



DURBAN UNIVERSITY OF TECHNOLOGY
INYUVESI YASETHEKWINI YEZOBUCHWEPHESHE
CREATIVE. DISTINCTIVE. IMPACTFUL.

**A Context-Aware Word Embedding Model for
Morphologically Rich Languages using Sesotho sa
Leboa as a Case Study**

by

Mosima Anna Masethe

22176153

Thesis submitted in fulfilment of the requirements of the

Doctor of Philosophy in Information Technology

in the

Department of Information Technology

FACULTY OF ACCOUNTING AND INFORMATICS

at the

Durban University of Technology

Supervisor: Professor Sunday O Ojo

31 March 2026

DECLARATION

I, Mosima Anna Masethe, declare that the thesis work entitled “*A Context-Aware Word Embedding Model for Morphologically Rich Languages using Sesotho sa Leboa as a Case Study*” submitted in fulfilment of the requirement for the Doctor of Philosophy in Information Technology, at Durban University of Technology, is my own original work and that was not previously submitted for a degree at another institution of higher learning. All sources used or quoted have been indicated and acknowledged through complete referencing.

Signature:

Date: 31 March 2026

Approval for Final Submission

Supervisor : Professor Sunday O Ojo

Signature:

Date: 31/03/2026

ACKNOWLEDGEMENTS

I would like to seize this opportunity to praise my Lord and Saviour Jesus Christ, who is the source of my existence and the creator of my being. He has provided me with the opportunity and excellent health to pursue this education. The following individuals were introduced to me as a result of His guidance:

- o Professor Sunday Ojo, whose prayerful, intelligent, dedicated provision and professional supervision were my motivation in this study.
- o Pastor Israel Nkodima, my spiritual father, the one who shepherds our souls: God hear his prayers, and his guidance and inspiration to us.
- o My husband and my research partner: Dr. Hlaudi Daniel; our princess Pholosho (Future Advocate); and our princess Paballo (Future Engineer), and Phethego (Future Farmer) without whose comprehension and collaboration, this investigation would not have been feasible.
- o “Mme waka ke Montedi”, Mrs. Chuene Dinah (my Grade 1 teacher) and my humble Daddy (Bread-winner), Mr. Mameta Hendrick, who instilled the spirit of study in my heart and instructed me to prioritize my relationship with the God of heaven in all of my endeavours.
- o Durban University of Technology that provided support in terms of resources to pursue this study and also granted me an opportunity to complete this research study, and
- o Sefako Makgatho Health Sciences University for granting me the support and opportunity to further my studies.
- o National Research Fund (NRF) for granting me a BAAP programme to concentrate on my PhD study and publications.

“Faith is the key to open all doors” by Bro. WM Branham

PUBLICATIONS FROM THIS THESIS

Published Journal Articles -DHET Accredited

- i. **M. A. Masethe**, S. O. Ojo and H. D. Masethe (2025). Optimising Contextual Embeddings for Meaning Conflation Deficiency Resolution in Low-Resourced Languages. *Computers* 2025, 14(9), 402; <https://doi.org/10.3390/computers14090402>
- ii. **M. A. Masethe**, S. O. Ojo and H. D. Masethe (2025). A Context-Aware Embedding Approach to Meaning Conflation Deficiency in Sesotho sa Leboa: Addressing Semantic Ambiguity. *Forum for Linguistic Studies* 7 (8), 845–867; <https://doi.org/10.30564/fls.v7i8.9831>
- iii. **MA Masethe**, HD Masethe, SO Ojo (2024). Context-Aware Embedding Techniques for Addressing MCD in Morphologically Rich Languages Word Embedding: A Systematic Review and Meta Analysis, *Computers* 2024, 13(10), 271; <https://doi.org/10.3390/computers13100271>

Published Conference Papers

- i. **M. A. Masethe**, S. O. Ojo and H. D. Masethe, "A Bibliometric Analysis of Embedding Techniques for Addressing Meaning Conflation Deficiency in Low-Resourced Languages," 2025 Conference on Information Communications Technology and Society (ICTAS), Durban, South Africa, 2025, pp. 1-6, doi: 10.1109/ICTAS64866.2025.11155866.
- ii. **Masethe, Mosima Anna** and Masethe, Hlaudi Daniel and Ojo, Sunday Olusegun and Owolawi, Pius A, Paraphrase Generation Model Using Transformer Based Architecture (January 4, 2024). Available at SSRN: <https://ssrn.com/abstract=4683780> or <http://dx.doi.org/10.2139/ssrn.4683780>
- iii. **Masethe, Mosima Anna** and Masethe, Hlaudi Daniel and Ojo, Sunday Olusegun and Owolawi, Pius A, Word Sense Disambiguation Pipeline Framework for Low Resourced Morphologically Rich Languages (November 23, 2022). Available at SSRN: <https://ssrn.com/abstract=4332896> or <http://dx.doi.org/10.2139/ssrn.4332896>

- iv. **M. A. Masethe**, H. D. Masethe and S. O. Ojo, "Context-Based Question Answering Using Large Language BERT Variant Models for Low Resourced SsaL Language," 2024 4th International Multidisciplinary Information Technology and Engineering Conference (IMITEC), Vanderbijlpark, South Africa, 2024, pp. 507-513, <https://doi.org/10.1109/IMITEC60221.2024.10850997>

ABSTRACT

Meaning conflation deficiency (MCD) is a major issue in natural language processing (NLP) to improve low-resource languages with complex morphologies. This research study tackles the MCD problem in word vector space (WVS) word embedding (WE) models, that struggle with polysemous words. The study explored optimization strategies and graph-based architectures to resolve sense ambiguities in resource-limited environments. A computational experimental design (CED) assessed traditional neural models (BiLSTM, NNLM, CNN), transformer models (BERT, RoBERTa, ELECTRA, BART, XLNet, ELMo), and graph-based approaches (domain-specific, integrated and hierarchical). Models were tested on a *sense-annotated SsaL corpus* with polysemous terms, focusing on ambiguous words. Experimental results showed significant performance gains through optimization. Optimised BART improved accuracy to 96% from 87%, and XLNet increased precision to 92% from 83%. CNN models with dropout regularization saw the largest F1-score improvement, rising from 69% to 98%, highlighting regularization's importance for semantic sparsity. McNemar's test ($p < 0.05$) confirmed the improvements as significant across all models. The model was optimised using morphology-aware preprocessing, context-sensitive attention, SMOTE to address class imbalance, and tailored dropout regularisation to reduce overfitting.

Graph-based approaches excelled, with the domain-specific graph achieving near-perfect scores (98% precision, recall, F1-score, accuracy) by integrating linguistic knowledge. The integrated graph maintained a strong performance (88% across metrics), utilizing sentence similarity, domain indicators, and clustering. The hierarchical graph, with moderate results (83-86%), highlighted the importance of graph granularity in disambiguation. ROC curves (AUC 0.90-1.00) and confusion matrices validated the effectiveness of graph methods in capturing disambiguation cues, outperforming distributed representations with limited data. The analysis showed performance differences among model families. BERT was the strongest baseline (87% accuracy, 60% F1-score), followed by Multilingual BERT (89% accuracy, 55% F1). Multilingual models were inconsistent—Multilingual RoBERTa only had 53% accuracy and 17% F1-score—indicating poor representation of Bantu

morphological structures. Models performed well on frequent senses such as “*location*” but struggled with infrequent senses such as “*time*”, highlighting issues with class imbalance and data scarcity. Ablation studies highlighted key architectural insights for low-resource settings. The encoder-only BART configuration achieved approximately 97% of the full model’s overall classification accuracy, showing that reduced architectural complexity resulted in only a minor loss in predictive performance, suggesting that computationally efficient implementations can preserve predictive effectiveness while reducing model complexity. XLNet’s permutation-based attention effectively handled long-distance dependencies in agglutinative languages. Techniques such as morphology-sensitive tokenization and focused fine-tuning minimized misclassification errors and enhanced the detection of infrequent word senses across all model categories. This research contributes significantly to computational linguistics and low-resource NLP by showing that MCD requires multi-faceted optimization beyond architectural choices. The domain-specific graph framework bridges data-driven neural methods with linguistic encoding, achieving state-of-the-art performance via domain knowledge and semantic representation. The study proves that transformer models can adapt to low-resource contexts with careful fine-tuning, highlighting limits in current multilingual models for morphologically complex African languages. Findings impact low resourced language technology beyond SsaL, offering a blueprint for word sense disambiguation in morphologically rich languages with limited data. Semantic processing is feasible with constrained resources, promoting linguistic equity in NLP. Limitations include dataset scale, computational resource constraints for hyperparameter tuning, and domain-specific graph construction needing manual lexicon development. Future research should broaden sense inventories, explore cross-lingual transfer learning from Bantu languages, develop few-shot learning for new senses, and enhance disambiguation in applications such as machine translation. Creating morphological tokenization for Bantu languages and multilingual models trained on African languages is crucial for advancing computational semantics low resourced communities.

TABLE OF CONTENTS

DECLARATION	ii
ACKNOWLEDGEMENTS.....	iii
PUBLICATIONS FROM THIS THESIS	iv
ABSTRACT.....	vi
LIST OF FIGURES	xvi
LIST OF TABLES.....	xix
LIST OF ABBREVIATIONS AND ACRONYMS	xxi
CHAPTER 1 : INTRODUCTION	1
1.1 Background and Motivation	1
1.2 Problem Statement.....	6
1.3 Thesis Statement	8
1.4 Research Questions	8
1.5 The Study Aim and Objectives:.....	9
1.6 Alignment of Research Questions, Objectives, Methods and Deliverables	10
1.7 Overview of Research Methodology	11
1.8 Thesis Contribution.....	11
1.9 Thesis Structure	12
CHAPTER 2 : THE SESOTHO sa LEBOA LINGUISTIC CONTEXT.....	16
2.1 Introduction.....	16
2.2 The SsaL Language.....	17

2.3 SsaL as a Morphologically Rich Language	20
2.3.1 SsaL morphological typology	22
2.3.2 SsaL words morphology	23
2.3.3 Morphological processes in SsaL	24
2.3.4 SsaL Morphemes	25
2.3.5 Lexical categories morphology	26
2.3.6 Derivational, reduplication, and infixes morphology	29
2.3.7 SsaL orthography	31
2.3.8 SsaL Sentence Structure	32
2.4 Ambiguities in the SsaL Language	34
2.4.1 Lexical ambiguity	35
2.4.2 Morphological ambiguity	35
2.4.3 Structural (syntactic) ambiguity	36
2.4.4 Semantic ambiguity	36
2.4.5 Pragmatic ambiguity	36
2.5 SsaL linguistic feature implications for WE or meaning representation models	37
2.5.1 Morphological complexity	37
2.5.2 Disjunctive orthography	37
2.5.3 Polysemy and ambiguity	38
2.5.4 Tokenization and vocabulary inflation	38
2.5.5 Implications for meaning representation	38
2.6 Polysemy Representation in Lexical Databases-	38
African WordNet	38

2.7 SsaL as a Low Resource Language and implication for WE Models.....	41
2.7.1 The low-resource challenge and dataset creation	41
2.7.2 Morphological richness and embedding performance	42
2.7.3 Cross-lingual transfer learning and shared embedding spaces	42
2.7.4 Context-aware and multi-prototype embeddings for disambiguation	42
2.8 Chapter Summary	43
CHAPTER 3 : STATE OF THE ART	44
3.1 Introduction.....	44
3.2 Meaning Conflation Deficiency (MCD).....	45
3.3 Word Embeddings (WE).....	49
3.3.1 Word2Vec	54
3.3.2 GloVe.....	55
3.3.3 FastText.....	55
3.4 Approaches addressing Word vector space (WVS) word embedding (WE) Meaning Conflation Deficiency (MCD)	55
3.4.1 Contextual Word Embedding (CWE).....	55
3.4.2 Embedding from Large Language Models (ELMo)	56
3.4.3 Text-to-Text Transfer Transformer (T5)	57
3.4.4 Generative Pre-Training (GPT)	57
3.4.5 ELECTRA.....	57
3.4.6 Bidirectional Encoder Representations for Transformers (BERT).....	57
3.5 Context-Aware Word Embeddings (CAWE)	59
3.5.1 PolyLM	63

3.5.2 XLNet	63
3.5.3 DistilBERT	64
3.5.4 RoBERTa.....	64
3.5.5 GBCN	64
3.5.6 DeBERTa.....	65
3.5.7 ALBERT	65
3.6 Relevant Studies on Sesotho sa Leboa (SsaL) within African NLP	68
3.6.1 Research Background on Morphologically Complex and Low-Resource African Languages	69
3.6.2 Situating Sesotho sa Leboa (SsaL) within African and Bantu Language Studies	69
3.6.3 Unique Morphological Challenges of Sesotho sa Leboa.....	70
3.6.4 Implications for Semantic Modelling and the Present Study.....	71
3.7 Systematic Literature Review (SLR).....	72
3.7.1 Research Strategy.....	72
3.7.2 Inclusion criteria and exclusion criteria.....	74
3.7.3 Data integration and statistical evaluation	76
3.7.4 Publication analysis and results	77
3.8 Meta-Analysis	77
3.8.1 Publication bias and meta-regression.....	82
3.9 Bibliometric Analysis for Word Embeddings.....	86
3.9.1 Bibliometric analysis approach.....	86
3.9.2 Country co-citation analysis	88
3.9.3 Co-authorship and collaborative networks	89

3.9.4 Keyword co-occurrence analysis	91
3.10 Descriptive Statistics of Primary Studies.....	93
3.11 Gaps in the Literature.....	98
3.12 Chapter Summary	99
CHAPTER 4 : THEORETICAL FRAMEWORK.....	101
4.1 Introduction.....	101
4.2 Distributional Semantics	101
4.3 Vector Space Models (VSMs)	105
4.4 Latent Semantic Analysis (LSA)	106
4.5 Neural Network-Based Models.....	106
4.6 Word Embedding Models	107
4.6.1 Word2Vec	107
4.6.2 Global Vectors for Word Representation (GloVe)	109
4.6.3 FastText.....	109
4.7 Context-Dependent Representations.....	109
4.7.1 Embeddings from Language Models (ELMo)	109
4.7.2 Bidirectional Encoder Representations from Transformers (BERT).....	110
4.7.3 Multi-sense embeddings and sense representation	112
4.8 Chapter Summary	113
CHAPTER 5 : METHODOLOGY	114
5.1 Introduction.....	114
5.2 Rationale for Adopting Computational Experimental Design.....	115
5.3 Research Methodological Framework	116

5.3.1 Corpus creation and annotation	117
5.3.2 Model construction and training	118
5.3.3 Evaluation and Validation.....	119
5.4 Algorithm Design.....	120
5.5 Data Collection and Annotation.....	122
5.6 Selection of Models for Experimentation	124
5.7 Performance Tuning of Algorithms	128
5.7.1 Optimized BART-based.....	128
5.7.2 Optimized ELMo	131
5.7.3 Optimized XLNet.....	133
5.7.4 Optimized CNN Using Dropout-Enhanced Variants.....	136
5.7.5 Bidirectional Long Short-Term Memory (Bi-LSTM)	138
5.7.6 Attention-based BiLSTM (A-BiLSTM)	140
5.7.7 Neural Network Language Model (NNLM)	144
5.7.8 BERT	145
5.7.9 Robustly Optimized BERT Approach (RoBERTa).....	146
5.7.10 Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA).....	148
5.7.11 Hierarchical graph.....	150
5.7.12 Domain-specific graph.....	150
5.7.13 Integrated graph	150
5.8 Chapter Summary	152
CHAPTER 6 : EXPERIMENTATION, RESULTS, AND DISCUSSIONS.....	154

6.1 Introduction.....	154
6.2 Experimentation Platform and Environment	155
6.3 Dataset Description.....	155
6.4 Exploratory Data Analysis (EDA).....	157
6.4.1 Word Frequency Distribution	158
6.4.2 Sense diversity	159
6.4.3 Class imbalance	159
6.4.4 Data Overview Summary Table.....	159
6.4.5 Data preparation and cleaning	166
6.5 Optimization and Fine-tuning.....	167
6.6 Experimental Results and Discussion.....	172
6.6.1 Experimental Results for Optimized BART, ELMo, XLNET and CNN	172
6.6.2 Experimental Results for BiLSTM, Optimized BiLSTM, NNLM, and Optimized NNLM	195
6.6.3 Experimental Results with BERT	211
6.6.4 Experimental results with Multilingual Bert.....	214
6.6.5 Experimental results with ELECTRA	216
ELECTRA.....	217
6.6.6 Experimental results with Multilingual ELECTRA	219
Multilingual ELECTRA.....	220
6.6.7 Experimental Results with RoBERTa	222
RoBERTa.....	223
6.6.8 Experimental results with Multilingual RoBERTa.....	225

Multilingual RoBERTa.....	225
ELECTRA.....	228
Multilingual ELECTRA.....	228
RoBERTa.....	228
Multilingual RoBERTa.....	228
6.6.9 Experimental results of graph-based models	231
6.7 Results Discussion	244
6.8 Chapter Summary	246
CHAPTER 7 : SUMMARY, CONCLUSIONS AND FUTURE WORKS.....	248
7.1 Summary	248
7.2 Conclusion	258
7.3 Future Work.....	260
REFERENCES	263

LIST OF FIGURES

Figure 2.1: The geographical distribution of SsaL	18
Figure 2.2:SsaL dialects.....	19
Figure 2.3: Types of Ambiguities in SsaL	35
Figure 3.1: Embeddings taxonomy	50
Figure 3.2: PRISMA flow diagram for the CAWE research	74
Figure 3.3: Forest plot for distribution of effect size of META summary of CAWE in Table 3-4.....	79
Figure 3.4: Forest plot with subgroups for distribution of effect size of META.....	81
Figure 3.5: Galbraith plot.....	82
Figure 3.6: Funnel plot.....	83
Figure 3.7: Bubble Plot per Publication year.....	84
Figure 3.8: Bubble plot for dataset	85
Figure 4.1: Distributional semantics and hypothesis	105
Figure 5.1: Research methodology framework.....	116
Figure 6.1: Frequent senses in Sense-Annotated SsaL Corpus.....	162
Figure 6.2:Distribution of Word Instances Counts	164
Figure 6.3:Distribution of unique senses per word.....	164
Figure 6.4: Box Plot comparison of word instances and unique senses	165
Figure 6.5:Word Cloud of frequent words on “Sense-Annotated SsaL Corpus”	166
Figure 6.6: Comparison of accuracy: Baseline versus Optimised	174
Figure 6.7: Comparison of F1-Scores: Baseline and Optimized Models	175
Figure 6.8: Precision Comparison Baseline vs. Optimised Model	176
Figure 6.9: Comparison of Recall: Baseline Model versus Optimised Model	177
Figure 6.10: Confusion matrix—ELMO.....	178
Figure 6.11: Confusion matrix—Optimized ELMO.....	179

Figure 6.12: Confusion matrix—BART	179
Figure 6.13: Confusion Matrix—Optimized BART.....	180
Figure 6.14:Confusion Matrix—XLNet	181
Figure 6.15: Confusion Matrix—Optimized XLNet.	181
Figure 6.16: Confusion matrix—CNN.	182
Figure 6.17: Confusion Matrix—Optimized CNN with dropout variants.....	182
Figure 6.18: ROC Curve—Optimized BART.	184
Figure 6.19: PCA of BART predictions	187
Figure 6.20:Confusion matrix BART with ablation	188
Figure 6.21: PCA encoder representation of BART	190
Figure 6.22: Multiclass ROC curve for BART ablation studies	191
Figure 6.23: Training vs. validation accuracy and loss for BART with ablation	192
Figure 6.24: Tokenized sentence length distribution.....	193
Figure 6.25: Comparison Baseline vs. Optimised Model.....	197
Figure 6.26: Confusion matrix BiLSTM	198
Figure 6.27: ROC curve BiLSTM	200
Figure 6.28: Optimized BiLSTM Confusion matrix.....	201
Figure 6.29: Optimized BiLSTM ROC curve.....	203
Figure 6.30: NNLM confusion matrix.....	205
Figure 6.31: NNLM ROC curve	206
Figure 6.32: Optimized NNLM Confusion Matrix.....	208
Figure 6.33: Optimized NNLM ROC Curve	210
Figure 6.34:Model Training Loss and Accuracy	212
Figure 6.35:BERT Confusion matrix.....	213
Figure 6.36: Multilingual BERT loss over epochs and accuracy over epochs	215
Figure 6.37: Multilingual BERT confusion matrix.....	216
Figure 6.38: ELECTRA loss over epochs and accuracy over epochs	218

Figure 6.39: ELECTRA Confusion matrix	219
Figure 6.40: Multilingual ELECTRA loss over epochs and accuracy over epochs	221
Figure 6.41: Multilingual ELECTRA confusion matrix.....	222
Figure 6.42: RoBERTa Loss over epochs and accuracy over epochs	223
Figure 6.43: RoBERTa confusion matrix	224
Figure 6.44: Multilingual RoBERTa loss over epochs and accuracy over epochs.....	226
Figure 6.45: Multilingual RoBERTa confusion matrix	227
Figure 6.46: Grouped metrics for transformer variance	229
Figure 6.47: The scatter plot	230
Figure 6.48: The F1-score ranking models	231
Figure 6.49: Performance Summary of Classification Report.....	233
Figure 6.50: Hierarchical graph confusion matrix	234
Figure 6.51: Hierarchical graph ROC curve	235
Figure 6.52: Domain-specific graph confusion matrix	236
Figure 6.53: Domain-specific graph ROC curve	237
Figure 6.54: Domain- specific graph	238
Figure 6.55: Domain-specific graph	240
Figure 6.56: Integrated graph confusion matrix	241
Figure 6.57: Integrated graph confusion matrix	242
Figure 6.58: Integrated graph cluster	244

