include emotional linear keyword matching, stemmed keyword matching, and the more advanced rule-based methods that consider contextuality and affinity-based methods such as CLSA, CPLSA, and DIM. Here, CLSA and CPLSA are categorical classifications based on LSA and PLSA. Additionally, we also compared with context-based emotion vector construction methods(Agrawal and An, 2012), namely context-based Wiki, context-based Guten, and context-based W-G. For machine learning methods, we finetuned the DistilBERT(Sanh *et al.*, 2019) model on the Emotion(Saravia *et al.*, 2018b) dataset. Collectively, study 2 compares Emotion AWARE with

ten similar techniques proposed in recent literature, using SemEval 2007, ISEAR and Fairy Tales datasets. For this, we incorporated the experiments included in previous work (Kim *et al.*, 2010; Agrawal and An, 2012). As presented in table 5.6, AWARE outperforms all methods for most combinations of datasets and emotions.

Method		LKM	SKM	CLSA	CPLSA	CNMF	DIM	Wiki (CB)	Guten (CB)	W-G (CB)	DistilBERT (emotions)	Emotion AWARE
Emotion	Dataset											
Anger	SemEval	0.071	0.119	0.112	0.244	0.278	0.175	-	-	-	0.252	0.371
	ISEAR	0.274	0.192	0.631	0.456	0.579	0.286	0.628	0.582	0.668	0.494	0.571
	Fairy Tales	0.213	0.193	0.512	0.313	0.651	0.392	-	-	-	0.575	0.708
Fear	SemEval	0.252	0.291	0.511	0	0.618	0.404	-	-	-	0.292	0.488
	ISEAR	0.393	0.451	0.071	0	0.056	0.351	0.592	0.536	0.574	0.673	0.602
	Fairy Tales	0.385	0.363	0.641	0	0.714	0.255	-	-	-	0.508	0.696
Joy	SemEval	0.203	0.258	0.402	0.254	0.648	0.712	-	-	-	0.698	0.772
	ISEAR	0.365	0.379	0.103	0.341	0.011	0.564	0.564	0.574	0.542	0.673	0.738
	Fairy Tales	0.632	0.621	0.727	0.436	0.741	0.255	-	-	-	0.772	0.741
Sadness	SemEval	0.226	0.282	0.337	0.378	0.475	0.253	-	-	-	0.281	0.554
	ISEAR	0.383	0.361	0.106	0.282	0.017	0.337	0.408	0.253	0.296	0.557	0.611
	Fairy Tales	0.462	0.449	0.642	0.371	0.761	0.242	-	-	-	0.655	0.741
Average	SemEval	0.188	0.237	0.341	0.219	0.505	0.386	-	-	-	0.382	0.546
	ISEAR	0.353	0.345	0.227	0.269	0.165	0.372	0.548	0.486	0.521	0.599	0.631
	Fairy Tales	0.419	0.406	0.629	0.279	0.716	0.285	-	-	-	0.627	0.722

Table 5.6 Comparison of results for four emotion assembles (anger, fear, sadness, joy)

Study 3: Elicitation of Four Emotion Assembles (disgust, surprise, trust, anticipation) Using GoEmotions and SemEval-2018

For the rare emotions of disgust, trust, anticipation, and surprise, we used GoEmotions and SemEval-2018 datasets and compared with stemmed keyword matching and the DistilBERT model finetuned with the Emotions dataset. Table 5.7 presents the results where AWARE outperforms all other methods across the four emotions.

Dataset	Method	Disgust	Surprise	Trust	Anticipation
GoEmotions	SKM	0.377	0.234	0.076	0.048
	DistilBERT (Emotions)	0.451	0.523	0.362	0.491
	Emotion AWARE	0.648	0.655	0.707	0.491
SemEval 18	SKM	0.252	0.084	0.147	0.068
	DistilBERT (Emotions)	0.515	0.185	0.175	0.215
	Emotion AWARE	0.826	0.515	0.439	0.289

Table 5.7 Comparison of results for four emotion assembles (disgust, surprise, trust, anticipation)

Study 4: Elicitation of 2, 8 and 14 emotion assembles in increasing granularity

This study demonstrates Emotion AWARE's ability to generate emotion assembles at diverse levels of granularity. Table 5.8 presents these granular emotion assembles for the same text. Only the emotions with non-zero scores are shown in this table. For instance, row 2 depicts a positive score in the two-emotion assemble, anticipation and trust as the detected emotions in the eight emotions assemble, and in the 14 emotions assemble, trust is further split into trust, acceptance, and admiration alongside the corresponding intensity scores.

ID	Sentence	Two emotion as-semble	8 emotion assem-ble (main emotions shown)	14 emotion profile (main emotions shown)
1	"That's a maryland fan comment. You should be ashamed of yourself."	Negative:0.937	disgust: 0.558	disgust_loathing: 0.299
			anger: 0.179	boredom: 0.259
			sadness: 0.169	anger: 0.179
2	"Our father will protect us <3"	Positive: 0.977	trust: 0.529,	sadness: 0.169
			anticipation: 0.268	trust: 0.371
				acceptance: 0.115
3	" But I'm sort of confused as to how the pic was taken, sorry if you said it but I'm a bit stupid and tired"	Negative: 0.797	surprise: 0.347,	admire: 0.043
			disgust: 0.211	interest_vigilance: 0.207
			sadness: 0.186	anticipation: 0.061
4	"Best bedtime Christmas story ever."	Positive: 0.998	trust: 0.376	distraction: 0.292
			joy: 0.337	amazement_surprise: 0.055,
			surprise: 0.285	boredom: 0.151
				disgust_loathing: 0.061
				sadness: 0.186
				admire: 0.308
				acceptance: 0.068
				joy_ecstasy: 0.337
				amazement_surprise: 0.285

Table 5.8 Demonstrating the Elicitation of 2, 8 and 14 Emotion Assembles in Increasing Granularity

Study 5: Emotion AWARE Adapted for the Finance Sector Using the PhraseBank Dataset

Domain adaptability is a core capability of Emotion AWARE. In studies 5 and 6, we demonstrate this capability for the financial and technology sector. For the financial sector, we used the PhraseBank dataset, which contains financial statements classified based on positive and negative emotions. Emotion AWARE was adapted to this domain by simply expanding the

vocabulary with 20 words each for positive and negative classes using the L&M financial emotion lexicon(LOUGHRAN and MCDONALD, 2011). Following the domain adaptation, two emotion assembles were generated and compared with the stemmed keyword matching technique, finetuned DistilBERT with the Emotion dataset and SentiWordNet. Emotion AWARE is used with both the default vocabulary and the extended vocabulary using L&M. Table 5.9 summarizes the results; notably, AWARE surpasses all methods across all metrics.

Approach	Positive			Negative			Average F1
	Prec.	Rec.	F1	Prec.	Rec.	F1	
Stemmed keyword Matching	0.405	0.649	0.499	0.228	0.523	0.318	0.412
SentiWordNet	0.307	0.511	0.383	0.833	0.698	0.763	0.572
DistilBERT (emotions)	0.831	0.937	0.871	0.813	0.593	0.690	0.782
Emotion AWARE (Default vocab)	0.922	0.853	0.886	0.699	0.829	0.759	0.821
Emotion AWARE (LnM + Default vocab)	0.903	0.879	0.891	0.765	0.805	0.785	0.843

Table 5.9 Comparison of results generated by Emotion AWARE adapted for the finance sector

Study 6: Emotion AWARE Adapted for the Technology Sector Using Senti4SD8 Dataset

Study 6 is the domain adaptation for the technology sector, where we used the Senti4SD dataset, which contains conversations from the Stackoverflow community classified by emotion. Similar to study 5, we evaluated the proposed approach with default vocabulary as well as extended vocabulary along with stemmed keyword matching, SentiWordNet, and finetuned DistilBERT. Here both positive and negative classes were extended with 20 words extracted using Emotion AWARE running on the training set. As shown in table 5.10, Emotion AWARE outperforms all other methods in this adaptability task.

Approach	Positive			Negative			Average F1
	Prec.	Rec.	F1	Prec.	Rec.	F1	
SKM	0.973	0.672	0.795	0.397	0.922	0.555	0.681
SentiWordNet	0.722	0.772	0.746	0.642	0.848	0.731	0.741
DistilBERT (emotions)	0.896	0.945	0.921	0.925	0.861	0.892	0.911
Emotion AWARE (Default vocab)	0.954	0.868	0.909	0.816	0.933	0.871	0.891
Emotion AWARE (LnM + Default vocab)	0.945	0.911	0.928	0.883	0.927	0.904	0.921

Table 5.10 Comparison of Results Generated by Emotion AWARE When Adapted for the Technology Sector

Study 7: Robustness of Emotion AWARE Across Intensifiers and Inhibitors

Intensifiers and inhibitors are subjectively used in emotion expressions, which means an emotion detection method must be robust to intensifiers and inhibitors, specifically in digitalised emotion expressions where physical cues are unavailable. We created a new dataset to demonstrate this robustness property of Emotion AWARE because state-of-the-art datasets used in related work are limited in including varying intensifiers and inhibitors. To construct this manually curated dataset, we selected a random subset of 80 sentences from the Fairy Tales dataset and introduced intensifiers and inhibitors to each sentence to generate an additional 160 sentences.

Sentence (Valence is bolted)	Modifier valence (increment / decrement)	Modifier intensity	Emotion Profiles (Only top emotions are shown)	Original sentence top Emotion's score (joy_ecstasy)
work was good for the first half	-	-	'admire': 0.229, 'joy_ecstasy': 0.339	0.339
work was incredibly good for the first half	Increment (+)	0.9	'amazement_surprise': 0.197, 'joy_ecstasy': 0.444	0.444
work was very good for the first half	Increment (+)	0.8	'admire': 0.216, 'joy_ecstasy': 0.426	0.426
work was quite good for the first half	Increment (+)	0.6	'admire': 0.194, 'joy_ecstasy': 0.425	0.425
work was considerably good for the first half	Increment (+)	0.6	'admire': 0.232, 'joy_ecstasy': 0.409	0.409
work was almost good for the first half	Decrement (-)	0.2	'admire': 0.202, 'joy_ecstasy': 0.316	0.316
work was kinda good for the first half	Decrement (-)	0.3	'joy_ecstasy': 0.244, 'admire': 0.255	0.244
work was sort of good for the first half	Decrement (-)	0.4	'joy_ecstasy': 0.206, 'acceptance': 0.214, 'admire': 0.262	0.206
work was barely good for the first half	Decrement (-)	0.7	'boredom': 0.164, 'disgust_loathing': 0.23, 'sadness': 0.433	0.054

Table 5.11 Emotion Profiles Based on Sentence Valence and Intensities

Table 5.11 demonstrates the evaluation of a single sentence using known intensifiers and inhibitors and their corresponding impact on the emotion score and emotion category. Here, the valence and intensity of modifiers are derived from prior work of VADER (Hutto and Gilbert, 2014). In case of incrementing or decrementing the modifier, the current top emotion's score will be increased or decreased with a factor of corresponding modifier intensity, as explained in equation 2. Then, the emotion profile will be normalized according

to equation 5.3. For this experiment, we used a sample sentence from the SemEval-2018 dataset. As depicted in table 5.11, the base sentence *"work was good for the first half"* is classified as joy_ecstasy with an intensity score of 0.339 and admire with a score of 0.229. In the subsequent rows, we added intensifiers and inhibitors with varied valence that modify the emotion expressed in the sentence. In descending order of table 5.11, the intensity score of the top emotion of the base sentence (joy_ecstasy) decreases. This illustrates that AWARE has correctly identified all modifiers attributed to emotion labels and varied intensity scores accordingly.

Version	Original	Intensified	Inhibited
Sentence	She is happy	She is extremely happy	She is partly happy
Most significant emotion	Joy	Joy	Joy
Most significant emotion score	0.54(x)	0.68(>x)	0.41(<x)

Table 5.12 Demonstrating Robustness across Intensifiers and Inhibitors

The manually curated dataset was used to evaluate Emotion AWARE, SentiWordNet, and stemmed keyword matching. Even though these approaches construct multi-facet emotion profiles, for this experiment, we have only considered the most significant emotion as it is the most impacted by such modifications. For instance, if the original sentence's most significant emotion is joy and has a score of x , it is expected that the intensified sentence's score of joy will be $>x$. In contrast, the inhibited sentence's score of joy will be $<x$ (table 5.12). Thus, we consider the original sentence's most significant emotion score in inhibited and intensified cases to see whether the approach has correctly identified modifiers. As the dataset consisted of 80 sentences, we calculated the mean of the most significant emotion's score as the evaluation metric. Here, DistilBERT (Emotion) is not included as it only provides labels (table 5.13).

Approach	Mean of most significant emotion's score			
	Original Sentences	Intensified Sentences	Inhibited Sentences	
Stemmed Keyword Matching	0.188	0.237	0.103	
SentiWordNet	0.122	0.107	0.107	
Emotion AWARE	0.561	0.907	0.413	

Table 5.13 Performance of Inhibitor and Intensifier Detection

As seen in the mean emotion scores, Emotion AWARE has increased from 0.346 in the intensified case and decreased from 0.161 in the inhibited case. This shows that AWARE has

correctly modified the emotions compared to the corresponding original sentences. Stemmed keyword matching has incorporated the modifiers to some extent but is bottlenecked with modifier capturing limitations. When considering SentiWordNet, none of the modifiers were detected, which has mitigated the scores even in intensified sentences.

Study 8: Robustness of Emotion AWARE in Negation Detection

Approach	F1 Score	
	Original sentences	Negated sentences
Improved Keyword Matching	0.435	0.462
SentiWordNet	0.799	0.297
DistilBERT (Emotions)	0.805	0.056
Emotion AWARE	0.818	0.841

Table 5.14 Results for Robustness of Emotion AWARE in Negation Detection

Similar to Study 7, we randomly selected 80 sentences from the Fairy Tales dataset and manually negated to create a new dataset of 80 negated sentences. We used negation terms such as ‘no’, ‘not’, and ‘never’ to reverse the emotions. We used this dataset to evaluate the robustness of Emotion AWARE with that of stemmed keyword matching, SentiWordNet and DistilBERT finetuned on the Emotion dataset. Table 5.14 presents mean F1 scores of emotion detection for original and negated sentences in this dataset. It is interesting to note that although SentiWordNet and DistilBERT show comparable accuracies to AWARE for the original sentences, they perform poorly for the negated sentences, unlike Emotion AWARE which scores 0.841 F1 score. We hypothesize that this observed behaviour is likely a result of the model’s tendency to prioritize emotion-specific terms while disregarding the presence of negating words within the sentences. The datasets used in Studies 7 and 8 will be made publicly available as a secondary outcome of this work. This dataset consisted of 320 sentences, 80 per original, negated, intensified, inhibited and optimal for modifier evaluation.

Study 9: Explainability of Emotion Assembles using Constituent Intensity Scores and Terms of Emotional Significance

Study 9 evaluates the explainability of the emotion assembles generated by the Emotion AWARE framework, using both intensity scores and terms that contribute to the detection of an emotion. Figure 5.5 illustrates this capability for a sample sentence randomly selected from the Fairy Tales dataset, “*How fortunate I am; it makes me so happy, it is such a pleasant thing to know that something can be made of me*”. The framework generates intensity peaks

for the terms “fortunate”, “happy” and “pleasant”, which distinguishes the contributing terms and their significance in the emotion assemble. These intensities are based on the w_dist in the Algorithm 1.

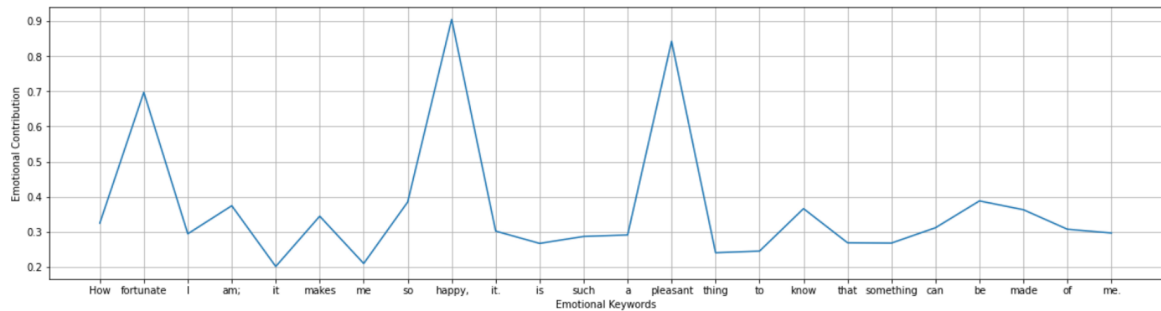


Fig. 5.5 Demonstrating Explainability of Emotion Assembles using Constituent Intensity

Figure 5.6 presents a further demonstration of explainability with emotion keyword extraction. Here, the positive and negative samples are randomly selected from the Fairy Tales dataset. We combined some samples to create a mixed sample. The colour scheme depicts emotional significance, where shades of green are for positive emotions and shades of red are for negative emotions. The intensity scores are depicted on the right side of the image, which further improves the explainability of the emotion assemble.

Example	Intensity scores	
Positive Sample	Term	Score
The son said: "Most gracious father, I will show her to you in the form of a beautiful flower," and he thrust his hand into his pocket and brought forth the pink, and placed it on the royal table, and it was so beautiful that the king had never seen one to equal it.	beautiful	0.74
	beautiful	0.73
	gracious	0.73
	Royal	0.66
	Equal	0.55
Negative Sample	Term	Score
The king ordered the man to be brought before him, and threatened with angry words that unless he could before the morrow point out the thief , he himself should be looked upon as guilty and executed .	Angry	0.78
	executed	0.75
	threatened	0.74
	thief	0.71
	guilty	0.68
Mixed Sample	Term	Score
"It was saying, 'You are so beautiful , I like you very much. 'Tweet, tweet,' sang the bird, as he flew out into the green woods, and Tiny felt very sad . The little prince was at first quite frightened at the bird. It was like a giant, compared to such a delicate little creature as himself. But when he saw Tiny , he was delighted , and thought her the prettiest little maiden he had ever seen.	delighted	0.90
	frightened	0.88
	delicate	0.85
	beautiful	0.81
	sad	0.76
	prettiest	0.71
	maiden	0.71

Fig. 5.6 Contributing Terms and Corresponding Intensity Scores for Emotion Explainability

The following table 5.15 summarizes the emotion keyword results for the entire fairy tales dataset. Here, for each sample in the dataset, the top emotion and top keyword are

extracted. The table contains each of the emotion categories: fear, anger, joy, surprise and sadness, along with the ten most frequent keywords per category. These keywords reflect the corresponding emotions, which further validate AWARE.

Emotion	Keywords with highest significance per category
Fear	‘frightened’, ‘terrified’, ‘afraid’, ‘uneasy’, ‘alarmed’, ‘fear’, ‘cried’, ‘trembling’, ‘anxious’, ‘trembled’
Anger	‘angry’, ‘stupid’, ‘angrily’, ‘cried’, ‘growled’, ‘annoyed’, ‘nasty’, ‘rage’, ‘fury’, ‘refused’
Joy	‘happy’, ‘pleased’, ‘joy’, ‘glad’, ‘merry’, ‘delighted’, ‘beautiful’, ‘rejoiced’, ‘happily’, ‘good’
Surprise	‘exclaimed’, ‘astonished’, ‘surprised’, ‘heavens’, ‘startled’, ‘shocked’, ‘bewildered’, ‘yelping’, ‘sudden’, ‘interrupting’
Sadness	‘wept’, ‘sad’, ‘grieved’, ‘cry’, ‘sorrowfully’, ‘poor’, ‘unhappy’, ‘mournfully’, ‘troubled’, ‘tears’

Table 5.15 Ten Most Frequent Keywords per Emotion Category in Fairy Tales Dataset

Emotion AWARE tackles the complexity and ambiguity of emotion detection through its four key capabilities: multi-granular emotion assemblies, adaptability, robustness, and explainability. Unlike other emotion detection models, this framework generates emotion assemblies by integrating heuristic-based prior knowledge with learned knowledge from fine-tuned language models. Empirical validation of Emotion AWARE was conducted through nine studies using advanced datasets that represent diverse human emotions. Studies 1–4 evaluated the detection of a range of emotions, from binary sentiment to a more granular classification of up to 14 emotions, including both common and rare emotions from Plutchik’s wheel. Emotion AWARE consistently outperformed fine-tuned DistilBERT, underscoring the significance of lexicon-based prior knowledge.

The framework’s adaptability was demonstrated in Studies 5 and 6, tailored for finance and technology domains, showing a 6% improvement in F1-score with an extended vocabulary compared to DistilBERT. Many contemporary models lack this domain adaptability, struggling with the nuances of emotion polarity across different contexts. For example, the term "unpredictable" may be positive in film reviews but negative in financial or HR contexts(e.g., “*the impact on share market indices is unpredictable*” or “*the employee response to governance in unpredictable*”). Emotion AWARE addresses these limitations by incorporating a small set of domain-specific terms and embeddings, allowing it to capture context and emotion more accurately.

Robustness was highlighted in Studies 7 and 8, where the framework effectively detected implied emotions, intensifiers, inhibitors, and negations, assigning intensity values

accordingly. In particular, Study 8 revealed that Emotion AWARE outperformed models like DistilBERT and SentiWordNet in negation detection. For instance, while DistilBERT could correctly detect emotions in positive sentences like *“I am truly glad to hear it!(joy)”* and *“I am truly sad to hear it!(sad)”*, it failed with the negated version *“I am truly not glad to hear it!”*, misidentifying the emotion as joy. Emotion AWARE’s heuristic approach to handling negations significantly improved accuracy in these cases. Study 9 demonstrated the framework’s explainability by unpacking emotion assemblies into contributing terms and corresponding intensity scores, offering transparency and rationale behind detected emotions.

The practical implications of Emotion AWARE are vast. Its robust, adaptable, and explainable emotion detection capabilities are highly valuable as emotional expression increasingly occurs in digital spaces. For instance, in complex, multi-stakeholder environments like long-term healthcare, Emotion AWARE can be customized for each stakeholder’s vocabulary. This enables the explanation of emotion assemblies and supports decision-making by converging or diverging the emotional profiles of all parties involved, fostering better consensus-building in human behavior analysis.

5.3 Ablation Studies of Emotion Aware

This section presents two practical use cases of the Emotion AWARE algorithm in healthcare and finance, demonstrating the value of neuro-symbolic architectures with features such as explainability, adaptability, and controllability. In the first use case, Emotion AWARE was applied to detect patient emotions and emotion transitions in digital healthcare settings, particularly during conversations between patients, carers, and healthcare practitioners. This was tested on a large real-world dataset of 60,000 call recordings from Cancer Information and Support Services, provided by Cancer Council Victoria, Australia. The results highlighted the algorithm’s effectiveness in identifying emotions and emotion transitions during telehealth interactions. In the second use case, the Emotion AWARE algorithm was used to analyze emotions within financial conference transcripts, focusing on emotional conflicts, agreements, and fluctuations between managerial and non-managerial participants, as well as their correlation with various market variables. This analysis utilized a dataset of 12,000 financial transcripts from over 500 organizations, covering the period from 2010 to 2020. The following sections will provide detailed explanations of these use cases.

5.3.1 Emotion Analysis of Cancer Council Victoria

Introduction

Recent advancements in Artificial Intelligence (AI) have significantly contributed to its widespread adoption in healthcare applications and services. The global pandemic has accelerated the integration of AI into telehealth, enhancing service quality and driving new models of care. AI is now extensively applied across various clinical, medical, and healthcare settings, including diagnosis, surgery, treatment, recovery, and patient-centered care. These innovations have transformed and optimized healthcare through improved efficiency and productivity. Additionally, there is a growing emphasis on the development of explainable and trustworthy AI applications, particularly in healthcare, to ensure reliability and transparency.

Cancer is a major contributor to social and economic burdens, with over 50% of Australians affected by cancer experiencing unmet informational and psychological support needs (Harrison *et al.*, 2009; Wang *et al.*, 2018b), particularly among the most disadvantaged groups who report higher levels of unmet needs. Despite advancements in diagnosis and treatment, disparities in cancer outcomes and survival are widening for these disadvantaged populations (Tervonen *et al.*, 2017). Given the complexity of cancer, patients require ongoing support and significantly more information than is typically provided during consultations (Ekberg *et al.*, 2014). Individuals impacted by cancer often experience elevated psychological distress throughout their lives (Boltong *et al.*, 2016). With rising survival rates, it is crucial to offer more efficient and continuous support to those affected. Addressing these unmet needs enhances care satisfaction, physical and mental health, reduces carer burden, and improves quality of life (QoL) and overall well-being (Lisy *et al.*, 2019).

Cancer Council Victoria (CCV), Australia, is a non-profit cancer charity dedicated to cancer research, prevention, advocacy, and support for individuals affected by any type of cancer. They employ experienced cancer nurses who serve as a central point of contact for those affected by cancer in the state, addressing unmet needs for information and psychological support. The staff are trained to provide accurate, timely, and relevant information, as well as emotional support. The most common reasons people contact CCV include seeking emotional and practical support, referrals to supportive care services, and information about cancer treatment, prevention, diagnosis, and general inquiries. CCV is a valuable resource for healthcare professionals and community organizations and serves as a public information hub for cancer-related issues, particularly for those in regional and remote areas with limited access to primary care. This helps ensure consistency with national quality standards in cancer care.