LIST OF TABLES

Table 1.1: Alignment of Research Questions, Objectives, Methods and Deliverables	10
Table 2.1: Morphological Processes in SsaL	25
Table 2.2: Morphological Category for SsaL	28
Table 2.3: Sentence Structure in SsaL	33
Table 3.1: Summary of Performance Benchmarks of Embeddings.....	66
Table 3.2: WE Approaches Key Features and Methods	67
Table 3.3: Overview of the results from the database search for WE Models	73
Table 3.4: Inclusion and Exclusion Criteria	76
Table 3.6: Meta-Regression per Publication Year.....	84
Table 3.7: Meta-Regression per Dataset.....	85
Table 3.8: Trim and Fill Meta-Analysis Results.....	86
Table 4.1: Key Concepts and Semantic Foundations	112
Table 5.1: Deployed Algorithms.....	121
Table 5.2: Selection Criteria	125
Table 5.3: Model Selection Rationale.....	126
Table 6.1: Dataset Description Summary	158
Table 6.2: Distribution of Ambiguous Lemmas, Frequencies, and Associated Sense Counts in the Sense-Annotated SsaL Corpus.....	160
Table 6.4: Summary of Results Baseline vs. Optimised.....	172
Table 6.5: Statistical Analysis Summary (McNemar’s results).....	183
Table 6.6: BART Ablation Studies	187
Table 6.7: Evaluation Metrics for BiLSTM, Optimized BiLSTM, NNLM, and Optimized NNLM	196
Table 6.8: Bert Performance Metrics.....	211
Table 6.9: Multilingual BERT Performance Metrics	214
Table 6.10: ELECTRA Performance Metrics.....	217

Table 6.11: Multilingual ELECTRA for MCD Performance Metrics	220
Table 6.12: RoBERTa Performance Metrics	223
Table 6.13: Multilingual RoBERTa Performance Metrics	225
Table 6.14: Performance Summary of Transformer Models	228
Table 6.15: Performance Summary of Classification Report	232

LIST OF ABBREVIATIONS AND ACRONYMS

ABSA	Aspect Based Sentiment Analysis
AI	Artificial Intelligence
ATE	Aspect Term Extraction
AWN	African Wordnet
BART	Bidirectional and Auto-Regressive Transformers
BERT	Bi-Directional Encoder Representation
Bi-LSTM	A Bidirectional Long Short-Term Memory
BN	Bayesian Network
BPE	Byte-Pair Encoding
CAWE	Context-Aware Word Embeddings
CBOW	Continuous Bag of Words
CNN	Convolutional Neural Network
CPT	Conditional Probability Tables
CRFs	Conditional Random Fields
CED	Computational Experimental Design
DS	Distributional Semantics
DT	Decision Trees
ELMo	Embedding from Large Language Models
ELECTRA	Efficiently Learning an Encoder that Classifies Token Replacements Accurately
FT	FastText
GloVe	Global Vectors for Word Representation
GPT	Generative pre-trained transformer
GPT	Generative Pre-Training
KBWE	Knowledge-Based Word Embedding

KNN	K-Nearest Neighbors
LLM	Large Language Model
MCD	Meaning Conflation Deficiency
MR	Meaning Representation
MRLs	Morphologically Rich Languages
NLMs	Neural Language Models
NLP	Natural Language Processing
NNLM	Neural Network Language Model
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
RoBERTa	Robustly Optimised BERT Pretraining Approach
SsaL	Sesotho sa Leboa
SMOTE	Synthetic Minority Over-sampling Technique
T5	Text-to-Text Transfer Transformer
WE	Word Embeddings
WEMs	Word Embedding Models
WSD	Word Sense Disambiguation
WVS	Word Vector Space
QA	Question and Answering

CHAPTER 1 : INTRODUCTION

1.1 Background and Motivation

Distributional semantics (DS) is a usage-based model of meanings, based on the assumption that the statistical distribution of linguistic items in context plays a key role in characterizing their semantic behaviour. DS is primarily an empirical method for the analysis of lexical meaning (Mitchell & Lapata, 2010; Lenci, 2018). This model is also known as vector space semantics, due to the use of vector space in modelling the lexical meaning. Vector-based models of word meaning have become increasingly popular in computational linguistics and natural language process (NLP). The appeal of these models lies in their ability to represent meaning simply by using distributional information under the assumption that words occurring within similar contexts are semantically similar (Mitchell & Lapata, 2010; Kartsaklis, 2014). DS offers both a vector-based model to represent meaning and computational methods to learn such representations from language data. DS vector space models are built from text corpora as samples of language usage, offering new ways to investigate the interplay between meaning and contexts (Mitchell & Lapata, 2010).

In NLP, conventional DS based word vector space (WVS) models of meaning representations, are generally called word embeddings (WE). WE are real-valued vector representation of words by embedding both semantic and syntactic meanings obtained from unlabelled corpora (Yaghoobzadeh & Schutze, 2016). Due to their strong generalisation capability, word embedding (WE) language modelling approaches have demonstrated effectiveness and flexibility in capturing and storing valuable syntactic and semantic information that can be transferred and integrated into downstream natural language processing tasks.

This success has been demonstrated in a wide range of NLP applications including machine translation, syntactic parsing, text classification, question answering, and retrieval systems. However, this is most visible in the NLP context of analytic or isolating resource rich and less morphologically rich languages (MRLs), such as English (Takala, 2016; Vasic & Brajkovic, 2018). The same cannot

be said of agglutinating African Bantu languages, such as South African SsaL, which are morphologically richer, and so, posing additional computational complexity challenge in word meaning representation for NLP. This could be a factor in the paucity of WE model benchmarks for these languages.

Word vector space (WVS) word embedding (WE) models are typically directed at representing words in isolation. However, evidence in sentential priming suggests that semantic similarity is more complex than simply a relation between isolated words (Mitchell & Lapata, 2010; Kartsaklis, 2014). Word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) is a semantic shortcoming identified in static vector-based word embeddings, in which a single vector encodes multiple meanings of a polysemous word without considering the context (Camacho-Collados & Pilehvar, 2018). WVS WE MCD arises from the complex interplay between morphology and meaning, particularly in morphologically rich languages. This limitation hinders the model's ability to correctly capture and convey the intended meaning of a word, often resulting in semantic misinterpretation. In morphologically rich languages, a single base form can yield numerous meanings depending on applied affixes and grammatical structure, which further amplifies ambiguity (Camacho-Collados & Pilehvar, 2018).

In linguistics, a word is said to be ambiguous when it has two or more meanings; otherwise, it is called monosemous. An ambiguous word can be polysemous or homonymous. Polysemous words have multiple related meanings, while homonymous words have meanings that are completely unrelated. According to the Cambridge Dictionary, homonymy refers to “a word that sounds the same (homophone) or is spelled the same (homograph) as another word but has a different meaning.” NLP focuses on homographs with different meanings. Each individual meaning of an ambiguous word is called a sense. A lexical resource that lists different senses of words is commonly referred to as a sense inventory. Examples include the Princeton WordNet (Miller, 1995), and its derivatives such as African WordNet (Griesel et al., 2020; Bosch & Griesel, 2017).

Based on the principle of economical versatility of words (Zipf, 1958), frequent words tend to have more senses, but most words in sense inventories are monosemous. For example, 83% of the 15k words in WordNet 3.0 are listed as monosemous (Camacho-Collados & Pilehvar, 2018). This is a constraining factor in using the WordNet and its African WordNet variant solely as a sense inventory knowledge base for NLP tasks and applications. The MCD problem also manifests in WordNet in its polysemous words' representation, with implications for the WVS WE model. WordNet represents the polysemous words by capturing their different meanings at lexical level, but without giving consideration regarding the polysemy classes to which they belong (Lin et al., 2002) This polysemous words representation in WordNet affects its usability as a WE model resource for NLP tasks and applications (Freihat, 2014).

A knowledge-based word embedding (KBWE) (Devlin, Chang, et al., 2019) model is a method for semantic representation that integrates lexical knowledge from structured semantic resources such as WordNet, BabelNet, or FrameNet into the word embedding space. This approach seeks to improve word vectors through better disambiguation, relational precision, and contextual understanding (Raganato et al., 2017). Unlike embeddings derived solely from distributional statistics that depend on word co-occurrence in large datasets, KBWE models blend traditional data-driven techniques with human-curated semantic information. This process explicitly embeds relationships such as synonymy, hypernymy, and hyponymy into word vectors, thereby boosting the interpretability and semantic accuracy of embeddings. Such an approach proves particularly beneficial for tasks such as WSD, assessing semantic similarity, and answering questions. This study intends to address this gap in the context of use of African WordNet as a sense inventory resource for the KBWE model (Devlin, Chang, et al., 2019).

Due to the inherent ambiguity present in all human languages, WVS WE models burdened by MCD lack the necessary semantic robustness to adequately represent word meanings in agglutinative Bantu African languages, in which ambiguities are fundamental to their morphological richness. This might

partially explain the findings of a recent audit on human language technology focused on South African languages, which indicates an absence of WSD research on these languages (Moors et al., 2018). To fill this gap, it is expedient that WVS WE models for this language family, adequately accommodate word ambiguities.

The context in which a word is applied must be regarded as pivotal to comprehending its meaning and ought to be integrated into WVS WE models to represent meaning accurately. This understanding underlies the development of contextual word embedding (CWE) models (Pasini et al., 2021; Scarlini et al., 2020; Smith, 2019) Such models deal with CWE, which have been employed effectively across several tasks in NLP, having proved to carry useful semantic information (Scarlini et al., 2020). Thus, they can contribute to enhancing performance and reducing complexity of NLP tasks such as WSD and application systems such as machine translation (Pasini et al., 2021).

WSD, the process of automatically identifying the meaning of a polysemous word in a sentence, is a fundamental task in NLP (Nasiruddin, 2013). Hence, enhancing its performance and reducing complexity, would have a spin-off effect on the downstream NLP tasks and applications. In dealing with the MCD problem in WVS WE models, a few approaches have been posited for modelling individual word senses. The main distinctive feature of these models is in their meaning modelling scheme and sources of the senses (Camacho-Collados & Pilehvar, 2018).

It is on this basis that the paradigm of sense distinctions is defined, in categorizing WE into unsupervised and knowledge-based models. Unsupervised WE models directly learn word senses from text corpora, while knowledge-based models exploit the sense inventory of lexical resources, as their main source for meaning representations. Unsupervised WE models are much related to word sense induction (WSI) or discrimination, sense distinctions being induced by analysing text corpora. WSI is the automatic identification of word senses by analysing the contexts in which they appear (Nasiruddin, 2013). In KBWE models, word senses are represented as defined in an external sense

inventory, such as the African WordNet, for a number of South African Bantu languages (Griesel et al., 2020), including SsaL.

KBWE assumes existence for some predefined word senses in a sense inventory, with the challenge that a corpus from which all possible word senses can be predefined, could be difficult to obtain. This result is the possibility of out-of-vocabulary problems in NLP tasks. The unsupervised WE models rely on progressive induction of senses from text corpora which may be non-exhaustive. This provides a rationale for a hybrid model that exploits the complementarities of the two WE model approaches, by combining the two. With such a hybrid model, one can start with an initial SI which can be updated as new word senses are discovered from text corpora through the unsupervised technique.

The hybrid model requires a novel integration of the two WE model approaches. Such a hybrid approach boils down to a semi-supervised approach, which could be more appropriate for under-resource Bantu African languages, such as SsaL. In this study the problem of WVS WE MCD was addressed using unsupervised and knowledge-based techniques of sense distinctions for the WE model. This is a shift from word level to sense level WE, to accommodate word sense representation.

The WE model was called context-aware word embedding (CAWE), because the context (or sense) of usage of a word was made integral to the word meaning representation. That is, the CAWE model will represent individual meanings of words –word senses – as independent representations. To address the complexities of morphologically rich languages (MRLs), this study considered the applicability of word embedding (WE) models for MRL settings by drawing on approaches previously proposed in non-African language contexts (Vasić & Brajković, 2018; Chaudhary et al., 2018), which explored how WE methods can be adapted to better represent morphologically rich languages. The finer-grained meaning representation that provides for representation at the character, morpheme, word, and sense (or lexeme) levels, in a semantic network of WE models was proposed. The purpose of the study is to develop a CAWE model that effectively addresses the MCD problem in WVS WE models.

1.2 Problem Statement

Word embedding models are integral components in NLP systems, heavily relied upon by semantic similarity and machine translation tasks to represent words within a continuous vector space. Nevertheless, traditional embeddings such as Word2Vec, GloVe, and FastText have a significant limitation termed word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD), in which the multiple meanings of a polysemous word are merged into a single representation. These models map discrete lexical items, $w \in V$ into, to a continuous vector space, $\mathbb{R}^d$, facilitating tasks such as similarity measurement, classification, and translation. Conventional embedding techniques (e.g., Word2Vec, GloVe) assign one vector per word, regardless of its contextual meaning, causing semantic ambiguity and diminished performance in downstream tasks, especially in morphologically rich languages (MRLs) like SsaL. Unlike analytic languages such as English, MRLs display substantial inflection, derivation, and agglutination, encapsulating complex grammatical, syntactic, and semantic information in single tokens (e.g., *ke a mo rata* — “I love him/her,” where the prefix *ke* indicates subject agreement, *a* signifies tense/aspect, and *mo-* denotes object concord) (Camacho-Collados & Pilehvar, 2018).

The significant morphological productivity of SsaL poses various computational challenges. These include constraints in corpus annotation; available SsaL corpora are typically limited, lack balance across different domains, and are deficient in detailed sense annotation. Despite advances in word embedding research, SsaL has no benchmark embedding model nor a standardized evaluation dataset. This absence of baselines prevents systematic experimentation and reproducibility, making it challenging to evaluate or compare embedding approaches for this morphologically rich language. Additionally, there is a deficiency in contextual modelling, in which current embeddings trained on these corpora fail to adequately capture morphosyntactic and semantic contexts. This is particularly problematic when shifts in meaning are induced by affixes, reduplication, or derivational morphology (Ansell et al., 2021).

Formally, given a vocabulary $V = \{w_1, w_2, \dots, w_n\}$, a corpus $C = \{S_1, S_2, \dots, S_m\}$, and a context window $\Omega(w_i)$ around word w_i , conventional embedding models develop a function for mapping.:

$$f: V \rightarrow \mathbb{R}^d$$

so that words with similar meanings possess comparable vector representations derived from co-occurrence statistics. However, this mapping presupposes a single vector for each word, overlooking that each word, w_i , can be associated with multiple senses $S(w_i) = \{s_1, s_2, \dots, s_k\}$, where $k \geq 2$. The meaning conflation emerges due to the objective function, which usually involves maximizing the log-likelihood:

$$\max_{\theta} \sum_{(w,c) \in D} \log P(c | w; \theta)$$

in which all contexts c for w are pooled together, thereby ignoring sense differences, where each symbol means as given below:

- θ represent parameters of the model
- D represents the annotated sense corpus
- w represent the target word
- c represent the context word

In SsaL, this shortcoming becomes apparent when contextually different meanings (for instance, *go ja* meaning “to eat,” compared with *ja* in *moja* denoting “eater”) are shared into a single compact vector, neglecting the morphological indicators provided by prefixes and suffixes. As a result, this amalgamation diminishes semantic accuracy, causing suboptimal performance in context-sensitive applications such as machine translation, question answering, and sense disambiguation (Camacho-Collados & Pilehvar, 2018).

This study proposes a CAWE Model for MRLs, utilizing SsaL as a case study, which addresses WVS WE MCD by incorporating morphological decomposition, contextual embeddings, and sense clustering techniques. The model introduced enhances the traditional embedding function to:

$$f': (w_i, \Omega(w_i), M(w_i)) \rightarrow \mathbb{R}^d$$

where $M(w_i)$ encodes morphological attributes such as the root, affixes, and derivational category, and $\Omega(w_i)$ depicts the syntactic and semantic framework. This hierarchical integration facilitates the projection of semantically unique meanings into distinct subspaces within the embedding manifold, enhancing both sense-specific representation and interpretability.

The primary objective of this research is to develop a sense-annotated and morphologically enriched corpus for SsaL, alongside creating a CAWE model that utilizes subword morphology, contextual dependencies, and sense-level guidance. The study aims to empirically demonstrate, through metrics such as accuracy, precision, recall, and F1-score, that the integration of linguistic structures and contextual learning effectively mitigates MCD in WVS WE models.

1.3 Thesis Statement

The Meaning Conflation Deficiency (MCD) problem in word vector space (WVS) word embedding (WE) models, in the context of a low resource morphologically rich language, such as SsaL, can be addressed through a context-aware word embedding (CAWE) model that is based on a combination of knowledge-based and unsupervised WE methods.

1.4 Research Questions

The main research question that was addressed in this research study is as follows:

“How can the MCD problem in WVS WE models be effectively addressed, in the context of a low resource morphologically rich language, such as SsaL, using a context-aware WE model that is based on a combination of knowledge-based and unsupervised WE methods?”

To answer this main research question the following research sub-questions are answered:

RQ1: What salient features of a morphologically rich low resource language, such as SsaL, should be considered in developing a CAWE model that effectively addresses the MCD problem in WVS WE models?

RQ2: What are the MCD limitations of existing WVS WE models in representing these features?

RQ3: How can a context-aware WE model that addresses these limitations be developed, using a combination of knowledge-based and unsupervised WE methods?

RQ4: How can the CAWE model effectively be evaluated for SsaL?

1.5 The Study Aim and Objectives:

Main Aim: The main aim of this research study was to develop and evaluate a WE model that enhances performance and reduces complexity of WSD application tasks, in the context of a low resource morphologically rich language, using a combination of knowledge-based and unsupervised WE methods.

The following study achieved this by implementing the following objectives:

RO1: Systematically determine salient features for low-resources morphologically rich language in developing a CAWE model that mitigates MCD problems in WVS WE models.

RO2: Investigate MCD limitations of existing WVS WE models in representing the determined features.

RO3: Develop a CAWE model that addresses these limitations using a combination of knowledge-based and unsupervised WE methods.

RO4: Experimentally evaluate the effectiveness of the CAWE model in the context of SsaL.

1.6 Alignment of Research Questions, Objectives, Methods and Deliverables

Table 1.1: Alignment of Research Questions, Objectives, Methods and Deliverables

Research Questions	Research Objectives	Research Methods	Research Deliverables
What salient features of a morphologically rich low resource language, such as SsaL, should be considered in developing a CAWE model that effectively addresses the MCD problem in WVS WE models?	Systematically determine salient features for low-resources morphologically rich language in developing a CAWE model that mitigates MCD problems in WVS WE models.	Systematic Literature review and analysis.	Salient morphology-aware representations.
What are the MCD limitations of existing WVS WE models in representing these features?	Investigate MCD limitations of existing WVS WE models in representing the determined features.	Systematic literature review and experimentation	MCD limitations in existing WVS WE models.
How can a context-aware WE model that addresses these limitations be developed, using a combination of knowledge-based and unsupervised WE methods?	Develop a context-aware WE model that addresses these limitations using a combination of knowledge-based and unsupervised WE methods.	Model development by experimentation.	An adapted corpus-based combination of knowledge-based and unsupervised model.
How can the CAWE model effectively be evaluated for SsaL?	Experimentally evaluate the effectiveness of the CAWE model in the context of SsaL.	Experimental evaluation using accuracy, precision, recall, and F1-score.	Experimental evaluation of the CAWE model.

1.7 Overview of Research Methodology

The study uses a hybrid approach combining knowledge-based and unsupervised techniques to address the meaning conflation deficiency (MCD) problem in word vector space (WVS) word embedding (WE) models. The knowledge-based component employs explicit linguistic information—such as domain-specific vocabularies and semantic links—to provide disambiguation clues and to mitigate the scarcity of training data. Structured knowledge is graphically represented, with nodes as sentences or words, and edges as semantic connections. Domain hubs group vocabulary into meaning-specific clusters, while pattern rules signal specific meanings. This tackles the key issue for statistical methods in low-resource settings: insufficient contextual diversity for learning distinct senses.

The unsupervised element leverages contextual embedding models such as BERT and BiLSTM to capture semantic patterns not covered by existing knowledge, offering flexibility beyond manual rules. These models are optimised with morphological awareness: subword tokenization respects boundaries, attention highlights key morphemes, dropout prevents overfitting, and class-balancing addresses sense frequency imbalances.

Integration occurs on various levels which are graph features, and neural embeddings form hybrid representations, while integrated graphs combine sentence and domain links for enhanced disambiguation. During optimization, knowledge-based cues guide neural training. This hybrid method addresses MCD in WVS WE models by creating sense-specific representations through explicit knowledge and derived patterns, particularly aiding infrequent senses. The technique is thoroughly assessed for its disambiguation accuracy and compared with other methods. Its effectiveness in word sense disambiguation (WSD); and advantages in machine translation are evaluated, showing benefits for SsaL NLP applications.

1.8 Thesis Contribution

This thesis presents seven interconnected contributions to the body of knowledge in natural language processing (NLP):

- The development of a context-aware word embedding model for addressing meaning conflation deficiency in word vector space embedding models.
- The development of a structured taxonomy of word embeddings for analysing and positioning static, contextual, and context-aware embedding paradigms.
- The development of a structured multi-phase research methodology framework for investigating and resolving meaning conflation deficiency in low-resourced morphologically rich languages.
- The development of a structured multi-phase research methodology framework for investigating and resolving meaning conflation deficiency in low-resourced morphologically rich languages.
- The development of a novel hybrid graph-based modelling approach for improving semantic representation and disambiguation in morphologically rich languages.
- An optimised suite of machine learning and deep learning algorithms for enhanced semantic disambiguation and contextual representation.
- Conduct of a meta-analysis and bibliometric mapping of word embedding models for understanding research trends, gaps, and methodological directions in meaning conflation deficiency studies.

1.9 Thesis Structure

The thesis is structured into seven chapters as follows:

Chapter 1 (Introduction) introduces the study by outlining the broader problem of MCD in WVS WE models, particularly its implications for low-resourced, morphologically rich languages such as SsaL. This chapter presents the research questions, objectives, significance, and the scope of the study within the context of computational linguistics.

Chapter 2 (The Sesotho sa Leboa Linguistic Context) This chapter studies SsaL, a South African language in the Sotho-Tswana Bantu subgroup, highlighting computational approaches and word

embeddings. With dialectal diversity and limited digital presence, this is vital for low-resource NLP. Its complex morphology, including agglutination and noun classes, complicates tasks such as lemmatization, requiring subword models. The chapter addresses morphemes and lexical categories, orthographic challenges, and syntactic dependencies for embedding models. Chapter 2 explores semantic ambiguity and the need for context-based word embeddings. Lexical databases such as African WordNet try to capture polysemous meanings but fall short. These challenges emphasize the need for morphologically informed, semantically sensitive models for African languages.

Chapter 3 (State of the Art) examined word embeddings and their role in addressing the MCD problem in WVS WE models. The chapter started with a theoretical overview of semantic representation models from frequency-based to neural embeddings. Traditional embeddings, for instance, Word2Vec and GloVe were shown to conflate word meanings, leading to WVS WE MCD. The chapter analysed challenges faced by SsaL, in which morphological complexity and polysemy exacerbate WVS WE MCD problem. CAWE models, offering dynamic representations, addressing MCD problem in WVS WE models by adjusting to context, thus enhancing performance and simplifying WSD tasks. The chapter integrated empirical findings through an SLR, following PRISMA standards, highlighting recent innovations in embeddings for low-resource languages. The review found research gaps, especially in African languages. A meta-analysis and bibliometric analysis with VOS viewer showed performance trends, and research patterns; and descriptive statistics illustrated study characteristics. This chapter links theory and empirical insight, providing a foundation for creating models sensitive to low-resource language subtleties.

Chapter 4 (Theoretical Framework) The chapter outlines the evolution from basic distributional methods to advanced neural and contextual models, focusing on balancing static and contextual embeddings and addressing polysemy. This chapter establishes a foundation for robust WSD systems for complex languages such as SsaL. By integrating corpus-based, subword-level, and CAWE models for Sesotho sa Leboa, the study merges distributional semantics and polysemy theory with evaluations of various embedding methods. This research introduces new datasets and embedding strategies for

SsaL, reinforcing NLP approaches that preserve semantic granularity, thus addressing a gap in computational linguistics for underrepresented African languages.

Chapter 5 (Methodology) This chapter details the methodology used to explore MCD in WVS WE models, a language with rich morphology and limited resources, using computational experiment design. Chapter 5 connects linguistic theory with practical models; and critiques word embeddings to suggest context-sensitive models for polysemous words. The research aims to reduce semantic ambiguity in low-resource settings, focusing on SsaL and creating an annotated corpus. The chapter describes data analysis, model architecture, training, embedding methods, and various models such as BART, RoBERTa, and BERT, along with evaluation metrics such as precision, recall, and F1-score. Statistical tests confirm improvements. Chapter 5 ends with graph-based algorithms, BERT variant comparisons, and ensures transparency, reproducibility, and ethical standards.

Chapter 6 (Experimentation, Results and Discussions) This chapter describes the experimental process to mitigate Word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD), ensuring reproducibility and comparison. Using Google Colab with a machine-learning stack, the study achieved transparent implementation and repeatability. A *sense-annotated SsaL corpus* was introduced, relevant for morphologically rich, low-resource contexts, considering vocabulary scale and sentence length. Exploratory data analysis informed model design, revealing a long-tailed distribution for ambiguous lemmas, significant polysemy, and skewed label prevalence. A summarized table highlighted ambiguous tokens, instance counts, and senses. The data pipeline improved training balance while retaining linguistic nuance. Experimental results compared baseline and optimised architectures, using diagnostic tools such as confusion matrices and ROC curves to assess both overall and class-specific performance. Optimization enhanced precision, recall, and F1-score, despite morphological complexity and sparse class distributions, with error analysis showing overlapping senses. Graph-based methods improved sense separability via relational structures, complementing sequence-based models. Findings highlight the (i) the need for preprocessing and addressing imbalance, yet their inadequacy on their own, (ii) effectiveness of architecture and graph enhancements

in reducing MCD in WVS WE models, and (iii) importance of pairing global metrics with class diagnostics to identify semantic confusions. These insights inform thesis conclusions and recommendations for Word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) advancement in SsaL and similar languages.

Chapter 7 (Summary, Conclusion, and Future Works) This chapter summarizes the objectives that have been accomplished, wraps up the thesis, and proposes potential paths for future research in related areas.

CHAPTER 2 : THE SESOTHO sa LEBOA LINGUISTIC

CONTEXT

2.1 Introduction

This chapter outlines the linguistic foundations of SsaL, to provide essential context for computational experiments in later chapters. As a morphologically rich bantu language, it poses challenges and opportunities for Word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) resolution. Understanding these complexities is vital for developing effective meaning representation models, sense embeddings, and disambiguation methods for low-resource languages. The chapter examines unique linguistic features of SsaL and their impact on NLP models. Section 2.2 introduces the language origins, classification, and sociolinguistic status, providing context for its use in research. Section 2.3 explains SsaL’s morphological richness affecting lexical representation. This section includes: 2.3.1 on morphological typology, and 2.3.2 on word structure. Section 2.3.3 covers morphological processes such as inflection, derivation, and compounding, while 2.3.4 differentiates between free and bound morphemes. Additionally, Section 2.3.5 reviews the morphology of lexical categories such as nouns, pronouns, verbs, adjectives, adverbs, and ideophones. Section 2.3.6 discusses significant morphological phenomena such as reduplication, derivation, and infixation for meaning representation. Section 2.3.7 covers orthographic standards, and Section 2.3.8 outlines sentence structures, including basic, complicated, and compound phrases. Section 2.4 focuses on ambiguities in SsaL, highlighting polysemy, synonymy, and homonymy in semantic ambiguity and meaning conflation in computer models. Section 2.5 addresses how these linguistic features affect word embedding (WE) models, particularly their impact on learning, contextualization, and sense differentiation. Section 2.6 reviews polysemous terms in digital lexicon resources such as African WordNet, pointing out the challenges in representing varied meanings for marginalized African languages. Section 2.7 discusses SsaL as a low-resource language, tackling constraints from limited annotated corpora, tools, and resources in NLP system development. These are relevant to

computational methods, for instance, transfer learning, multilingual modelling, and corpus-based sense annotation, later discussed. Section 2.8 summarizes the chapter's findings and leads into the methodology chapter, in which computational models are evaluated by the linguistic principles detailed. This chapter bridges language theory and computational application, ensuring the creation of linguistically aware, culturally sensitive, and technically sound NLP models for low resourced African languages.

2.2 The SsaL Language

Sesotho sa Leboa (SsaL), also known as Sepedi or Northern Sotho, is a Bantu language native to South Africa. It is among the nation's 11 officially recognized languages. Figure 2.1 shows evolution of SsaL, which is influenced by a complex interplay of factors, such as cross-cultural exchanges, population movements, and the dynamic sociocultural environment of the region. Further research is needed to fully understand the various influences that have shaped this language over time; and to uncover the nuances of its linguistic development (Mabela & Ditsele, 2024). In South Africa, SsaL is designated an official language, which enhances its utilization in government, education, and media. This decision functions as a vital indicator of identification for the SsaL-speaking populations, linking speakers to their cultural history and customs (Joffe, 2016). SsaL serves as the medium of instruction in early education, particularly in areas where it is predominantly spoken, prior to the shift to English in higher grades (Zerbian, 2007; Rakgogo, 2024).

Figure 2.1 further displays the linguistic basis of SsaL, facilitating an understanding of its usage, status, and societal significance in the regions where it is spoken. Speakers of SsaL engage in their cultural heritage and traditions through this language, which serves as a significant marker of identity for the communities. Moreover, researchers assert that the boundaries of the SsaL region are defined by an imaginary line that passes through Springs, Middelburg, Groblersdal, Lydenberg, and Sabie (Joffe, 2016). The line then continues north of the Sabie River to Bushbuckridge, extending to Messina, then to the border with Bechuanaland, and finally southward through the Potgietersrus district (which is now known as Mokopane), Modimolle, and concludes in Tshwane (Joffe, 2016).

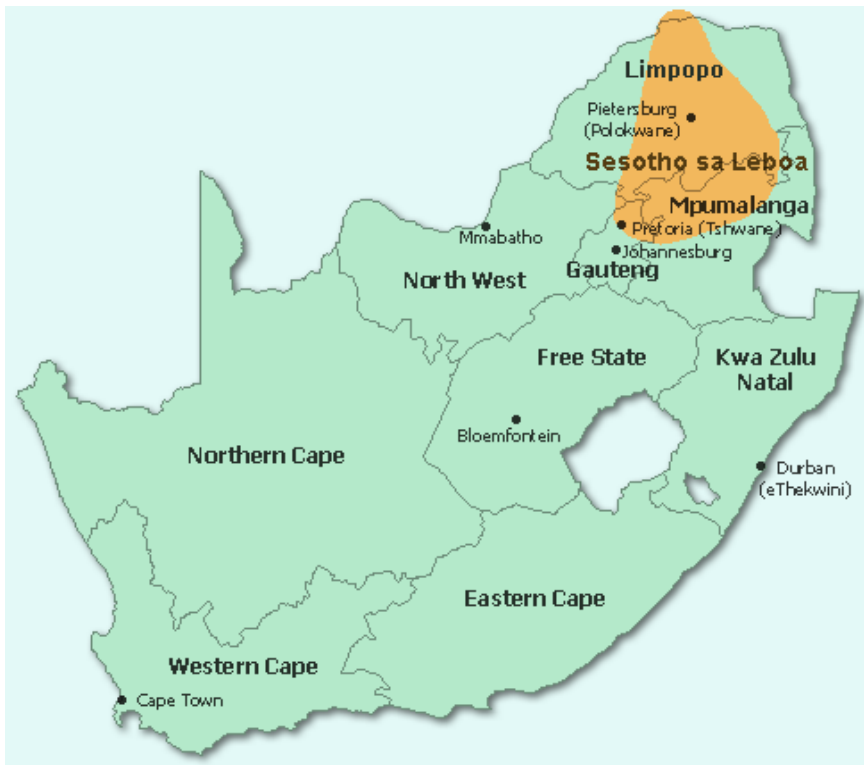


Figure 2.1: The geographical distribution of SsaL (Joffe, 2016)

The origin of the acronym SsaL, according to members of the academic community, shown in Figure 2.2, demonstrates that it is a member of the Sotho subgroup found within the north-eastern area of Bantu languages, which are spoken by millions of people (Rakgogo & Zungu, 2022). It is further stated that it is of the utmost importance to identify the sociolinguistic roots of the labels of SsaL, in addition to the classification of its dialectal characteristics by other academics (Rakgogo & Zungu, 2022).

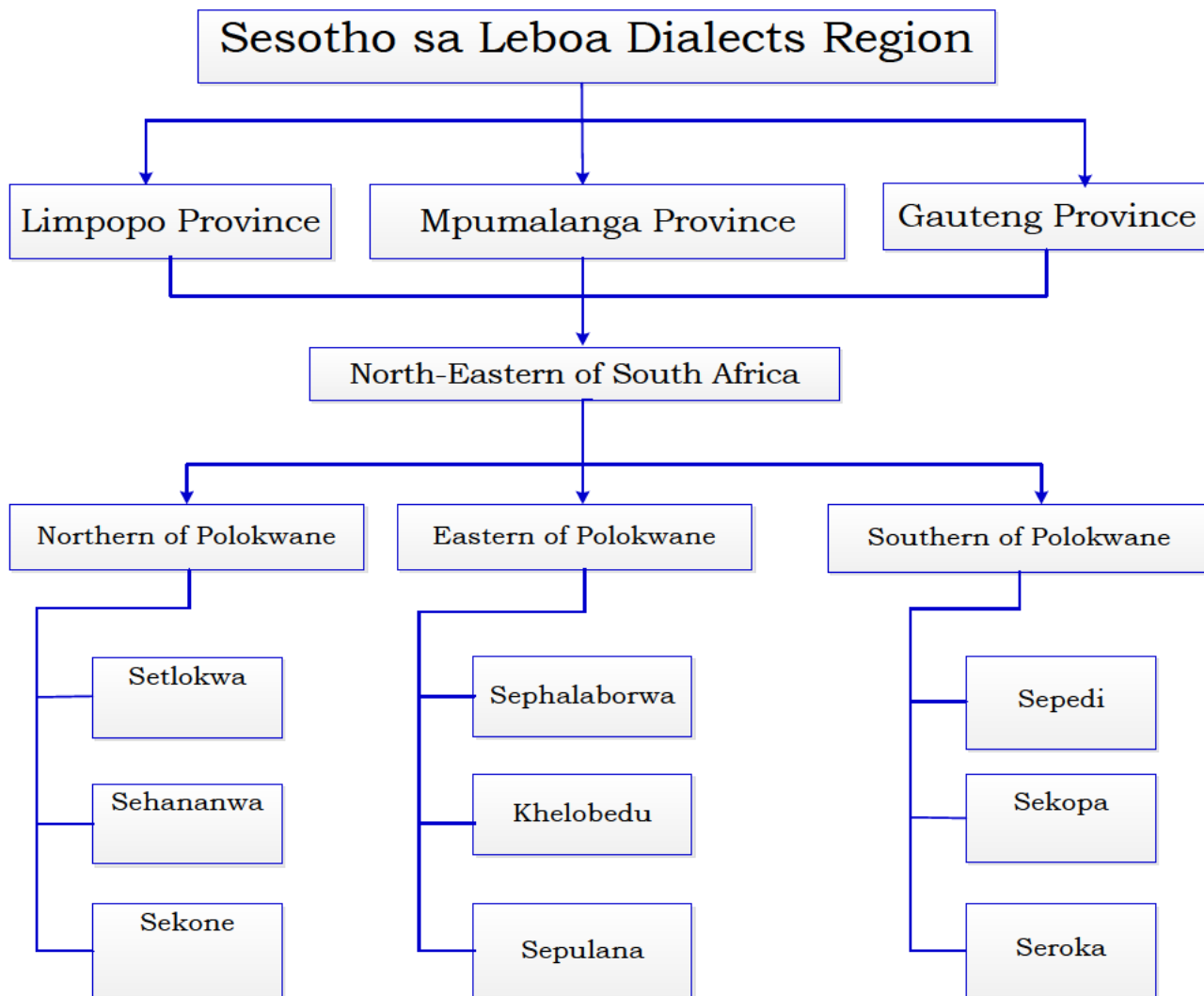


Figure 2.2:SsaL dialects (Collins et al., 2021)

Sotho is the collective name for Sesotho sa Leboa (Northern Sotho), Sesotho (Southern Sotho), and Setswana (Tswana): Sotho is the word that encompasses all three regions (Collins et al., 2021) (Rakgogo, 2016). The Sotho language group forms the second-largest current Bantu language group, after the Nguni language group (Collins et al., 2021). Census results from 2012 indicate that the Sotho language is spoken by 24.7% of South Africa's total population (Statistics South Africa, 2012). The research work by Rakgogo & Zungu (2021) argues that, from a conventional point of view, these three languages are grouped together on the basis of similar linguistic traits. However, they are also regarded as separate and independent languages, because they possess distinctive qualities that significantly divide them from one another.

While SsaL encompasses a range of dialects, researcher (Mojela, 2016) argues that efforts to document and analyse these variations may not be the most effective approach. Standardizing the language, rather than preserving dialectal differences, could lead to greater language unity and could therefore facilitate the development of comprehensive language resources (Mojela, 2016; Rakgogo, 2024) . Proponents of this view believe that mapping the geographical distribution of dialects and investigating sociolinguistic factors may not be a priority; the focus should be on establishing a single, widely accepted standard form of the language (Mabela & Ditsele, 2024).

The language SsaL belongs to the Sotho-Tswana branch of the Niger-Congo family of languages (Greenberg, 1949; Bendor-Samuel, 1989; Good, 2017; Joffe, 2016), which is described in Figure 2.2 as an umbrella word that incorporates several different dialects. These dialects include Pedi, Lobedu, Tlokwa, Kone, Pulana, Hananwa, Kopa, Phalaborwa, and Roka. There are approximately 6.2 million people who speak SsaL, according to the population census that was conducted in 2022 (Statistics South Africa, 2022).

2.3 SsaL as a Morphologically Rich Language

The language SsaL is well-regarded for its morphological richness, owing to its intricate and productive morphological system, which is a characteristic trait of many Bantu languages (Nurse & Gérard, 2003; Poulos & Louwrens, 1994). In this language, meaning is frequently embedded within the structure of words by concatenating various morphemes, leading to long and information-rich words, especially in the context of verbs. This agglutinative morphology facilitates the formation of words by combining morphemes sequentially, each serving a specific grammatical or semantic role. For example, the verb *ke a go rata* (“I love you”) is broken down into *ke* (1st person singular subject), *a* (present tense marker), *go* (2nd person object), and *rata* (verb root “to love”). This morphological framework encapsulates both syntactic and semantic relations that would typically necessitate full phrases or clauses in more analytic languages such as English.

A significant element adding to this morphological intricacy is the noun class system, comprising over 18 grammatical noun categories that govern morphology agreement throughout the sentence structure.

Every noun is part of a specific class, typically signalled by a prefix, which affects the form of adjectives, pronouns, verbs, and quantifiers associated with it (Poulos & Louwrens, 1994; Greenberg, 1949). As an illustration, the noun *mosadi* (“woman”) falls into Noun Class 1, with its adjective agreement forming *mosadi o mobotse* (“the beautiful woman”), where *o mobotse* aligns in class and number. This thorough system of concordial agreement is syntactically necessary and intricately woven into the morphosyntactic structure of the language.

Moreover, SsaL employs extensive verbal morphology via both inflectional and derivational means. Inflectional markers convey tense, aspect, mood, negation, and polarity, such as in *ke tla mo bona* (“I will see him”), in which *tla* marks the future tense, *mo* serves as the object concord, and *bona* represents the verb root (“see”). In contrast, derivational processes facilitate the semantic broadening of base verbs through causative, applicative, reciprocal, and passive affixes (Louwrens, 1991). For example, the verb *ruta* (“teach”) can be transformed into *rutana* (“teach each other”) with the reciprocal suffix *-ana*, or into *rutiša* (“cause to teach”) using the causative suffix *-iša*. Beyond verbal intricacies, nominal morphology involves affixation to express pluralization, possessive structures, and diminutives. The noun *ngwana* (“child”) becomes *bana* (“children”) in the plural, signifying a shift in noun class from singular to plural, along with a class change in concord from Class 1 to Class 2.

The morphological characteristics, especially the extensive use of affixation, agreement in concord, and morphophonology that is non-concatenative, create notable difficulties for computational processing tasks such as word segmentation, POS tagging, and Word Sense Disambiguation (WSD) (Pretorius et al., 2005). In languages with limited resources such as SsaL, the lack of extensive annotated datasets and the unavailability of morphological analysers make these challenges more pronounced, necessitating the development of NLP models informed by linguistic principles to effectively generalize across different morphological forms.

Therefore, a comprehensive grasp of the morphological structure of SsaL is crucial, not just for linguistic analysis, but also for the advancement of sophisticated language technologies. This includes the creation of tools such as morphological segmenters, machine translation systems, and word embedding models, which are essential in alleviating challenges related to semantic ambiguity and

polysemy, particularly concerning issues related to WVS WE MCD (Ansell, 2020; Camacho-Collados & Pilehvar, 2018).

2.3.1 SsaL morphological typology

SsaL serves as a typical example of the agglutinative morphological typology commonly found in numerous Bantu languages, in which sequential morpheme affixation conveys grammatical meaning (Nurse & Gérard, 2003). These morphological structures, characterized by concatenative morphology, strong class agreement, and layered derivation, pose challenges for NLP tasks. Tools need to handle a wide range of inflectional variations, substantial morpheme productivity, and structures driven by agreement (Mathe & Eiselen, 2021).

Agglutinative language is characterized by the linear arrangement of morphemes, each expressing a distinct grammatical function (Nhlapo, 2015). Its words often consist of numerous bound morphemes, resulting in a high morpheme-to-word ratio (Mojapelo, 2009). This feature signifies that meanings are frequently constructed by appending multiple prefixes, infixes, and suffixes to a root, each adding a unique grammatical or semantic detail (Thaba et al., 2020). Unlike fusional languages, which use affixes to communicate several grammatical categories at once, or isolating languages, which rely mainly on free morphemes, understanding the individual morphemes and their combination rules in an agglutinative language is vital for grasping a word's entire meaning (Febrian et al., 2022).