However, emotionality, a crucial aspect of an information source like this, has not been thoroughly explored. In this case study, we aim to bridge this gap by utilizing the Emotion AWARE framework to detect patient emotions and emotion transitions based on telehealth call recordings. We evaluated the framework on 60,000 call recordings provided by CCV,

and the results demonstrate its effectiveness in detecting and profiling patient emotions and their transitions.

Methodology

The proposed AI framework is illustrated in figure 5.7. It receives two types of input data from CCV call centres, 1) call recordings as audio files and associated metadata that is unstructured data and 2) data entry and notes input by the nurse that is structured data, optionally, there could also be un/structured data provided by the caller such as an end-of-call feedback survey or later correspondence. The framework consists of several modules namely speech-to-text module, Neuro-symbolic thematic pipeline, Emotion AWARE module. These components collaboratively work to produce several outputs that contribute towards service quality improvement use cases at different levels, such as call classifiers, predictive acoustic models, and call allocation optimization schedules, training and auditing. A primary aggregate output that encompasses and feeds into all such use cases is a collection of analytics dashboards (as shown in figure 5.7) that draw actionable insights from the unified data model to address specific challenges and/or generate value that directly serves organisational and research objectives. Ethics approval was granted by CCV [HREC 2107] for this study.

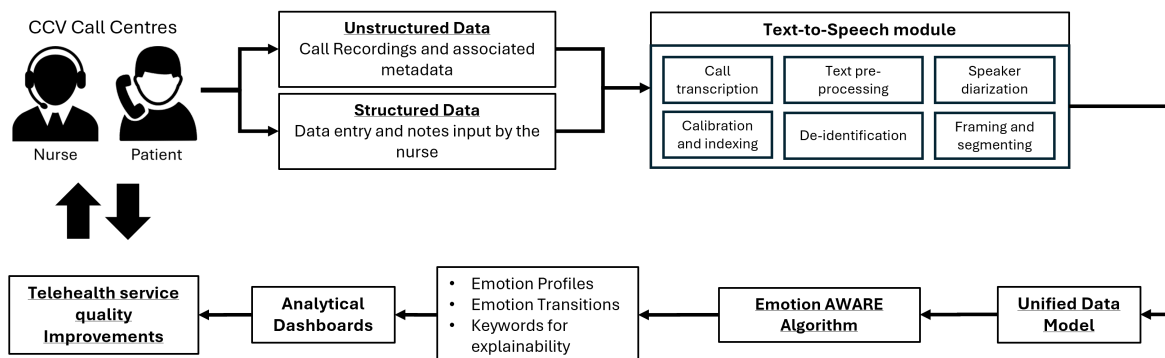


Fig. 5.7 Utilising neuro-symbolic architectures to improve telehealth services in Cancer Council Victoria

Speech-to-text module begins with the transcription of call recordings from speech to text, noting the spoken words are a conversation between two people, caller and nurse. Recent advancements in deep learning have significantly improved speech-to-text transcription with commercial cloud services reporting close to 100% accuracies and very low processing latencies. For this purpose we used Azure speech to text cognitive service based on the initial benchmark tests conducted (Microsoft Azure, 2024). The transcription output from Azure contains a detailed breakdown of the conversation between nurse and patient, including the spoken text, duration of the recording, individual phrases with caller identification and phrase level timestamps. Calibration and indexing are most effective when there is a large volume

of call recordings. Aggregate and summary metrics from this corpus can be used to validate and verify missing data, anomalous content and other irregularities such as long/short call durations, long periods of silence and levels of acceptable noise. Following transcription, the text files are subject to de-identification checks to ensure ethical and privacy standards that remove all mentions of names, addresses, date of birth as well as any other custom checks for identifying information.

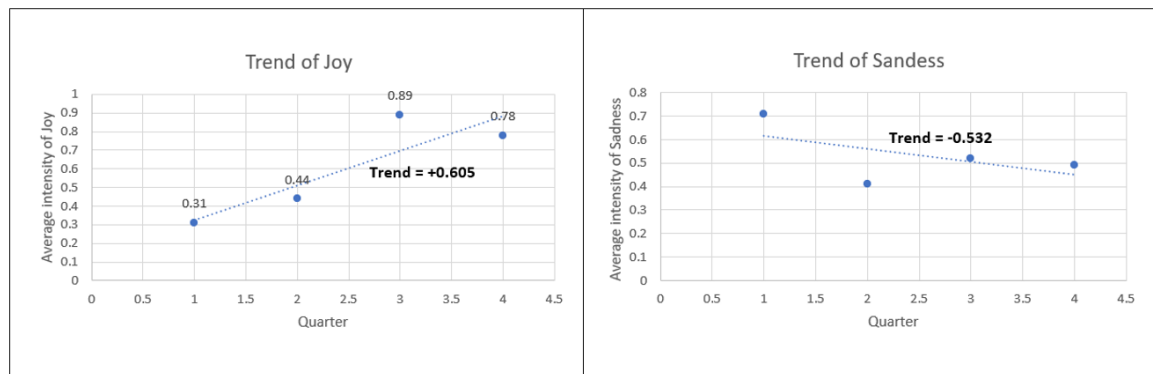


Fig. 5.8 Emotion trend calculation in the call transcripts

The pipeline also contained standard text pre-processing techniques such as stop-word removal, lemmatization, keyword extraction, and n-gram extraction. Speaker diarisation is also completed as part of this stage where a Gaussian Mixture Model is used to model each speaker, given the nurses are a finite set of speakers, the process can be simplified using additional parameters. Finally framing and windowing are used to prepare the audio component for further downstream techniques such as using machine learning to classify intense emotion expressions, discussion vs disagreement and long periods of silence. Outputs from these components will be forwarded to the unified data model. Unified data model consist of several types of data prepared to be processed by AI based modules. Firstly, structured data collected from the caller (such as age, demographics, clinical information, distress score etc) and the nurse notes which can be a mix of numerical, categorical and free-text variables. Secondly, unstructured data that has been preprocessed in text-to-speech module, and thirdly, any domain specific information such as a medical or clinical ontology that can be used as a reference point.

This unified data model will be then forwarded to the AI modules namely Emotion AWARE and neuro-symbolic thematic pipeline. Emotion AWARE will detect the 8 basic emotions in these transcriptions consisting of four negative; anger, fear, sadness, disgust, and four positive; joy, anticipation, trust and surprise. Emotion AWARE was used for the generation of emotion profiles with intensities at different levels of granularity within the

patient communication. The profile begins at the sentence level, and then dialogs which are instantiated by either party, followed by segments of the call and finally, the call itself. Having intensities recorded at multiple granularities ensures the telehealth service can be evaluated and improved at each level. For instance, the difference between the emotion profile at the start of the call and the end of the call signifies both the value generated by the healthcare practitioner and the value received by the caller.

In addition to the emotion profiles, emotion transitions were modelled for the call durations. Here, per sentence, emotions were extracted and converted into a time series. The entire conversation is split into k segments, typically $k=4$, as quarters are translatable for evaluation and training. An emotion algebra is computed for the emotion profile of each segment that can be formulated as a trend line across the segments. Thereby, an increasing trend indicates the emotion intensity has increased over the duration of the call and vice versa. This transition can be used in real-time to address patient distress, while in the long-term, it can be summarised into an evaluation metric as well as used for training new recruits and building up a knowledge base of common issues.

Given a time series $E = \{e_1, e_2, \dots, e_n\}$, split into four quarters, the average emotion score for each quarter is:

$$\bar{e}_j = \frac{1}{Q} \sum_{i=(j-1)Q+1}^{jQ} e_i, \quad j = 1, 2, 3, 4 \quad (5.4)$$

To fit a line $y = mx + c$ to the points $(1, \bar{e}_1), (2, \bar{e}_2), (3, \bar{e}_3), (4, \bar{e}_4)$, the slope m is:

$$m = \frac{4 \sum_{i=1}^4 i \bar{e}_i - \sum_{i=1}^4 i \sum_{i=1}^4 \bar{e}_i}{4 \sum_{i=1}^4 i^2 - (\sum_{i=1}^4 i)^2} \quad (5.5)$$

The intercept c is:

$$c = \frac{1}{4} \sum_{i=1}^4 (\bar{e}_i - mi) \quad (5.6)$$

Experiments and Results

The proposed AI framework was implemented as an API-based data lakehouse platform using Python for AI models and Microsoft SQL Server (MSSQL) for the data lakehouse. We then applied the framework to 60,000 call recordings provided by CCV call centres. All calls were specifically focused on the provision of telehealth support for cancer care, and the callers included patients, carers, family, the general public and healthcare providers. Figure 3 depicts the distribution of calls for four years, 2018 to 2021. This represents the complexity of detecting emotion at the sentence level of each call and further justifies the

need to determine emotion transition as a metric to evaluate the effectiveness of a call or calls in a large corpus.

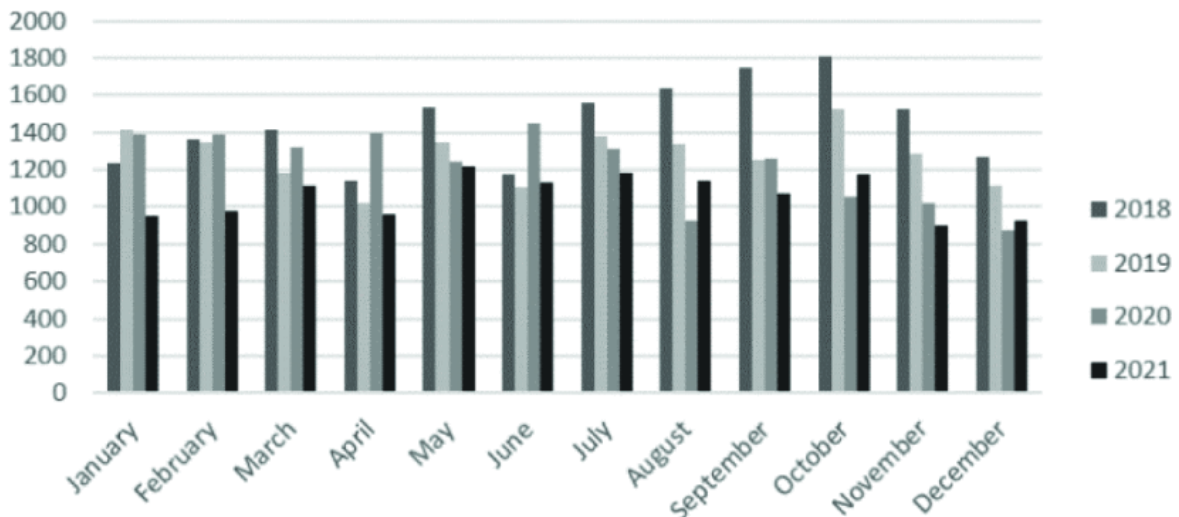


Fig. 5.9 Call volumes by month from 2018-2021

Our descriptive analysis further revealed that breast, bowel, and lung cancer are the top three cancer types that contact the service and require support. Moreover, the newly diagnosed patients are among the highest to seek support via CCV support service, along with the patients who are in the advanced stages of cancer. The emotion profiles highlight that trust was the most prominent positive emotion expressed at the call level, but many callers also expressed disgust and sadness. See figure 5.10.



Fig. 5.10 Call volumes by positive and negative emotions for the years 2018-2021

In terms of call duration, a high volume of calls less than 15 minutes expressed sadness, whereas longer calls over 15 minutes contained more expressions of disgust (figure 5.11).

The emotion transition analysis revealed that trust and joy emotions demonstrated the overall largest increase, while the four negative emotions all decreased over the call duration. This is an initial indication that the Cancer Council call service provision is effective in

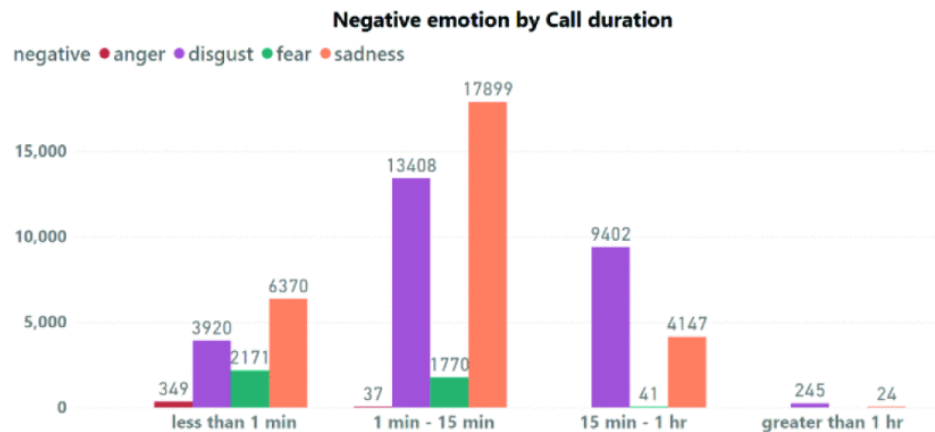


Fig. 5.11 Emotion profiles by Call duration and four negative emotions

alleviating the emotional distress at the start of the session. It was also noted that surprise and anticipation decreased significantly; again, this can attributed to the service quality of CCV. Drawing on the multi-granular composition of the emotion transition, we compared the emotion transition for nurse vs caller (figure 5.12). This comparison revealed that sadness and sympathy expressed by the nurse increased over the span of the call while the intensity of trust expressed by the caller increased over the call duration. This insight highlights the impact of sharing the burden of grief and the importance of training and counselling for nurses in the frontline of cancer care.

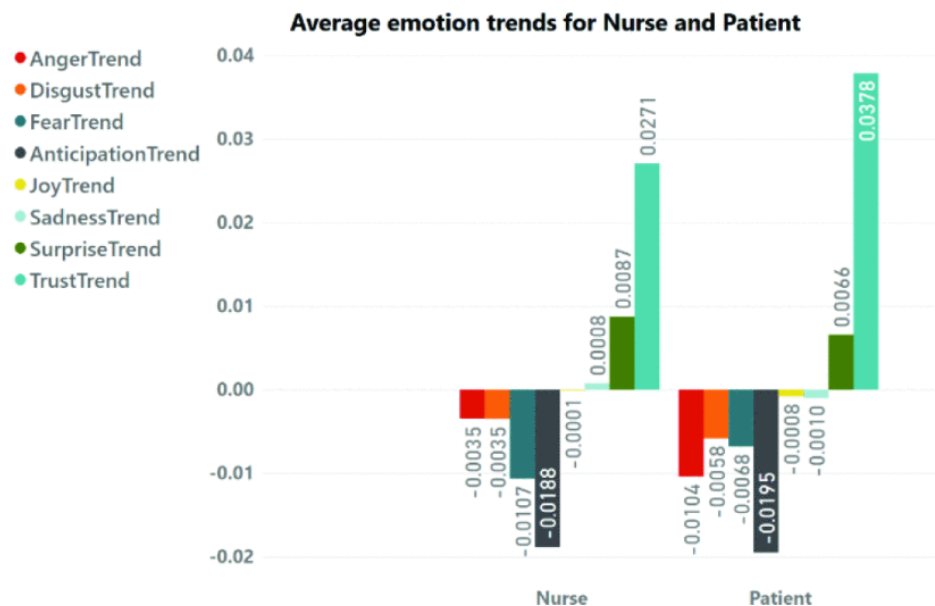


Fig. 5.12 Average emotion trend variation for the patient and the nurse

In this ablation study, we have proposed a novel AI-based framework for the detection of patient emotions and emotion transitions based on call recordings of telehealth services utilizing the Emotion AWARE algorithm. We have evaluated the framework on a large corpus of 60,000 calls and the results confirm its capability and output, which can also be visualized and interrogated by healthcare practitioners using interactive dashboards. These capabilities contribute towards service quality improvement as well as the SROI goals of the research program.

5.3.2 Analysing the Role of Manager Emotions on Stock Returns

Introduction

Humans are driven not only by logic but also by emotions. Drawing insights from social effects of emotions, in this ablation study we examine how investors react to manager emotions portrayed during earnings calls. Prior research provides evidence that disclosure tone (Zhang *et al.*, 2012; Price *et al.*, 2012), managerial affective states (MAYEW and VENKATACHALAM, 2012), deceptive speech markers (Larcker and Zakolyukina, 2012), and predetermined scripts (Lee *et al.*, 2015) are informative features of earnings calls. Researchers agree that earnings calls convey important value-relevant information. During these calls, managers have an opportunity to explain quarterly results in a relatively unconstrained manner, provide more details on firm past and future performance, and address analysts' questions (Matsumoto *et al.*, 2011). We focus on manager emotions conveyed in the earnings conference calls. More specifically, we investigate whether manager emotions convey relevant information to market participants and examine the association between emotions and investor response. Here we define emotions as discrete, intense, but temporary experiences elicited by a particular stimulus that often include action sequences (Repenning *et al.*, 2021).

We depart from the premise that emotions are part of all human actions. They can be helpful, harmful, influence behaviors or guide decisions. Emotions matter because individuals are not isolated 'emotional islands' (Barsade and Gibson, 2007). Instead, emotions bring attachments between actors; take different forms (e.g., anger, sadness, joy, surprise) and exhibit varying degrees of intensity (Baxter *et al.*, 2019). The psychology literature has established that human emotions exert influence on the receivers' behavior, generating social effects (van Kleef and Côté, 2022). There is consensus in this literature that "emotional expressions influence observers' affective reactions, inferential processes, and behaviors across various domains" (van Kleef and Côté, 2022).

Emotions may lead managers to express themselves and behave in ways that will affect market actors' perceptions regarding the potential future value of the company. We argue that these perceptions will drive investors' trading decisions, which will be reflected in the

firm's stock returns. Drawing insights from the theory of emotions as social information, we hypothesize that investors react to the managers' emotions exhibited in the earnings calls. Scholarly consensus has been built around two social consequences of emotions. First, emotion profiles entail reactions (Sutton, 1991). For example, managers conveying negative emotions may cause the same negative emotions among investors. Second, observed emotions and/or emotion changes may be taken as cues of deeper motives (van Kleef *et al.*, 2004). This consequence suggests that investors could use negative emotions portrayed by the CEO as warning signals or red flags to stay away from the investment opportunity. Alternatively, investors could be misled by managers' positive emotions even if the firm is underperforming. We, therefore, explore how investors react to the emotion profiles exhibited by managers at the earnings calls. Prior literature suggests that stock returns capture the average change in investors' beliefs following an event (Ball and Brown, 1968; Bamber *et al.*, 2010). Hence, we use cumulative abnormal returns measured over different windows around the earnings call date to capture investor reaction.

Methodology

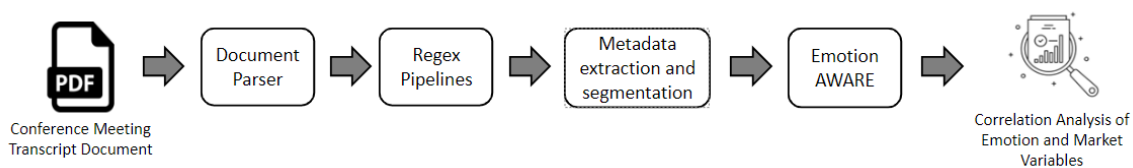


Fig. 5.13 Pipeline for analyzing the impact of manager emotions on stock returns

Pipeline for the managerial emotion analysis is provided in the figure 5.13. Here as the dataset sample we collect quarterly earnings conference call transcript data from EIKON Refinitiv Thomson Reuters for S&P 500 firms during the period 2010-2020. The database maintains a comprehensive set of transcripts for earnings calls with relevant details, including date, unique company identifiers, and verbatim transcripts. It also separates the conference call remarks by participant, section, and order of the comments. We focus on S&P 500 firms because it allows us to obtain a sample with a reasonable number of conference call transcripts. We exclude firms operating in financials (SIC 6000-6999) and utilities (SIC 4900-4949) industries. We also exclude observations from exchange codes 10, 11 and 12. We obtained 6,875 transcripts, which are available in PDF format, with the following information: company name, earnings quarter, conference call date, corporate participants (e.g., CEO, CFO) and other participants (e.g., analysts), and the Presentation and Q&A sections with each participant's name, position, and related text of their talk.

The pipeline begins by parsing the documents to extract their text content. Once extracted, the text is processed using a set of regex expressions to extract key metadata and segment the transcript into different sections. The extracted metadata includes the company name, conference quarter and date, as well as a list of participant names and roles. During segmentation, the transcript is divided into presentation and Q&A sections, and each participant's subsegments are separately collected for further emotional analysis. For the emotion analysis, the Emotion AWARE(Gamage *et al.*, 2024) algorithm is applied, utilizing an 8-primary-emotion setting to analyze the emotions present. We used the Presentation and the Q&A sections of the earnings conference calls to identify managers' emotions. The presentation section generally contains a small number of larger text segments by the CEO and Other Managers, whereas, in the Q&A section, there are many smaller text segments from the CEO, Other Managers and Analysts. We extract emotion profiles on these text segments by splitting each segment spoken by an individual into sentences, and we construct emotion profiles for each sentence and then normalized the emotion profiles for all sentences of a participant in each section for further analysis.

For the correlation analysis, we obtained stock return quarterly data from CRSP(Center for Research in Security Prices, 2024), financial data from Compustat(S&P Capital IQ, 2024), and analyst forecast information from I/B/E/S(Refinitiv, 2024). We matched firms based on company identifiers. After removing observations for missing financial variables and stock prices, we have a final sample of 6,165 firm-quarter observations. Table 5.16 and table 5.17 show the sample distribution by quarters and by industry. We do not observe significant industry clustering. Our sample contains a relatively great number of firms from the manufacturing and construction industries.

Table 5.16 Sample distribution by quarters

Quarters	Number Obs.	Percentage
2010Q1-2011Q4	1,154	18.65
2012Q1-2013Q4	1,228	19.84
2014Q1-2015Q4	1,230	19.87
2016Q1-2017Q4	1,288	20.81
2018Q1-2020Q4	1,317	21.28
Total	6,217	100.00

Table 5.17 Sample distribution by industry

SIC Code	Industry Type	Number Obs.	Percentage
1-999	Agriculture, Forestry, Fishing	4	0.06
1000-1999	Mining, Building	402	6.50
2000-2999	Construction	1,342	21.32
3000-3999	Manufacturing	1,906	30.76
4000-4899	Transportation, Communication, Electric, gas	462	7.45
5000-5999	Wholesale, Retail	865	13.95
7000-7999	Hotels, Services	918	14.76
8000-8999	Services	288	4.65
9000-9999	International, Non-operating	30	0.45
Total		6,217	100.00

We employ the following empirical regression model to examine the market reaction to manager emotions:

$$\begin{aligned}
 CAR_{i,t} = & \alpha + \beta_1 EMOTION_{i,t} + \beta_2 SIZE_{i,t} + \beta_3 MBTS_t + \beta_4 RETVOL_{i,t} \\
 & + \beta_5 UNEXP_EARN_{i,t} + \beta_6 LEV_{i,t} + \beta_7 LAG_{i,t} \\
 & + \beta_8 FORE_NUM_{i,t} + \beta_9 FORE_DISP_{i,t} + \beta_{10} LOSS_D_t \\
 & + \text{Year effects} + \text{Industry effects} + \varepsilon_i
 \end{aligned} \tag{5.7}$$

CAR is the cumulative abnormal return three days around the conference call date (-1, +1) that starts a day preceding the conference call and ends a day following it. The abnormal returns are calculated based on the Fama-French three-factor model calibrated on the estimation window of 90 days from t_{-120} to t_{-31} days. We calculate emotions using Emotion AWARE for each manager and for each one of the eight primary emotions (anger, fear, sadness, disgust, joy, anticipation, trust, surprise). These emotions can also be grouped into positive emotions (joy, anticipation, trust, surprise) or negative emotions (anger, fear, sadness, disgust).

Following prior studies, Equation 5.7 controls for other quantitative and qualitative information available to investors around the earnings announcement date and is associated with the market reaction to the call (Lee *et al.*, 2015). First, we control for size, growth, and risk, which have been shown to be related to market returns. Firm size (SIZE) is the natural logarithm of the market value of equity at the end of the current quarter; firm growth (MBTS) is the market-to-book ratio, calculated as the market value of shareholders' equity at the end of the current quarter scaled by the book value of equity; and firm risk is the return volatility

(RETVOL), measured as the standard deviation of daily stock returns over the 125 trading days prior to the earnings announcement. We also control for the market reaction to current-period earnings by including the magnitude of unexpected earnings (UNEXP_EARN), with expectations based on the last summary consensus median analyst forecast prior to the earnings conference call. Leverage (LEV), measured by total liabilities divided by total assets, controls for increased information demand when firms are experiencing financial distress. We capture the uncertainty of firm operations (FORE_DISP) using the standard deviation of analysts' earnings forecasts scaled by the most recent stock price. The model also includes the reporting lag (LAG); the number of annual earnings forecasts (FORE_NUM); and an indicator variable equals one if firm i reported negative income before taxes in at least one of the most recent four quarters preceding firm i 's interview at time t , and zero otherwise (LOSS_D).