This characteristic presents both challenges and opportunities for natural language processing, as it necessitates robust morphological analysis tools capable of segmenting words into their constituent morphemes to facilitate accurate parsing and semantic interpretation (Dlamini, Jembere, et al., 2020). This is particularly relevant for tasks such as machine translation, in which word-for-word translation can often fail due to the compact information density within single agglutinated forms (Dlamini, Jembere, et al., 2020). This extensive morphological complexity means that word embedding models developed for less inflected languages are often inadequate for SsaL, necessitating specialized approaches that leverage sub-word units or morpheme-level representations to capture the full semantic and syntactic nuances (Dlamini et al., 2021). This is further exacerbated by many words particularly in morphologically rich languages, do not appear frequently enough in training corpora to allowing for

adequate generalization of word-based models (Dlamini et al., 2021). Therefore, explicitly modelling morphemes can significantly reduce the out-of-vocabulary rate and enhance the performance of natural language processing tasks in SsaL (Dlamini, Jembere, et al., 2020).

2.3.2 SsaL words morphology

Lexical structures are fundamentally organized around stems, with prefixes and suffixes serving to denote grammatical categories (Nurse & Gérard, 2003). Verb morphology exhibits considerable complexity, encapsulating subject agreement, object agreement, tense, aspect, and mood within a singular verb form (Ebbers, 2009; Pretorius et al., 2005). Similarly, nominal morphology is characterized by an intricate system of noun classes, each linked with distinct concordial agreements that permeate the sentence structure (Thaba et al., 2020). This complex system of morphological agreement requires advanced computational methodologies for precise parsing and generation, any deviations possibly resulting in substantial semantic and syntactic inaccuracies (Dlamini, Jembere, et al., 2020).

The intricacy of SsaL's morphology highlights the essential role of sub-word analysis, particularly in tasks such as generating word embeddings. Traditional models focusing solely on whole words frequently struggle to grasp the subtle meanings embedded within individual words (Dlamini et al., 2021). Consequently, there is a pressing need for specialized morphological parsers capable of breaking down complex words into their basic morphemes, thereby enabling a finer detail in linguistic representation (Kelly et al., 2014). The dynamic nature of these morphological processes leads to an expansive vocabulary, challenging natural language processing models that depend on fixed word lists (Dlamini, Jembere, et al., 2020; Kristó, 2006). This active word creation significantly complicates the development of lexical resources, because it often results in data sparsity due to the infrequent appearance of many complex forms in text datasets (Fesseha et al., 2021). As a result, computational approaches for SsaL should transcend superficial word forms to include an in-depth comprehension of the morphological structures at play, facilitating improved generalization, and decreasing data sparsity in language models (Mathe & Eiselen, 2021). This focus on morphology is especially crucial for

devising accurate pronunciation models, morphologically rich similar words typically following consistent pronunciation patterns (Ebbers, 2009).

The morphology of SsaL words is influenced by a strong interaction among roots, affixes, and noun class systems that organize word formation across various grammatical categories. Every word is built upon a root, which serves as a fundamental semantic element, and is enhanced by both derivational and inflectional morphemes (Ebbers, 2009). For instance, starting from the root *-rut-*, a variety of words can be formed: *o ruta* (“to teach”), *morutegi* (“teacher”), *thuto* (“education”), *moithuti* (“learner”), showcasing recursive derivations achieved using affixation. In verb derivation, several suffixes, or “extensions”, alter both valency and meaning. Examples include the passive “-*wa*”, applied “-*ela*”, causative “-*iša*”, and reciprocal “-*ana*”, among others. In SsaL, meaningful morphological structures are present in proper names as well; a study of personal and place names showed the relationship between morphological forms and their semantic meanings (Mojapelo, 2009). Furthermore, observational syntax indicates that derivational and inflectional morphemes usually adhere to a predictable order—lexical root → derivation → inflection—which supports Greenberg’s typological hierarchy (Nhlapo, 2015). Morphological complexity highlights the language ability for construction; and poses specific challenges for computational analysis, in which accurately identifying morpheme boundaries and differences is essential for precise morphological parsing and for following NLP tasks (Du Toit & Puttkammer, 2021; Mathe & Eiselen, 2021).

2.3.3 Morphological processes in SsaL

SsaL showcases a diverse set of morphological processes, including inflection, derivation, and compounding, which together enhance the expressiveness and grammatical intricacy of the language (Kristó, 2006). Inflection involves altering a word’s form to indicate grammatical distinctions such as tense, person, number, or negation, without changing the word’s essential meaning. In SsaL, for instance, the verb root *-rata* (“love”) is transformed into *ke a go rata* (“I love you”) by inflecting with stacked prefixes: *ke-* (1st person), *a-* (present tense), and *go-* (2nd person object). This exemplifies a

typical agglutinative pattern in which individual morphemes are arranged in fixed positions within the verb structure. Derivational morphology expands or shifts the meaning of a base word by affixation or morphological modification. A documented study examines the sequence of morphotactic ordering in derivation and inflection. This study employs Greenberg’s universal principle which states that derivational morphemes come before inflectional morphemes (Nhlapo, 2015). For example, from the root -rut- (“teach”), words such as *morutegi* (“teacher”), *thuto* (“education”), and **moithuti** (“learner”) can be derived, illustrating a process of derivational layering that changes both lexical category and meaning. Compounding merges two or more words into a unified semantic entity, typically changing the meaning while preserving grammatical properties. Studies on SsaL compounds reveal that they operate morphosyntactically as single nouns, retaining agreement patterns, headedness, and number concord (Thaba et al., 2020). For instance, *molao* (“law”) + *motheo* (“foundation”) combines to create *molaotheo* (“constitution”)—a lexical item acting as a noun with compound semantics. The morphological processes, which include inflectional variation, derivational creativity, and compound formation, are essential to the grammar of SsaL. These processes hold considerable importance for computational models, particularly for tasks such as morphological analysis, word segmentation, and word sense disambiguation.

Table 2.1: Morphological Processes in SsaL

Process	Function	Example
Inflection	Adjust grammatical features	<i>ke a go rata</i> (“I love you”)
Derivation	Creates new lexemes via affixation	<i>morutegi, thuto, moithuti</i>
Compounding	Merges words into a single semantic unit	<i>molaotheo</i> (‘constitution’)

2.3.4 SsaL Morphemes

SsaL morphology is defined by the interplay between free and bound morphemes, highlighting its agglutinative characteristics and the structural tendencies of Bantu languages. Free morphemes are standalone units with meaning that do not require attachment to other morphemes. For instance, in

seeta (“shoe”), *eta* (“lead”) is a free morpheme, maintaining its meaning independently; whereas *se-* acts as a bound prefix. Similarly, the root */rek-/* (“buy/sell”) appears as a free morpheme in terms such as *rek-a* (“buy”). Free morphemes also include particles, for instance, *ke* (agentive, “by”) in the construction *longwa ke mpša* (“bitten by a dog”), performing independent syntactic functions. On the other hand, bound morphemes cannot stand alone and must attach to a root. These consist of prefixes (e.g., *mo-* in *mo-rek-i* “buyer” and *se-* in *seeta*), suffixes (e.g., *-gadi* for feminine, *-ana/-nyana* for diminutive, *-eng* for locative), and bound roots such as *-tho* in *motho* (“person”).

In SsaL, verbal morphology exemplifies the extensive use of bound morphemes through both derivational and inflectional processes. For example, derivational suffixes such as *-is-* (causative) in *rek-is-a* (“to cause to sell”) and *-el-* (applicative) in *rek-el-a* (“to sell for”) modify the meaning or argumentative structure of verbs. Inflectional morphemes such as the perfective *-ile* in *batsheg-ile* (“they have laughed”) indicate grammatical distinctions without changing the word’s class. These morphological processes give rise to polymorphemic word forms, in which various bound morphemes merge with a free morpheme to convey complex meanings. The frequent use of bound morphemes, including noun class prefixes and verbal suffixes, demonstrates that SsaL’s morphological system is systematic, productive; and aligns with the broader typological characteristics of the Bantu language family.

2.3.5 Lexical categories morphology

2.3.5.1 Nouns (*Maina*)

The SsaL language categorizes nouns into a system of noun classes, indicated by prefixal morphemes, which influence the agreement with verbs, adjectives, pronouns, and other modifiers (Nhlapo, 2015). The format follows: prefix + stem (+ suffix). Examples include simple nouns such as *motho* “person,” as well as derived nouns such as *mohumagadi* “queen,” stemming from the root *huma* with the augmentative suffix *-gadi*. Additionally, deverbatives are formed from verbs, for instance, *modiredi* or

mošomedi translating to “worker” from the verbs *direla* or *šomela* “work for”, Locative nouns belonging to classes 16 to 18 are used as adverbs, such as *motseng*, which means “in the village.”

2.3.5.2 Pronouns (*Mabitšo a bontšhi*)

Pronouns encompass various morphological categories that align with noun class: emphatic pronouns highlight the referent as independent forms (e.g., *yena* “he/she”). Possessive pronouns express possession (*gago* “yours”, *ya bona* “theirs”). Quantitative pronouns denote amount (*bohle* “all”). Personal pronouns denote person, number, and occasionally class (*ke* “I”, *re* “we”).

2.3.5.3 Verbs (*Mediri*)

Verbs constitute the central element of predicate structures and are morphologically rich, frequently encoding tense, aspect, mood, and voice. The fundamental structure is composed of: subject concord + tense/aspect morpheme + verb root + (derivational suffixes) + final vowel. Derivational suffixes include: causative (*-is-*), applicative (*-el-*), passive (*-w-/-iw-*), and reciprocal (*-an-*). Inflectional morphemes indicate tense (*-ile* perfect), aspect (progressive *-a*), and mood (subjunctive, imperative).

2.3.5.4 Adjectives (*Melaodi*)

Adjectives match the head noun in noun class through an adjective concord. Structure: concord + adjective stem (e.g., *mo-golo* “big” for Class 1, *me-golo* for Class 4). Sources: Certain adjectives originate from class 14 abstract nouns (e.g., *bohlale* “cleverness” functioning as an adjective).

2.3.5.5 Adverbs (*Mantšu a tlaleletšo*)

Adverbs indicate manner, time, place, or intensity. Common examples include: *bjale* for “now” and *maabane* for “yesterday”. Adverbs are also frequently formed through locative nouns such as *motseng*, meaning “in the village”, by reduplicating words such as in *bjalebja* for “quickly”, or by utilizing locative particles such as *ga-*, as seen in *GaSekhukhune*.

2.3.5.6 Prepositions (*Go swana le dipartikele tša boemedi*)

Although SsaL lacks prepositions such as those found in English, the language employs specific particles which serve analogous purposes: *ka* (instrumental “with”), *ke* (indicating agency as “by”), *mo* (denoting location as “in”), and *go* (used for both infinitive and locative purposes as “to/at”).

2.3.5.7 Conjunctions (*Dikgokaganyi*)

Conjunctions serve to connect clauses, phrases, or words. Coordinating conjunctions include “*le*” (“and”) and “*goba*” (“or”). Subordinating conjunctions are exemplified by “*gore*” (“that/so that”) and “*ge*” (“when/if”).

2.3.5.8 Interjections (*Mantšu a ikemetšego*)

Independent expressions that convey emotions, directives, or reactions include: *Aowa!* (indicating “No!”), *Eeh!* (signifying “Yes!”), and *Ao!* (expressing surprise).

2.3.5.9 Particles (*Dipartikele*)

Particles perform grammatical roles without changing form: Agentive: *ke* “by”. Connective: *le* “and/with”. Copulative: *ke* “is/are”. Hortative: *a* “let” (as in *A re yeng* “Let’s go”). Question: *na* (used for yes/no inquiries).

2.3.5.10 Ideophones (*Mantšu a boikutlo*)

Expressive words that evoke vivid sensory experiences, often mimic sounds: *phaa!* (swift motion), *thuu!* (strong impact), *gwii!* (heavy fall). These can stand alone or be woven into sentences to enhance their dramatic impact.

Table 2.2: Morphological Category for SsaL

Category	Morphological Features	Example(s)	Notes
Nouns	Prefix + stem + (suffix); noun class concord	<i>motho</i> “person”, <i>motseng</i> “in the village” Sentence: <i>Motho o dula motseng.</i> (The person lives in the village.)	Locatives (Classes 16–18) often function as adverbs
Pronouns	Class-agreement forms; emphatic, possessive, quantitative, personal	<i>yena</i> “he/she”, <i>gago</i> “yours”, <i>bohle</i> “all” Sentence: <i>Yena o tla tla ka moso.</i> (He/She will come tomorrow.)	Agreement with noun class is obligatory

Verbs	Subject concord + TAM + root + derivational suffix(es) + FV	<i>ba-rek-iš-a</i> “they cause to sell” Sentence: <i>Ba-rek-i š-a diaparo mmarakeng.</i> (They are selling clothes at the market.)	Causative, applicative, passive, reciprocal
Adjectives	Adjective concord + stem	<i>mo-golo</i> “big” (Cl.1), <i>me-golo</i> (cl.4) Sentence: <i>Ngwana o mo-golo o ya sekolong.</i> (The big child is going to school.)	May derive from Class 14 nouns
Adverbs	Basic or derived; locative noun forms, reduplication, particles	<i>bjale</i> “now”, <i>bjalebjale</i> “quickly” <i>Re kopane gona bjalebjale.</i> (We met just now/quickly.)	Locative <i>ga-</i> often used
Prepositions/Particles	Invariable particles	<i>ka</i> “with”, <i>ke</i> “by”, <i>mo</i> “in” Sentence: <i>O ngwala ka pene.</i> (He/She writes with a pen.)	Often multifunctional
Conjunctions	Coordinating/subordinating	<i>le</i> “and”, <i>goba</i> “or”, <i>gore</i> “that” Sentence: <i>Ke tla reka dijo le meetse.</i> (I will buy food and water.)	No agreement required
Interjections	Standalone forms	<i>Aowa!</i> “No!”, <i>Eeh!</i> “Yes!” Sentence: <i>Aowa! Ga go bjalo.</i> (No! It is not like that.)	Express emotion or reaction
Particles	Grammatical markers	<i>ke</i> (copulative), <i>na</i> (question) Sentence: <i>Na o tla ya mmošong?</i> (Will you go to the meeting?)	Often overlap with prepositions
Ideophones	Vivid, sensory, onomatopoeic	<i>phaa!</i> (sudden), <i>thuu!</i> (impact) Sentence: <i>Koloi e ile ya ema phaa!</i> (The car stopped suddenly!)	Adds expressive force
Nouns	Prefix + stem + (suffix); noun class concord	<i>motho</i> “person”, <i>motseng</i> “in the village” Sentence: <i>Motho o dula motseng.</i> (The person lives in the village.)	Locatives (class 16–18) often function as adverbs

2.3.6 Derivational, reduplication, and infixes morphology

2.3.6.1 Inflectional morphology

In SsaL, inflectional morphology is crucial to the language’s grammar, especially for indicating tense, aspect, mood, number, and agreement of noun classes. Unlike derivational processes that generate new words or change their lexical categories, inflection maintains the grammatical role while preserving the fundamental meaning of the root word (Kelly et al., 2014; Taljard et al., 2008). For instance, verbs in SsaL are inflected by adding prefixes and suffixes for tense, aspect, and mood. The verb “-rata” (to

love) becomes “*ke a go rata*” (I love you) by using the prefix “ke” to denote the first-person singular subject and “a” to indicate the present continuous aspect. Nouns are similarly inflected to reflect singular and plural forms, such as “*ngwana*” (child) versus “*bana*” (children), keeping class agreement through concord prefixes in noun-adjective and noun-verb structures.

2.3.6.2 Derivational Morphology

This derivational morphology transformation shifts the word from a verb to a noun and reflects social roles in its structure. Derivational suffixes such as “-i”, “-o”, or “-go” are commonly used to create abstract nouns or agentive forms. These morphological processes enhance the language’s lexicon by enabling the creation of new words via consistent morphotactic patterns, supporting its agglutinative nature and potential for vocabulary growth (Idris et al., 2022; Shafi et al., 2023).

2.3.6.3 Compounding morphology

In SsaL, compounding involves merging two or more lexical bases to create a novel lexical entity, typically with a unified meaning. This process can be categorized as either endocentric or exocentric, contingent upon the syntactic head and the semantic clarity of the compound (Kristó, 2006; Mojapelo, 2009). For instance, the compound “*setšhaba sa molao*” (nation of law or legal community) arises when the noun “*setšhaba*” (*nation*) and the noun phrase “*sa molao*” (*of law*) are combined to describe a nation governed by legal principles. Although its productivity is lower than affixation, compounding illustrates the language’s conceptual evolution and achieves socio-cultural naming roles, especially in identity expressions, professional jargon, and geographic names (Masethe et al., 2024a; Kristó, 2006).

2.3.6.4 Reduplication morphology

In SsaL, reduplication is a morphological technique employed for a variety of functions, such as denoting plural forms, adding emphasis, indicating repetitive actions, and conveying intensification (Shafi et al., 2023; Mojapelo, 2009). For instance, the verb “-bina” (to sing) can be duplicated as “-bina-bina” to signify singing that is ongoing or repeated. Similarly, adjectives and adverbs can be

reduplicated to stress emphasis: “*monyenyane*” (small) transforms into “*monyenyane-monyenyane*” (very small). This process can involve full reduplication, in which the entire stem is repeated, or partial reduplication, when only a portion of the stem is duplicated, based on phonological and morphological rules. Reduplication is a vibrant feature in both everyday and poetic language; and contributes to rhythm and prosody in oral traditions (Shafi et al., 2023; Mojapelo, 2009).

2.3.6.5 Infixation morphology

Infixation in SsaL is not commonly observed; however, it does appear in certain morphological settings, particularly in expressive or ideophonic forms. Instead of attaching a morpheme to either the start or conclusion of a word (as seen in affixation), infixation entails the placement of a morphological element inside the root (Botha, 2014). For instance, some ideophones might include phonological changes within the root to convey expressive meanings or sensory experiences. Nevertheless, infixation is not a primary morphological strategy; and is typically regarded as part of phonological changes or irregular morphophonemic patterns rather than a fully developed grammatical process (Botha, 2014).

2.3.7 SsaL orthography

SsaL utilizes a standardized orthographic framework that is predominantly phonemic, indicating a strong link between its spoken and written forms. This orthographic system is built upon the Latin script and incorporates specific conventions to capture the language’s phonological characteristics, including implosives, nasals, and tonal variations, although tone is not typically indicated in standard writing (Pretorius & Pretorius, 2009). As per the Pan South African Language Board (PanSALB) and specifications in Government Gazette 22223 from 21 April 2001, SsaL orthography excludes diacritics and special symbols to maintain simplicity and accessibility for literacy initiatives and formal education.

The word structure employs a disjunctive system, which separates prefixes (like subject concords) and roots into independent orthographic units (Modiba, 2024). For instance, the verbal phrase “*ke a go*

rata” (“I love you”) is displayed as three distinct words, highlighting the grammatical elements rather than a single inflected entity (Modiba, 2024). Although this disjunctive orthography enhances syntactic clarity, it poses difficulties for computational tasks such as part-of-speech tagging and word sense disambiguation (Idris et al., 2022). Despite strides towards standardization, variations in regional dialects and code-switching with Afrikaans or English can sometimes affect spelling conventions. The standard orthography is continually evolving due to educational, linguistic, and technological influences, particularly in the creation of digital resources for the low-resource languages of South Africa (Makutoane, 2022; Osborn, 2010; Modiba, 2024; Dlamini, et al., 2020).

The alphabet of SsaL adheres to the standards of the Bantu language family, with a system of consonants, vowels, and tonal distinctions (Dlamini, et al., 2020). Although it adheres to the known standards of the Bantu language family, disputes and issues about the standardization of the writing system persist. The existing orthography fails to sufficiently represent the subtleties of the language’s varied phonetic inventory, especially the tonal system, which is essential for expressing meaning (Yigezu et al., 2021). Diacritical markings, while effectively representing complex phonological structures, can be burdensome, and may impede the widespread use of the writing system, particularly in digital and technical environments (Rahab & Mothapo, 2019). Additionally, the reliance on Latin-based letters, a legacy of colonial influences, has been criticized by some language purists who advocate for the development of a more uniquely SsaL-based orthography that better reflects the language’s cultural and linguistic identity. Its orthography, or writing system, is mostly based on the Latin alphabet (Modiba, 2024).

2.3.8 SsaL Sentence Structure

In SsaL, simple sentences are formed by a single independent clause featuring one subject-predicate combination. These sentences convey a single, complete thought, without the use of subordinate or coordinate elements. For instance, *Ngwana o a sepela* (“The child is walking”) merges the noun subject *Ngwana* with the predicate *o a sepela*; *Ke a ithuta* (“I am studying”) pairs the subject pronoun *Ke* with

a verb phrase; and *Koloi e eme* (“The car stopped”) links the noun *Koloi* with the predicate *e eme*. Each example remains syntactically uncomplicated, articulating just one principal action or state.

Complex sentences consist of one independent clause and at least one dependent clause, in which the subordinate clause is introduced by a subordinating conjunction. For example, in *Ke tla dula fao ge pula e a na* (“I will stay there when it rains”), the primary clause *Ke tla dula fao* is enhanced by *ge pula e a na* (“when it rains”). Similarly, in *Ba ile go reka ka ge dijo di fedile* (“They went to buy because the food was finished”), *ka ge* introduces a causal clause. In *Ke nyaka go bona ge bana ba fihla* (“I want to see when the children arrive”), *ge* introduces a temporal clause. These constructions permit the main concept to be detailed with relationships of time, cause, or condition.