Empirical Results

Samples of emotions extracted are provided in table 5.18. For example, in the following CEO statement: *"We are very, very pleased with Galveston Bay"* the keyword is 'pleased'. The top emotion detected by our algorithm is joy, with a 0.95 score, followed by anticipation, trust, and surprise, with scores of 0.03, 0.01 and 0.01, respectively. We do not identify other emotions in this statement.

Table 5.18 Emotion Classification of CEO Statements during the Q&A

Emotions	CEO Statements during the Q&A	Keyword	Emotion Dictionary
Anticipation	What I can tell you is that, of course, we expect the momentum of Luxury and Professional Beauty to continue in 2018.	expect	{ 'Surprise': 0.01, 'Anticipation': 0.63, 'Trust': 0.10, 'Joy': 0.22, 'Sad': 0.02, 'Fear': 0.04 }
Anger	There's been a tremendous amount of frustration from the technician to get himself organized.	frustration	{ 'Sad': 0.18, 'Fear': 0.10, 'Anger': 0.69, 'Disgust': 0.04 }
Joy	We are very, very pleased with Galveston Bay.	pleased	{ 'Surprise': 0.01, 'Anticipation': 0.03, 'Trust': 0.01, 'Joy': 0.95 }
Surprise	It is very, very unusual for us to not get the win on product.	unusual	{ 'Surprise': 0.42, 'Anticipation': 0.04, 'Trust': 0.01, 'Joy': 0.14, 'Sad': 0.12, 'Fear': 0.11, 'Anger': 0.11, 'Disgust': 0.06 }

Section	Stat	Joy	Antici.	Trust	Surpr.	Anger	Sad	Fear	Disg.
CEO Presentation	Mean	0.559	0.136	0.133	0.031	0.014	0.038	0.021	0.013
	SD	0.142	0.036	0.038	0.011	0.009	0.022	0.013	0.008
CEO Q&A	Mean	0.519	0.104	0.134	0.034	0.026	0.071	0.036	0.024
	SD	0.129	0.028	0.036	0.011	0.012	0.030	0.016	0.011
Managers' Pres	Mean	0.524	0.134	0.149	0.036	0.022	0.061	0.037	0.020
	SD	0.077	0.023	0.030	0.009	0.009	0.022	0.013	0.008
Managers' Q&A	Mean	0.531	0.108	0.140	0.035	0.028	0.076	0.039	0.027
	SD	0.076	0.017	0.022	0.007	0.009	0.021	0.012	0.008

Table 5.19 Emotion Statistics for Different Sections of the Transcripts

Table 5.19 presents the summary statistics. Here, we report the emotional scores of the CEO and the management participants during each of the sections of the conference call. In the Presentation section, CEOs and Other Managers display more positive than negative emotions. For example, CEOs present earnings disclosures using emotion words associated to Joy, Trust, and Anticipation. Nonetheless, other managers exhibit higher scores of negative emotion words related to sadness and fear than CEOs. In the QA section, positive (negative) emotions tend to decrease (increase) for both management participants. When CEOs and Other Managers answer analysts' questions, they continue to use joy emotions, but they also exhibit higher scores for anger and disgust emotions.

Table 5.20 Dependent and Control variables of the analysis

Variable	Mean	SD	P25	P50	P75
CAR	0.001	0.059	-0.032	0.002	0.036
SIZE	9.691	1.058	8.995	9.555	10.321
MBTS	-1.337	0.854	-1.731	-1.271	-0.809
RETVOL	0.018	0.009	0.012	0.016	0.020
UNEXP_EARN	-0.001	0.029	-0.001	0.000	0.002
LEV	0.611	0.167	0.496	0.604	0.732
LAG	60.935	15.617	54.000	61.000	67.000
FORE_NUM	18.159	7.144	13.000	18.000	22.000
FORE_DISP	0.120	0.258	0.030	0.050	0.120
LOSS_D	0.029	0.168	0.000	0.000	0.000

In table 5.20, we present the statistics for the dependent and control variables. The sample firms have an average (median) quarterly cumulative abnormal return (CAR) of 0.001 (0.002). The mean (median) firm has a size of 9.69 (9.55) and firm growth of -1.33 (-1.27). The average of return volatility and unexpected earnings is 0.018 and -0.001, respectively.

Firm leverage is 61% on average, and the reporting lag is 61 days. Our sample firms have an average of 18 annual earnings forecasts, a forecast dispersion of 12%, and 2.9% of them have reported an earnings loss.

Table 5.21 Correlation matrix of selected variables with statistical significance levels

	CAR	CEO_PR_P	CEO_PR_N	CEO_QA_P	CEO_QA_N	SIZE	MBTS	RETVOL	UNE_ERN
CAR	1.000	0.035*	-0.072**	0.027**	-0.025	0.001	-0.055**	-0.014	0.055***
CEO_PR_P	0.035*	1.000	0.090***	0.077***	0.373***	0.031*	-0.017	0.001	-0.007
CEO_PR_N	-0.072**	0.090***	1.000	-0.020	0.395***	-0.077**	0.100***	0.032*	0.016
CEO_QA_P	0.027**	0.077***	-0.020	1.000	0.302***	-0.140**	-0.025*	0.055***	0.018
CEO_QA_N	-0.025	0.373***	0.395***	0.302***	1.000	0.000	0.065**	-0.229***	0.056**
SIZE	0.001	0.031*	-0.077**	-0.140**	0.000	1.000	-0.266**	0.192**	0.046**
MBTS	-0.055**	-0.017	0.100***	-0.025*	0.065**	-0.266**	1.000	-0.014	-0.074***
RETVOL	-0.014	0.001	0.032*	0.055***	-0.229***	0.192**	-0.014	1.000	-0.001
UNE_ERN	0.055***	-0.007	0.016	0.018	0.056**	0.046**	-0.074***	-0.001	1.000

Table 5.21 presents the correlation coefficients between the variables in our study. We find that CEO positive emotions, such as joy and trust, are positively correlated with CAR, while negative emotions, such as sadness and fear, are negatively correlated with CAR. Interestingly, the correlations seem stronger in the Presentation section compared to the QA section. For example, the correlation coefficient between CAR and CEO_PRES_POSITIVE is 0.035 (significant at 5%), whereas the correlation coefficient between CAR and CEO_QA_POSITIVE is only 0.027 (significant at 10%). Moreover, we correlate our eight emotion variables and find mostly positive and significant correlations between them. In untabulated results, we find that for the CEO, anger has a positive correlation with joy (16%), sadness (91%), and disgust (86%). These correlation results are similar to those reported by Akansu et al. (2017), using a sample of CEOs in Fortune 500 companies, which indicate that anger has a positive and significant correlation with happiness (36%) and with disgust (22%). Consistent with their study, we also report a positive correlation between joy and sadness of 21%. While these strong correlations between positive and negative emotions might be surprising, research in psychology suggests that mixed emotions are increasingly prevalent when individuals face difficult decisions with conflicting goals or objectives (Kahneman, 1992). CEO negative emotions are strongly correlated with small firm size, high firm growth and high firm risk.

5.4 Chapter Summary

This Chapter explains the application of neuro-symbolic architectures in emotion elicitation with the novel Emotion AWARE algorithm. The chapter starts by highlighting the significance of emotions as an aspect of language comprehension and analysing the existing computational methods of emotion elicitation in the literature. Then, it describes the Emotion AWARE methodology in detail, along with the empirical evaluations with nine experiments with several standard benchmark datasets. The chapter concludes by discussing two ablation

studies of Emotion AWARE, highlighting the practical usage of the framework in real-world use cases.

Chapter 6

Neuro-symbolic AI for Natural Language Interactions

6.1 Chapter Introduction

We are steadily moving towards a digitalized world where digital interactions are becoming more frequent than human interactions. Today, we are surrounded by various digital devices and applications, and the interaction with digital interfaces is becoming increasingly integral to daily life. However, with the complexity of user requirements growing alongside technological innovations, these human-computer systems are advancing rapidly. Novel computer systems cover a wider range of use cases, features, and components. This expansion is reshaping how we interact with technology, making it more embedded in our everyday activities and pushing the boundaries of what digital systems can achieve in various aspects. With the growing complexities of human-computer systems, masking their complexities and configurations to provide simple and user-friendly access to the end users is a challenging task. In this premise, the ability to seamlessly communicate with systems through natural language is a substantial advancement, enhancing both user experience and operational efficiency.

The development of chatbots or intelligent agents has become an increasingly discussed topic today, especially with the emergence of LLMs. Generally, chatbots can help enhance human-computer interaction by facilitating various tasks ranging from retrieving information from knowledge bases to automating tasks based on natural language commands. When considering the literature, chatbot development has evolved and improved over the time. In the early stages, these systems were limited to simple string-matching techniques based on the predefined QnA datasets. In these systems, once a user query is presented, it finds the most

matching questions in the knowledge base and emits the corresponding response. Although these heuristic approaches have higher controllability and transparency, they were limited by pre-defined knowledge and scope of language variations. In contrast, recent chatbots have avoided these limitations with the language understanding and generation capabilities of LLMs to maintain more natural interactions with computer systems. However, lack of controllability, explainability and dependencies on vast datasets that may not include the most recent information limits their practicality and adaptability in dynamic environments.

In this context, neuro-symbolic AI can be highlighted as a transformative approach to consolidate systematic, rule-based reasoning capabilities of symbolic AI with the learnability of complex neural architectures. This hybrid methodology enhances the system's ability to understand and generate language more neutrally while improving the explainability and controllability of AI decisions, a crucial factor in sensitive applications. In this chapter, I introduce Cooee, a neuro-symbolic chatbot architecture specialized in energy environments. Cooee implementation will be discussed in three main components: integration of structured data spaces, integration of unstructured multi-modal data spaces, and multi-agentic implementation for task automation. All three of these systems are experimented and deployed in La Trobe Energy Analytics Platform - LEAP.

6.2 La Trobe Energy Analytics Platform - LEAP

La Trobe University in Victoria, Australia, is a multi-campus, multi-functional tertiary education institution dedicated to sustainability and innovation. The La Trobe Energy AI/Analytics Platform (LEAP) is a pioneering initiative aiming to achieve net zero carbon emissions across its campuses by 2029. With a substantial investment of \$75 million, LEAP seeks to reduce reliance on the Victorian state power grid's brown coal-based energy, increase the use of renewable energy from rooftop photovoltaic panels, and implement battery storage solutions. The platform employs advanced AI and analytics methodologies to optimize energy consumption and generation, representing a comprehensive approach to energy management.

Developed as a "Living AI Lab" project by the Centre for Data Analytics and Cognition (CDAC), LEAP incorporates a range of innovative technological components. These include an all-data lakehouse, a data quality framework, an AI and automation engine, and multi-stakeholder analytics dashboards. This architecture provides an intelligent and cohesive abstraction to assist in cataloging, unifying, and managing the complexities of microgrids. The efficient management of energy is crucial, especially in the face of volatile global energy markets and the increasing emphasis on sustainable energy use, coupled with legislative frameworks aimed at reducing carbon footprints.

AI-based smart systems are transforming the future of automation and operational intelligence, and energy management is no exception. The proliferation of strategies for smart microgrids, alongside the emergence of the Internet of Things (IoT) and Edge Computing, enables fine-grained analysis and understanding of energy consumption and disbursement within organizations. Smart grids (SGs) facilitate intelligent energy distribution and management, considering the behaviors and actions of all stakeholders. LEAP's cloud-edge architecture leverages AI and data analytics for microgrid energy optimization and net zero carbon emissions, providing a sustainable and efficient energy management framework for La Trobe University.

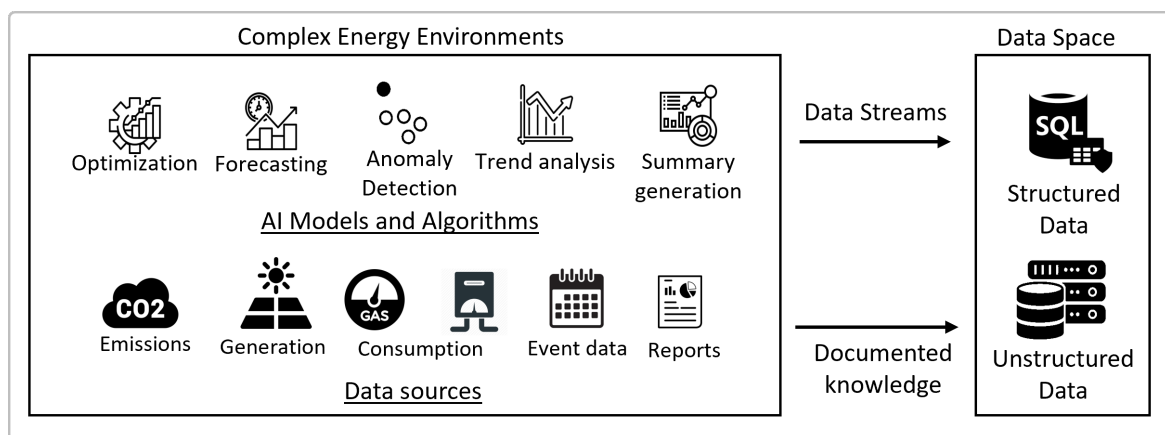


Fig. 6.1 La Trobe energy analytics platform(LEAP) Data Space

This figure 6.1 provides an overview of the data space in the LEAP platform. Energy data streams encompass a wide range of elements, including consumption measures such as gas and electricity, emissions such as CO₂, and generation measures such as renewable energy sources. They also factor in events that affect energy usage, such as weather conditions, climate variations, and scheduled works or outages. The extraction of latent representations from these multimodal energy data streams, combined with the optimization of grid operations and automated control of sources, has led to complex and diverse AI use cases in the energy sector. This also contributes to a real-time and highly dynamic data environment.

Information and data in these spaces, gathered from various sources, are primarily stored in two forms: structured and unstructured. Structured data, typically housed in relational databases, consists mainly of numerical values related to the consumption, generation, and emissions of different entities within the environment. Unstructured data encompasses text and images, including documented knowledge such as periodic summary reports and notifications about events and outages, as well as other documented information such as definitions, processes, and FAQs. The efficient management and analysis of both structured

and unstructured data enable LEAP to enhance energy optimization and achieve its sustainability goals. Given the complex and dynamic nature of the data space, it is crucial to build systems that can interact with these data spaces and utilize this data for retrieval, analysis, and decision-making processes.

6.3 Interacting with Dynamic Structured Data Spaces Using Natural Language

6.3.1 Introduction

The structured data space consisted of multiple data sources. Each of these sources represents a different aspect of the energy environment. These sources collect data on three main dimensions of energy: emissions, generation, and consumption. Emissions data measures carbon dioxide and other greenhouse gas emissions, while generation data tracks energy produced from various sources, including renewable sources such as solar. Consumption data provides information about the energy usage of various entities and buildings in the university. This raw data is organized and stored in relational databases, resulting in multiple tables with high-granularity data streams that usually represent time-series data. These databases allow for efficient querying and analysis, forming the backbone of the structured data infrastructure. In addition, these raw data undergo several machine learning algorithms, covering different secondary tasks such as optimization, forecasting and anomaly detection on a periodical basis in the energy environment. These tasks generate secondary and latent representations of the raw energy data, which capture deeper insights and patterns not immediately visible in the raw formats.

With all these sources, currently, LEAP structured data space consists of over 1 billion data points, streaming at 15-minute intervals to capture the most granular level of detail. These energy measures are spanned across 6 different campuses of La Trobe University in Victoria. The Melbourne Bundoora campus, being the largest, generates the majority of the data streams. In addition to the Bundoora campus, the data space includes the Melbourne city campus and four regional campuses in Bendigo, Albury-Wodonga, Mildura, and Shepparton.

While the LEAP platform presents advanced analytical dashboards that offer significant insights and summaries of the structured data, retrieving specific measures can sometimes require navigating multiple pages with filtering. This can be a challenging task, especially for users unfamiliar with the system. Additionally, although these graphs and visuals effectively summarize and provide data overviews, they are configured to show a certain view of data, which limits its customizability. Thus, the capability to interact with the structured

data space using natural language is a valuable feature that enhances user-friendliness and customizability. By allowing users to retrieve various energy-related features of different entities across diverse temporal durations through natural language queries, the platform can significantly improve accessibility and flexibility.

Before the LLM era, traditional chatbots primarily functioned on predefined question-answer pairs or defined routines. Most of these techniques were based on string-matching methods, which map user questions to pre-constructed answers. These chatbots are usually fit for basic tasks like answering FAQs or handling well-defined routines such as placing orders or making reservations. However, when we are interacting with dynamic environments that require communicating with data sources to generate answers, it is not possible to leverage these static approaches. For instance, a query like, *“What are the total carbon emissions of building X in the last 24 hours, and how does this compare with the organizational baseline?”* requires contextualization to generate an accurate, real-time response. Conventional chatbots based on intent classification struggle in such scenarios because they cannot construct responses dynamically.

On the other hand, chatbots based on LLMs use prompt engineering to generate SQL statements from user queries. While this approach offers more flexibility, it also poses significant challenges. LLMs can hallucinate, generating syntactically incorrect or harmful queries, which is particularly risky when interacting with databases. Therefore, granting LLMs too much flexibility is not advisable. However, LLMs can still be leveraged in controlled ways to enhance chatbot capabilities. Given the closed energy management domain and the limited variations of possible SQL queries, neuro-symbolic architectures are ideally suited for this scenario. These architectures combine the strengths of symbolic AI and neural networks, providing a robust solution for generating accurate and contextual responses in real-time.

6.3.2 Cooee-internal - Neuro-symbolic Architectures for Interacting with Structured Energy Spaces

Cooee-internal is a conversational agent implemented within the LEAP platform to handle energy-related queries by communicating with the structured data space for developer use. The high-level architecture of Cooee-internal is illustrated in figure 6.2. Heterogeneous internal and external data sources feed multimodal energy data streams into both standard data storage (linear) and AI algorithms and models (latent). The AI module builds latent representations, predictions, associations and segments of energy consumption, generation, and management. Cooee consists of 4 main modules, namely Chatbot Interface, Natural

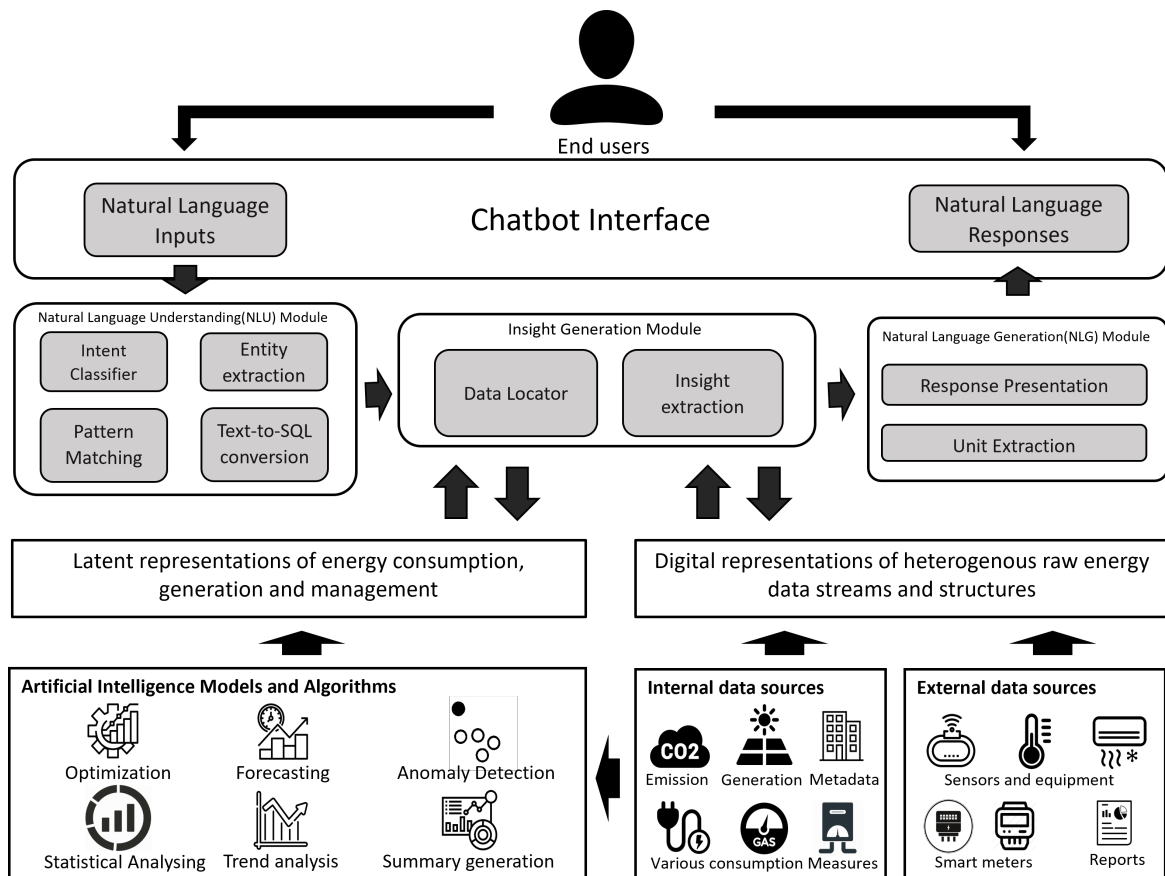


Fig. 6.2 High-level architecture of Cooe internal chatbot

Language Understanding (NLU) Module, Insight Generation Module, and Natural Language Generation (NLG) Module.

Chatbot Interface Module

The Chatbot Interface Module is the front end for all end-user interactions within the LEAP platform. This module includes a chat popup and a Q&A dashboard (figures 6.3 and 6.4). Users can type in their queries or click through recommended questions to find answers and ask follow-up questions. The responses from Cooe can have three main types: single values, a list of values, and detailed tabular responses. The main intention of the chat popup implementation is to provide easy access to users from anywhere within the LEAP platform. The Q&A dashboard is designed to handle complex tabular responses, offering a more comprehensive Q&A experience.

Building	Total Energy Consumption (kWh)	Estimated Total Cost (AUD)	Peak Energy Consumption (kWh)
Library	122794.88	13093.46	10141.08
Donald Whitehead Building	65242.06	6897.97	2364.19
LIMS 1	158577.52	16661.66	7902.79

Your query is interpreted as : SELECT Building, Total Energy Consumption, Estimated Total Cost, Peak Energy Consumption WHERE Building IN ('Library','Donald Whitehead Building','LIMS 1') from 2021-04-01 to 2021-04-30

Try Asking...

- What is the maximum consumption in library in last month?
- What is the total energy consumption of donald whitehead today?
- What is the estimated cost of whitehead this year?
- What is the solar generation of library and lims1 april 2021?
- What is the lowest consumption in library in 2021?
- What is the average consumption of library in this year?

Fig. 6.3 Cooee-internal Q&A Dashboard Implementation

ASK COOEE

Hi! Its COOEE here..How can I help you?

LEAP Bot • 12:56 PM

what is the peak energy consumption of library in 2020 April?

User • 12:56 PM

10209.7 kWh

Your query is interpreted as SELECT PEAK ENERGY CONSUMPTION of LIBRARY in 2020-04

LEAP Bot • 12:56 PM

Enter your message here

Fig. 6.4 Cooee-internal Chat Pop-up Implementation

Natural Language Understanding(NLU) Module

All the user queries will be passed through the Natural Language Understanding (NLU) Module, which is directly connected to the chatbot interface module. It utilizes both LLMs and symbolic approaches to fully interpret user inputs. The NLU Module comprises four components: an entity extractor, an intent classifier, a pattern matcher, and a text-to-SQL converter. The Entity Extractor is responsible for identifying key entities within user questions, regardless of their position in the query. In LEAP, the entity recognition component of Cooee was trained to extract three types of entities, namely units/buildings(BLD), attributes (ATR),

and temporal information (TT). Table 6.1 provides further details on these entities. Additionally, in complex cases, such as groups of entities present, it can capture these information phrases and decouple them to identify specific information.

Table 6.1 Entity Recognition Classes in NLU Module

Entity Class	Description
BLD	Buildings and units of the university (110 items). Ex: Donald Whitehead Building, Library, LIMS1 Building.
ATR	Energy-related attributes in the LEAP platform. Examples include Maximum-peak-consumption, estimated-total-cost, energy-consumption year-to-date (YTD), etc.
TT	Temporal information. Can be relative or absolute. <i>Relative</i> : last week/month/year, last Friday, yesterday, etc. <i>Absolute</i> : 2022/03/17, April 2021

In order to implement the entity extractor module, we utilized the BERT model(Devlin *et al.*, 2019), which is a lightweight encoder-based LLM followed by a rule-based rectification process to guard rail the outputs. The BERT model was initially fine-tuned with a sequence-to-sequence classification for entity recognition, assigning labels to each token or word processed by the model. We used a custom-made dataset based on LEAP for fine-tuning, which included 7,500 automatically constructed and annotated questions with ATR, BLD, and TT tags using templates. The fine-tuning process utilized the BERT model with a token classification head from the transformers library. The sentence size was 128, and padding and truncation were applied appropriately. The learning rate was set as 1e-05. Figure 6.5 illustrates an instance of the entity extraction component.

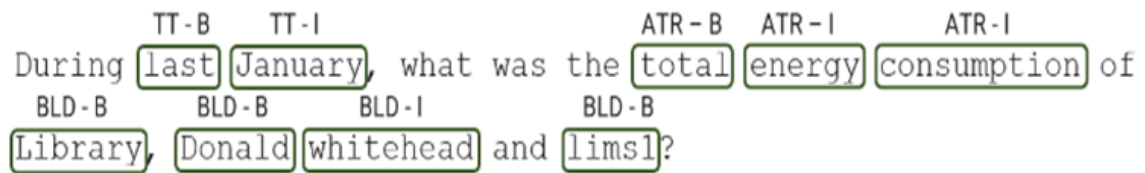


Fig. 6.5 Instance of Entity Recognition in a Sample query

After recognizing the entities, they were forwarded to a rule-based rectification process to match the exact values in the structured data space. Temporal entities were processed through Python's datetime packages and other rules to normalize them into a standard format suitable for SQL queries. For attributes and entity names, an embedding-based matcher was used to map them to the exact entities present in the structured data space. Predefined lists

for entities and attributes were utilized along with BERT embeddings, and the extracted entity was replaced by the most closely matching predefined entity based on cosine similarity. Additionally, some entities and attributes were mapped using commonly used synonyms. This approach provides both good controllability and flexibility, allowing the system to capture different forms of entities, attributes, and temporal features in the query processing task.

Cohee uses intent classification to forward different types of queries to corresponding processing units. This is based on the extracted entities. Currently, it supports two types of intentions: simple data retrieval (list and single value responses) and complex data retrieval (tabular responses) queries. The pattern-matching component and text-to-SQL module utilize rule-based methods to generate accurate and syntactically correct SQL statements. Based on the intent information and extracted entities, queries are compared against a set of regexes to find the patterns in the question. If the user query matches with a regex, then it will be mapped to a template, which can be processed by the Text-to-SQL module. For an instance, "What is the [attribute] of [building] in [Period]?" will be mapped to *SELECT SUM([attribute]) FROM TABLE WHERE building = '[building]' AND DATETIME IN [Period]*. If the natural query does not match with the regexes, based on the entity extraction data, it will automatically map to a template.

This rule-driven nature ensures that SQL statements adhere to specified patterns, safeguarding the system by preventing the creation of unexpected or harmful SQL statements. As this component directly communicates with the LEAP production database, maintaining strict control over the SQL generation process is crucial. Additionally, leveraging LLMs for entity extraction ensures that a wide range of entities and their various forms in user queries are accurately mapped to the corresponding columns and features in the relational data space. Collectively, integrating symbolic and connectionist approaches enhances system productivity while ensuring trustworthiness.

Table 6.2 Regex classes and their corresponding standard templates

Regex type	Question patterns	SQL patterns
Querying about building attributes	#Sample pattern a: What is the [attribute] of [building] in [Period]? Ex: What is the total energy consumption of health sciences building in 2022?	SELECT SUM([attribute]) FROM TABLE WHERE building == [building] and DATETIME IN [period]
Continued on next page		

Regex type	Question patterns	SQL patterns
Querying about building and attribute aggregations	#Sample pattern a: What is the [aggregator] [feature] in [Period]? #Sample pattern b: Name the building with [aggregator] [feature] in [Period]? Ex: What is the maximum total estimated cost in this year? Name the building with lowest energy consumption in last month?	SELECT [aggregator]([attribute]) FROM TABLE SELECT [aggregator]([attribute]) FROM TABLE WHERE DATETIME IN [period] SELECT [building] FROM TABLE WHERE attribute == [aggregator]([attribute]) AND DATETIME IN [period]
Entity comparison	#Sample pattern a: What are the buildings with [comparator] [value] of [feature] in [Period]? #Sample pattern b: What are the buildings with [attribute] [comparator] [value] in [Period]? Ex: What are the buildings with more than 2000 kWh of energy consumption this week? How many buildings with less 25000AUD of total cost last month?	SELECT [building] FROM TABLE WHERE attribute [comparator] [attribute] AND DATETIME IN [period] SELECT COUNT([building]) FROM TABLE WHERE attribute [comparator] [attribute] AND DATETIME IN [period]

Insight Generation Module

This module communicates with the complex data space and generates insights in response to user queries. Once the natural language queries are converted to standard SQL statements, these can be processed by this module. The first task of this module is to map structured statements to corresponding latent and linear data representations. This is done by comparing target attributes in the queries and data table columns using fuzzy string matching. Once it locates the correct subset of data in the space, it will extract insights that consider parameters in the query, such as temporality, aggregations, and attributes. The module then responds to the natural language generation module with the corresponding data frames.

Natural Language Generation(NLG) Module

This module is responsible for building the responses. It uses information extracted from the NLU component and then introduces context and syntax to the answers extracted to provide a complete response. Additionally, it will create an interpretation of how Cooe understood the user query, which increases the transparency of the underlying process. The module also performs unit extraction to provide more comprehensive answers. Furthermore, the intent of the query is considered when crafting responses. For simple queries, responses are provided

in text format, while for tabular queries, the results are formatted as HTML tables to enhance readability.

Experiments

While the Cooee-internal chatbot offers important features such as explainability, controllability, and trustworthiness when communicating with production data spaces in the LEAP platform, we manually evaluated it using various use cases. However, assessing overall features such as controllability and explainability is not straightforward. Therefore, we focused on comparing the text-to-SQL performance of Cooee-internal with comparable approaches available at the time of its implementation.

Table 6.3 Comparing Cooee QnA capability with TAPAS and Text-to-SQL

Setting	TAPAS	T5-text-to-SQL	Cooee-QnA
With 92 buildings, 708 question-answer pairs	1.27%	60.25%	95.93%
With 30 buildings, 278 question-answer pairs	30.21%	57.91%	96.77%
With 12 buildings, 140 question-answer pairs	65.71%	62.14%	94.28%

For this experiment, we used the tabular Q&A method and sequence-to-sequence classification method, both based on small-scale LLMs similar to the BERT model, which has demonstrated high accuracy in benchmark tests. The evaluation utilized a data table containing six energy attributes for 92 buildings, where each row corresponds to a building and columns represent different attributes. We automatically constructed the question-answer pairs for the evaluation dataset using regex patterns (Gamage *et al.*, 2022). The experiment was split into three parts with increasing sizes of the data table to investigate performance scalability. For the table question answering method, we used TAPAS (Herzig *et al.*, 2020) with the Hugging Face table question answering pipeline. TAPAS (Herzig *et al.*, 2020) is a tabular Q&A model based on small-scale LLM, which is finetuned on multiple datasets (Zhong *et al.*, 2017; Pasupat and Liang, 2015). TAPAS receives as input a data table along with a set of natural language questions. The major limitation of TAPAS is its decreasing accuracy as the volume of data expands, and its limit of all data in a single table. With the sequence-to-sequence classification method, first, the natural language queries will be converted to SQL statements and then used to query the data tables (Guo *et al.*, 2019b; Xu *et al.*, 2017). In this way, the approach is independent of data table length or the number of tables. This is

based on another small-scale LLM named T5(Raffel *et al.*, 2019), which has been fine-tuned on annotated datasets(Yu *et al.*, 2018; Hazoom *et al.*, 2021). However, this method is not extensible or customizable and is also impacted by the lack of relevant annotated datasets.

The experiment was split into three parts with increasing sizes of the data table to investigate performance scalability. The results of this comparative evaluation are reported in table 6.3. As seen in the results, Cooee outperforms both query approaches by a considerable margin, a 35.68% margin for 92 buildings, 38.86% margin for 30 buildings and 32.14% for 12 buildings. These results confirm that although state-of-the-art approaches provide higher accuracies with benchmark datasets, they underperform in practical use cases. As elaborated in the results, TAPAS exhibits low accuracy with larger data tables. With a data table of 92 rows, the accuracy is 1.27%. The T5-text-to-SQL method is consistent across all three experiments but produces far lower accuracies compared to Cooee. Furthermore, in scenarios with multiple data tables, this method is less accurate as table mapping is a challenging task without entity extraction.

Improving Contextualized Question Answering

To enable chatbots to communicate in a human-like manner demands a deep natural language understanding as well as the capability to infer the meaning from the context(Caldarini *et al.*, 2022). As humans, we comprehend the language effortlessly as we often make several assumptions about the receiver's level of understanding (Zhou *et al.*, 2022; Kosinski, 2023). As a result, in human communication, language can often be incomplete and ambiguous, omitting some explicit details and letting the receiver fill these gaps. This makes human communication a more straightforward and efficient process. However, when humans communicate with computer systems, this nature can pose several challenges as these systems do not have the same level of background knowledge and inference capability as humans(Naous *et al.*, 2020; Liebrecht and van Hooijdonk, 2020). Here in these experiments, we focus on three main challenges that chatbots encounter while enabling more natural and human-like communication: Ambiguity in the language, handling incompleteness and resolving co-reference. To resolve these challenges in Cooee, we leveraged the recent decoder-based LLM GPT3.5 turbo(Brown *et al.*, 2020) merged with symbolic approaches to build contextualized Q&A pipelines.

Identify queries that require resolving the context

Initially, when a query is presented, it is passed to the NLU module for entity recognition and pattern identification, regardless of whether it is a contextual question. To provide accurate energy information, three dimensions are required: the energy attribute (ATR), the entity or building being referred to (BLD), and the temporal value expressing a datetime or a

period (TT). Queries that include all these dimensions can be independently answered by Cooee-internal. In cases where these dimensions are vague or incomplete and need context to be resolved, the queries are identified after entity recognition and pattern-matching tasks and forwarded to a specialized module. This module leverages contextual information to fill in the missing details.

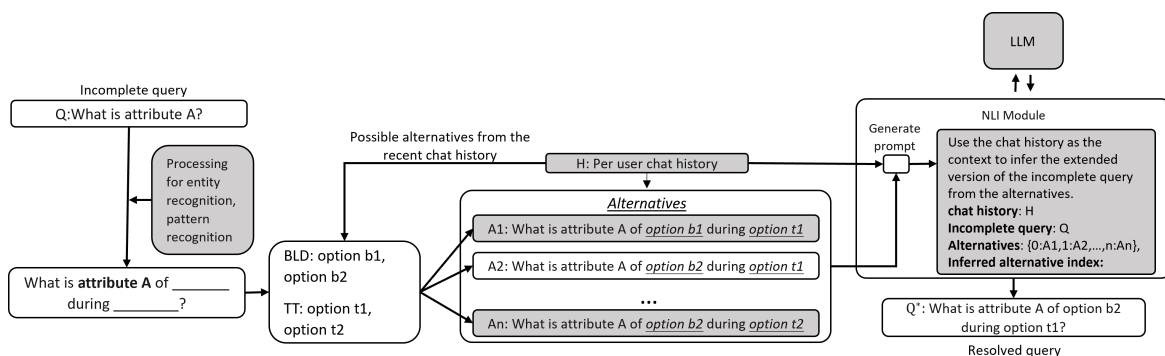


Fig. 6.6 Resolving Contextualized questions using LLMs

Utilizing GPT for Contextual question answering(CQA)

Here it will use chat history as an additional input to acquire the context. In order to facilitate this, per-user chat history will be recorded and stored in a database. In addition to that, it will have a Natural language inference(NLI) module which will communicate with ChatGPT to resolve questions that require contextual knowledge. Here firstly it will identify the missing or coreferring entities and extract suitable options from the recent chat history for them. Afterwards, these options will be used to construct complete alternative versions of the incomplete query.

Constructed alternatives, along with the chat history, will be used in the NLI module to populate an input prompt that can be forwarded to the GPT3.5 turbo Model. Here, the prompt contains a template to fill in with chat history, incomplete queries, and alternative choices. Furthermore, the prompt consisted of an output indicator to facilitate concise response generation. To ensure the deterministic behaviour of the model, the temperature parameter is set to a lower value, such as 0.2. LLM will select the best complete version of the query from the given options and emit the index of it back to the NLU module. This will be chosen as the rectified version of the input query and proceed to the Insight generation and NLG modules for answer extraction. A more elaborated flow diagram of how Cooee resolves vague and incomplete questions is provided in figure 6.6.

Experiments

This section explains three use cases of contextualized question answering with Cooee to resolve ambiguities, rectify incomplete questions, and co-reference resolution.

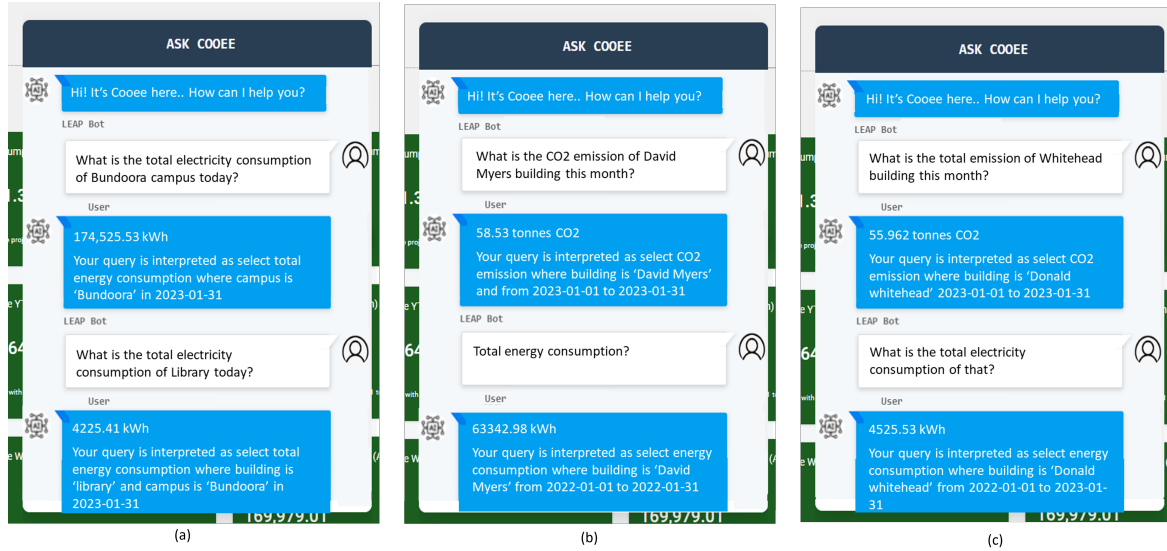


Fig. 6.7 Contextualized question answering examples

1) Resolve ambiguities

The retrieval of energy information from different entities in the LEAP environment can be ambiguous at times. For instance, when Cooee is functioning in multi-campus circumstances, asking the question "What is the energy consumption of the library today?" can be vague as it does not clearly say which campus library that is referred to. Such situations are typical in human communication and should be resolved to provide more accurate results to the user. During the chat, once the user has referred to a specific campus name, we cannot expect the user to include the campus name in every subsequent query. Additionally, in such cases, all dimensional data is available in the query "energy consumption"(ATR), "library"(BLD) and "today"(TT) which we can consider as a complete question. So here, we use prior chat history to resolve the ambiguous dimension. As the first step, this ambiguous entity will be mapped to n number of disambiguated dimensions such as "library(campus A), library(campus B), library(campus C)" extracted from the chat history. Then, the ambiguous dimension will be replaced with the options to construct n different disambiguated alternative queries. Then, each of these resulting queries, along with the user's chat history, will be forwarded to the NLI module to construct a prompt and get the query with the highest entailment probability via GPT. An example is shown in figure 6.7(a); the first user queries about the "Bundoora campus", which is one of the campuses in the LEAP environment. The subsequent query asked about the electricity consumption of the library which is an

ambiguous entity. During the contextualization, that is resolved and mapped to the library of the Bundoora campus to provide accurate answers.

2) Rectify incomplete questions

In human-to-human communication, some information is often intentionally omitted to make the conversation more simple. Since humans can infer such missing information by considering the context, comprehension can be achieved effortlessly even in the presence of incomplete information. But in conversational AI when each user query is treated as a discrete input, deducing omitted information posits a significant challenge. For an instance, the user first queries *"What is the attribute A1 of building B1 today?"* and gets the answers and then poses a follow-up question asking *"What is attribute A2?"*, it can be confusing to the chatbot, as the follow-up question does not contain all information to process the query.

Here the second query contains only an attribute value but does not contain building or date-time information. In such cases, similar to resolving ambiguities, user chat history is used along with natural language inference to infer missing entities to build a complete user query. Here possible options for each of the entities will be taken from the chat history. In case of two missing entities and n_1 and n_2 possibilities for each of those, $n_1 \times n_2$ complete alternatives will be considered and the one with the highest entailment probability will be chosen with the ChatGPT following the same pattern as explained in the figure 6.6 and forwarded to text-to-SQL module for further processing. Figure 6.7(b) depicts a scenario of answering an incomplete question. As the first question user queries about CO2 emissions of the David Myers building during this month and gets answers. The subsequent question is posed about energy consumption without explicitly specifying a building or a time frame. Thus, in this case, those dimensions are resolved by considering the chat history.

3) Coreference resolution

In human dialogues, it's a common feature to employ co-referential expressions. These expressions often take the form of pronouns and are utilized to refer to entities that have been previously mentioned in the conversation. In such cases, identifying such pronouns and mapping them to corresponding entities is an important task. To achieve this we extended the entity extraction module to facilitate pronoun detection and map them to correct entities with the help of regex patterns. Then similar to previous use cases these entities will be replaced with the most probable entailment entities considering the chat history. An example of co-reference resolution is shown in figure 6.7(c). As with previous cases, here, first, it queries about emissions of Whitehead building this month and gets the results. Afterwards, it queries electricity consumption with the co-referring entity "that". Here, the subsequent

query matches the pattern of "What is [ATR] of [BLD]? and the pronoun is in the place of BLD. Thus, it will be replaced with possible options from the chat context as in the previous cases, and the corresponding building name with the highest probability will be chosen.

6.4 Interacting with Multi-modal Unstructured Data Spaces using Natural Language

6.4.1 Introduction

Unstructured data plays a crucial role in many environments. Unlike structured data, unstructured data comes from a variety of sources and exists in diverse formats, making it challenging to interact with uniformly. The primary types of unstructured data include text and images. Despite its lack of structure, unstructured data holds valuable information essential for numerous use cases.

In LEAP, we have access to various documents, including annual reports, periodical summary reports, operational and maintenance logs, and incident and outage reports. These data are important in various knowledge retrieval, decision-making as well as reasoning activities in the energy spaces. Moreover, as a university actively working towards net zero emissions by 2030, La Trobe University has numerous net zero initiative projects across all campuses. Most of this information is available on the university's blogs and website. This data is of interest to the general public, including students, staff, and other stakeholders interested in the university's initiatives. Additionally, users without a technical background might need explanations regarding specific terminologies and processes related to energy information. This data, found in these unstructured sources, can be carefully used to answer questions. On the other hand, unstructured knowledge is a valuable resource for identifying reasons behind various behaviours in energy usage. Event and outage notifications can be particularly useful for developers in such cases.

There are several approaches in the literature to retrieve knowledge from unstructured data sources. The most common approach is to utilize string-matching techniques to map documents or input queries to most similar unstructured text sources. Methods like FAQ chatbots or sequence-to-sequence classification using encoder-decoder language models represent some of the closest attempts at utilizing unstructured knowledge for Q&A. Other techniques, such as regular expressions, keyword and topic extraction, and document clustering methods like TF-IDF, cover certain aspects of information gathering but fall short of enabling intelligent Q&A based on these sources. As a result, it becomes challenging to fully leverage these methods for meaningful interactions in unstructured spaces.

Recent advances in generative AI, particularly decoder-based LLMs, show great potential due to their human-like language generation capabilities. These models can be further utilized to interact with unstructured knowledge spaces. However, while these systems excel at general conversations and knowledge retrieval from their vast pre-trained databases, they do not have access to specific data sources within our unstructured knowledge spaces. To address this, we can augment these decoder models with the missing information using neuro-symbolic pipelines, enabling them to be efficiently utilized for our specific purposes. With the Cooee-external chatbot implementation, we aim to leverage neuro-symbolic approaches to augment LLMs and utilize them to interact with unstructured data in energy spaces for various tasks.

6.4.2 Cooee-external - Neuro-symbolic Architectures for Interacting with Unstructured Energy Spaces

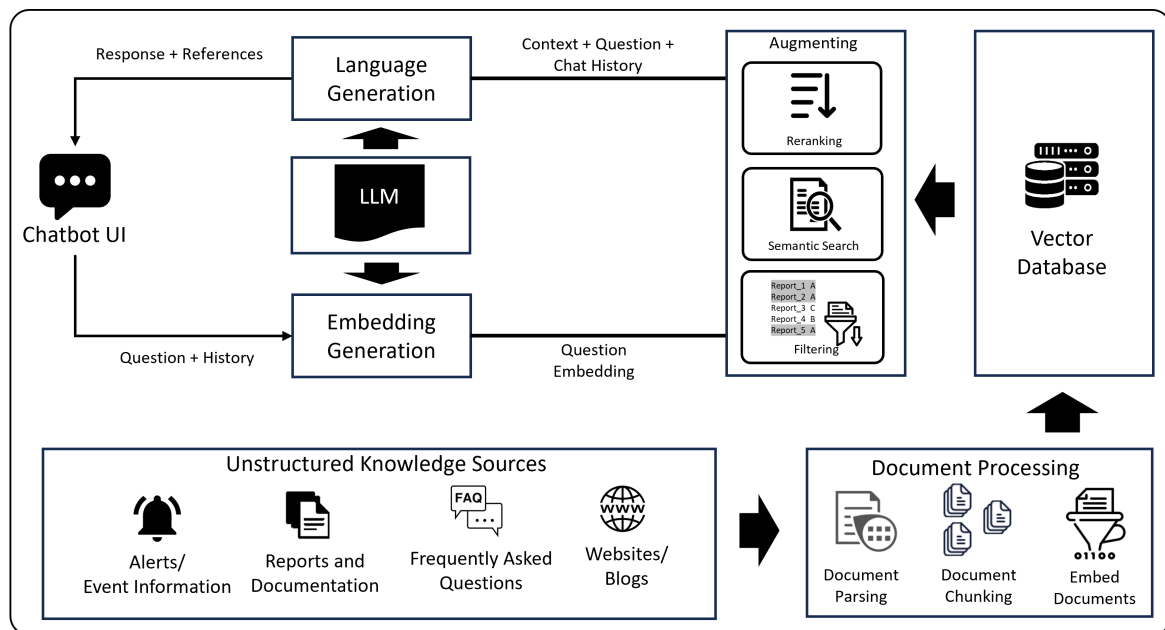


Fig. 6.8 Cooee External Framework for interacting unstructured data spaces

Figure 6.8 showcases the proposed Cooee-external chatbot framework. This chatbot utilizes the LEAP environment's unstructured multi-modal data space for information retrieval and decision-making processes. This architecture is based on the retrieval-augmented generation (RAG) approach, which utilizes both symbolic and connectionist methodologies in NLC. This framework leverages the pre-trained LLM, GPT-3.5 Turbo, for both the language generation and embedding generation processes. Symbolic methods ensure the correct

context is augmented to the language generation process by utilizing semantic similarity matching, reranking, and metadata filtering techniques.

As shown in figure 6.8, unstructured data from various sources such as alerts, event information, reports, documentation, FAQs, and websites/blogs will be forwarded to the document processing module. This module involves several steps to transform raw data into suitable formats for efficient retrieval. The process begins with document parsing, which extracts the text content and images from these sources. Following this, document chunking breaks down large documents into smaller, manageable chunks. This step is essential for processing the information more efficiently and ensuring the embeddings generated are precise and relevant. Finally, it uses the LLM to convert the chunked text segments to numerical vectors using embedding construction. For an efficient semantic search and retrieval process, these embeddings will be then stored in vector databases.

Once the vector database is populated with document embeddings, the system employs several augmenting techniques to refine the retrieval process. Semantic search uses the question embedding, generated from the user's query and history, to perform a search within the vector database, retrieving documents that are semantically similar to the query. The reranking then adjusts the order of these retrieved documents based on their relevance to the query, ensuring that the most pertinent information is prioritized. Additionally, the filtering applies specific criteria to refine the list of retrieved documents, enhancing the accuracy and relevance of the information presented to the user.