Compound sentences connect two or more independent clauses of equal grammatical importance, typically using coordinating conjunctions. In the sentence “*Ba ya mmarakeng, gomme ba reka dijo*” (“They are going to the market, and they are buying food”), the word “*gomme*” connects two actions. “*Ke tla ithuta, eupša ke tla iketla pele*” (“I will study, but I will rest first”) contrasts two clauses with “*eupša*.” “*Re jele dijo, gape ra nwa le tee*” (“We ate food, and we also drank tea”) uses “*gape*” to add an additional action. In each instance, the clauses could stand independently; however, the conjunction establishes a relationship between them.

Table 2.3: Sentence Structure in SsaL

Sentence Type	SsaL Example	English Translation	Structure / Notes
Simple	<i>Ngwana o a sepela.</i> <i>Ke a ithuta.</i> <i>Koloi e eme.</i>	The child is walking. I am studying. The car stopped.	Single independent clause: subject (<i>Ngwana</i>) + predicate (<i>o a sepela</i>). Subject pronoun (<i>Ke</i>) + predicate (<i>a ithuta</i>). Subject (<i>Koloi</i>) + predicate (<i>e eme</i>).
Complex	<i>Ke tla dula fao ge pula e ena.</i> <i>Ba ile go reka ka ge dijo di fedile.</i>	I will stay there when it rains.	Independent clause (<i>Ke tla dula fao</i>) + dependent clause (<i>ge pula e e na</i>).

	<i>Ke nyaka go bona ge bana ba fihla.</i>	They went to buy because the food had been consumed. I want to see when the children arrive.	Independent clause (<i>Ba ile go reka</i>) + dependent clause (<i>ka ge dijo di fedile</i>). Independent clause (<i>Ke nyaka go bona</i>) + dependent clause (<i>ge bana ba fihla</i>).
Compound	<i>Ba ya mmarakeng, gomme ba reka dijo.</i> <i>Ke tla ithuta, eupša ke tla iketla pele.</i> <i>Re jele dijo, gape re nwile tee.</i>	They are going to the market, and they are buying food. I will study, but I will rest first. We ate food, and we also drank tea.	Two independent clauses joined by coordinating conjunction (<i>gomme</i>). Two independent clauses joined by <i>eupša</i> (“but”). Two independent clauses joined by <i>gape</i> (“and also”).

2.4 Ambiguities in the SsaL Language

In SsaL, the inherent complexities within its morphology, syntax, lexis, and pragmatics lead to ambiguity, as explored in Figure 2.4, which investigates the primary types of ambiguities that occur.

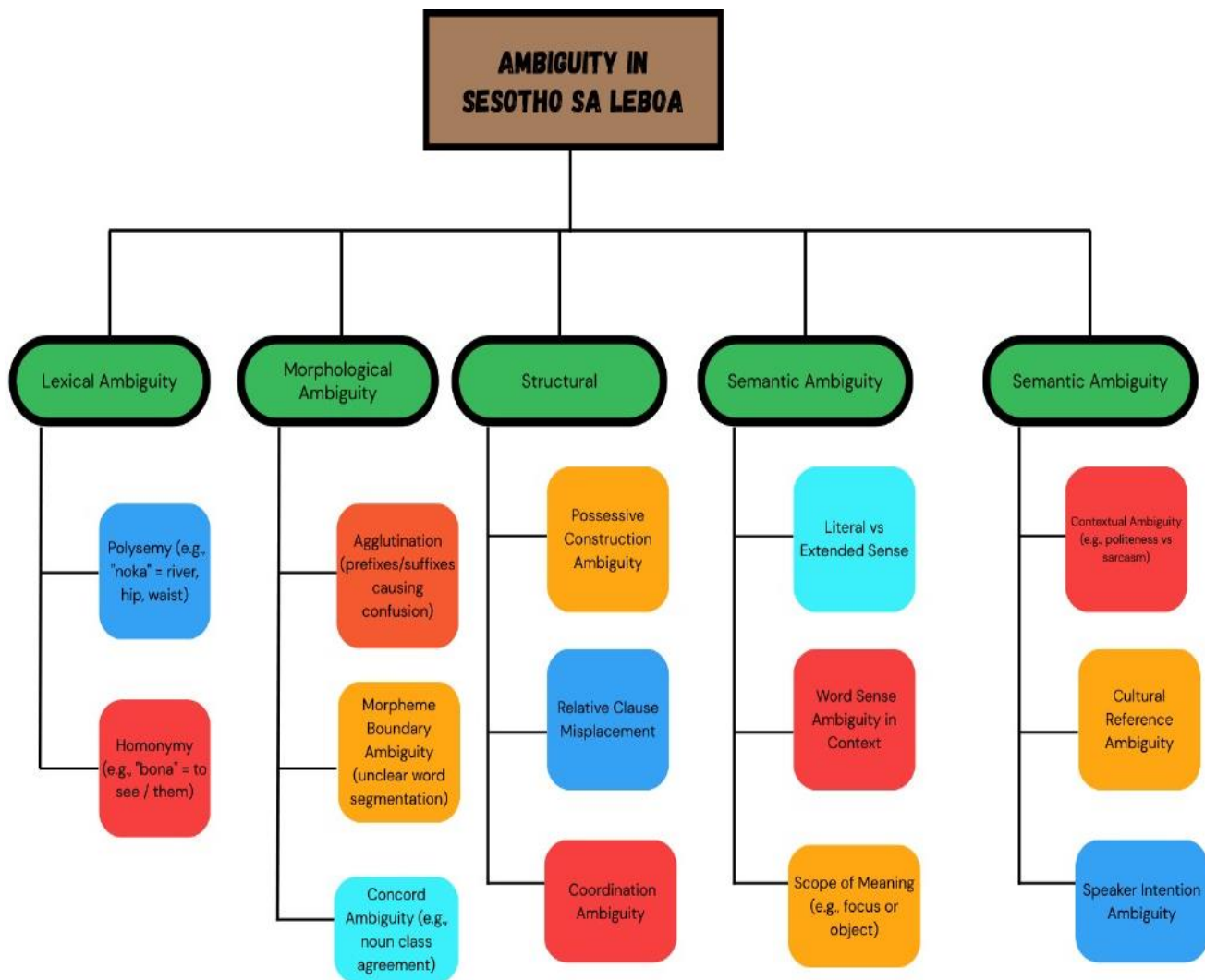


Figure 2.3: Types of Ambiguities in SsaL (Mojapelo, 2007; Mojapelo, 2009; Chokoe, 2000)

2.4.1 Lexical ambiguity

Lexical ambiguity occurs when a word form (lexeme) possesses multiple interpretations. In SsaL, this is frequent, due to both polysemy (multiple related meanings) and homonymy (distinct meanings) (Chokoe, 2000). For instance, “*noka*” can denote river, hip, or waist, contingent on the context. Similarly, “*bona*” can represent either see (verb) or them (pronoun). These differences pose challenges for both human comprehension and computational models, particularly in resolving meaning from context. Chokoe (2000) discusses these ambiguities as cases of “*saying the unmeant*,” emphasizing how such lexemes can hinder effective communication if not clarified via context.

2.4.2 Morphological ambiguity

SsaL is characterized by its rich morphology and agglutinative nature, employing numerous affixes to convey grammatical and semantic nuances. This can result in ambiguity when such affixation yields

multiple possible interpretations (Mojapelo, 2009). For instance, the phrase "*ke a mo rata*" can mean "I love him/her," but when spoken rapidly or without context, it might be misheard or mistaken for other similar morphologically rich derived structures. Verb prefixes, such as *ke*, *o*, and *ba*, along with noun class concords, add to the complexity of overlapping forms. As a result, ambiguity in morpheme boundaries and class concords may lead to misunderstandings among native speakers as well as challenges for NLP systems.

2.4.3 Structural (syntactic) ambiguity

This kind of ambiguity arises due to a sentence structure that can be interpreted in multiple syntactic ways. For example, "*Monna o rata ngwana wa mosadi*" might mean either "The man loves the woman's child" or "The man loves a child (who is a woman)." In this case, the use of possessive constructions and noun class references causes syntactic confusion when not clarified by punctuation or contextual cues (Chokoe, 2000; Modiba, 2024).

2.4.4 Semantic ambiguity

Semantic ambiguity occurs when phrases or clauses can be logically or conceptually interpreted in multiple ways. Such ambiguity intersects with lexical and structural ambiguity but places more emphasis on interpretation than on form (Deerwester et al., 2015; El-alamy et al., 2022). For example, the expression "*Ba tšeere kgwele ya bona*" can be translated to mean "They took their ball" or, when viewed metaphorically, "They took their chance." This phrase, while excluding figurative interpretations, still retains literal semantic ambiguity. This is particularly true when it comes to body parts (e.g. *leoto*, *hlogo*) or objects (e.g. *kgwele*) used in non-traditional contexts.

2.4.5 Pragmatic ambiguity

Pragmatic ambiguity arises from contextual factors, in which the interpretation relies on situational, social, or cultural signals. For instance, the phrase "*O seke wa bolela bjalo!*" may be interpreted as "Don't say that!" indicating a rebuke, or "Don't speak like that!" suggesting a correction, depending on the tone, context, or the relationship between the speaker and listener. The difficulty lies in that, without prosodic or cultural indicators, pragmatic nuances might be missed, notably in written text or

computerized language processing. In SsaL, ambiguity spans several linguistic aspects—from morphology and syntax to pragmatics and semantics. These ambiguous elements present challenges in fields such as education, lexicography, and computational linguistics, particularly for systems designed for word sense disambiguation (WSD) or machine translation in lesser-studied African languages (Bakx, 2006; Chokoe, 2000).

2.5 SsaL linguistic feature implications for WE or meaning representation models

Grasping the linguistic intricacies of SsaL is crucial for constructing word embedding models (WEMs) and for enhancing meaning representation (MR) in natural language processing systems. This morphologically rich and low resourced African language poses some specific challenges to both traditional and current contextual embedding methods.

2.5.1 Morphological complexity

SsaL features agglutinative morphology, characterized by the attachment of various morphemes—both inflectional and derivational—to a root word to express grammatical roles, including tense, aspect, mood, negation, and noun class alignment. For example, verbs such as *ke a go rata* (“I love you”) merge markers for subject, mood, and object into one word. This results in a proliferation of vocabulary and challenges with data sparsity in conventional embeddings such as Word2Vec, which depend on frequency-based co-occurrence patterns.

2.5.2 Disjunctive orthography

The language employs a disjunctive writing system, meaning that a phrase that might constitute a single grammatical word in other languages is represented as multiple orthographic words. For instance, *o tla mmona* (“you will see him”) is divided into auxiliary, subject, object, and verb components. This impacts tokenization and context window modelling in WEMs, leading models that do not account for this disjunctiveness to mistakenly process parts of a unified concept as separate tokens.

2.5.3 Polysemy and ambiguity

The SsaL language demonstrates significant polysemy and lexical ambiguity. For illustration, the word “*noka*” may refer to a “river,” “hip,” or “seasoning,” necessitating word sense disambiguation (WSD) to determine the appropriate meaning within a given context. Static embeddings fall short because they allocate only one vector per word. Meanwhile, contextual models such as BERT offer potential solutions, yet they demand extensive annotated corpora, which are unfortunately scarce for SsaL.

2.5.4 Tokenization and vocabulary inflation

Due to the morphological complexity and disjunctive orthography inherent in SsaL, tokenizers designed for English or Indo-European languages inadequately represent this language. Byte-Pair Encoding (BPE) and Workpieces tokenizers generate fragmented morphemes, which result in expanded vocabularies and suboptimal subword modelling. This adversely affects performance in downstream tasks and compromises the integrity of semantic representations.

2.5.5 Implications for meaning representation

Instability in embeddings: Pronounced inflection results in uncommon word forms that have unreliable vectors. Contextual inconsistencies: Disjunctive tokens hinder the model’s capacity to learn semantically coherent context windows. Loss in morpho-syntax: WEMs that disregard grammatical structure struggle to generalize across semantically similar tokens that differ in affixes.

2.6 Polysemy Representation in Lexical Databases-African

WordNet

Polysemy is a linguistic phenomenon wherein one word form possesses multiple meanings or senses. In the domain of NLP, correctly representing polysemous terms is vital for tasks such as word sense disambiguation (WSD), machine translation, and information retrieval. Lexical databases, such as WordNet, offer foundational tools for modelling semantic relationships and have been expanded to

various languages, including African languages, through initiatives such as African WordNet (AfWN). Within the context of SsaL, addressing polysemy in such databases poses challenges both linguistically and computationally. WordNet structures words into synsets—groups of synonyms each denoting a specific sense—and links semantic relationships like hypernymy (is-a), hyponymy (kind-of), and meronymy (part-of). However, for polysemous terms in low-resource, morphologically rich languages (MRLs), sense boundaries are more fluid, necessitating refined representation strategies (Camacho-Collados & Pilehvar, 2018; Mathe & Eiselen, 2021).

The SsaL language displays significant levels of polysemy, stemming from its agglutinative word structure, tonal shifts, and context-specific applications. Likewise, “bona” may function as a verb (to see) or as a demonstrative pronoun (those), based on varying syntactic or morphological indicators. Contrary to analytic languages in which semantic boundaries are usually well-defined and supported by extensive corpora, polysemy in SsaL is dynamically influenced by the speaker’s intention, the usage domain (such as agriculture, religion, or anatomy), and morpho-phonological structures. This presents challenges in aligning these senses within lexical databases that were initially designed for Indo-European languages.

African WordNet (AfWN) is an initiative of the Global WordNet Association, also using insights from multilingual projects such as NLU-UBUNTU, to adapt the WordNet architecture for several Bantu languages, among them isiZulu, isiXhosa, Shona, and SsaL. Adopting the Princeton WordNet’s synset alignment framework, the African languages’ sense entries generally correspond to existing English synsets, facilitating cross-lingual integration. Within this framework, synsets comprise a lemma, gloss, and example sentence. In SsaL, lemmas with multiple meanings are catalogued multiple times, each associated with a distinct synset. The AfWN includes bilingual glosses and may also incorporate grammatical and morphological elements, for instance, noun classes and derivation patterns.

Various challenges have been recognized in the representation of polysemy within AfWN for SsaL. Firstly, there is inadequate contextual annotation: AfWN glosses and example sentences often lack the

full context needed to distinguish SsaL meanings. The boundaries of word senses are influenced by factors such as noun class concords, verb derivations, and tone, which are scarcely annotated in current resources. Secondly, there is a misalignment of semantic categories, SsaL polysemy sometimes including cultural concepts that are difficult to translate into English. Terms for kinship, spatial extensions, or farming metaphors might be compressed into a single English gloss, thereby losing semantic richness.

Limited lexicographic depth arises when existing synset entries do not clearly differentiate between sense sub-types (e.g., *noka* as either a natural feature or a body part). This poses challenges in applications such as machine translation in which precise domain-specific meanings are crucial. Lastly, morphological complexity is an issue, with agglutinative forms such as *go-ruta* (to teach), *thuto* (education), *moruti* (teacher/pastor), *baruti* (pastors/teachers) generating derivational ambiguity that has yet to be systematically documented by AfWN. This hampers the ability to capture groups of meanings from a single root morpheme.

Approaches to improve polysemy representation include incorporating multi-prototype word embeddings into AfWN, enabling the representation of a lemma's multiple meanings through separate vectors. By clustering the usage contexts from the SsaL corpora, one can derive senses in a data-driven manner. To ensure that culturally and linguistically meaningful distinctions are preserved, crowdsourced Gloss Expansion involves community validation of senses and glosses, particularly from native speakers (FAA et al., 2012). Connecting AfWN with corpus-based WSD systems by utilizing semi-supervised WSD methods on automatically annotated corpora aids in confirming and increasing the variety of distinct senses.

Morphologically informed lexicon mapping enhances synsets by adding details such as noun class, stem/root origin, derivational group, and domain labels (e.g., anatomy, geography, spirituality), thus improving the mapping accuracy of polysemous terms. Figure 2.5 demonstrates the polysemy of SsaL,

a complex linguistic aspect that requires meticulous representation in lexical databases such as African WordNet. Although AfWN provides an essential framework, it needs substantial improvements to handle morphological diversity, cultural meanings, and context distinctions. The combining of lexicographic expertise with corpus-based and neural representation methods can enhance the semantic grasp and digital interpretability of SsaL polysemy, thereby supporting equitable and comprehensive NLP progress for African languages. The type of mapping can be extended to encompass hundreds of polysemous terms within the African WordNet for SsaL.

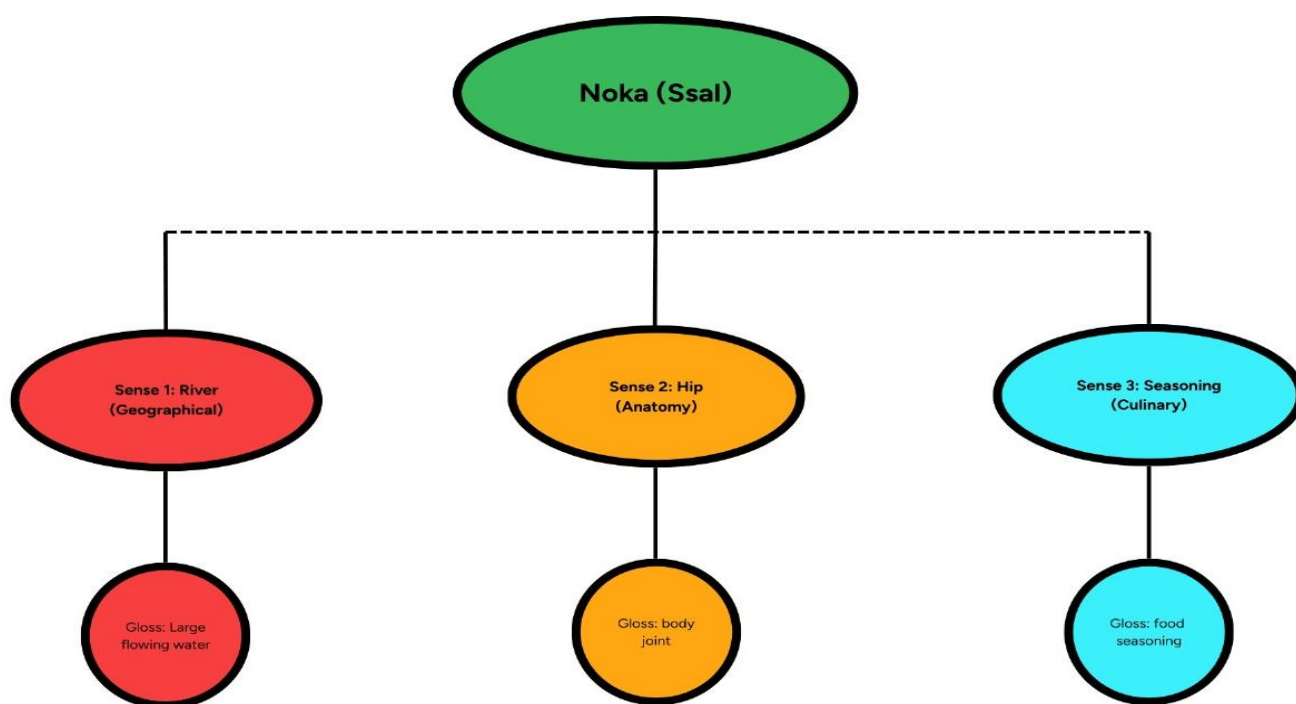


Figure 2.4: Polysemous terms for SsaL

2.7 SsaL as a Low Resource Language and implication for WE Models

2.7.1 The low-resource challenge and dataset creation

SsaL faces a notable shortage of curated linguistic resources, significantly hindering the advancement of effective word embedding models. Marivate, Sefara, Chabalala (2021) emphasize this problem by detailing their efforts to compile small-scale news datasets for SsaL and Setswana, using news headlines as the core corpus. Marivate, Sefara, Chabalala (2021) further tested data augmentation

techniques to enhance classification outcomes, highlighting the urgent need for more extensive datasets in low-resource settings. This situation demonstrates that foundational embedding and NLP systems frequently operate with very constrained, domain-specific text inputs, typically lacking extensive linguistic diversity (Marivate et.al, 2022).

2.7.2 Morphological richness and embedding performance

SsaL is characterized by its extensive use of agglutination, which involves frequent morphological changes such as inflection, derivation, and class agreement. In a study by (Eiselen & Gaustad, 2023), three South African languages, including SsaL, were compared, to assess the influence of training data volume on embedding quality. The study demonstrated that relatively small datasets, consisting of approximately 300,000 tokens, are capable of producing effective embeddings through the use of FLAIR models, which are designed for language identification, analysis, and recognition. These models achieved performance levels similar to those trained on significantly larger datasets, regardless of morphological complexities. This suggests that opting for appropriate model architectures, such as those based on characters or context, can mitigate the challenges posed by limited data by learning from a small but representative sample (Makgatho, Marivate, Sefara, & Wagner, 2021).

2.7.3 Cross-lingual transfer learning and shared embedding spaces

In response to the limited availability of monolingual data, (Makgatho, Marivate, Sefara, & Wagner, 2021) developed monolingual embeddings for Setswana and SsaL, subsequently aligning them within a unified cross-lingual space through vector mapping. This method facilitated transfer learning and enhanced the consistency of semantic representation across linguistically similar languages. Cross-lingual embedding techniques such as these provide a promising strategy for SsaL, utilizing richer languages such as Setswana as resource contributors, to improve semantic alignment and generalization in contexts with limited data (Makgatho, Marivate, Sefara, & Wagner, 2021).

2.7.4 Context-aware and multi-prototype embeddings for disambiguation

Chronis & Erk (2020) highlight the efficacy of context-aware embeddings and multi-prototype models in combatting meaning conflation. This is particularly crucial for languages with rich morphology and

multiple word meanings, such as SsaL, which require nuanced semantic distinctions. These adaptable embeddings modify word representations according to the context, facilitating successful management of words such as “*gola*,” which can have different meanings depending on context. Despite acknowledging obstacles such as data scarcity and high computational demands, the above authors assert that context-aware strategies greatly improve word sense disambiguation and semantic accuracy (Chronis & Erk, 2020; Tian et al., 2014).

2.8 Chapter Summary

This chapter examined SsaL, focusing on building computational approaches and efficient word embeddings for this South African language. Part of the Sotho-Tswana Bantu subgroup, SsaL features significant dialectal diversity and oral traditions, with limited digital representation, making it key for low-resource NLP research. The rich morphology of SsaL, including agglutination and noun class systems, complicates tasks such as lemmatization, requiring subword models. The chapter examines morphemes and lexical categories, emphasizing orthographic challenges and syntactic dependencies crucial to embedding models. This chapter discusses semantic ambiguity and the need for word embeddings to discern meanings through context. Lexical databases such as African WordNet attempt to capture polysemous meanings but remain insufficient. Despite challenges due to limited resources and tools, these constraints highlight the need for morphologically rich informed and semantically sensitive models for African languages.

CHAPTER 3 : STATE OF THE ART

3.1 Introduction

This chapter delivers a thorough examination of the current literature and technological progress addressing the Meaning Conflation Deficiency (MCD) problem in Word vector space (WVS) word embedding (WE) models. This issue is particularly challenging in linguistically complex, resource-limited languages such as SsaL. Initially, Section 3.2 defines and discusses the issue of meaning conflation deficiency, a major drawback of traditional vector-based models that combine various senses into one embedding. Section 3.3 surveys the development of word embeddings, focusing on both static and contextual methodologies. Section 3.4 focuses on strategies to combat word vector space (WVS) word embedding (WE), meaning conflation deficiency (MCD), considering multi-sense as well as context-aware techniques. Section 3.5 builds on this by analysing the rise of CAWE including BERT and ELMo, and their effectiveness in resolving meaning across different linguistic settings. To maintain methodological consistency, Section 3.6 expands the review to include more contemporary research on Bantu and other African languages and further discuss how their morphological and resource-related challenges compare with those of SsaL. The section strengthens the review by including previous SsaL linguistic and computational benchmarks so that the proposed model is more clearly positioned within the existing body of work. This section improves both the contextual grounding and the validation of the study. Section 3.7 details the Systematic Literature Review (SLR) approach undertaken, adhering to PRISMA standards. In Section 3.8, an exhaustive meta-analysis reviews key research, assessing embedding methods, natural language processing tasks, corpus utilization, and reported outcomes. Concurrently, Section 3.9 conducts a bibliometric study to illustrate worldwide research trends, collaborative efforts, and keyword interactions over time. Concluding the analysis, Section 3.10 offers descriptive statistics of the primary studies, providing quantitative insight into current literature's scope and shortcomings. These sections collectively meld qualitative and quantitative data to spotlight current trajectories, identify research voids, and direct the future evolution of successful word embedding solutions for low-resourced languages.

3.2 Meaning Conflation Deficiency (MCD)

MCD, a problem of vector-based WE models, refers to a semantic shortcoming identified in static vector-based word embeddings, in which a single vector encodes multiple meanings of a polysemous word without considering the context (Camacho-Collados & Pilehvar, 2018). For instance, terms like “bank” that denote both a financial institution and a riverbank are represented by a singular embedding in models such as Word2Vec and GloVe. This overly simplistic semantic representation fails to discern detailed sense distinctions, which are essential for meaning-sensitive NLP tasks such as WSD, Machine Translation, Information Retrieval, and Sentiment Analysis. This limitation of single-vector representations leads to the propagation of semantic ambiguity in downstream tasks, thereby diminishing their effectiveness (Pilehvar & Navigli, 2014; Reisinger & Mooney, 2010).

The word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD), occurs when a single word has several meanings (polysemy), however, conventional word embedding techniques (such as Word2vec or GloVe) allocate a single, static vector to each word, irrespective of its context (Huang et al., 2025; Jayakody et al., 2023). This results in the loss of complex semantic information, complicating the differentiation between many meanings of a term (Jayakody et al., 2023). This " WVS WE MCD" in traditional word embeddings, in which a single vector often represents multiple distinct meanings of a word, directly impacts the precision and interpretability of downstream applications (Ansell et al., 2021). It poses a significant challenge in the context of morphologically rich complex languages, characterized by intricate derivation and inflection systems (Ansell, Bravo-Marquez and Pfahringer, 2021).

The significant morphological diversity complicates the association between form and meaning, exacerbating WVS WE MCD. Form-meaning links are clearer in simpler languages, while morphologically rich languages need a deeper understanding of their relationships. This complexity affects language production and interpretation and makes it challenging for natural language

processing models to handle the range of morphological patterns in these languages (Pilehvar, 2019b ; Cer et al., 2018).

This research aims to elucidate the precise mechanisms by which morphological richness intensifies the difficulties related to meaning conflation, via an examination of the consequences of WVS WE MCD. Ultimately, it aims to improve understanding of language processing in various linguistic contexts. In the context of morphologically rich languages (MRLs) (Pilehvar, 2019b) the intricate word structures significantly increase the difficulty of language processing. MRLs exhibit extensive use of inflection, derivation, and compounding, which alter word forms according to grammatical features. In contrast, languages with simpler morphological structures—such as English—tend to convey meaning through syntax and word order rather than through affixation or word form changes. While polysemy is a challenge in all languages, it tends to be more pronounced in MRLs due to their complex morphological variations (Wang et al., 2021; Rahab & Mothapo, 2019).

NLP models face distinct challenges when morphological complexity intersects with Word Vector Space Word Embedding Meaning Conflation Deficiency (WVS WE MCD). Multi-sense conflation occurs when different meanings or senses are compressed into a single semantic representation, posing significant obstacles—especially in languages where word forms carry rich grammatical and semantic information. The challenge extends beyond recognizing these morphological forms to include understanding their interrelations and their contributions to the word's overall meaning within a context. In the realm of NLP for MRLs, the interplay between morphology and WVS WE MCD is critical. The intricacies added by inflection, derivation, and compounding heighten the chances of mistaking different meanings of the same root word.

In reality, word embeddings operate efficiently, despite the polysemous nature of many words, a persistent theoretical dilemma (Pilehvar and Collier, 2016). Words convey various concepts to different individuals, so portraying these words as one point in semantic space is ineffective; this issue

arises from a “shortcoming in meaning conflation,” in which polysemy alters word embedding due to semantic ambiguity (Pilehvar and Collier, 2016). The almost similar vector space concepts “left” and “erroneous” may possess different meanings. Word embedding algorithms may find it challenging to ascertain the significance of an ambiguous word’s meaning. Semantic modelling, a key area in NLP for robots, often treats word types as single entities in semantic space, posing challenges (Ghasemi & Keyvanrad, 2021).

Many methods of depicting semantics have one major drawback: they regard every kind of word as a singular representation in semantic space, hence blending all the conceivable meanings of the word. Recent research has focused much on this issue, which concerns instructing words to have many interpretations for every meaning. The sense models created, however, are incorrect for word senses that do not occur very frequently or do not relate to any sense array. By isolating word representations according to a semantic network’s complete knowledge, some academics (Pilehvar and Collier, 2016) have developed a solution to these issues. Researcher Pilehvar (2019b) claims that studies on a complex neural network-based reverse dictionary system reveal that a basic strategy may significantly reduce meaning conflation (Pilehvar, 2019b). Making use of massive data sets and information sources, the existing approaches to modelling words, word senses, and ideas create dense vector representations that show patterns of meaning in a constrained area. Although many methods work well in just one language, they may be improved with the availability of cross-linguistic data. Researchers have posited a hypothesis that highlights multilingualism in concept vector representations independent of language, keeping unique links between ideas in order to overcome this problem. Word meaning disambiguation, semantic word similarity, and bilingual or cross-lingual models are all outdone by the method’s high-coverage models. For languages that do not have many resources, this works wonders. Sense embeddings are fascinating because of their flexibility and ability to retain relevance, even when they have constraints. According to (Camacho-Collados & Pilehvar, 2018) knowledge-based sense embeddings have greatly enhanced word sense induction (WSI), even helping to disambiguate words more.

Semantic uncertainty can be solved with a neural network-driven backward dictionary system that does not need many changes (Pilehvar, 2019b). This method allows people to describe things and gain terms that go with them, which makes it possible to complete actions such as idea search or reverse vocabulary. In this case, a tool used for cutting grass should lead to “sickle” (*selepe*). These kinds of methods may be especially helpful for writers and translators who have trouble readily remembering words, or who need help clarifying ideas (Tian et al., 2014; Yang and Mao, 2016; N. Zhang et al., 2020). With a neural network-based backward dictionary for SsaL, appropriate words could be found automatically, based on descriptions. This could make the language much easier to use.

Languages with much grammar but few resources, such as SsaL, suffer from WVS WE MCD. The processes of compounding, derivation, and inflection provide words in these languages a great deal of morphological flexibility. More accurate word descriptions, such as context-aware and multi-prototype embeddings, are required for these languages. There are several vectors that may stand in for a single word, with each of those vectors conveying a distinct meaning or sense (Tian et al., 2014; Yang and Mao, 2016; Zhang et al., 2020). This immediately targets the MCD problem in WVS WE models by eliminating a muddled procedure for storing words with multiple meanings. The embedding approach assigns a unique vector to each category due to the multi-faceted nature of words. Because of this, no two words have the same meaning, and those meanings may shift based on context. Each word's meaning determines which vector is best used in a phrase (Won, Lee and Kang, 2020).

Cross-lingual word embeddings allow one to learn multiple word representations without direct bilingual signal. This makes it possible to learn languages with either few or many resources. Cross-lingual word embeddings use projections to connect pre-trained embeddings to a shared hidden space. Most of these methods are based on orthogonal transformation, which assumes that language vector spaces are isomorphic. This need is not always met, although, especially so in languages with much grammatical richness. Researchers (Biesialska and Costa-Jussà, 2020) have proposed a self-supervised

method to augment the alignment of unsupervised language word embeddings (Biesialska and Costa-Jussà, 2020). The suggested method makes it possible for cross-lingual embeddings to be better aligned. It preserves length and center while aligning word vectors and their translations, achieving superior results in bilingual language generation compared to current methods (Biesialska and Costa-Jussà, 2020).

3.3 Word Embeddings (WE)

The diagram in Figure 3.1 is a semantic representation—from static word embeddings (Word2Vec, GloVe, FastText) (Rodrigues da Silva and Caseli, 2021b ; Ansell, 2020c; Ilie et al., 2021b), which assign a single vector per type and consequently exhibit WVS WE MCD. Furthermore, Figure 3.1 illustrates contextual embeddings (ELMo, BERT/GPT/T5/ELECTRA) (Li, Lei and Ji, 2022; N. Li et al., 2021; Loureiro and Jorge, 2019), which generate token-in-context vectors and thereby partially mitigate conflation, culminating in a category labelled context-aware word embedding (e.g., “BERT+BiLST+CNN”, XLNet, RoBERTa, ALBERT, DeBERTa, PolyLM) that purportedly resolves word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) through task-specific architectures or cross-sentence fusion. This concept encapsulates the principal notion: more context often results in less sense conflation.

Nevertheless, two cautions are significant: (i) Numerous models such as (XLNet, RoBERTa, ALBERT, DeBERTa) are conventional contextual language models; these models attain genuine "context-awareness" only when enhanced with explicit mechanisms (e.g., co-attention, document-level fusion, multi-prototype/sense goals). (ii) The empirical research indicates residual conflation for infrequent meanings, domain shifts, code-switching, and morphologically rich languages; even robust language models may conflate meanings when the context is brief or ambiguous. The replication of ELMo in both the centre and right boxes suggests this overlap: the same encoder may still merge meanings unless accompanied by sense-aware supervision, morphology-aware tokenization, or discourse-level context. A comprehensive assessment in Figure 3.1 indicates that static techniques are

inferior to contextual and task-aware/context-fused approaches in word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD)robustness; nonetheless, thorough, sense-level examination (WSD or WSI, rare-sense recall, OOV stress testing) is essential prior to asserting that WVS WE MCD has been "resolved" (Li, Lei and Ji, 2022; N. Li et al., 2021; Loureiro and Jorge, 2019).

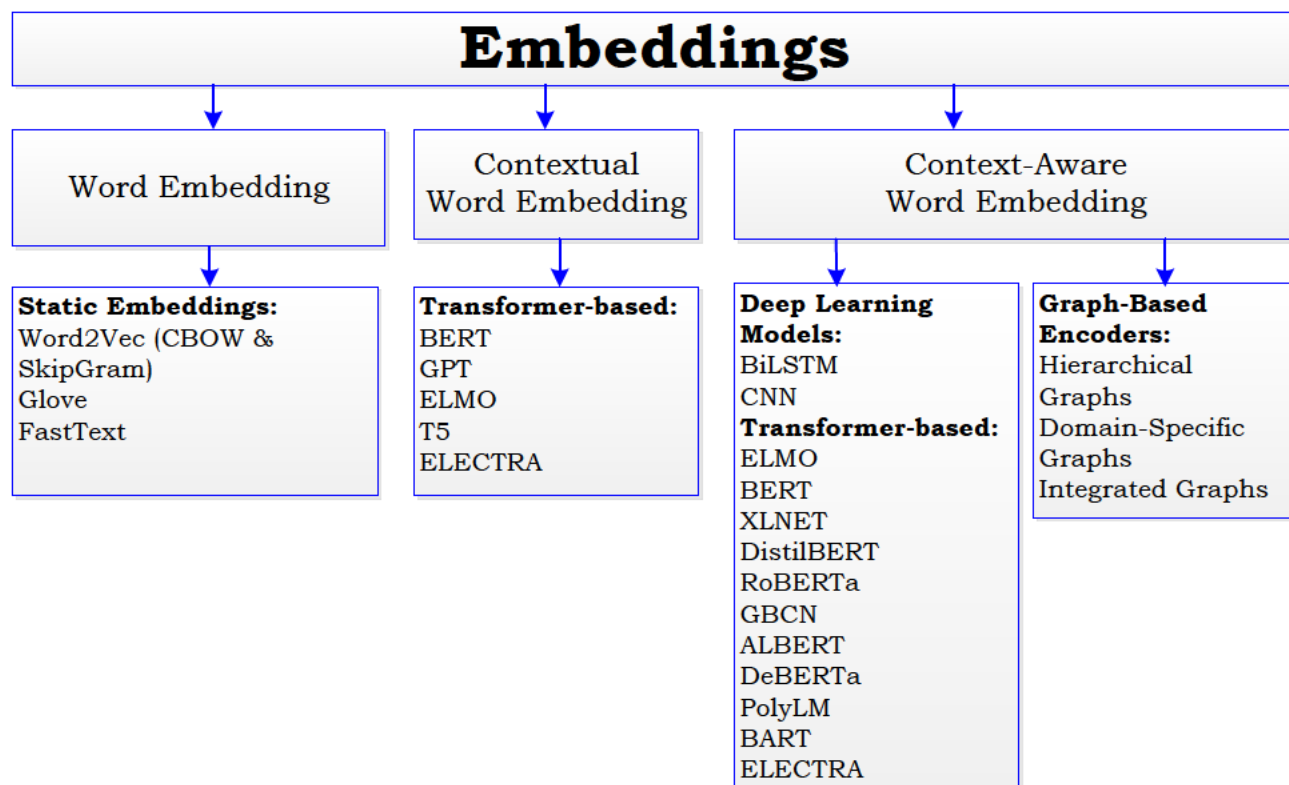


Figure 3.1: Embeddings taxonomy

Word embeddings map words to a fixed-size vector and use massive text dataset word co-occurrence patterns to establish semantic links, with WE's word sign being constant and encoding words without context. Simple GloVe or Word2Vec models have allowed this to occur (da Silva and Caseli, 2020a). Word embedding ignores multi-meaning words and creates one vector graphic per word. In NLP, sensation models surpass Word2Vec, GloVe, and FastText (Rodrigues da Silva and Caseli, 2021a).

Static WEs such as Word2Vec, GloVe, Wang2Vec, and FastText map words to vectors without addressing multi-meaning words. Word embeddings are controversial because words have several meanings. Similarly to "*mafahla*" in SsaL, homophones have several meanings. Their pronunciation

and writing match the language. Interestingly, word embeddings display *mafahla* (lungs, twins) interpretations per one mathematical symbol. Ansell (2020) refers to the concept of bewilderment shortage. Indeterminacy influences word embedding, and analysing word similarities through cosine distance highlights vector patterns. However, research has not extensively explored how meaning conflation scarcity impacts word-based NLP systems in low-resource languages such as SsaL (Ansell, 2020).

Static vectors (Word2Vec, GloVe, FastText) provide a single representation for each type; hence, polysemy merges meanings, distorting neighbourhoods and impairing WSD/WiC performance (Camacho-Collados & Pilehvar, 2018). Static word embeddings exhibit WVS MCD, merging several meanings into a singular vector, hence adversely affecting sense-sensitive tasks in morphologically rich, low-resourced languages (MRLs) (Camacho-Collados & Pilehvar, 2018). Static embeddings assign a fixed vector to each word regardless of its usage context, often leading to performance limitations in tasks requiring fine-grained semantic understanding (Ansell et al., 2021; Daniel Loureiro, 2019). One significant limitation of earlier models such as Word2Vec and FastText is that they assign a single vector representation to each word, thereby overlooking the polysemous nature of words and the multiple meanings they can convey (Silva & Caseli, 2021). This static representation fails to differentiate between varying senses of a word, which is particularly problematic for ambiguous words in which a single form can have several distinct meanings, thereby limiting their utility in disambiguation tasks (Thurnbauer et al., 2023).

Camacho-Collados & Pilehvar, (2018) proposed two main strategies to counter WVS WE MCD: (1) multi-prototype or sense embeddings, derived from corpora or aligned with lexical inventories, and (2) contextualized embeddings that generate a token-specific vector based on sentence context. Contextual models are attractive because they integrate end-to-end with downstream fine-tuning and require no external sense inventory (Devlin, M.-W. Chang, et al., 2019).

This MCD challenge in WVS WE models has prompted the advancement of sense embeddings, which aim to represent each token with a distinct embedding for each of its potential meanings, often by applying clustering algorithms to contextualized embeddings to derive representative sense vectors (Wang et al., 2025). This challenge has led to the development of multi-prototype/sense embeddings, which aim to represent each token with a distinct embedding for each of its potential meanings, often by applying clustering algorithms to contextualized embeddings to derive representative sense vectors (Wang et al., 2025). This method addresses the word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) by enabling the representation of distinct semantic usages, although it can still be limited by the granularity and coverage of the clustering algorithms (Pilehvar & Camacho-Collados, 2020).

The advancement of contextual word embeddings, which generate dynamic representations based on a word's surrounding context, has significantly alleviated many issues associated with static word embeddings; and provides a reliable means for capturing the semantic and syntactic properties of word usage in context (Pilehvar & Camacho-Collados, 2020). However, even with these advancements, contextualized embeddings can still exhibit variance across different contexts, potentially disrupting semantic consistency for synonyms, and necessitating further research into their stability and reliability (Wang & Zhang, 2023a).

To circumvent this word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) issue, context-aware word embedding models have emerged, allowing for dynamic word representations that adapt to the surrounding linguistic context, thereby addressing the MCD in WVS WE by enabling words to be represented by an infinite set of possible embeddings (Ansell et al., 2021). The development of context-aware word embedding models has become crucial for mitigating the challenges posed by the WVS WE MCD in word meaning representation. Such embedding enhances the performance and reduces the complexity of the WSD task and its applications (Geline et al., 2024).

A more advanced approach involves dynamic contextual word embeddings, which generate representations that vary not only with linguistic context but also with temporal and social dimensions,

thereby capturing more intricate semantic dynamics (Hofmann et al., 2021). These embeddings, building upon the strengths of contextualized models, offer a richer representation by modelling time and social space jointly, making them particularly attractive for diverse NLP tasks involving complex semantic variability (Hofmann et al., 2021). Despite these advancements, the integration of sense representations into downstream applications is not always straightforward, often necessitating careful consideration of suitable evaluation benchmarks, which have historically been lacking (Pilehvar, 2019).

This deficiency highlights the need for models capable of generating distinct word representations based on contextual cues, effectively addressing the meaning conflation problem prevalent in static embedding approaches (Horn, 2017). This inherent limitation of single-prototype representations implies that a word’s varied meanings cannot be adequately captured, particularly when these meanings diverge significantly (Huang et al., 2012). This MCD problem in WVS WE models impedes accurate semantic modelling: the inability to distinguish between different senses of a word can hinder the performance of NLP systems (Camacho-Collados & Pilehvar, 2018). To overcome this, models that consider dynamic polysemy aim to represent a word’s meaning as a distribution over multiple senses rather than a single fixed vector (Camacho-Collados & Pilehvar, 2018). This approach seeks to assign specific embeddings for each sense of a word, which is particularly useful for disambiguating both contrastive and nuanced senses (Trask et al., 2015).