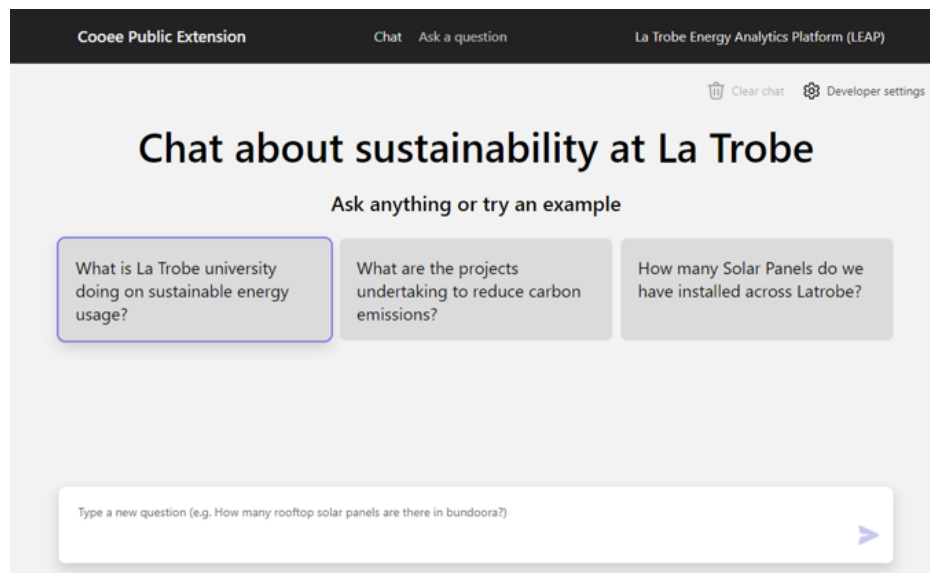


Fig. 6.9 Cooee-external Implementation (a)

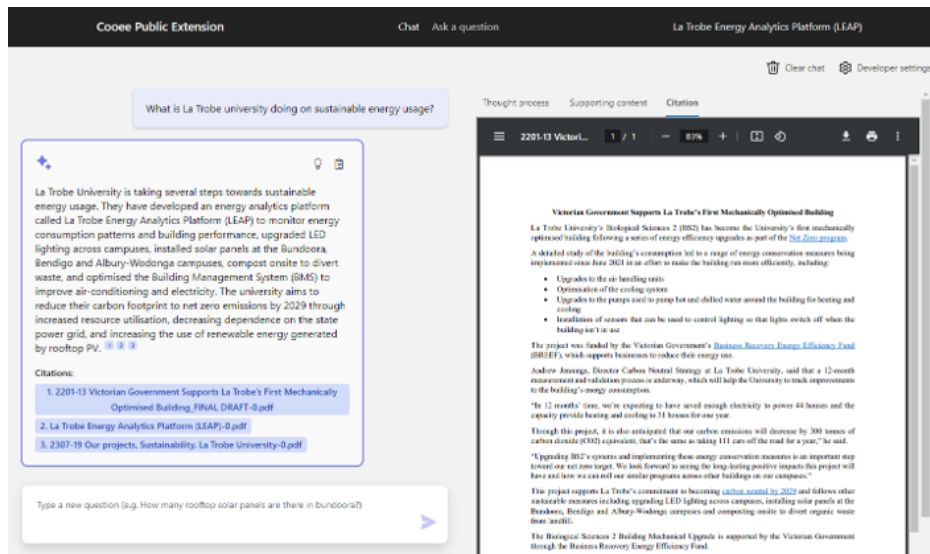


Fig. 6.10 Cooee-external Implementation (b)

LLM works as the core component of the chatbot, generating natural language responses leveraging on the constructed context to the questions. This allows it to provide factually correct responses, as the required knowledge to answer these questions is provided to the LLM, thereby minimizing the risk of hallucinations. These generated responses, along with relevant references, are then presented to the user through the Chatbot UI. By effectively combining these approaches, the Cooee-external chatbot framework achieves a robust, adaptable, and explainable system capable of handling a wide range of user queries with high accuracy and relevance. Cooee-external implementation is illustrated in figure 6.9 and figure 6.10.

6.4.3 Advancing the Retrieval Process

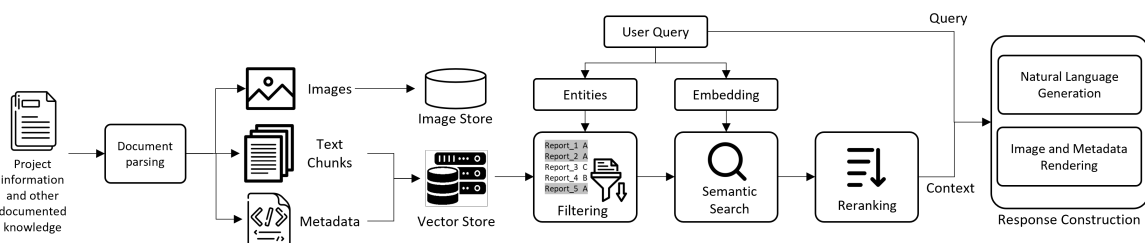


Fig. 6.11 Advanced retrieval augmented generation techniques for documented knowledge retrieval

Figure 6.11 describes the advanced retrieval techniques utilized with rule-driven approaches. Several advancements, such as image retrieval, re-ranking, and metadata filtering, were used to enhance the default retrieval-augmented architecture with decoder-based LLMs.

Images included in documents add significant value, especially in annual reports, blogs, and website articles, which often contain information-rich images that help convey ideas more effectively. As mentioned earlier, when public users inquire about the status of various net zero initiatives, including reference images with the corresponding answers can be very meaningful. Therefore, in the proposed neuro-symbolic pipeline, we incorporate this as an additional feature. During document parsing, any images present in the documents will be separately extracted and stored in a dedicated image repository. The corresponding text chunks will be tagged with references to these images. During the question-answering process, if these tagged text chunks are used to answer questions, the corresponding images will also be retrieved and presented to the end user. This approach ensures more effective and informative responses.

In addition to that, re-ranking and metadata filtering are used to optimize the retrieval process by reducing semantic search time. The unstructured data space of LEAP contains usage summary profiles of the main university buildings, which are essential for various reasoning processes in energy behaviors. These summaries are periodically created and added to the unstructured data space. They provide snapshots of each building's energy behavior, featuring key metrics such as total daily energy consumption, the hour of peak energy usage, week-to-date consumption, and the average consumption for the week. Additionally, event information will be periodically added to this unstructured space to explain the reasons behind different energy behaviors. This information will be stored alongside various other unstructured data in the space.

When a query is presented, metadata filtering is used to optimize the search process by reducing the need to compare the query with all types of data points. Each type of data point is tagged with its type, allowing the system to route the query to the appropriate document collection. This eliminates unnecessary searches and ensures the accuracy of the results. Furthermore, to ensure responses are accurate and up-to-date, re-ranking is performed after semantic search based on the dates of the information pieces. This means that, in addition to embedding similarity, the recency and type is also considered to determine the most relevant document chunks.

6.4.4 Hierarchical Vector Spaces for Optimized Semantic Search

With traditional linear vector search in a growing database, the time complexity increases proportionally to the size of the database records. LEAP's unstructured data space consists

of a large volume of documents that need to be persisted in the vector stores, and the vector database is periodically updated with data such as summary reports and insights. This growing data presents unique challenges. To address these challenges, we implemented the vector database in a hierarchical manner to reduce query times, lower computational costs, and improve the contextual relevance of the chatbot responses.

In linear vector search, the question embedding must be compared with each embedding in the document store, resulting in a time complexity of $t(n) = O(n)$, where n is the number of chunks in the vector store. Even with metadata filtering, which reduces the scope of vector search, the challenge remains as each type of document can still grow significantly. To construct the hierarchical vector store, we used the constrained k-means algorithm (A. Demiriz P. S. Bradley, 2000) to cluster the document embeddings into groups. We ensured that the groups were of roughly equal sizes to optimize searching. Each cluster is then represented by the average embedding of the document embeddings within the group.

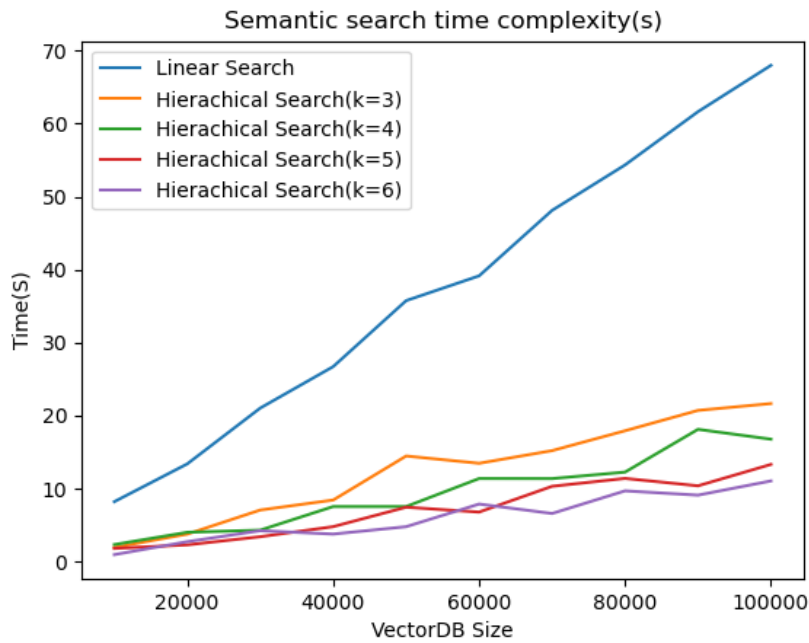


Fig. 6.12 Time complexity comparison of linear and hierarchical vector search with growing DB sizes

When semantic search is performed, the system first compares the question with the cluster embeddings to select the most suitable group. It then continues with the comparisons within the selected group. This approach reduces the time complexity to approximately $y = m(x/k + k)$, where m is a constant, k is the number of roughly evenly sized clusters, and y and x represent the time and the size of the vector database, respectively.

Figure 6.12 showcases an experiment conducted using a simulated dataset of 100,000 chunks with varying k values to evaluate the optimization of a hierarchical vector store. Additionally, we assessed the quality of responses using the hierarchical vector store compared to linear search in question answering. For this experiment, we used a sample containing 40 document chunks and 40 questions, with one question per chunk, to create a Q&A dataset.

During the hierarchical step, these chunks were clustered into four groups with the size of 8-12. To evaluate the quality of the generated responses, we employed the Ragas evaluation framework (Es *et al.*, 2023) which provides several key automatic metrics grounded in LLM judgments. The following metrics were used in our evaluation:

- **Faithfulness:** Measures how well the generated answer is factually supported by the retrieved context. Each claim in the answer is checked against the context, and the score is the proportion of supported claims:

$$\text{Faithfulness Score} = \frac{\text{Number of claims supported by the retrieved context}}{\text{Total number of claims in the response}}$$

A higher score indicates fewer hallucinations and stronger factual consistency.

- **Answer Relevancy:** Measures how well the generated answer aligns with the user input. It is computed by generating N artificial questions from the response, calculating the cosine similarity between the user input embedding (E_o) and each generated question embedding (E_{g_i}), and averaging the results:

$$\text{Answer Relevancy} = \frac{1}{N} \sum_{i=1}^N \frac{E_{g_i} \cdot E_o}{\|E_{g_i}\| \|E_o\|}$$

Higher scores indicate better semantic alignment and reduced irrelevant content.

- **Context Recall:** Measures the proportion of claims in the reference answer that are supported by the retrieved context. The reference is treated as a proxy for the ground truth, and each claim is checked against the retrieved content:

$$\text{Context Recall} = \frac{\text{Number of claims in the reference supported by the retrieved context}}{\text{Total number of claims in the reference}}$$

Higher scores indicate that more of the necessary information was successfully retrieved.

- **Context Match (Context Precision):** Measures the proportion of retrieved chunks that are relevant to answering the question. It is computed as the mean of the precision@ k

values across the retrieved chunks:

$$\text{Context Precision@K} = \frac{\sum_{k=1}^K (\text{Precision@k} \times v_k)}{\text{Total number of relevant items in the top } K \text{ results}}$$

where:

$$\text{Precision@k} = \frac{\text{true positives@k}}{\text{true positives@k} + \text{false positives@k}}$$

Here, K is the number of retrieved chunks, and $v_k \in \{0, 1\}$ is the relevance indicator for the chunk at rank k . Higher scores indicate more focused and relevant retrieval.

These metrics are computed using LLM-in-the-loop judgment, enabling scalable and automatic evaluation without human annotation. They collectively assess factuality (faithfulness), answer quality (relevancy), and retrieval quality (context recall and precision). As shown in table 6.4, these metrics exhibit comparable measures in both cases for this sample of question answering. However, the time savings are significantly higher in larger experiments, as reflected in the plot (figure 6.12). This ensures enhanced performance while maintaining quality comparable to the linear search baselines.

Measurement	Linear Search	Hierarchical Search
Faithfulness	0.720	0.684
Answer Relevancy	0.989	0.984
Context Recall	0.913	0.833
Context Match (%)	86.661	83.331
Execution Time (s)	8.617	6.746

Table 6.4 Comparison of the response quality in linear and hierarchical search

6.5 Leverage Neuro-symbolic Architectures to Implement Multi-agentic Behaviours in Energy Environments

6.5.1 Introduction

As discussed in the preceding sections, neuro-symbolic AI approaches offer a powerful means to retrieve information from both structured and unstructured data by integrating LLMs with various rule-driven methods. Beyond information retrieval, these neuro-symbolic architectures can be leveraged to automate tasks performed periodically or frequently by humans, reducing the significant manual effort traditionally required for such processes. To achieve this, we can design agents capable of executing well-defined tasks by leveraging

LLMs alongside constrained rule sets, as well as providing access to specific tools and data sources. This setup enables agents to perform tasks precisely while adhering to structured guidelines and effectively utilizing relevant data.

6.5.2 Multi-agentic Framework for Energy Task Automation

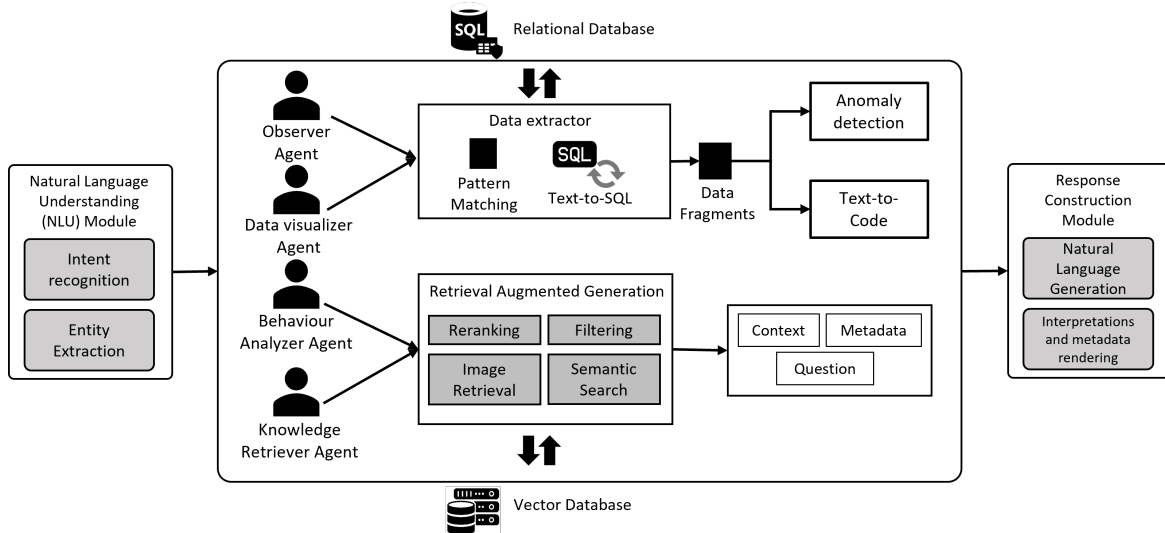


Fig. 6.13 Multi-Agent Chatbot Architecture for Decision Support in Net-Zero Emissions Energy Systems

The proposed architecture for the multi-agent implementation of decision support in net-zero emissions energy systems is illustrated in figure 6.13. This architecture incorporates four specialized expert agents designed to automate key functions within energy systems. These agents are the Observer, tasked with monitoring energy environments to detect significant highlights and anomalies; the Behavior Analyzer, leveraging energy patterns and event data for fact-driven behavioural analysis; the Visualizer, creating visual interpretations of energy data for enhanced understanding; and the Knowledge Retriever, adept at extracting and compiling documented information along with images for comprehensive data retrieval. The multi-agent chatbot architecture consists of three primary components: Natural Language Understanding (NLU) Module, Chatbot Core and Response Construction Module. The following subsections describe the composition and functionality of each component.

Natural Language Understanding Module

The Natural Language Understanding (NLU) Module directly receives queries from the user interface, where they undergo initial processing. Its core function is to interpret these queries and extract key information for further processing steps. The main duty here is to

extract the intent from the user input and redirect the query to the correct destination. Intent recognizer employs the zero-shot classification capability of the LLMs. It comprises a prompt template that takes the input query as a parameter and outputs the ID of the agent best suited to handle the query. To enhance precision, the template incorporates sample questions for each intent. The structure of the intent recognizers prompt template is outlined in figure 6.14. In addition to the intent recognizer, NLU module also contains the entity extractor which we had in the Cooee-internal implementation. Here entity information will be extracted only if the intent matches with the Observer and Visualizer agents as these agents will communicate with the structured data space.

```

"""Check whether the given query belongs to which
intent.
1. Description - Description of the intent.
   Sample Questions - [q11,q12,q13]
2. Description - Description of the intent.
   Sample Questions - [q21,q22,q23]
...
The answer should be only [1,2..]. Do not give any
explanations, only the number of the intent.
query :{query}
Answer: """

```

Fig. 6.14 Prompt template for the intent recognizer component

Chatbot Core

Chatbot core receives the user queries along with extracted entities and intent information. It consists of four specialized agents, each tasked with executing specific functions within the energy environment, leveraging both structured and unstructured data. Depending on the identified intent of the query, the appropriate agent is selected to carry out the task and deliver the results to the response construction module. The four agents within the Chatbot Core are the Observer, Data Visualizer, Behavior Analyzer, and Knowledge Retriever. Detailed descriptions of these agents and their specific functions are provided in table 6.5.

Here the Observer and the Data Visualizer rely on a data extractor component to obtain relevant data segments from the relational database for their respective tasks. This component utilizes the entity information derived from the NLU module. It fills in regex templates to formulate SQL statements, retrieving the necessary data fragments. This approach ensures the generation of accurate SQL statements, offering an advantage over the free-form SQL generation typically seen with language models.

The Observer is tasked with monitoring the energy environment to identify any noteworthy anomalies or highlights that need to be brought to the attention of responsible personnel. By default, it detects irregularities in generation and consumption metrics against baseline models and pre-set thresholds over the most recent seven-day period. Upon specific request,

Table 6.5 Expert Agents in the Multi-Agent System for Net-Zero Emissions

Agent	Description
Observer	The Observer specializes in monitoring the energy environment to identify any anomalies or critical highlights that should be brought to the attention.
Data Visualizer	The Data Visualizer's role involves creating graphical representations of energy-related attributes that align with user instructions.
Behaviour Analyzer	The Behavior Analyzer is capable of examining and detailing the behaviour of a specific building over a defined period.
Knowledge Retriever	The Knowledge Retriever addresses queries that fall outside the scope of the other three agents, providing detailed explanations and answers to general questions.

it can also focus its observations on selected building(s) and specified time frames. The output provided by the Observer includes the names of any buildings exhibiting anomalies, along with the relevant dates of these occurrences.

The Data Visualizer is designed to create graphical representations of energy-related attributes. It leverages the text-to-code generation capabilities of language models, specifically generating Python code using Plotly(Plotly Technologies Inc., 2023) as per user instructions. Afterwards, the code will be executed using the Python engine, and the visualizations are then forwarded to the response construction module. For constructing these plots, the Data Visualizer uses data fragments extracted from the structured database, combined with user-specified parameters like the types of plots desired.

The Behavior Analyzer and Knowledge Retriever are built upon RAG to source the most relevant documented knowledge. Besides comparing the embeddings of user queries with document chunks during general semantic searching, the RAG component employs additional methods such as reranking, filtering, and image retrieval techniques to enhance the retrieval process. Furthermore, when multiple document stores exist within the energy environment, these agents utilize a merged retrieval approach to enhance the efficacy and accuracy of information gathering. The behaviour analyser can analyze the behaviour of a given building during a given temporality. Analyzing the energy behavior of a specific building necessitates a focused overview of various energy attributes, rather than accessing granular energy measures in multiple tables. Therefore, utilizing summary energy reports from the energy platform is a more effective strategy. In addition, the corresponding energy environment may experience various events and outages that may affect energy behaviours. These can range from planned outages, like equipment repairs, to unplanned malfunctions. Additionally, different functions and events within the environment can lead to significant

shifts in energy usage. The Behavior Analyzer employs this knowledge to provide a thorough explanation of the building's behaviour during the analyzed period.

The Knowledge Retriever's focus is on delivering responses to user inquiries through knowledge retrieval. These queries encompass a range of topics, including but not limited to, defining entities, explaining processes, and retrieving information about projects and other documented knowledge. The Knowledge Retriever employs advanced retrieval methods such as image retrieval and reranking. During document parsing, when images are present, they are extracted and stored in an image store and corresponding image references will be tagged with the metadata of the text chunks. Furthermore, the agent re-ranks project information based on temporal tags to ensure users receive the most updated information.

Response Construction Module

The Response Construction Module integrates outputs from all four agents to formulate responses to user queries. It employs LLM to generate these responses in natural language. In scenarios involving retrieval-augmented tasks, the module also presents document and image references to enrich the response with relevant contextual information.

6.5.3 Experiments

This architecture was empirically evaluated for the decision support needs of the net-zero emissions energy system, the La Trobe Energy AI Platform (LEAP). As discussed earlier, LEAP consists of rich structured and unstructured data spaces, encompassing relational databases and vector data spaces. To effectively utilize this resource in our chatbot, we created three separate document stores from the unstructured knowledge: general document store, summary reports store, and event notifications store. These stores were designed specifically to support the RAG agents, namely the Behavior Analyzer and Knowledge Retriever.

The general document store was established using a compilation of frequently asked questions (FAQs) along with their answers, definitions of energy-related terms, and other pertinent energy process information. Furthermore, as La Trobe University actively works towards achieving net-zero emissions by 2029, there is considerable interest in the progress of its sustainability projects. Information about these ongoing projects is also integrated into the general vector store, making it readily accessible to the Knowledge Retriever agent.

The summary reports store was developed using daily energy summaries of the 10 main buildings, automatically extracted from the LEAP platform. This store provides snapshots of each building's energy behaviour, featuring key metrics such as total daily energy consumption, the hour of peak energy usage, week-to-date consumption, and the average consumption for the week, among others. To facilitate effective filtering operations,

News/Outages document store
<pre> {"content": " "University Library Welcomes New Batch with Orientation Program" The university library recently held an orientation program for incoming Master of Science students, introducing them to the vast resources and research facilities available. "source": "News_Outages", "date": "2023/11/17"}, {"content": "Unplanned outage There was an unplanned outage that affected the University library, Donald Whitehead Building and LIMS1.", "source": "News_Outages", "date": "2023/11/16"}, {"content": "Planned outage There was an Air condition repairing LIMS1, Library and Donald Whitehead Building.", "source": "News_Outages", "date": "2023/11/12"} </pre>
Summary document store
<pre> {"content": "" Energy consumption Month to date(kWh) : 115,545 Energy consumption Week to date(kWh) : 4,135 Average daily energy consumption of the week(kWh):4,135 Energy consumption of the day(kWh) : 1,110 Peak consumption hour: 8-9 AM"", "source": "Summary", "date": "2023/11/18"}, {"content": "" Energy consumption Month to date(kWh) : 111,545 Energy consumption Week to date(kWh) : 5,735 Average daily energy consumption of the week(kWh):4,208 Energy consumption of the day(kWh) : 5,120 Peak consumption hour: 11-12 PM"", "source": "Summary", "date": "2023/11/17"}, </pre>

Fig. 6.15 Instances from event notifications and summary reports stores for RAG

the building name and date are used as metadata in these documents. The event notification store is constructed by compiling notifications about various events and outages. This information is sourced directly from the Infrastructure and Operations Unit and Information Services at the university, ensuring up-to-date and relevant data for comprehensive energy management and analysis. Figure 6.15 showcases some instances from these document stores.

For the chatbot's development, we employed GPT-3.5 turbo (Brown *et al.*, 2020) from OpenAI as the primary LLM, accessed via their API. In language generation tasks, we opted for higher temperature settings (1) to encourage diversity in the outputs, whereas for text-to-code and text-to-SQL tasks, we chose lower temperature (0.5) to align with the deterministic nature. Other parameters were retained at their default settings. The RAG component employs the text-embedding-ada-002 embedding model for extracting embeddings, using

cosine distance as the metric for similarity. The LangChain framework was selected for the implementation of these functionalities(LangChain Team, 2023).

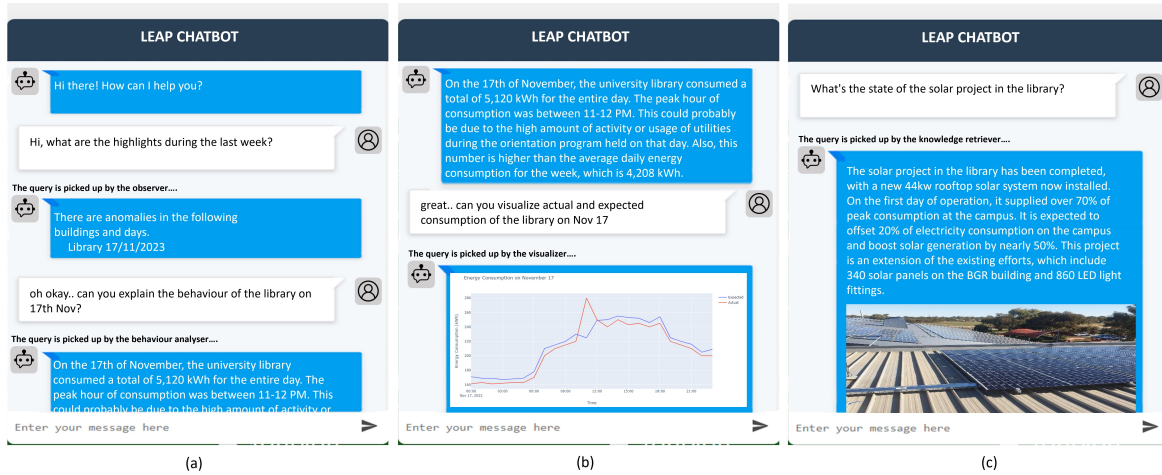


Fig. 6.16 Chatbot workflow for decision support in net zero energy systems

Figure 6.16 presents an illustrative conversation scenario from the chatbot. This is an everyday interaction within the LEAP platform that is conducted manually. However, the introduction of the chatbot has significantly streamlined this process. This scenario demonstrates the chatbot's capability to identify key highlights and underlying reasons, facilitating the generation of actionable insights for end users. In the conversation illustrated in figure 6.16 (a) and (b), the chatbot employs the Observer, Behavior Analyzer, and Visualizer agents to formulate responses to the user's queries.