This situation directly addresses the limitation of fixed sense numbers by allowing for a more flexible and accurate representation of word meaning contextually in which it is used, thus capturing the dynamic evolution of word meanings over time (Hofmann et al., 2021). Such models can either generate a fixed number of senses based on an external sense inventory, or dynamically create new senses as needed through non-parametric clustering (Pilehvar & Camacho-Collados, 2020). However, relying on external lexicons for predefined sense distinctions can limit a model’s ability to handle out-of-vocabulary words; and may not align perfectly with the sense distinctions observed in specific text corpora (Camacho-Collados & Pilehvar, 2018).

Furthermore, the conflation deficiency in traditional word embeddings, which represents multiple meanings with a single vector, not only compromises semantic accuracy but also violates the triangle inequality in Euclidean spaces, further diminishing the utility of such models in fine-grained semantic tasks (Camacho-Collados & Pilehvar, 2018). Addressing this, context-aware word embedding models have emerged as a promising solution to mitigate the WVS WE MCD by generating distinct representations for words based on their surrounding context, thus enhancing performance and reducing complexity in word sense disambiguation tasks (Camacho-Collados & Pilehvar, 2018). This approach is particularly beneficial because it integrates sentential information directly into the word representations, helping to model polysemy more effectively (Wang & Zhang, 2023b). This dynamic approach stands in stark contrast to static embeddings, which assign a fixed representation irrespective of the surrounding words, thereby failing to capture the fluidity of word meanings across different contexts (Pilehvar & Camacho-Collados, 2020).

3.3.1 Word2Vec

Word2Vec is a fundamental neural network model used for generating word embeddings. In vector space, output vectors are structured to place commonly used circumstances close together. This suggests that context may reveal a word's meaning. Word2Vec has two word embedding models. The CBOW and Skip-Gram models play a role here. These models are grounded in the Distributional Hypothesis, which suggests that words with similar meanings tend to appear in similar contexts. The CBOW model specifically requires a unigram, which represents an n-gram consisting of a single continuous phrase. This model averages the contexts of all words in the selection box to gain the context of the word. Skip-grams demonstrate word meaning in sentences. We can maintain the context of distant words in the series and deduce the context of a given word (Ilie et al., 2021a). Equations (3.1) and (3.2) show the skip-gram model's goal (Ansell, 2020):

$$J(c, o, u, v) = \log \vartheta(u_c^T v_o) + \sum_{i=1}^k E_{w_i \sim p_n} [\log (1 - \sigma(u_c^T v_{w_i}))] \quad (3.1)$$

where input vector $u \in \mathbb{R}$ and output vector $v \in \mathbb{R}$, $i \in V$, given that c is the centre word, o is a word in the context of c , k is the number of negative samples per true context word, and the sigmoid function σ is given by

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (3.2)$$

3.3.2 GloVe

The GloVe algorithm uses unsupervised learning to embed words by creating a term co-occurrence matrix. The combinatorial multitude of circumstances gives this matrix huge dimensions (Ilie et al., 2021b).

3.3.3 FastText

Word2Vec is augmented by FastText, which employs CBOW and Skip-Gram. The first improvement addresses the drawbacks of not considering word location in the context pane. The recommended technique learns word position representations to weight context terms (Ilie et al., 2021b).

3.4 Approaches addressing Word vector space (WVS) word embedding (WE) Meaning Conflation Deficiency (MCD)

3.4.1 Contextual Word Embedding (CWE)

Models change word embeddings based on phrase context. The words around these embeddings modify them dynamically, often created using BERT, GPT, or ELMo language models. CWE word representations fluctuate with phrase context. Enhanced models employing deep learning and transformer architectures, such as BERT or GPT, utilize CWE to generate multiple embeddings for the same word depending on its context (Zhang et al., 2020b). The context of the target word continually changes contextualized representations. This describes context sensitivity. ELMo, BERT, and GPT-2 provide contextualized embeddings, or deep neural language models. A variety of downstream NLP tasks may be modelled on these models. Language models synthesize these context-dependent word representations, which depend on the entire input sentence, to generate phrase embeddings (Zhang et al., 2020b).

WE approaches that acquire word representations only from surrounding words are context-level. Recently trained context-level models are computationally efficient. If inference extracts a vectorized word representation from an individual word, research has concentrated on the training data for models. Word embedding's inability to display many word meanings is one of its main drawbacks. New models called contextual word embeddings have been developed. These models are better because their representations change with the circumstances (Vusak, Kuzina and Jovic, 2021). To overcome WVS WE MCD, modern embedding methods map each word to a vector at every point in time. The researchers (Hu, Li and Liang, 2019) propose a deep contextualized embedding-based sense representation and monitoring system to address what, when, and how word-meaning evolves. Results show that the framework enhances word-change identification, and captures fine-grained word senses (Hu, Li and Liang, 2019).

3.4.2 Embedding from Large Language Models (ELMo)

Character-based ELMo generates word embeddings that are not contextualized. It utilizes convolutional filters, two stacked highway layers, and a linear transformation. These representations let ELMo create two language models using BiLSTM networks: One model estimates the probability of past tokens and the other forecasts future tokens (Vusak, Kuzina and Jovic, 2021; Ansell, 2020). ELMo's core feature is estimating the next word in a sentence. After training on a large dataset, the model recognizes linguistic patterns. A two-layer bidirectional LSTM language model was trained using the Billion Word Benchmark dataset built on a context-independent character CNN layer (Rodrigues da Silva and Caseli, 2021b; Ansell, 2020) The development of contextualized models such as ELMo and BERT has been a major NLP advancement. Transfer learning is used in NLP to create contextualized models. Contextualized models generate contextual word embeddings from input word sequences, for instance, phrases. Word embeddings convey word meanings in isolation, whereas contextual word embeddings incorporate linguistic processes beyond the word level (Ansell, 2020).

3.4.3 Text-to-Text Transfer Transformer (T5)

As reported by Katsarou, Rodríguez-gálvez and Shanahan (2022) T5 and multilingual T5 (mT5) are used to translate contextualized embeddings into the Word2Vec space, providing a transformer-specific solution for computation. Prior to utilizing this method, contextualized embeddings were converted to Word2Vec. Transformer models are being increasingly applied in practical scenarios (Katsarou, Rodríguez-gálvez and Shanahan, 2022).

3.4.4 Generative Pre-Training (GPT)

GPT model uses a left-to-right language modeling approach. Instead of LSTM layers, the GPT model extracts features using the transformer architecture. This design excels in many NLP tasks. GPT can manage long-term text dependencies by keeping more organized memory, creating a more complete model (Vusak, Kuzina and Jovic, 2021; Ushio *et al.*, 2021).

3.4.5 ELECTRA

Few studies explore word context utilising contextualized word embedding (transformers) rather than context-free methods (Balakrishnan et al., 2022). Transformers and deep neural networks assess ensemble learning models' identification of meaningful and nonsensical Twitter catastrophic tweets (Balakrishnan et al., 2022b). Bi-LSTM gave simpler transformers such as ELECTRA and Talking-Heads Attention F-scores between 80% and 84%, making these better than BERT. Our results indicate that simple current transformers may be used without significant computer work (Balakrishnan et al., 2022a; Song and Huang, 2021; Li, Lei and Ji, 2022).

3.4.6 Bidirectional Encoder Representations for Transformers (BERT)

Unlike GPT, BERT utilizes a bidirectional transformer architecture. It also leverages two unsupervised pre-training objectives, as noted by Vusak, Kuzina and Jovic (2021). Specifically, BERT is a deep, multi-layered transformer encoder trained concurrently on masked language modeling and next-sentence prediction tasks (Rodrigues da Silva and Caseli, 2021b). N. Li et al. (2021) investigate whether BERT's contextualized embeddings can create noun-sense word vectors for knowledge-based

completion. The research found that contextualized embeddings of masked word instances create better vectors than BERT’s static vectors and other models (N. Li et al., 2021; Ushio et al., 2021).

A contextualized word embedding model such as BERT (Ansell, 2020c) can be conceptualized as a two-component formulation, as outlined in equations (3.3) and (3.4). It consists of an embedding matrix $E \in \mathbb{R}^{|V| \times d}$ and a contextualization function C , which together transform an input embedding sequence $x_1, x_2, \dots, x_n \in \mathbb{R}^d$ into a sequence of output representations $y_1, y_2, \dots, y_n \in \mathbb{R}^d$.

$$P(w|y) = [\text{softmax}(E_y + b)]_w \quad (3.3)$$

$$= \frac{e^{e_w^T y + b_w}}{\sum_{j \in V} e^{e_j^T y + b_j}} \quad (3.4)$$

These static word-level representations often favour a word’s most common meaning in that they are unaffected by context. Because language models such as BERT, which are pretrained on massive amounts of text, are generally accessible, contextualized embeddings have grown in popularity. Contextual word embeddings capture polysemous words’ characteristics (Zhang et al., 2020b). Unsupervised sense representations, unlike word embeddings, help scholars identify multiple meanings of a word through contextual semantics. Researchers focus on deep-brain language models and sensory embeddings. The sense embedding model (Sense2vec) beat traditional word embeddings in syntactic and semantic parallelism tests, proving that the language resource may improve Portuguese NLP tasks (Rodrigues da Silva and Caseli, 2021b). In the semantic textual similarity challenge, researchers tested pre-trained deep neural language models (ELMo and BERT) in feature-based and fine-tuning transfer learning. After fine-tuning, the multilingual and Portuguese BERT language models outperformed the ELMo model in accuracy tests (Rodrigues da Silva and Caseli, 2021b).

Zhang et al., (2020a) recommend a knowledge-based method for acquiring word meaning representations for later challenges. The paper specifically proposes utilizing sememes, the smallest fundamental meaning components, to leverage meaning-based data from existing utilizing human

understanding to develop global context vectors and achieve effective soft word sense disambiguation for words with multiple meanings. This study integrates a contextual attention mechanism within the Skip-gram model framework to generate sense-specific embeddings. External evaluations and analyses reveal that these embeddings effectively mitigate polysemy and improve performance in various downstream applications. Moreover, the intrinsic experimental results go beyond standard word similarity benchmarks, capturing nuanced and varied interpretations of meanings with greater precision (Zhang et al., 2020b). Due to the meaning conflation weakness in standard word embeddings, several research have modified neural language models (NLMs) to produce more sense-specific word embeddings. The research study of (Loureiro and Jorge, 2019) uses context2vec to create sense embeddings as a recent contextual word embedding approach (Loureiro and Jorge, 2019). The contextualized word embedding approach still suffers from meaning conflation. Hereafter, the jumbled meanings may be harder to separate (Ansell, 2020c).

3.5 Context-Aware Word Embeddings (CAWE)

Context-aware word embeddings (CAWE) are advanced word representation techniques in Natural Language Processing (NLP) that capture the meaning of a word depending on the context in which it appears. This might involve syntax, semantics, or subject matter, and most of the time, CAWE adds more information or parts to make people more aware of their surroundings. Such often uses more complicated environmental data or models than what is used in CWE. Systems that describe words, such as context2vec or ELMo, change how they show sentence or document settings on the fly. State-of-the-art natural language processing results can be found in newly created deep neural models for Portuguese language. The lexical similarity tasks in the ASSIN dataset were used by researchers (Li, Lei and Ji, 2022) to evaluate the differences between deep neural language models and traditional word embeddings for both Brazilian and European Portuguese.

Tests indicated that the ELMo language model was more accurate than all other models not trained on Portuguese. ELMo performed better alone and with standard word embeddings after reducing the dataset's word count before training (Li, Lei and Ji, 2022). Traditional methods often approach this as

a text categorization problem based on machine learning. Nonetheless, many of these methods depend on manually created and literature-derived components. To demonstrate CADEN—a CAWE network developed for analyzing financial texts, Li et al. (2021) simultaneously retain both global and local contextual information. This approach enables the model to capture and incorporate attitude-aware user information, thereby enhancing its overall performance.

A topic that remains unresolved is the normalization of word embeddings to obtain context-sensitive information. First, context-aware dynamic word embedding (CDWE) has the capability to take into account the historical usage of a word, and specialized meaning to address these challenges. Loureiro and Jorge (2019) suggest a CDWE-based attention-based convolutional neural network for extracting aspect terms (ATE). The network uses an attention approach to assign distinct values to each encoding to recall prior meanings (Loureiro and Jorge, 2019).

Yang et al. (2019) suggested using context-aware large language models. This system includes a Bi-LSTM layer, which is tasked with learning long-range dependencies, thereby generating deep contextual representations. Yang et al. (2019) found that the proposed structure functioned well in subsequent positions. In datasets utilizing the feature representations from the proposed framework, both Random Forest and Support Vector Machines outperformed four baseline models. This indicates that the framework, combined with ELMo, is effective in identifying vulnerable data flow patterns and supporting detection efforts (Yang et al., 2019). Additionally, (Kavatagi and Rachh, 2021; Loureiro, Mário and Camacho-collados, 2022) employed a Bidirectional Long Short-Term Memory network with Conditional Random Fields (BiLSTM + CRF), using labeled sensitive information across multiple categories to predict conversation outcomes (Kavatagi and Rachh, 2021; Loureiro, Mário and Camacho-collados, 2022). Conditional random fields are capable of categorizing and arranging sequential data. The combination of BiLSTM and CRF, augmented by BERT embeddings and WordShape features, achieves an F1 score of 96.73%, representing the optimal model configuration (Kavatagi and Rachh, 2021; Loureiro, Mário and Camacho-collados, 2022).

Contextual embeddings such as BERT and ELMo generally outperform others. Using ELMo embeddings with logistic regression proved to be the most effective configuration, achieving strong accuracy rates of 90% on Twitter and 94% on YouTube. Conversely, Word2vec and GloVe, well-known word embeddings, encapsulate a word's meaning into a fixed vector, irrespective of context. In contrast, BERT and ELMo have the ability to produce varied word representations for identical terms based on its location within a sentence. Due to their training on extensive datasets, these models can significantly streamline numerous NLP challenges (Li et al., 2020; Ilie et al., 2021b).

Word2Vec fuzzy logic approach, GloVe, and BERT embeddings achieve high accuracy across multiple datasets (Li et al., 2020)(Ilie et al., 2021b; Vusak, Kuzina and Jovic, 2021). This demonstrates that the model understands complex sarcasm. The BERT approach improves pretrained models faster by improving smaller, domain-specific datasets, which may enhance outcomes for various aims. The advantages of this strategy outweigh its high computational power requirements. ELMo can grasp context and create alternative models for homonyms, which appear the same but have different meanings. Because it reads the whole input line, ELMo understands words better than competing models (Li *et al.*, 2020).

ELMo works because Bidirectional LSTM (BiLSTM) is taught using language modelling on massive volumes of text. BiLSTM combines forward and backward LSTM sections to evaluate texts and to anticipate word patterns. This makes a two-way language model that improves language learning using word embeddings and BiLSTM processing suitable for various NLP applications (Li et al., 2020; Loureiro, Mário and Camacho-collados, 2022). The study article (Vusak, Kuzina, and Jovic, 2021) cleverly groups Turkish project concepts using word embedding, grouping, and categorization. The proposed model extracts keywords from Turkish project bid titles and summaries using FastText, BERT, and TF/IDF word-embedding. Classification and grouping extracts concepts from text. More than 30,000 complicated noun-verb uncertainty instances have been generated by scientists (Martin, Yang and Hsu, 2022).

Taggers within 1% of each other on the WSJ have 57% to 75% success rates for this problem set (Masethe et al., 2024). After incorporating contextual word embeddings and tailored training data, our top tagger has an accuracy of 89%, a 14% absolute (or 52% relative) increase. The following stage with this improved tagger reduces text-to-speech synthesis homograph disambiguation mistakes by 28% (Masethe, Masethe and Ojo, 2024) compared with the prior best-trained model.

In CAWE like BERT, mBERT, and ELMo, word representations are adjusted dynamically based on the context that surrounds them (N. Li et al., 2021). This approach surpasses static word vectors by allowing each word's embedding to change depending on its usage within a phrase. Models like BERT achieve a complete understanding of a word by examining the words that surround it, analyzing the text from both directions: left-to-right and right-to-left. Unlike static word embeddings, Context-Aware Word Embeddings (CAWE) produce distinct vector representations for the same word depending on the sentence in which it appears. Take the Sesotho sa Leboa word “gola”, for instance, which can mean either “to grow” or “to become large” based on usage.

In the sentence “*Ngwana o gola ka pela*” (The child grows quickly), CAWE interprets “gola” as referring to biological or developmental growth. Conversely, in “*Dilo di gola kudu*” (Things are getting bigger), the same word refers to increasing in size or magnitude. These nuanced interpretations are captured by generating context-specific embeddings, meaning the model accurately differentiates between the two senses of “gola”. By doing so, CAWE effectively addresses the WVS WE MCD problem (N. Li et al., 2021), ensuring that each word is represented according to its true contextual meaning every time it appears.

BERT, mBERT, and ELMo are CAWE models that change word representations (S. Li et al., 2021). The method outperforms static word vectors because phrase position determines word embedding. BERT can read both ways and interpret context-dependent words. CAWE create phrase-specific word vectors. “Gola” means “grow” and “become large.” Including “gola” to suggest development helps “*Ngwana o gola ka pela*” (The child grows swiftly). “*Dilo di gola kudu*” means “things are growing” with “gola”. Context-aware embeddings allow “gola” sentence meanings. The model always

understands “*gola*” to naturally solve WVS WE MCD using the context-sensitive technique (S. Li et al., 2021).

3.5.1 PolyLM

To avoid the WVS WE MCD, various strategies add word meanings to embeddings. Contextualized task-specific embedding approaches outperform these algorithms on word-sense-induction (WSI) tasks (Ansell, Bravo-Marquez and Pfahringer, 2021). Acquiring sense embeddings presents challenges because these embeddings add complexity to language models. PolyLM outperforms previous sense embedding algorithms while having six times fewer factors than the latest specialized WSI approach (Ansell, Bravo-Marquez and Pfahringer, 2021). Experts created PolyLM (Ansell, Bravo-Marquez and Pfahringer, 2021), a comprehensive contextualized sense embedding model (Ansell, 2020c) to reduce meaning conflation. PolyLM learns sense embeddings and contextualization via masked language modelling (Ansell, Bravo-Marquez and Pfahringer, 2021). PolyLM is capable of performing word sense NLP tasks without prior training. By solely leveraging sense embeddings, PolyLM effectively eliminates WVS WE MCD. Due to varying contexts, word meanings can change. As a result, a text component is represented by both a list and a concept. According to the initial finding of PolyLM, the probability of word (w) appearing in context (c) is equivalent to the combined probabilities of its individual components. (Ansell, 2020a). As seen in Equation (3.5):

$$P(w|c) = \sum_{s \in S_w} P(s|c) \quad (3.5)$$

3.5.2 XLNet

XLNet's distinctive pre-training objective is based on the Transformer-XL framework., focuses on Permutation Language Modeling (PLM) (Yang et al., 2019). One of the benefits of PLM is that it does not rely on masked tokens, which makes the pre-training phase more similar to the fine-tuning process. Additionally, the dataset used to train XLNet is significantly larger than the one used for BERT., incorporating a wide range of web content from various sources along with the BERT dataset. Instead of using WordPiece, XLNet uses Sentence Piece, an open-source tokenization method akin to WordPiece (Yang et al., 2019). XLNet leverages the Transformer-XL architecture to achieve

bidirectional learning by calculating probabilities over all possible permutations of factorization orders (Vusak, Kuzina and Jovic, 2021).

3.5.3 DistilBERT

Kavatagi and Rachh (2021) proposed a linguistically informed approach for identifying hate speech within Twitter data. Their model integrates Distil-BERT, a lightweight context-aware transformer, with a Support Vector Machine (SVM) classifier to enhance detection performance. The study's findings reveal that applying a context-aware word embedding (CAWE) framework significantly boosts classification accuracy compared to traditional methods. DistilBERT, a streamlined version of the versatile BERT, is 40% more compact, operates 60% quicker, and possesses 97% of BERT's language comprehension capabilities. The distillation process allows for effective training of this versatile language model (Kavatagi and Rachh, 2021).

3.5.4 RoBERTa

RoBERTa is clearly mentioned as an improved version of BERT. It was found that including the Next Sentence Prediction (NSP) task during pre-training actually reduced performance in the tested experimental settings; thus, RoBERTa focused solely on the Masked Language Modeling (MLM) task during its pre-training phase. In addition, the model is optimized using various hyperparameters—such as increased batch sizes—to improve its effectiveness on downstream tasks. Like XLNet, RoBERTa is pretrained on large-scale corpora primarily consisting of internet-based text. For tokenization, RoBERTa utilizes byte-level Byte Pair Encoding (BPE), which makes it more complex to extract token-specific embeddings, since it requires the explicit encoding of spaces (Loureiro, Mário Jorge and Camacho-Collados, 2022; Ushio et al., 2021).

3.5.5 GBCN

In their research, Li et al. (2020) point out that BERT's input setup, limited to sequences of words, fails to incorporate extra contextual information. To address this issue, they introduce a new technique called GBCN. This approach refines and strengthens BERT's effectiveness in aspect-based sentiment analysis by incorporating a gating mechanism alongside context-aware aspect embeddings. The input

is processed separately through a context-enhanced embedding layer and the BERT model, resulting in enhanced overall performance. Using standard BERT settings, the model achieves an F1 score of 88.0, which increases to 92.9 when the enhanced context-aware embeddings (CAWE) are applied (Li et al., 2020).