In this example, the user initially inquires about any significant events in the past week. This query is directed to the Observer agent, based on its intent, which identifies an anomaly on November 17th at the university library. Subsequently, the user requests an analysis of the library's behaviour on that day. The Behavior Analyzer then provides a detailed energy summary for the library on the specified date, along with event information, offering a factual explanation of the day's activities and their implications. Finally, the user asks the chatbot to provide a visual comparison of the actual and expected energy consumption, to which the chatbot, through the Visualizer agent, responds with an appropriate graphical representation. This interaction exemplifies how the chatbot simplifies complex data analysis and presentation for the end user. Figure 6.16 (c) illustrates another scenario where the Knowledge Retriever agent is employed. In this case, a user inquires about the completion status of the library's rooftop solar project. The chatbot responds by detailing the project's current progress and further enriches the explanation with relevant image references, offering a clearer and more detailed understanding of the project status.

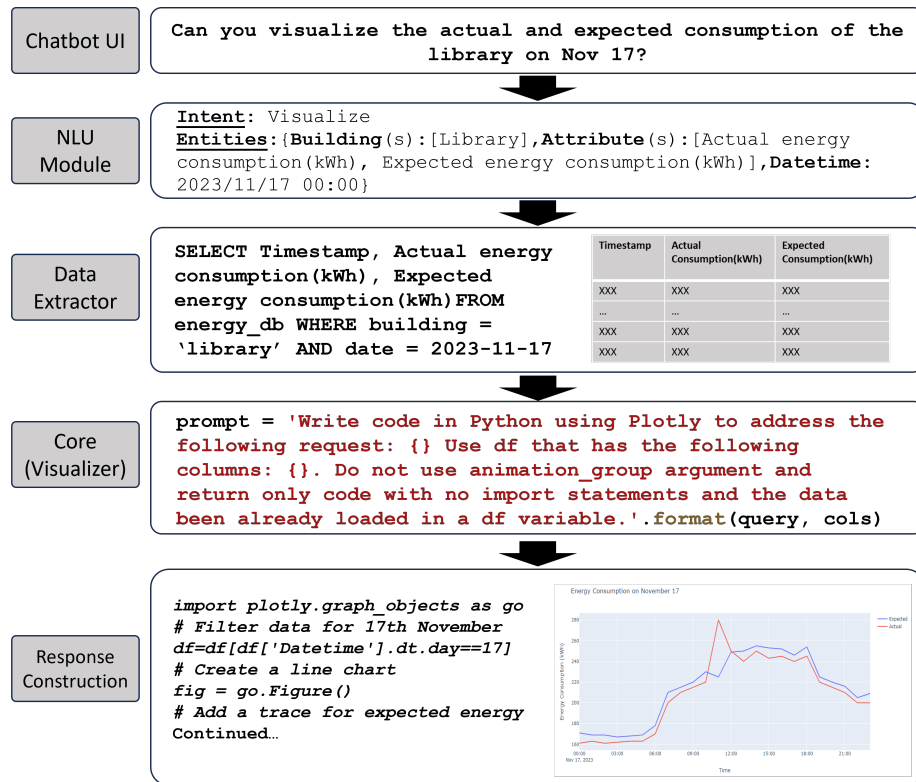


Fig. 6.17 End-to-End Workflow of the Visualizer Agent in the Chatbot System

Figure 6.17 further showcases the end-to-end workflow of the visualizer agent in the chatbot system. The process starts with the user's query submission through the Chatbot UI, followed by the NLU Module's role in identifying the intent and extracting and standardizing entity data. The subsequent phase involves the Data Extractor, which utilizes this information to formulate SQL queries that retrieve the necessary data segments for the visualization. Afterwards, it will be forwarded to the visualization prompt which assembles the Plotly code to produce the visual display, ultimately delivered back to the user via the Chatbot UI.

6.6 Chapter Summary

This chapter discusses the application of neuro-symbolic architectures for enhancing the natural language interactions. It explains about the implementation of the Cooee - energy AI chatbot for complex energy environments. Chapter discusses about three integrations of Cooee in La Trobe Energy Analytics Platform (LEAP) namely Cooee-internal for interacting with structured dataspace, Cooee-external for interacting with multi-modal unstructured dataspace and multi-agentic implementation for the task automation. These neuro-symbolic

methodologies significantly enhance interaction, decision-making, and automation in energy management, establishing a robust paradigm for leveraging AI in complex systems.

Chapter 7

Conclusion

This chapter concludes the thesis with a summary of research contributions, limitations of the current research and directions for future research. A conclusive summary of the research is presented in Section 7.1, detailing the key contributions of Neuro-symbolic architectures for natural language comprehension. Section 7.2 describes how those contributions made it possible to address the key research questions. This chapter and the thesis conclude in Section 7.3 with a discussion of the advances in the field of NLC and evolving needs in the fast-changing AI landscape.

7.1 Summary of the contributions

Chapter 2 presents a comprehensive overview of the significant role of natural language data in modern computer systems and discusses how various computational models attempt to process language intelligently. It further examines the evolution of traditional NLP systems, highlighting the transition from explicit, rule-based methods to data-intensive neural models. Furthermore, Chapter 2 discusses the emergence of LLMs, detailing their architectures, behaviours, and the numerous advantages they offer across various applications. The chapter also provides a comprehensive analysis of the limitations inherent in these systems, including their increasing complexity and unsustainable growth trends, lack of explainability, and lack of compositional language understanding, which often lead to fundamental errors. The chapter concludes by widening the viewpoints on the need for more effective approaches to address these limitations while also maintaining the benefits of the models.

Chapter 3 analyses human language comprehension, explaining several techniques that enable effective and efficient language comprehension in humans. Then, it compares these human comprehension strategies to the computational methods of language comprehension with a particular focus on neuro-symbolic NLP methods. The chapter explores symbolic and

connectionist paradigms in detail, outlining the strengths and limitations of each approach and underscores the necessity of their integration. Finally, it introduces a neuro-symbolic framework for natural language comprehension, offering a structured guide for developing NLP solutions that synergistically combine the advantages of both symbolic and connectionist methods.

Comprehensive studies and experiments that were conducted to evaluate the proposed neuro-symbolic framework for natural language comprehension are described in detail in Chapters 4, 5, and 6. Chapter 4 describes empirical evaluations in two main sections: the experiments enhancing word representations and semantic intelligence systems constructed with hybrid architectures. The word representation-enhancing experiments explain two main implementations for visually grounding word embeddings and emotion infusion of word representations. Chapter 4 further showcases applications leveraging the proposed framework, including Extremism Detection in Languages, identifying and analyzing extremist language patterns to prevent radicalization and promote safer communication environments; Figurative Language Detection, recognizing and interpreting idiomatic and figurative expressions; enhancing the system's ability to understand complex language usage and improving overall language comprehension; and Detecting Deceptive Content, developing systems to detect fake news and other forms of deceptive content, helping to maintain the integrity of information and combat misinformation.

Chapter 5 explains the application of neuro-symbolic architectures in emotion elicitation with the novel Emotion AWARE algorithm. The chapter starts by highlighting the significance of emotions as an aspect of language comprehension and analysing the existing computational methods of emotion elicitation in the literature. Then, it describes the Emotion AWARE methodology in detail, along with the empirical evaluations with nine experiments with several standard benchmark datasets. The chapter concludes by discussing two ablation studies of Emotion AWARE, highlighting the practical usage of the framework in real-world use cases. The first use case, Emotion AWARE was applied to detect patient emotions and emotion transitions in digital healthcare settings, particularly during conversations between patients, carers, and healthcare practitioners. In the second use case, the Emotion AWARE algorithm was used to analyze emotions within financial conference transcripts, focusing on emotional conflicts, agreements, and fluctuations between managerial and non-managerial participants, as well as their correlation with various market variables.

Chapter 6 discusses the application of neuro-symbolic architectures for enhancing natural language interactions. It explains about the implementation of the Cooee - energy AI chatbot for complex energy environments. The chapter discusses about three integrations of Cooee in La Trobe Energy Analytics Platform(LEAP), namely Cooee-internal for interacting

with structured dataspace, Coee-external for interacting with multi-modal unstructured dataspace and multi-agentic implementation for task automation. These neuro-symbolic methodologies significantly enhance interaction, decision-making, and automation in energy management, establishing a robust paradigm for leveraging AI in complex systems.

In summary, we have highlighted the importance of neuro-symbolic architectures as a viable alternative to the data-intensive, connectionist trends of LLMs in natural language comprehension systems while resolving their limitations. The proposed approaches address various aspects of language comprehension, showcasing their applicability across diverse use cases. We have demonstrated the practical benefits of hybrid systems that combine explainability and controllability by integrating small-scale language models with rule-based methods. This approach eliminates the need to continually expand LLMs to achieve more sophisticated capabilities. Overall, these advancements in neuro-symbolic AI bridge the gap between human and computational language comprehension, enabling the development of effective, efficient, and sustainable systems.

7.2 Addressing the research questions

This section describes how the above contributions have addressed the research questions presented in the Thesis.

- **RQ1: What is the role of current computational methods in enabling machines to intelligently process natural language?**

The first research question explores the computational methods currently employed to comprehend and process natural languages in computer systems. This question is addressed in Chapter 2, which offers an in-depth analysis of both traditional and modern techniques, with a particular focus on LLMs and their role in natural language comprehension. The chapter also addresses the following sub-questions.

- (a) **Has the evolution of computational methods over time led to significant advancements in language comprehension capabilities?**

The evolution of computational methods has indeed led to significant advancements in language comprehension capabilities. As reviewed in Section 2.3, the field progressed from simple rule-based systems, which relied heavily on hand-crafted symbolic rules, to advanced neural models that leverage large datasets and learn representations automatically. This evolution highlights key milestones where connectionist methods, such as neural networks, overcame the limitations

of symbolic approaches by effectively utilizing data as a valuable resource. These developments have resulted in more adaptable and accurate natural language processing systems.

(b) **Which underlying mechanisms and key advantages enable large language models (LLMs) to outperform previous approaches and become the current state-of-the-art in natural language processing?**

The underlying mechanisms and key advantages that establish large language models (LLMs) as the state-of-the-art in natural language processing are extensively examined in Sections 2.4 and 2.5. Section 2.4 describes the emergence of LLMs based on the transformer architecture, detailing how self-attention mechanisms and large-scale pretraining enable these models to capture complex language patterns. It also categorizes the three main types of LLMs and explains their functioning and training processes. Section 2.5 highlights the advantages of LLMs, such as their ability to perform a wide range of downstream language comprehension and generation tasks with high accuracy. Their growing adoption across various applications demonstrates capabilities once thought unattainable by traditional methods, making them pivotal in advancing NLP technology.

(c) **What current trends and limitations of large language models (LLMs) reveal significant research gaps that require further investigation?**

Section 2.6 explores the current trends and limitations of large language models (LLMs), identifying three primary challenges: increasing model complexity leading to unsustainable resource demands, limited explainability of model decisions, and a lack of compositional behaviors necessary for robust language understanding. These limitations are shown to create clear and pressing research gaps, highlighting areas where future work is needed to develop more efficient, transparent, and cognitively aligned NLP models.

• **RQ2: What cognitive and linguistic principles from human language comprehension can inform the design of effective AI systems for natural language comprehension?**

As we improve computational models, our ultimate goal is to build more human-like systems. However, examining the fundamental errors that current state-of-the-art computational models make, as well as the challenges inherited with current trends, prompts us to reconsider whether these approaches truly align with human cognitive behaviours. Thus this research question investigates what makes human language comprehension an effortless and efficient process. It is addressed in Chapter 3, which

examines the foundational elements of human language comprehension from the perspective of cognitive psychology.

- (a) **To what extent do the principles of compositionality and systematicity enable humans to comprehend language more efficiently and effectively?**

Compositionality and systematicity significantly enable humans to comprehend language efficiently and effectively by allowing them to balance limited memory capacity with inferential processes. This balance facilitates the generation and understanding of infinite language variations from a limited set of basic elements and rules. This is discussed in Sections 3.3.1 and 3.3.2, which explore these principles as key strategies in human language comprehension.

- (b) **How significantly do humans leverage other modalities to ground the language comprehension process?**

Humans significantly leverage other modalities to ground language comprehension by associating linguistic symbols with real-world sensory and contextual references. This multimodal grounding creates a richer and more holistic understanding of language, beyond purely textual information. In contrast, large language models (LLMs) often rely solely on text modality, which limits their ability to fully comprehend language in context. This is discussed in Section 3.1.3, which analyzes grounded language learning techniques in humans and compares them with LLM limitations.

- (c) **To what extent does the construction-integration theory explain systematic and hierarchical aspects of language understanding?**

The construction-integration theory significantly explains how humans develop systematic and hierarchical mental models during language comprehension, enabling the organization of information at multiple levels for effective understanding. This is explored in Section 3.1.4, which provides a detailed examination of the theory and its role in hierarchical language processing.

- **RQ3: How can neuro-symbolic architectures be leveraged to bridge the gap between human and computational methods for natural language understanding?**

This research question explores neuro-symbolic architectures as a potential approach to address the current unsustainable trends and limitations of LLMs, as well as their divergence from human methods of comprehension. This question is further examined in Chapter 3. Chapter 3 also addresses the following sub-questions.

(a) **To what extent do traditional symbolic methods provide advantages and face limitations in language processing?**

Traditional symbolic methods in language processing offer key advantages such as controllability, adaptability, and explainability, allowing for transparent and manageable system behavior without the resource-intensive demands and unsustainability often seen in large neural models. These features make symbolic approaches valuable for building trustworthy and interpretable language systems. However, their limitations include difficulties in scaling to complex rule sets and an inability to effectively learn or converge with large-scale data, which restricts their flexibility and performance in handling the diversity of natural language. This balance of strengths and weaknesses is discussed in Section 3.3.1, which highlights how integrating symbolic methods with other approaches can help create systems that better mirror human language comprehension while overcoming some of these traditional limitations.

(b) **Which strategies can effectively enhance the performance and applicability of large language models (LLMs) within the connectionist paradigm?**

Strategies to enhance the use of LLMs as the state-of-the-art approach within the connectionist paradigm include techniques such as prompt engineering, transfer learning, and the implementation of guardrails to improve their effectiveness in developing natural language comprehension systems. These strategies build on the strengths of connectionist NLP methods, which leverage learnability and access implicit aspects of comprehension that symbolic methods often cannot reach, while addressing their weaknesses. This topic is discussed in detail in Sections 3.3.2 and 3.3.4 of the thesis.

(c) **Could neuro-symbolic architectures, by merging symbolic reasoning with connectionist learning, create more robust systems that align with human cognitive strategies?**

This sub-question is examined in Section 3.3.3, which outlines several existing neuro-symbolic approaches discussed in the literature and highlights their missing elements. Section 3.3.5 then introduces the proposed neuro-symbolic framework for natural language comprehension, demonstrating how techniques from both symbolic and connectionist paradigms can be integrated to build robust systems that more closely align with human cognitive processes. It further emphasises how such systems can incorporate features like sustainability, controllability, and explainability, ensuring safe and trustworthy deployments.

- **RQ4: How can neuro-symbolic architectures be utilized to develop practical applications for sustainable development?**

This research question is thoroughly examined in Chapters 4, 5, and 6, which showcase various practical applications of systems developed using the proposed neuro-symbolic framework. These chapters also address the following sub-questions.

- (a) **How significantly can incorporating implicit elements like emotion and visual knowledge improve the robustness of word representations in natural language processing?**

Integrating implicit aspects such as visual grounding and emotion infusion can significantly enhance the robustness of word representations, improving their effectiveness in various natural language processing tasks like classification, clustering, and semantic similarity. This is addressed in Section 4.2, which presents these two approaches—visual grounding (Section 4.2.1) and emotion infusion (Section 4.2.2)—and evaluates their impact through multiple experiments.

- (b) **To what extent can neuro-symbolic architectures effectively capture and analyze complex language comprehension aspects such as figurative, deceptive, and extremist language usage?**

Neuro-symbolic architectures can effectively capture complex language comprehension aspects—including figurative, deceptive, and extremist language by integrating explainability, controllability, and efficient use of LLMs. This is discussed in Sections 4.3, 4.4, and 4.5, which explore the application of neuro-symbolic architectures in three key facets of language comprehension: disinformation detection, idiom detection, and extremism detection in natural languages. Each system was empirically validated, demonstrating features such as explainability, controllability, and the ability to utilize small-scale models like BERT instead of more complex LLMs. These approaches were further applied to social media analysis use cases, highlighting their practical significance and effectiveness.

- (c) **How effectively do neuro-symbolic architectures facilitate robust, explainable, and controllable emotion elicitation in natural language processing systems?**

This is addressed in Chapter 5, which introduces the novel algorithm Emotion AWARE: An Artificial Intelligence Framework for Adaptable, Robust, Explainable, and Multi-Granular Emotion Analysis. The algorithm was evaluated through nine experiments to assess its key features, including emotion elicitation, adaptability, robustness, and explainability. Additionally, the chapter presents two

ablation studies of Emotion AWARE, showcasing its practical applicability in real-world use cases within the healthcare and financial domains.

- (d) **To what extent can neuro-symbolic architectures improve natural language interactions in critical energy infrastructures by enabling more effective, multi-modal, and automated communication systems?**

Chapter 6 focuses on the application of neuro-symbolic architectures to enhance natural language interactions. It details the development of Cooee, an AI chatbot designed for complex energy environments. The chapter explores three integrations of Cooee within the La Trobe Energy Analytics Platform (LEAP): Cooee-Internal, for interacting with structured data spaces; Cooee-External, for handling multi-modal unstructured data spaces; and a multi-agent implementation for task automation.

7.3 Future directions

The current trends in natural language comprehension, dominated by LLMs, are widely regarded as the state-of-the-art. However, as discussed in this thesis, these methods diverge significantly from human-like language comprehension strategies, relying on fundamentally different mechanisms. These approaches face numerous challenges, including growing complexity, unsustainable resource requirements, limited controllability, and explainability, as well as fundamental errors that violate principles like compositionality. Additionally, the emphasis on LLMs has led to often overlooking traditional NLP approaches, which better align with human comprehension strategies. Current methods attempt to address the entire spectrum of language comprehension solely through LLMs, leading to these limitations. This highlights the critical need for neuro-symbolic architectures in the design of natural language comprehension systems, as validated in this thesis through various experiments.

In the coming years, generative AI technologies are expected to evolve further, following these trends and integral to everyday life. To address the challenges posed by current LLM-based methods while retaining their strengths, it is essential to incorporate these models into neuro-symbolic pipelines. The use of rule-driven guardrails in such architectures can enhance controllability and explainability, ensuring reliable and trustworthy deployments. Collaborative efforts between industry, academia, and policymakers will be vital for establishing guidelines and standards that govern the safe and ethical use of generative AI.

As the demand for LLMs continues to grow, it becomes critical to strengthen their integration with neuro-symbolic approaches for efficient and effective deployments that benefit a more significant segment of society. The current landscape, where a few key

organizations dominate with their access to vast data resources and sophisticated computing infrastructure, has concentrated power and control. Neuro-symbolic strategies provide an alternative by enabling the development of sustainable, small-scale LLMs enhanced with rule-driven methodologies and guided by human expertise. This decentralization not only reduces reliance on resource-intensive models but also improves the transparency and accountability of model development.

To advance this paradigm, increased research efforts are necessary, alongside the formulation of policies and guidelines to support the development of neuro-symbolic architectures. These efforts will help mitigate the unsustainable trends of current practices while fostering systems that are controllable, explainable, and sustainable, ultimately ensuring their widespread benefits.

Bibliography

- A. Demiriz P. S. Bradley, K.P.B. (2000) Constrained k-means clustering. *Microsoft Research* pp. 1–8
- Abdullah, M., AlMasawa, M., Makki, I., Alsolmi, M. and Mahrous, S. (2020) Emotions extraction from arabic tweets *International Journal of Computers and Applications* **42**, pp. 661–675
- Acheampong, F.A., Nunoo-Mensah, H. and Chen, W. (2021) Transformer models for text-based emotion detection: a review of bert-based approaches *Artificial Intelligence Review* **54**, pp. 5789–5829
- Acheampong, F.A., Wenyu, C. and Nunoo-Mensah, H. (2020) Text-based emotion detection: Advances, challenges, and opportunities *Engineering Reports* **2**
- Adamson, T., Lett, E., Glick, J., Garrison-Desany, H.M. and Restar, A. (2022) Experiences of violence and discrimination among lgbtq+ individuals during the covid-19 pandemic: a global cross-sectional analysis *BMJ Global Health* **7**(9)
- Adikari, A. and Alahakoon, D. (2021) Understanding citizens' emotional pulse in a smart city using artificial intelligence *IEEE Transactions on Industrial Informatics* **17**, pp. 2743–2751
- Adikari, A., Gamage, G., de Silva, D., Mills, N., Wong, S.M.J. and Alahakoon, D. (2021a) A self structuring artificial intelligence framework for deep emotions modeling and analysis on the social web *Future Generation Computer Systems* **116**, pp. 302–315
- Adikari, A., Nawaratne, R., Silva, D.D., Ranasinghe, S., Alahakoon, O. and Alahakoon, D. (2021b) Emotions of covid-19: Content analysis of self-reported information using artificial intelligence *Journal of Medical Internet Research* **23**, p. e27341
- Adoma, A.F., Henry, N.M. and Chen, W. (2020) Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition in: *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (IC-CWAMTIP)* pp. 117–121
- Agrawal, A. and An, A. (2012) Unsupervised emotion detection from text using semantic and syntactic relations *Proceedings - 2012 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2012* pp. 346–353
- Agrawal, A., An, A. and Papagelis, M. (2018a) Learning emotion-enriched word representations in: *Proceedings of the 27th International Conference on Computational Linguistics*, (Eds.) E.M. Bender, L. Derczynski and P. Isabelle pp. 950–961 Association for Computational Linguistics, Santa Fe, New Mexico, USA

- Agrawal, A., An, A. and Papagelis, M. (2018b) Learning emotion-enriched word representations in: *Proceedings of the 27th International Conference on Computational Linguistics* pp. 950–961
- Ahmad, S., Asghar, M.Z., Alotaibi, F.M. and Khan, S. (2020) Classification of poetry text into the emotional states using deep learning technique *IEEE Access* **8**, pp. 73865–73878
- Ahmed, H., Traore, I. and Saad, S. (2017) Detecting opinion spams and fake news using text classification *Security and Privacy* **1**, p. e9
- AI, S. (2024) Stable diffusion <https://stability.ai/stable-diffusion> accessed: 2024-10-20
- Ajao, O., Bhowmik, D. and Zargari, S. (2019) Sentiment aware fake news detection on online social networks in: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 2507–2511
- Alduailej, A. and Alothaim, A. (2022) Araxlnet: pre-trained language model for sentiment analysis of arabic *Journal of Big Data* **9**, p. 72
- Alfi, M., Kaghazgaran, P., Caverlee, J. and Morstatter, F. (2018) Measuring the impact of isis social media strategy *MIS2. Marina Del Ray, California*
- Ali, R.H., Pinto, G., Lawrie, E. and Linstead, E.J. (2022) A large-scale sentiment analysis of tweets pertaining to the 2020 us presidential election *Journal of Big Data* **9**, p. 79
- Allcott, H. and Gentzkow, M. (2017) Social media and fake news in the 2016 election *Journal of Economic Perspectives* **31**(2), p. 211–36
- Alm, C.O. (2009) *Affect in Text and Speech* VDM Verlag Dr. Müller (January 9, 2009)
- Alsawidan, N. and Menai, M.E.B. (2020) A survey of state-of-the-art approaches for emotion recognition in text *Knowledge and Information Systems* **62**, pp. 2937–2987
- Angskun, J., Tipprasert, S. and Angskun, T. (2022) Big data analytics on social networks for real-time depression detection *Journal of Big Data* **9**, p. 69
- Anthropic (2023) Claude 3 opus <https://docs.anthropic.com/en/docs/intro-to-claude> accessed: 2023-05-20
- Asaad, R. and Ali, R. (2019) Back propagation neural network(bpnn) and sigmoid activation function in multi-layer networks *Academic Journal of Nawroz University* **8**, p. 216
- Asia, M. (2024) The emergence of large language models (llms) accessed: 2024-11-20
- Avramidis, E., Macketanz, V., Strohriegel, U., Burchardt, A. and Möller, S. (2020) Fine-grained linguistic evaluation for state-of-the-art machine translation pp. 346–356 Association for Computational Linguistics
- Awad, M. and Fraihat, S. (2023) Recursive feature elimination with cross-validation with decision tree: Feature selection method for machine learning-based intrusion detection systems *Journal of Sensor and Actuator Networks* **12**(5)

- Awan, I. (2017) Cyber-extremism: Isis and the power of social media *Society* **54**(2), pp. 138–149
- Aziz, R.H.H. and Dimililer, N. (2021) Sentixgboost: enhanced sentiment analysis in social media posts with ensemble xgboost classifier *Journal of the Chinese Institute of Engineers, Transactions of the Chinese Institute of Engineers, Series A* **44**, pp. 562–572
- Baccianella, S., Esuli, A. and Sebastiani, F. (2010) {S}enti{W}ord{N}et 3.0: An enhanced lexical resource for sentiment analysis and opinion mining European Language Resources Association (ELRA)
- Ball, R. and Brown, P. (1968) An empirical evaluation of accounting income numbers *Journal of Accounting Research* **6**(2), p. 159
- Bamber, L.S., Barron, O.E. and Stevens, D.E. (2010) Trading volume around earnings announcements and other financial reports: Theory, research design, empirical evidence, and directions for future research *SSRN Electronic Journal*
- Barsade, S.G. and Gibson, D.E. (2007) Why does affect matter in organizations? *Academy of Management Perspectives* **21**(1), p. 36–59
- Barsalou, L.W. (2008) Grounded cognition *Annual Review of Psychology* **59**(1), p. 617–645
- Bastings, J., Ebert, S., Zablotkaia, P., Sandholm, A. and Filippova, K. (2022) "will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification
- Baxter, J., Carlsson-Wall, M., Chua, W.F. and Kraus, K. (2019) Accounting and passionate interests: The case of a swedish football club *Accounting, Organizations and Society* **74**, p. 21–40
- Beck, I.L., McKeown, M.G. and Kucan, L. (2013) *Bringing words to life, second edition* Guilford Publications, New York, NY 2nd ed.
- Bender, E.M. and Koller, A. (2020) Climbing towards NLU: On meaning, form, and understanding in the age of data in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Eds.) D. Jurafsky, J. Chai, N. Schluter and J. Tetreault pp. 5185–5198 Association for Computational Linguistics, Online
- Bengio, Y., Simard, P. and Frasconi, P. (1994) Learning long-term dependencies with gradient descent is difficult *IEEE Transactions on Neural Networks* **5**(2), pp. 157–166
- Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A.C., Korbak, T. and Evans, O. (2024) The reversal curse: Llms trained on "a is b" fail to learn "b is a"
- Bierwisch, M. (2015) Generative grammar in: *International Encyclopedia of the Social Behavioral Sciences (Second Edition)*, (Ed.) J.D. Wright pp. 871–878 Elsevier, Oxford second edition ed.
- Bisk, Y., Zellers, R., Le bras, R., Gao, J. and Choi, Y. (2020) Piqa: Reasoning about physical commonsense in natural language *Proceedings of the AAAI Conference on Artificial Intelligence* **34**(05), p. 7432–7439

- Black, S., Gao, L., Wang, P., Leahy, C. and Biderman, S. (2021) Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow
- Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017) Enriching word vectors with subword information *Transactions of the Association for Computational Linguistics* **5**, pp. 135–146
- Boltong, A., Ledwick, M., Babb, K., Sutton, C. and Ugalde, A. (2016) Exploring the rationale, experience and impact of using cancer information and support (cis) services: an international qualitative study *Supportive Care in Cancer* **25**(4), p. 1221–1228
- Bracci, S., Ritchie, J.B., Kalfas, I. and Op de Beeck, H.P. (2019) The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks *Journal of Neuroscience* **39**(33), pp. 6513–6525
- Bright, J. (2018) Explaining the emergence of political fragmentation on social media: The role of ideology and extremism *Journal of Computer-Mediated Communication* **23**(1), pp. 17–33
- Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020) Language models are few-shot learners
- Bruni, E., Tran, N.K. and Baroni, M. (2014) Multimodal distributional semantics *Journal of Artificial Intelligence Research* **49**, p. 1–47
- Budiharto, W. and Meiliana, M. (2018) Prediction and analysis of indonesia presidential election from twitter using sentiment analysis *Journal of Big Data* **5**, p. 51
- Buyl, M., Rogiers, A., Noels, S., Dominguez-Catena, I., Heiter, E., Romero, R., Johary, I., Mara, A.C., Lijffijt, J. and Bie, T.D. (2024) Large language models reflect the ideology of their creators
- Bötticher, A. (2017) Towards academic consensus definitions of radicalism and extremism *Perspectives on Terrorism* **11**(4), pp. 73–77
- Cacciari, C. (2014) Processing multiword idiomatic strings: Many words in one? *The Mental Lexicon* **9**, pp. 267–293
- Caldarini, G., Jaf, S. and McGarry, K. (2022) A literature survey of recent advances in chatbots *Information* **13**(1)
- Calefato, F., Lanubile, F., Maiorano, F. and Novielli, N. (2018) Sentiment polarity detection for software development pp. 128–128 ACM
- Calvillo, J., Brouwer, H. and Crocker, M.W. (2021) Semantic systematicity in connectionist language production *Information* **12**(8)
- Cambria, E., Li, Y., Xing, F.Z., Poria, S. and Kwok, K. (2020) Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis pp. 105–114 ACM

- Cao, J., Li, J., Yin, M. and Wang, Y. (2022) Online reviews sentiment analysis and product feature improvement with deep learning *ACM Transactions on Asian and Low-Resource Language Information Processing*
- Center for Research in Security Prices (2024) Crsp us stock database available from: <https://www.crsp.org/>
- Chatterjee, A., Narahari, K.N., Joshi, M. and Agrawal, P. (2019) Semeval-2019 task 3: Emocontext contextual emotion detection in text pp. 39–48 Association for Computational Linguistics
- Chen, X., Wee, C., Leong, B., Flor, M. and Klebanov, B.B. (2020) Go figure! multi-task transformer-based architecture for metaphor detection using idioms: Ets team in 2020 metaphor shared task pp. 235–243
- Chérargui, M.A. (2012) Theoretical overview of machine translation in: *International Conference on Web and Information Technologies*
- Chersoni, E., Hollenstein, N., Jacobs, C., Oseki, Y., Prévot, L. and Santus, E., (Eds.) (2022) *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* Association for Computational Linguistics, Dublin, Ireland
- Chitic, R. (2020) Real or fake? accessed: 2024-08-03
- Chung, J., Gulcehre, C., Cho, K. and Bengio, Y. (2015) Gated feedback recurrent neural networks in: *Proceedings of the 32nd International Conference on Machine Learning*, (Eds.) F. Bach and D. Blei **Proceedings of Machine Learning Research**, vol. 37 pp. 2067–2075 PMLR, Lille, France
- Clark, H.H. (1996) *Using Language* Cambridge University Press
- Cole, D. (2024) The chinese room argument *The Stanford Encyclopedia of Philosophy* substantive revision Wed Oct 23, 2024
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P. (2011) Natural language processing (almost) from scratch *Journal of machine learning research* **12**(Aug), pp. 2493–2537
- Corti, L., Zanetti, M., Tricella, G. and Bonati, M. (2022) Social media analysis of twitter tweets related to asd in 2019–2020, with particular attention to covid-19: topic modelling and sentiment analysis *Journal of Big Data* **9**, p. 113
- CrowdFlower (2024) Crowdfower sentiment analysis and emotion in text <https://www.crowdflower.com/data/sentiment-analysis-emotion-text/> accessed: 2024-10-20
- Daneman, M. and Merikle, P.M. (1996) Working memory and language comprehension: A meta-analysis *Psychonomic Bulletin amp; Review* **3**(4), p. 422–433
- De Gibert, O., Perez, N., García-Pablos, A. and Cuadros, M. (2018) Hate speech dataset from a white supremacy forum *arXiv preprint arXiv:1809.04444*

- Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990) Indexing by latent semantic analysis *Journal of the American Society for Information Science* **41**(6), pp. 391–407
- Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G. and Ravi, S. (2020) Goemotions: A dataset of fine-grained emotions
- Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L. (2009) Imagenet: A large-scale hierarchical image database in: *2009 IEEE Conference on Computer Vision and Pattern Recognition* pp. 248–255
- Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* **1**, pp. 4171–4186
- Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* **1**, pp. 4171–4186
- Dong, X., Wang, Y., Yu, P.S. and Caverlee, J. (2023) Probing explicit and implicit gender bias through llm conditional text generation
- Doshi-Velez, F. and Kim, B. (2017) Towards a rigorous science of interpretable machine learning
- Du, M., Liu, N. and Hu, X. (2019) Techniques for interpretable machine learning
- Du, P. and Nie, J.Y. (2018) Mutux at semeval-2018 task 1: Exploring impacts of context information on emotion detection pp. 345–349 Association for Computational Linguistics
- Dunietz, J., Burnham, G., Bharadwaj, A., Rambow, O., Chu-Carroll, J. and Ferrucci, D. (2020) To test machine comprehension, start by defining comprehension pp. 7839–7859 Association for Computational Linguistics
- Dziri, N., Lu, X., Sclar, M., Li, X.L., Jiang, L., Lin, B.Y., Welleck, S., West, P., Bhagavatula, C., Le Bras, R., Hwang, J., Sanyal, S., Ren, X., Ettinger, A., Harchaoui, Z. and Choi, Y. (2023) Faith and fate: Limits of transformers on compositionality in: *Advances in Neural Information Processing Systems*, (Eds.) A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine vol. 36 pp. 70293–70332 Curran Associates, Inc.
- Ekberg, K., McDermott, J., Moynihan, C., Brindle, L., Little, P. and Leydon, G.M. (2014) The role of helplines in cancer care: Intertwining emotional support with information or advice-seeking needs *Journal of Psychosocial Oncology* **32**(3), p. 359–381
- ElevenLabs (2024) Elevenlabs beta <https://beta.elevenlabs.io/> accessed: 2024-10-20
- Es, S., James, J., Espinosa-Anke, L. and Schockaert, S. (2023) Ragas: Automated evaluation of retrieval augmented generation

- Esuli, A. and Sebastiani, F. (2006) Sentiwordnet: A publicly available lexical resource for opinion mining European Language Resources Association (ELRA)
- Fadel, A., Al-Ayyoub, M. and Cambria, E. (2021) Justers at semeval-2020 task 4: Evaluating transformer models against commonsense validation and explanation pp. 535–542 International Committee for Computational Linguistics
- Faizan, Z., Fatima, G. and Parveen, S. (2023) Radicalization among students with disabilities: Different ways government can support schools and teachers in preventing violence and extremism *Sustainable Business and Society in Emerging Economies* **5**(3), pp. 141–154
- Feldman, A. and Peng, J. (2013) Automatic detection of idiomatic clauses pp. 435–446 Springer Berlin Heidelberg
- Filieri, R., Acikgoz, F., Ndou, V. and Dwivedi, Y. (2021) Is tripadvisor still relevant? the influence of review credibility, review usefulness, and ease of use on consumers' continuance intention *International Journal of Contemporary Hospitality Management* **33**, pp. 199–223
- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E. (2001) Placing search in context: the concept revisited in: *Proceedings of the 10th International Conference on World Wide Web WWW '01* p. 406–414 Association for Computing Machinery, New York, NY, USA
- Fodor, J.A. and Pylyshyn, Z.W. (1988) Connectionism and cognitive architecture: A critical analysis *Cognition* **28**(1), pp. 3–71
- Foglia, L. and Wilson, R. (2013) Embodied cognition *Wiley Interdisciplinary Reviews: Cognitive Science* **4**
- Frazer, R. (2023) Marketing against extremism: Identifying and responding to moral disengagement cues in islamic state terrorist propaganda *Journal of Public Policy & Marketing* **42**(1), pp. 36–55
- Gallaher, C. (2021) Mainstreaming white supremacy: a twitter analysis of the american 'alt-right' *Gender, Place & Culture* **28**(2), pp. 224–252
- Gamage, G., De Silva, D., Mills, N., Alahakoon, D. and Manic, M. (2024) Emotion aware: an artificial intelligence framework for adaptable, robust, explainable, and multi-granular emotion analysis *Journal of Big Data* **11**(1)
- Gamage, G., Mills, N., Rathnayaka, P., Jennings, A. and Alahakoon, D. (2022) Cooee: An artificial intelligence chatbot for complex energy environments in: *2022 15th International Conference on Human System Interaction (HSI)* pp. 1–5
- Ganesh, B. and Bright, J. (2020) Countering extremists on social media: Challenges for strategic communication and content moderation
- Gernsbacher, M. (1999) Comprehension: A paradigm for cognition *American Scientist* **87**
- Gernsbacher, M.A. and Kaschak, M.P. (2003) Neuroimaging studies of language production and comprehension *Annual Review of Psychology* **54**(1), p. 91–114

- Ghanbari-Adivi, F. and Mosleh, M. (2019) Text emotion detection in social networks using a novel ensemble classifier based on parzen tree estimator (tpe) *Neural Computing and Applications* **31**, pp. 8971–8983
- Giachanou, A., Rosso, P. and Crestani, F. (2019) Leveraging emotional signals for credibility detection in: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR'19* p. 877–880 Association for Computing Machinery, New York, NY, USA
- GitHub (2024) Github copilot <https://github.com/features/copilot/> accessed: 2024-10-20
- Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative adversarial networks
- Graesser, A.C., Singer, M. and Trabasso, T. (1994) Constructing inferences during narrative text comprehension. *Psychological Review* **101**(3), p. 371–395
- Gule, L. (2019) Hate and identity: a social philosophical attempt to understand extremism
- Guo, C., Cao, J., Zhang, X., Shu, K. and Yu, M. (2019a) Exploiting emotions for fake news detection on social media *ArXiv* **abs/1903.01728**
- Guo, J., Zhan, Z., Gao, Y., Xiao, Y., Lou, J.G., Liu, T. and Zhang, D. (2019b) Towards complex text-to-SQL in cross-domain database with intermediate representation in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Eds.) A. Korhonen, D. Traum and L. Màrquez pp. 4524–4535 Association for Computational Linguistics, Florence, Italy
- Hardeniya, T. and Borikar, D.A. (2016) Dictionary based approach to sentiment analysis - a review *International Journal of Advanced Engineering, Management and Science* **2**(5)
- Harris, Z.S. (1954) Distributional structure
- Harrison, J.D., Young, J.M., Price, M.A., Butow, P.N. and Solomon, M.J. (2009) What are the unmet supportive care needs of people with cancer? a systematic review *Supportive Care in Cancer* **17**(8), p. 1117–1128
- Hasan, M., Rundensteiner, E. and Agu, E. (2019) Automatic emotion detection in text streams by analyzing twitter data *International Journal of Data Science and Analytics* **7**, pp. 35–51
- Hazoom, M., Malik, V. and Bogin, B. (2021) Text-to-sql in the wild: A naturally-occurring dataset based on stack exchange data *CoRR* **abs/2106.05006**
- Herzig, J., Nowak, P.K., Müller, T., Piccinno, F. and Eisenschlos, J. (2020) TaPas: Weakly supervised table parsing via pre-training in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Eds.) D. Jurafsky, J. Chai, N. Schluter and J. Tetreault pp. 4320–4333 Association for Computational Linguistics, Online
- von Heusinger, K., Maienborn, C. and Portner, P., (Eds.) (2019) *Semantics - Foundations, History and Methods* De Gruyter Mouton, Berlin, Boston
- Hill, F., Reichart, R. and Korhonen, A. (2015a) SimLex-999: Evaluating semantic models with (genuine) similarity estimation *Computational Linguistics* **41**(4), pp. 665–695

- Hill, F., Reichart, R. and Korhonen, A. (2015b) SimLex-999: Evaluating semantic models with (genuine) similarity estimation *Computational Linguistics* **41**(4), pp. 665–695
- Hitzler, P., Ebrahimi, M., Sarker, M.K. and Stepanova, D. (2024) Neuro-symbolic ai and the semantic web *Semantic Web* **15**(4), pp. 1261–1263
- Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory *Neural Comput.* **9**(8), p. 1735–1780
- Hofmann, T. (2013) Probabilistic latent semantic analysis *CoRR* **abs/1301.6705**
- Huang, Y.H., Lee, S.R., Ma, M.Y., Chen, Y.H., Yu, Y.W. and Chen, Y.S. (2019) Emotionx-idea: Emotion bert – an affectional model for conversation
- Hutto, C. and Gilbert, E. (2014) Vader: A parsimonious rule-based model for sentiment analysis of social media text *Proceedings of the International AAAI Conference on Web and Social Media* **8**, pp. 216–225
- Idioms, T. (2024) The idioms - largest idioms dictionary <https://www.theidioms.com/> accessed: 2024-10-20
- Ilyas, A., Shahzad, K. and Malik, M.K. (2022) Emotion detection in code-mixed roman urdu - english text *ACM Transactions on Asian and Low-Resource Language Information Processing*
- Ishaku, B., Aksit, S. and Maza, K.D. (2021) The role of faith-based organizations in counter-radicalization in nigeria: The case of boko haram *Religions* **12**(11)
- Jain, V.K., Kumar, S. and Fernandes, S.L. (2017) Extraction of emotions from multilingual text using intelligent text processing and computational linguistics *Journal of Computational Science* **21**, pp. 316–326
- Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T. and Sayed, W.E. (2023) Mistral 7b
- Jiang, P., Sonne, C., Li, W., You, F. and You, S. (2024) Preventing the immense increase in the life-cycle energy and carbon footprints of llm-powered intelligent chatbots *Engineering* **40**, p. 202–210
- Jolliffe, I. (2011) *Principal Component Analysis* pp. 1094–1096 Springer Berlin Heidelberg
- Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L. and Levy, O. (2020) Spanbert: Improving pre-training by representing and predicting spans *Transactions of the Association for Computational Linguistics* **8**, pp. 64–77
- Jurafsky, D. and Martin, J.H. (2024) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* 3rd ed. online manuscript released August 20, 2024
- Jwa, H., Oh, D., Park, K., Kang, J.M. and Lim, H. (2019) exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert) *Applied Sciences* **9**(19)

- K., A., P., D., Abraham, S.S., L., L.V. and Gangan, M.P. (2022) Readers' affect: predicting and understanding readers' emotions with deep learning *Journal of Big Data* **9**, p. 82
- Kahneman, D. (1992) Reference points, anchors, norms, and mixed feelings *Organizational Behavior and Human Decision Processes* **51**(2), p. 296–312
- Kaliyar, R.K., Goswami, A. and Narang, P. (2021) Fakebert: Fake news detection in social media with a bert-based deep learning approach *Multimedia Tools and Applications* **80**(8), p. 11765–11788
- Kang, X., Shteingardt, V., Wang, Y. and Dori, D. (2025) Neuro-conceptual artificial intelligence: Integrating opm with deep learning to enhance question answering quality
- Kendeou, P., van den Broek, P., White, M.J. and Lynch, J.S. (2009) Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of Educational Psychology* **101**(4), p. 765–778
- Khattak, A., Asghar, M.Z., Khalid, H.A. and Ahmad, H. (2022) Emotion classification in poetry text using deep neural network *Multimedia Tools and Applications* **81**, pp. 26223–26244
- Kim, S.M., Valitutti, A. and Calvo, R. (2010) Evaluation of unsupervised emotion models to textual affect recognition pp. 62–70 Association for Computational Linguistics
- Kintsch, W. (1998) *Comprehension: A Paradigm for Cognition* Cambridge University Press, Cambridge, UK
- Klebanov, B.B., Burstein, J. and Madnani, N. (2013) Sentiment profiles of multiword expressions in test-taker essays: The case of noun-noun compounds *ACM Trans. Speech Lang. Process.* **10**
- van Kleef, G.A. and Côté, S. (2022) The social effects of emotions *Annual Review of Psychology* **73**(1), p. 629–658
- van Kleef, G.A., De Dreu, C.K.W. and Manstead, A.S.R. (2004) The interpersonal effects of anger and happiness in negotiations. *Journal of Personality and Social Psychology* **86**(1), p. 57–76
- Klinger, R., De Clercq, O., Mohammad, S.M. and Balahur, A. (2018) IEST: WASSA-2018 Implicit Emotions Shared Task Technical report
- Kolluri, N., Liu, Y. and Murthy, D. (2022) Covid-19 misinformation detection: Machine-learned solutions to the infodemic *JMIR Infodemiology* **2**(2), p. e38756
- Kosinski, M. (2023) Evaluating large language models in theory of mind tasks
- Kumar, S. and Zymbler, M. (2019) A machine learning approach to analyze customer satisfaction from airline tweets *Journal of Big Data* **6**, p. 62
- Kumar, Y., Saini, S., Sharma, H., Payal, R. and Mishra, A. (2022) Feedback investigation on twitter dataset using classification approaches in: *Proceedings of International Conference on Recent Trends in Computing*, (Eds.) R.P. Mahapatra, S.K. Peddoju, S. Roy, P. Parwekar and L. Goel pp. 251–262 Springer Nature Singapore, Singapore

- Lambert, N., Castricato, L., von Werra, L. and Havrilla, A. (2022) Illustrating reinforcement learning from human feedback (rlhf) *Hugging Face Blog* <https://huggingface.co/blog/rlhf>
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. and Soricut, R. (2019) Albert: A lite bert for self-supervised learning of language representations
- Landauer, T.K. and Dumais, S.T. (1997) A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* **104**(2), p. 211–240
- LangChain Team (2023) Langchain: Building applications with llms through composability <https://github.com/langchain-ai/langchain> accessed: 2024-10-29
- Larcker, D.F. and Zakolyukina, A.A. (2012) Detecting deceptive discussions in conference calls *SSRN Electronic Journal*
- Lazaridou, A., Marelli, M. and Baroni, M. (2017) Multimodal word meaning induction from minimal exposure to natural text *Cognitive Science* **41**(S4), p. 677–705
- Lazer, D.M.J., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S.A., Sunstein, C.R., Thorson, E.A., Watts, D.J. and Zittrain, J.L. (2018) The science of fake news *Science* **359**(6380), p. 1094–1096
- Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-based learning applied to document recognition *Proceedings of the IEEE* **86**(11), pp. 2278–2324
- Lee, J., Chang, R.W., Ehrlich-Jones, L., Kwoh, C.K., Nevitt, M., Semanik, P.A., Sharma, L., Sohn, M.W., Song, J. and Dunlop, D.D. (2015) Sedentary behavior and physical function: Objective evidence from the osteoarthritis initiative *Arthritis Care Research* **67**, pp. 366–373
- Lenadora, D., Gamage, G., Haputhanthri, D., Meedeniya, D. and Perera, I. (2020) Exploratory analysis of a social media network in sri lanka during the COVID-19 virus outbreak *CoRR* **abs/2006.07855**
- Li, T.M.H., Chau, M., Wong, P.W.C. and Yip, P.S.F. (2012) *A Hybrid System for Online Detection of Emotional Distress* pp. 73–80
- Li, X., Gao, W., Feng, S., Wang, D. and Joty, S. (2021) Span-level emotion cause analysis by bert-based graph attention network pp. 3221–3226 ACM
- Liebrecht, C. and van Hooijdonk, C. (2020) Creating humanlike chatbots: What chatbot developers could learn from webcare employees in adopting a conversational human voice in: *Chatbot Research and Design*, (Eds.) A. Følstad, T. Araujo, S. Papadopoulos, E.L.C. Law, O.C. Granmo, E. Luger and P.B. Brandtzaeg pp. 51–64 Springer International Publishing, Cham
- Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L. (2014) Microsoft coco: Common objects in context in: *Computer Vision – ECCV 2014*, (Eds.) D. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars pp. 740–755 Springer International Publishing, Cham

- Lisy, K., Langdon, L., Piper, A. and Jefford, M. (2019) Identifying the most prevalent unmet needs of cancer survivors in australia: A systematic review *Asia-Pacific Journal of Clinical Oncology* **15**(5)
- LITKOWSKI, E. and RENKEN, M. (2013) Trusting what you're told: How children learn from others, by paul I.harris. the belknap press of harvard university press, cambridge, ma, usa, 2012. 253 pp. isbn 978-0-674-06572-7. *Science Education* **97**(5), p. 797–799
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019) Roberta: A robustly optimized bert pretraining approach
- LOUGHRAN, T. and MCDONALD, B. (2011) When is a liability not a liability? textual analysis, dictionaries, and 10-ks *The Journal of Finance* **66**, pp. 35–65
- Lu, Z., Afridi, I., Kang, H.J., Ruchkin, I. and Zheng, X. (2024) Surveying neuro-symbolic approaches for reliable artificial intelligence of things *Journal of Reliable Intelligent Environments* **10**(3), p. 257–279
- Malo, P., Sinha, A., Korhonen, P., Wallenius, J. and Takala, P. (2014) Good debt or bad debt: Detecting semantic orientations in economic texts *Journal of the Association for Information Science and Technology* **65**, pp. 782–796
- Manchunhui (2020) Us election 2020 tweets accessed: 2024-07-27
- Mandera, P., Keuleers, E. and Brysbaert, M. (2017) Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation *Journal of Memory and Language* **92**, pp. 57–78
- Marcus, G. (2020) The next decade in ai: Four steps towards robust artificial intelligence
- Martin, L. and Pu, P. (2014) Prediction of helpful reviews using emotions extraction *Proceedings of the AAAI Conference on Artificial Intelligence* **28**
- Matamoros-Fernández, A. and Farkas, J. (2021) Racism, hate speech, and social media: A systematic review and critique *Television & New Media* **22**(2), pp. 205–224
- Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P. and Mukherjee, A. (2020) Hatexplain: A benchmark dataset for explainable hate speech detection *CoRR* **abs/2012.10289**
- Matsumoto, D., Pronk, M. and Roelofsen, E. (2011) What makes conference calls useful? the information content of managers' presentations and analysts' discussion sessions *The Accounting Review* **86**(4), p. 1383–1414
- MAYEW, W.J. and VENKATACHALAM, M. (2012) The power of voice: Managerial affective states and future firm performance *The Journal of Finance* **67**(1), p. 1–43
- Merkx, D., Frank, S. and Ernestus, M. (2022) Seeing the advantage: visually grounding word embeddings to better capture human semantic knowledge in: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, (Eds.) E. Chersoni, N. Hollenstein, C. Jacobs, Y. Oseki, L. Prévot and E. Santus pp. 1–11 Association for Computational Linguistics, Dublin, Ireland

- Merriam-Webster (2024) Merriam-webster thesaurus <https://www.merriam-webster.com/thesaurus/> accessed: 2024-08-26
- Miao, X., Oliaro, G., Zhang, Z., Cheng, X., Jin, H., Chen, T. and Jia, Z. (2023) Towards efficient generative large language model serving: A survey from algorithms to systems
- Microsoft Azure (2024) Azure AI Services accessed: 2024-06-17
- MidJourney (2024) Midjourney <https://www.midjourney.com/> accessed: 2024-10-20
- Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013a) Efficient estimation of word representations in vector space
- Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013b) Efficient estimation of word representations in vector space *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013c) Distributed representations of words and phrases and their compositionality in: *Advances in neural information processing systems* p. 3111
- Moghaddam, F.M. (2005) The staircase to terrorism: A psychological exploration. *American Psychologist* **60**(2), p. 161–169
- Mohammad, S.M. and Turney, P.D. (2013) Crowdsourcing a word-emotion association lexicon **29**(3), pp. 436–465
- Mondal, S., Das, S. and Vrana, V.G. (2023) How to bell the cat? a theoretical review of generative artificial intelligence towards digital disruption in all walks of life *Technologies* **11**(2)
- Muennighoff, N., Tazi, N., Magne, L. and Reimers, N. (2023) Mteb: Massive text embedding benchmark
- Mukherjee, P., Badr, Y., Doppalapudi, S., Srinivasan, S., Sangwan, R. and Sharma, R. (2021) Effect of negation in sentences on sentiment analysis and polarity detection *Procedia Computer Science* **185**, pp. 370–379 publisher Copyright: © 2021 Elsevier B.V.. All rights reserved.; 2021 Complex Adaptive Systems Conference ; Conference date: 16-06-2021 Through 18-06-2021
- Muzny, G. and Zettlemoyer, L. (2013) Automatic idiom identification in wiktionary in: *Conference on Empirical Methods in Natural Language Processing*
- Nallaperuma, D., De Silva, D., Alahakoon, D. and Yu, X. (2018) Intelligent detection of driver behavior changes for effective coordination between autonomous and human driven vehicles in: *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society* pp. 3120–3125
- Naous, T., Hokayem, C. and Hajj, H. (2020) Empathy-driven Arabic conversational chatbot in: *Proceedings of the Fifth Arabic Natural Language Processing Workshop* pp. 58–68 Association for Computational Linguistics, Barcelona, Spain (Online)

- Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V., Vainbrand, D., Kashinkunti, P., Bernauer, J., Catanzaro, B., Phanishayee, A. and Zaharia, M. (2021) Efficient large-scale language model training on GPU clusters *CoRR abs/2104.04473*
- Nawaratne, R., Bandaragoda, T., Adikari, A., Alahakoon, D., De Silva, D. and Yu, X. (2017) Incremental knowledge acquisition and self-learning for autonomous video surveillance in: *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society* p. 4790–4795 IEEE
- Neumann, P. (2013) The trouble with radicalization *International Affairs* **89**, pp. 873–893
- Newman, M.L., Pennebaker, J.W., Berry, D.S. and Richards, J.M. (2003) Lying words: Predicting deception from linguistic styles *Personality and Social Psychology Bulletin* **29**(5), pp. 665–675 pMID: 15272998
- Nida, H., Mahira, K., Mudasir, M., Ahmed, M.M. and Mohsin, M. (2019) *Automatic emotion classifier* vol. 713 pp. 565–572
- Nijhawan, T., Attigeri, G. and Ananthakrishna, T. (2022) Stress detection using natural language processing and machine learning over social interactions *Journal of Big Data* **9**, p. 33
- Olan, F., Jayawickrama, U., Arakpogun, E.O., Suklan, J. and Liu, S. (2022) Fake news on social media: the impact on society *Information Systems Frontiers* **26**(2), p. 443–458
- Olshannikova, E., Olsson, T., Huhtamäki, J. and Kärkkäinen, H. (2017) Conceptualizing big social data *Journal of Big Data* **4**, p. 3
- OpenAI (2023) Gpt-4 technical report
- OpenAI (2024) Chatgpt <https://chat.openai.com/> accessed: 2024-10-20
- O’Hanlon, R., Altice, F.L., Lee, R.K.W., LaViolette, J., Mark, G., Papakyriakopoulos, O., Saha, K., Choudhury, M.D. and Kumar, N. (2024) Misogynistic extremism: A scoping review *Trauma, Violence, & Abuse* **25**(2), pp. 1219–1234 pMID: 37272372
- Park, J. and Myaeng, S.h. (2017) A computational study on word meanings and their distributed representations via polymodal embedding in: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Eds.) G. Kondrak and T. Watanabe pp. 214–223 Asian Federation of Natural Language Processing, Taipei, Taiwan
- Park, J.H., Xu, P. and Fung, P. (2018) Plusemo2vec at semeval-2018 task 1: Exploiting emotion knowledge from emoji and hashtags pp. 264–272
- Partee, B.H. (1984) Compositionality in: *Varieties of Formal Semantics*, (Eds.) F. Landman and F. Veltman pp. 281–312 Foris Publications, Dordrecht

- Pasupat, P. and Liang, P. (2015) Compositional semantic parsing on semi-structured tables in: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (Eds.) C. Zong and M. Strube pp. 1470–1480 Association for Computational Linguistics, Beijing, China
- Pelosi, S. (2020) Semantically oriented idioms for sentiment analysis. a linguistic resource for the italian language *Advances in Intelligent Systems and Computing* **1151 AISC**, pp. 1069–1077
- Pennington, J., Socher, R. and Manning, C. (2014) GloVe: Global vectors for word representation in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Eds.) A. Moschitti, B. Pang and W. Daelemans pp. 1532–1543 Association for Computational Linguistics, Doha, Qatar
- Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018a) Deep contextualized word representations in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (Eds.) M. Walker, H. Ji and A. Stent pp. 2227–2237 Association for Computational Linguistics, New Orleans, Louisiana
- Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018b) Deep contextualized word representations *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* **1**, pp. 2227–2237
- Petrov, G.T.S., Dean, J. and Vinyals, O. (2024) Gemini: A family of highly capable multi-modal models
- Pinheiro Cinelli, L., Araújo Marins, M., Barros da Silva, E.A. and Lima Netto, S. (2021) *Variational Autoencoder* pp. 111–149 Springer International Publishing, Cham
- Plotly Technologies Inc. (2023) Plotly: A python graphing library <https://plotly.com/python/> accessed: 2023-10-29
- Plutchik, R. (1991) *The emotions* University Press of America
- Plutchik, R. (2001) *The Nature of Emotions* vol. 89
- Poria, S., Cambria, E., Bajpai, R. and Hussain, A. (2017) A review of affective computing: From unimodal analysis to multimodal fusion *Information Fusion* **37**, p. 98–125
- Poria, S., Gelbukh, A., Cambria, E., Hussain, A. and Huang, G.B. (2014) Emosenticspace: A novel framework for affective common-sense reasoning *Knowledge-Based Systems* **69**, pp. 108–123
- Prabagar, K., Srikanthabala, K., Loganathan, N., Silva, D.D., Gamage, G., Rathnayaka, P., Perera, A.S. and Alahakoon, D. (2022) Investigating covid-19 vaccine messaging in online social networks using artificial intelligence pp. 1–6 IEEE
- Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N.A. and Lewis, M. (2023) Measuring and narrowing the compositionality gap in language models

- Price, S.M., Doran, J.S., Peterson, D.R. and Bliss, B.A. (2012) Earnings conference calls and stock returns: The incremental informativeness of textual tone *Journal of Banking and Finance* **36**(4), p. 992–1011
- Priyanka and Sinha, R.M. (2014) A system for identification of idioms in hindi 2014 7th *International Conference on Contemporary Computing, IC3 2014* pp. 467–472
- Putro, M.D., Nguyen, D.L. and Jo, K.H. (2020) A dual attention module for real-time facial expression recognition in: *IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society* pp. 411–416
- Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N. and Huang, X. (2020) Pre-trained models for natural language processing: A survey
- Rabeya, T., Ferdous, S., Ali, H.S. and Chakraborty, N.R. (2017) A survey on emotion detection: A lexicon based backtracking approach for detecting emotion from bengali text pp. 1–7 IEEE
- Radford, A. and Narasimhan, K. (2018) Improving language understanding by generative pre-training
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J. (2019) Exploring the limits of transfer learning with a unified text-to-text transformer *CoRR* **abs/1910.10683**
- Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P. (2016) Squad: 100, 000+ questions for machine comprehension of text in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* Association for Computational Linguistics
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M. (2022) Hierarchical text-conditional image generation with clip latents *arXiv preprint arXiv:2204.06125*
- Ramnath, S., Nema, P., Sahni, D. and Khapra, M.M. (2020) Towards interpreting bert for reading comprehension based qa pp. 3236–3242 Association for Computational Linguistics
- Ranasinghe, S., Gamage, G., Moraliyage, H., Mills, N., McCaffrey, N., Bucholc, J., Lane, K., Cahill, A., White, V. and Silva, D.D. (2022) An artificial intelligence framework for the detection of emotion transitions in telehealth services pp. 1–5 IEEE
- Refinitiv (2024) I/b/e/s (institutional brokers' estimate system) database available from: <https://www.lseg.com/en/data-analytics/financial-data/company-data/ibes-estimates>
- Repenning, N., Löhlein, L. and Schäffer, U. (2021) Emotions in accounting: A review to bridge the paradigmatic divide *European Accounting Review* **31**(1), p. 241–267
- Rotaru, A.S. and Vigliocco, G. (2020) Constructing semantic models from words, images, and emojis *Cognitive Science* **44**(4), p. e12830
- Rozental, A. and Fleischer, D. (2018) Amobee at semeval-2018 task 1: Gru neural network with a cnn attention mechanism for sentiment classification pp. 218–225 Association for Computational Linguistics

- Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning representations by back-propagating errors *Nature* **323**, pp. 533–536
- Russin, J., McGrath, S.W., Williams, D.J. and Elber-Dorozko, L. (2024) From frege to chatgpt: Compositionality in language, cognition, and deep neural networks
- Saffar, A.H., Mann, T.K. and Ofoghi, B. (2023) Textual emotion detection in health: Advances and applications *Journal of Biomedical Informatics* **137**, p. 104258
- Salton, G. (2017) Representations of idioms for natural language processing: Idiom type and token identification, language modelling and neural machine translation *Doctoral*
- Salton, G., Ross, R., Kelleher, J.D., Salton, G.D. and Ross, R.J. (2014) An empirical study of the impact of idioms on phrase based statistical machine translation of english to brazilian-portuguese *Conference papers*
- Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D. and Gadepally, V. (2023) From words to watts: Benchmarking the energy costs of large language model inference pp. 1–9
- Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019) Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter
- Saravia, E., Liu, H.C.T., Huang, Y.H., Wu, J. and Chen, Y.S. (2018a) Carer: Contextualized affect representations for emotion recognition pp. 3687–3697 Association for Computational Linguistics
- Saravia, E., Liu, H.C.T., Huang, Y.H., Wu, J. and Chen, Y.S. (2018b) CARER: Contextualized affect representations for emotion recognition in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (Eds.) E. Riloff, D. Chiang, J. Hockenmaier and J. Tsujii pp. 3687–3697 Association for Computational Linguistics, Brussels, Belgium
- Saxena, P. and Paul, S. (2020) Epie dataset: A corpus for possible idiomatic expressions *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **12284 LNAI**, pp. 87–94
- Sazzed, S. and Jayarathna, S. (2021) Ssentia: A self-supervised sentiment analyzer for classification from unlabeled data *Machine Learning with Applications* **4**, p. 100026
- Scherer, K.R. and Wallbott, H.G. (1994a) Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology* **66**, pp. 310–328
- Scherer, K.R. and Wallbott, H.G. (1994b) Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology* **66**(2), p. 310
- Schoeninger, G. (2023) Arxiv dives - a mathematical framework for transformer circuits - part 1 accessed: 2024-10-31

- Schreuder, R., D'Arcais, G.B.F. and Glazenborg, G. (1985) Semantic decomposition and word recognition in: *Meaning and the lexicon*, (Eds.) G.A.J. Hoppenbrouwers, P.A.M. Seuren and A.J.M.M. Weijters pp. 108–114 Foris Publications
- Seal, D., Roy, U.K. and Basak, R. (2020) *Sentence-Level Emotion Detection from Text Based on Semantic Rules* vol. 933 pp. 423–430
- Shahane, S. (2024) Twitter sentiment dataset accessed: 2024-08-26
- Shaikh, O., Zhang, H., Held, W., Bernstein, M. and Yang, D. (2023) On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning
- Shen, W., Chen, J., Quan, X. and Xie, Z. (2020) Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition
- Shrivastava, K., Kumar, S. and Jain, D.K. (2019) An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network *Multimedia Tools and Applications* **78**, pp. 29607–29639
- Shu, K., Mahudeswaran, D., Wang, S., Lee, D. and Liu, H. (2020) Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media *Big Data* **8**(3), pp. 171–188 pMID: 32491943
- Silva, D.D., Ranasinghe, W., Bandaragoda, T., Adikari, A., Mills, N., Iddamalagoda, L., Alahakoon, D., Lawrentschuk, N., Persad, R., Osipov, E., Gray, R. and Bolton, D. (2018) Machine learning to support social media empowered patients in cancer care and cancer treatment decisions *PLOS ONE* **13**, p. e0205855
- Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P. and Satoh, S. (2019) Spotfake: A multi-modal framework for fake news detection in: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)* pp. 39–47
- Sintsova, V. and Pu, P. (2016) Dystemo *ACM Transactions on Intelligent Systems and Technology* **8**, pp. 1–22
- Smith, S. and O'Hare, A. (2022) Comparing traditional news and social media with stock price movements; which comes first, the news or the price change? *Journal of Big Data* **9**, p. 47
- Sohangir, S., Wang, D., Pomeranets, A. and Khoshgoftaar, T.M. (2018) Big data: Deep learning for financial sentiment analysis *Journal of Big Data* **5**, p. 3
- Sowa, J. (1984) *Conceptual Structures: Information Processing in Mind and Machine The Systems Programming Series*
- S&P Capital IQ (2024) Compustat database available from: <https://www.spglobal.com/marketintelligence/en/campaigns/compustat>
- Spenader, J. and Blutner, R. (2007) Compositionality and systematicity *Sigecom Exchanges* pp. 163–174

- Staiano, J. and Guerini, M. (2014) Depechemood: A lexicon for emotion analysis from crowd-annotated news *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference 2*, pp. 427–433
- Strapparava, C. and Mihalcea, R. (2007) {S}em{E}val-2007 task 14: Affective text pp. 70–74 Association for Computational Linguistics
- Strapparava, C. and Valitutti, A. (2004) {W}ord{N}et affect: an affective extension of {W}ord{N}et European Language Resources Association (ELRA)
- Strobel, H., Gehrmann, S., Behrisch, M., Perer, A., Pfister, H. and Rush, A.M. (2018) Seq2seq-vis: A visual debugging tool for sequence-to-sequence models
- Suglia, A., Konstas, I. and Lemon, O. (2024) Visually grounded language learning: A review of language games, datasets, tasks, and models *Journal of Artificial Intelligence Research* **79**, p. 173–239
- Suhasini, M. and Srinivasu, B. (2020) *Emotion Detection Framework for Twitter Data Using Supervised Classifiers* vol. 1079 pp. 565–576
- Suresh, K. (2021) Covid vaccine tweets <https://www.kaggle.com/datasets/kaushiksuresh147/covidvaccine-tweets> accessed: 2024-07-26
- Sutskever, I., Vinyals, O. and Le, Q.V. (2014a) Sequence to sequence learning with neural networks in: *Advances in Neural Information Processing Systems*, (Eds.) Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence and K. Weinberger vol. 27 Curran Associates, Inc.
- Sutskever, I., Vinyals, O. and Le, Q.V. (2014b) Sequence to sequence learning with neural networks
- Sutton, R.I. (1991) Maintaining norms about expressed emotions: The case of bill collectors *Administrative Science Quarterly* **36**(2), p. 245
- Syakur, M.A., Khotimah, B.K., Rochman, E.M.S. and Satoto, B.D. (2018) Integration k-means clustering method and elbow method for identification of the best customer profile cluster *IOP Conference Series: Materials Science and Engineering* **336**, p. 012017
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) Rethinking the inception architecture for computer vision in: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 2818–2826
- Tervonen, H.E., Aranda, S., Roder, D., You, H., Walton, R., Morrell, S., Baker, D. and Currow, D.C. (2017) Cancer survival disparities worsening by socio-economic disadvantage over the last 3 decades in new south wales, australia *BMC Public Health* **17**(1)
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. and Lample, G. (2023) Llama: Open and efficient foundation language models
- Trischler, A., Ye, Z., Yuan, X. and Suleman, K. (2016) Natural language comprehension with the epireader

- Tzacheva, A., Ranganathan, J. and Mylavarapu, S.Y. (2019) *Actionable Pattern Discovery for Tweet Emotions* p. 46–57 Springer International Publishing
- Tzacheva, A., Ranganathan, J. and Mylavarapu, S.Y. (2020) *Actionable Pattern Discovery for Tweet Emotions* vol. 965 pp. 46–57
- Udochukwu, O. and He, Y. (2015) *A Rule-Based Approach to Implicit Emotion Detection in Text* vol. 9103 pp. 197–203
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) Attention is all you need
- Verma, P.K., Agrawal, P., Amorim, I. and Prodan, R. (2021) Welfake: Word embedding over linguistic features for fake news detection *IEEE Transactions on Computational Social Systems* **8**(4), pp. 881–893
- Vijayvergia, A. and Kumar, K. (2021) Selective shallow models strength integration for emotion detection using glove and lstm *Multimedia Tools and Applications* **80**, pp. 28349–28363
- Vosoughi, S., Roy, D. and Aral, S. (2018) The spread of true and false news online *Science* **359**(6380), pp. 1146–1151
- Wallon, H. (1972) *The emotions* vol. 1 University Press of America
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S. (2018a) GLUE: A multi-task benchmark and analysis platform for natural language understanding in: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, (Eds.) T. Linzen, G. Chrupała and A. Alishahi pp. 353–355 Association for Computational Linguistics, Brussels, Belgium
- Wang, T., Molassiotis, A., Chung, B.P.M. and Tan, J.Y. (2018b) Unmet care needs of advanced cancer patients and their informal caregivers: a systematic review *BMC Palliative Care* **17**(1)
- Wang, W., Chen, L., Thirunarayan, K. and Sheth, A.P. (2012) Harnessing twitter "big data" for automatic emotion identification in: *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing* pp. 587–592
- Wang, W.Y. (2017) "liar, liar pants on fire": A new benchmark dataset for fake news detection in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (Eds.) R. Barzilay and M.Y. Kan pp. 422–426 Association for Computational Linguistics, Vancouver, Canada
- Wang, X., Zhao, H., Yang, T. and Wang, H. (2020) Correcting the misuse: A method for the chinese idiom cloze test pp. 1–10
- Wang, Y., Feng, S., Wang, D., Yu, G. and Zhang, Y. (2016) *Multi-label Chinese microblog emotion classification via convolutional neural network* vol. 9931 LNCS pp. 567–580
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Le, Q. and Zhou, D. (2022) Chain of thought prompting elicits reasoning in large language models *CoRR* **abs/2201.11903**

- Williams, L., Bannister, C., Arribas-Ayllon, M., Preece, A. and Spasić, I. (2015) The role of idioms in sentiment analysis *Expert Systems with Applications* **42**, pp. 7375–7385
- Winata, G.I., Madotto, A., Lin, Z., Shin, J., Xu, Y., Xu, P. and Fung, P. (2019) Caire hkust at semeval-2019 task 3: Hierarchical attention for dialogue emotion classification pp. 142–147 Association for Computational Linguistics
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q. and Rush, A. (2020) Transformers: State-of-the-art natural language processing pp. 38–45 Association for Computational Linguistics
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q. and Rush, A.M. (2019) Huggingface’s transformers: State-of-the-art natural language processing
- Wu, C.H., Chuang, Z.J. and Lin, Y.C. (2006) Emotion recognition from text using semantic labels and separable mixture models *ACM Transactions on Asian Language Information Processing* **5**, pp. 165–183
- Wu, S. and Dredze, M. (2019) Beto, bentz, becas: The surprising cross-lingual effectiveness of bert *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* pp. 833–844
- Wu, Z. and Jiang, Y. (2019a) Disentangling latent emotions of word embeddings on complex emotional narratives in: *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II* p. 587–595 Springer-Verlag, Berlin, Heidelberg
- Wu, Z. and Jiang, Y. (2019b) Disentangling latent emotions of word embeddings on complex emotional narratives in: *CCF International Conference on Natural Language Processing and Chinese Computing* pp. 587–595 Springer
- Xiao, J. (2019) Figure eight at semeval-2019 task 3: Ensemble of transfer learning methods for contextual emotion detection pp. 220–224 Association for Computational Linguistics
- Xu, X., Liu, C. and Song, D. (2017) Sqlnet: Generating structured queries from natural language without reinforcement learning *CoRR* **abs/1711.04436**
- Yang, F., Pentyala, S.K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E.D., Ji, S. and Hu, X.B. (2019) Xfake: Explainable fake news detector with visualizations in: *The World Wide Web Conference WWW ’19* p. 3600–3604 Association for Computing Machinery, New York, NY, USA
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. and Cao, Y. (2023) React: Synergizing reasoning and acting in language models

- Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z. and Zhang, Y. (2024) A survey on large language model (llm) security and privacy: The good, the bad, and the ugly *High-Confidence Computing* **4**(2), p. 100211
- Yu, D., Yang, B., Liu, D., Wang, H. and Pan, S. (2023) A survey on neural-symbolic learning systems *Neural Networks* **166**, pp. 105–126
- Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., Zhang, Z. and Radev, D.R. (2018) Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task *CoRR* **abs/1809.08887**
- Yusifova, G.I. (2013) Syntactic features of english idioms *International Journal of English Linguistics* **3**, p. p133
- Zarieczna, N., Bertlich, T., Večkalov, B. and Rutjens, B.T. (2023) Spirituality is associated with covid-19 vaccination scepticism *Vaccine* **41**(1), p. 226–235
- Zeng, Z., Zhang, S., Ren, L., Lin, H. and Yang, L. (2021) Senti-bsas: A bert-based classification model with sentiment calculating for happiness research pp. 272–277 ACM
- Zhang, J.L., Davis, A.K., Ge, W. and Matsumoto, D.A. (2012) The effect of managerial “style” on the tone of earnings conference calls *SSRN Electronic Journal*
- Zhang, Y., Jin, R. and Zhou, Z.H. (2010) Understanding bag-of-words model: a statistical framework *International Journal of Machine Learning and Cybernetics* **1**, pp. 43–52
- Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D. and Du, M. (2023) Explainability for large language models: A survey
- Zhong, V., Xiong, C. and Socher, R. (2017) Seq2sql: Generating structured queries from natural language using reinforcement learning *CoRR* **abs/1709.00103**
- Zhou, X., Nie, Y. and Bansal, M. (2022) Distributed NLI: Learning to predict human opinion distributions for language reasoning in: *Findings of the Association for Computational Linguistics: ACL 2022* pp. 972–987 Association for Computational Linguistics, Dublin, Ireland
- Zhou, X. and Zafarani, R. (2020) A survey of fake news: Fundamental theories, detection methods, and opportunities *ACM Comput. Surv.* **53**(5)
- Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P.F. and Irving, G. (2019) Fine-tuning language models from human preferences *CoRR* **abs/1909.08593**
- Škvorec, T., Gantar, P. and Robnik-Šikonja, M. (2020) Mice: Mining idioms with contextual embeddings *Knowledge-Based Systems* **235**