3.5.6 DeBERTa

The core design of transformers is integral to DeBERTa, distinguishing it from BERT through its unique self-attention mechanism. The third model was developed using a pre-trained DeBERTa language model, which is based on the BERT transformer architecture, enhancing efficiency via the separation of attention mechanisms and boosting interpretability (Martin, Yang and Hsu, 2022). In DeBERTa, the location and token encodings are split into two separate vectors through a two-vector method. This enables the attention layers to differentiate between each encoding vector and derive new insights from them. The researchers also applied advanced decoding techniques to identify the precise and comparative positions for all words in a sentence. Due to these improvements, DeBERTa often surpasses BERT in performance, excelling in various natural language processing tasks (Martin, Yang and Hsu, 2022).

3.5.7 ALBERT

In their studies, Loureiro, Mário Jorge and Camacho-Collados (2022), along with Lan et al. (2020), suggested ALBERT as a more effective alternative to the standard BERT model. This "lite" version of BERT includes fewer parameters compared to the traditional BERT structure. Even with efficiency improvements like cross-layer parameter sharing, ALBERT maintains the same design as BERT. It boosts performance in subsequent tasks by replacing the NSP with the more challenging Sentence Order Prediction (SOP) objective. Both XLNet and ALBERT were developed using comparable datasets, utilizing sentence piece for tokenization. The testing speed is similar to that of BERT, and there are only a few versions available. ALBERT comes in several versions (Loureiro, Mário Jorge and Camacho-Collados, 2022; Lan et al., 2020).

Kumar and Kumar (2024) integrated WSD and Seq2Seq models to represent abstract information. The Seq2Seq framework effectively utilized LSTM components, embedding representations, and time-distributed fully connected layers for summarization. This approach required expanding vocabulary, embeddings, and LSTM units. Through extensive training with pre-processed data, the model learned to condense information, thereby capturing more context and detail from the input words, which enhanced the accuracy and appeal of its descriptions (Kumar and Kumar, 2024). Table 3.1 illustrates how successfully the authors handled WVS WE MCD.

Table 3.1: Summary of Performance Benchmarks of Embeddings

Embedding Category	Model	Metric Reported	Performance (%)
Traditional Word Embedding (WE)	FastText	Accuracy	89%
Contextual Word Embedding (CWE)	ELMo	Accuracy	89%
	ELECTRA	F1-Score	84%
Context-Aware Word Embedding (CAWE)	DistilBERT	Accuracy	85%
	DistilBERT	F1-Score	84%

CAWE is based on contextual word embeddings as outlined by Li, Lei, and Ji (2022), Li et al. (2021), and Loureiro and Jorge (2019), with the objective of capturing intricate linguistic properties and connections at the levels of words, phrases, and documents. As NLP technology advanced, contextual embeddings expanded to include not just words, but also phrases, sentences, and entire passages. This allows the reader to understand the broader context, beyond mere words. After undergoing extensive pretraining on large-scale datasets, CAWE models such as BERT, GPT, RoBERTa, and T5 are further fine-tuned to perform effectively on specific downstream tasks. These models are capable of learning various language patterns for application in numerous NLP tasks, thus enhancing their versatility. Transformers use self-attention mechanisms to identify relationships within complete sentences, thereby enabling CAWE (Li, Lei and Ji, 2022; N. Li et al., 2021; Loureiro and Jorge, 2019).

In summary, developing a key to address MCD in WVS WE models is significantly challenging due to the lack of vital language resources and annotated data that most WE methods need. CAWE methods

are specifically designed to handle multiple word senses and aid in semantic disambiguation suitable for WVS WE MCD. This is accomplished by combining both local and global context with deeper linguistic structures, relying heavily on lexical resources, sense-tagged corpora, and large text datasets. However, these resources are limited for SsaL. Additionally, the morphological complexity, agglutinative nature, and disjunctive orthography of SsaL present further obstacles to these methods.

Traditional Word Embeddings, such as Word2Vec (CBOW & SkipGram), GloVe, and FastText, are promising because of their non-reliance on annotated data. Nevertheless, their effectiveness is largely contingent upon access to vast, varied, and representative collections of unannotated texts, which presents a challenge for SsaL. Additionally, the agglutinative structure and intricate morphology of SsaL might complicate the precise representation of word associations and semantic relationships within these models. Context-aware word embeddings, for instance BERT, GPT, ELMo, T5, and ELECTRA, are capable of utilizing both annotated and non-annotated data, which makes them potentially viable for SsaL WE. Nonetheless, acquiring an adequate amount of high-quality annotated data to begin the process of generating dynamic vectors based on context can be challenging due to the scarcity of linguistic resources for SsaL. Table 3.2 provides a summary of the main features of the different WE approaches and methods.

Table 3.2: WE Approaches Key Features and Methods

Approach	Methods / Models	Key Features
Context-Aware Word Embedding (CAWE)	ELMo, BERT + BiLSTM + CNN, XLNet, DistilBERT, RoBERTa, GBCN, ALBERT, DeBERTa, PolyLM	Explicitly handles multiple word senses, models semantic disambiguation; suitable for WVS WE MCD; combines local and global context with deeper linguistic structure
Contextualized Word Embedding	BERT, GPT, ELMo, T5, ELECTRA	Dynamic vectors depending on context; encodes sentence-level semantics; attention-based or LSTM-based
Traditional Word Embedding	Word2Vec (CBOW & SkipGram), GloVe, FastText	Static vectors (context-independent); one vector per word; fails to handle polysemy

Due to the limited resources of SsaL, this research utilized a CAWE approach to construct an enhanced word embedding model for SsaL. Choosing this method specifically addresses various word meanings and handles semantic disambiguation, ideal for WVS WE MCD, by merging local and global contexts with advanced linguistic structures. This approach was then combined with existing lexical resources for SsaL, aiming to improve the challenges resulting from resource scarcity.

3.6 Relevant Studies on Sesotho sa Leboa (SsaL) within African NLP

Natural language processing research on African languages has expanded in recent years, but most African languages still remain under-resourced in terms of annotated corpora, benchmark datasets, and task-specific tools. As a Southern Bantu language, it shares typological features with other regional languages such as Setswana, which also exhibits significant morphological complexity and has been largely overlooked in NLP research despite its official status in several countries (Brown & Marivate, 2024). This complexity, coupled with its low resource availability, necessitates a focus on linguistically-informed NLP methodologies that can effectively capture its unique grammatical structures (Adebara & Abdul-Mageed, 2022). For instance, the presence of two distinct orthographies further compounds the resource scarcity for Sesotho, requiring the development of foundational language resources to enable even basic text readability metrics (Sibeko, 2023). These foundational resources are crucial for developing robust sentiment analysis models, as direct transfer from high-resource languages often proves insufficient given the distinct linguistic characteristics of Sesotho sa Leboa (Mabokela et al., 2025). This is particularly salient given the observed difficulties in sentiment analysis for morphologically rich African languages, where subtle changes in tone or affixation can significantly alter semantic meaning (Bhatia et al., 2023). Furthermore, Sesotho sa Leboa presents significant challenges in computational semantics due to high levels of lexical ambiguity, where numerous terms possess multiple meanings, complicating the identification of proper lexical categories and word senses (Masethe et al., 2025).

3.6.1 Research Background on Morphologically rich and Low-Resource African Languages

This situation is prevalent across many African languages, where efforts to develop robust natural language processing systems are frequently hampered by a lack of annotated data and specialized tools (Keinan & Hacothen-kerner, 2023; Mabokela et al., 2025). This deficit is often compounded by sociopolitical factors, such as national language policies prioritizing foreign languages, and an absence of standardized orthographies, further hindering the development of comprehensive NLP resources (Amol et al., 2024). Consequently, the reliance on pre-trained models developed for high-resource languages often yields suboptimal performance when applied to African languages due to significant linguistic divergences (Amol et al., 2024). This necessitates research into language-adaptive fine-tuning and the development of multilingual systems that can better generalize across under-represented languages, accounting for nuanced cultural and linguistic specificities (Amol et al., 2024). Specifically, addressing meaning conflation deficiency in morphologically rich, low-resource languages like Sesotho sa Leboa is critical, as polysemy and contextual ambiguity frequently degrade model precision in word sense disambiguation tasks (Masethe et al., 2025). For instance, efforts to apply sentiment analysis techniques developed for high-resource languages to African languages frequently encounter performance degradation due to these complexities (Muhammad et al., 2023; Sani, 2024). This is further complicated by challenges such as semantic misalignment and cultural variation in sentiment expression, which require specialized approaches beyond direct transfer of existing models (Lamin & Aziz, 2025). Such challenges highlight the need for comprehensive sentiment analysis frameworks specifically tailored for low-resource African languages, which not only account for linguistic intricacies but also pave the way for adaptable and robust models (Aliyu et al., 2024).

3.6.2 Situating Sesotho sa Leboa (SsaL) within African and Bantu Language Studies

The intricate agglutinative morphology and complex agreement systems necessitate advanced linguistic analysis to accurately segment and categorize lexical units, diverging significantly from

typical practices in Indo-European languages (Adebara & Abdul-Mageed, 2022). These characteristics mandate the development of specialized computational linguistic frameworks that can effectively model the rich internal structure of words and their interactions within a sentence, rather than relying on methodologies optimized for less morphologically rich languages (Aliyu et al., 2024). This is particularly critical as many machine translation and sentiment analysis systems, predominantly trained on high-resource languages like English and French, struggle to capture the cultural and linguistic nuances, including idioms and contextual cues, inherent in languages such as Sesotho sa Leboa (Malinga et al., 2024). This necessitates the creation of multilingual lexicons and specialized models capable of discerning language-specific sentiment scores and addressing mixed-language contexts to overcome the scarcity of accurately labeled datasets in these complex linguistic landscapes (Malinga et al., 2021; Malinga et al., 2021). Furthermore, the development of robust sentiment analysis models for Sesotho sa Leboa must account for its unique syntactic structures and the potential for dialectal variation, which can further complicate automated text processing (Margani & Ndugu, 2023).

3.6.3 Unique Morphological Challenges of Sesotho sa Leboa

This scarcity is particularly acute for Part-of-Speech tagging, where the development of accurate taggers is hindered by the limited availability of annotated datasets, in contrast to more resourced languages (Dibitso et al., 2022). This limitation often necessitates the use of semi-supervised or unsupervised learning techniques for morphological analysis, which can partially mitigate data scarcity but introduce additional complexities in validation and error propagation. The disjunctive orthography of Sesotho sa Leboa, where prefixes and other bound morphemes are often written separately from their stems, further complicates tokenization and lexical unit demarcation, making traditional word-based NLP approaches less effective (Simon et al., 2025). This contrasts with conjunctive orthographies and poses a significant hurdle for creating consistent tokenization schemes that preserve morphological integrity (Simon et al., 2025). For instance, differentiating between negation affixes and negation particles becomes challenging when tokens are treated uniformly without considering their morphological roles (Faaß, 2022).

3.6.4 Implications for Semantic Modelling and the Present Study

This is because standard byte-pair encoding and similar subword tokenization methods frequently fail to align with true morpheme boundaries, resulting in semantically incoherent or ambiguous token representations that hinder effective language model training (Jimoh & Nikolov, 2025). Consequently, the development of morphologically aware tokenization strategies becomes critical for improving the representational capacity and overall performance of language models in Sesotho sa Leboa (Jimoh & Nikolov, 2025; Jimoh & Nikolov, 2025; Shahid et al., 2025). Evidence from other morphologically rich, low-resource languages such as Slovak indicates that morphology-aware tokenization significantly enhances model performance across various NLP tasks, suggesting a similar benefit for Sesotho sa Leboa (Držík & Kapusta, 2026). This insight highlights the necessity of developing tokenization approaches that prioritize linguistic fidelity, optimizing segmentation to preserve meaningful units like roots and valid morphemes, thereby enhancing model effectiveness for low-resource languages (Bayram et al., 2025; S et al., 2022). Such approaches can mitigate the "token tax" by reducing the overall number of tokens required to represent linguistic information, thereby improving computational efficiency and semantic integrity for downstream tasks (Lundin et al., 2025). This is critical because existing tokenization approaches, often based purely on frequency statistics, tend to ignore the underlying semantic structure of text, leading to suboptimal encoding for morphologically rich languages (Lundin et al., 2025). Therefore, developing tokenization schemes that are sensitive to the morphological structure of Sesotho sa Leboa, perhaps through hybrid methods combining statistical and linguistic rules, is paramount for advancing NLP capabilities in this domain (Bayram et al., 2025). This necessitates a shift from purely data-driven tokenization to linguistically informed approaches that can better capture the intricate morphological structures of Sesotho sa Leboa (Arnett & Bergen, 2024). Indeed, morphologically aligned tokenization has been shown to improve performance for natural language understanding tasks, particularly when dealing with the compositional morphology of agglutinative languages (Bayram et al., 2025; Maksymenko & Turuta, 2025).

3.7 Systematic Literature Review (SLR)

An SLR constitutes a thorough and transparent approach to discovering, assessing, and synthesizing research evidence pertaining to a particular research question or domain. SLR's main goals start with determining the current state by recognizing the work that has been accomplished in the area, which includes existing models, methodologies, and datasets (Verner et al., 2012). Uncovering research gaps mean identifying shortcomings, underexplored domains, or methodological inconsistencies in prior studies (Snyder, 2019). Justifying research contributions includes illustrating the importance of the new research by contrasting it with previous works to underscore its novelty. Steering theoretical and experimental frameworks means aiding in methodological decisions through the critical synthesis of earlier empirical findings. The SLR in this study centres on the MCD problem in WVS WE models, thereby laying a comprehensive empirical and theoretical groundwork for research concerning SsaL. The following sub-questions shaped the investigation:

RQ1: What are the salient features of a morphologically rich low resource language, such as SsaL, should be considered in developing a CAWE model that effectively addresses the MCD problem in WVS WE models?

RQ2: What are the existing CAWE models that effectively address the MCD problem in WVS WE models?

3.7.1 Research Strategy

A systematic literature review was conducted across five major electronic databases: IEEE Xplore, SpringerLink, Web of Science, Google Scholar, and Scopus as shown on Table 3.3. This multi-database approach was designed to ensure a broad and unbiased evidence base. The search strategy utilized predefined keywords combined with Boolean operators, including terms such as “meaning conflation deficiency”, “context-aware word embedding”, “sense conflation”, “context conflation”, “entity conflation”, and “contextual word embedding”. Where available, field-specific tags and filters were applied within each database to increase search precision and relevance. The search was carefully

structured to enhance sensitivity, capturing both including studies not typically indexed in conventional repositories, to ensure comprehensive peer-reviewed and grey literature, coverage of the topic.

Table 3.3: Overview of the results from the database search for WE Models

Database	No. of Results	Search Phrase Used	Remarks
Scopus	269	“Meaning conflation deficiency” OR “context-aware word embedding” OR “sense conflation” OR “context conflation” OR “entity conflation” OR “contextual word embedding”	Broad coverage across disciplines; includes peer-reviewed journals and conference proceedings.
SpringerLink	10	Same as above	Primarily focused on computer science, engineering, and applied technology domains.
IEEE Xplore	35	Same as above	Rich source for engineering and technology-related publications.
Google Scholar	56	Same as above	Offers quick access to a wide variety of academic and grey literature.

Recognizing limited coverage of WVS WE MCD, a systematic search was undertaken across numerous databases and additional sources. As shown in Figure 3.2, the process yielded 403 records, of which 25 were retained for quantitative synthesis after rigorous screening. The study selection is depicted in a PRISMA flow diagram in Figure 3.2.

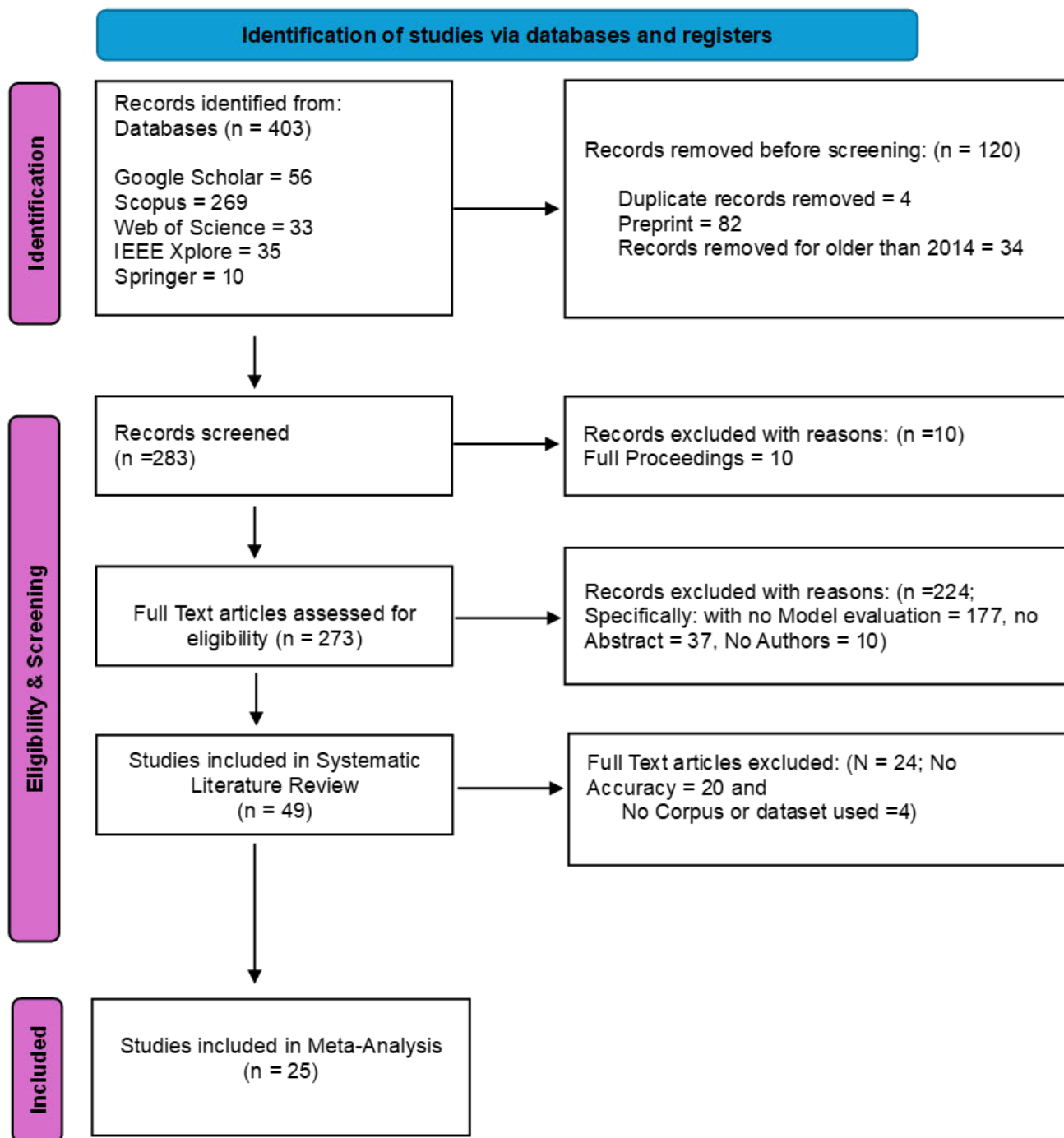


Figure 3.2: PRISMA flow diagram for the CAWE research (Hl'adek et al., 2016)

3.7.2 Inclusion criteria and exclusion criteria

The parameters for what to include and exclude are crucial for determining how relevant word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) research is to a review or selection of publications. Study type affected paper selection criteria. Even were the phrases carefully selected for a specific purpose, they did not clearly state what was included. Searching for hundreds of related articles was beneficial in the first phase. Buzzwords appeared in source titles, summaries,

content, and names. Table 3.4 shows the main phrases and extensive suggestions for what to include and what not to include. To ensure comprehensive coverage, works from multiple languages were included in the word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) project. SsaL, Amharic, and Xhosa were classified as low-resource languages, while English and Chinese represented high-resource languages. Research emphasis was placed on Finnish, Turkish, and Bantu due to their complex morphological structures.

Static word embedding models like Word2Vec, GloVe, and FastText generate fixed representations that cannot distinguish between different meanings of a polysemous word, resulting in WVS WE MCD. To assess whether word representations can adapt based on contextual usage, transformer-driven models—such as ELMo, BERT, mBERT, XLM-R, and other CAWE approaches—are examined for their effectiveness in mitigating this limitation. This research study explores the integration of static and CAWE, producing “hybrid models” that shed light on how both methods together may diminish MCD. Research efforts focus on tackling the central MCD challenge in WVS WE models, with WSD strategies considered suitable because correctly identifying a word sense addresses MCD.

To ensure validity and academic quality, only peer-reviewed journal or conference paper research were included. F1 score, accuracy, recall, and other word sense disambiguation task-specific metrics such as the SemEval WSD test may assess a model’s multi-class discrimination (MCD) performance. Studies on transformer-based embeddings and word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) during the past five to seven years demonstrate improvement. Studies focusing on high-resource languages that do not provide transferable methods or insights for low-resource settings may be excluded. Research without defined assessment criteria was discarded. This excludes unproven ideas and procedures, emphasizing those that have worked in practice.

Assessing entire WVS WE MCD papers demands careful consideration of what is either included or excluded. The analysis features the current state of the field, focusing on embedding-based techniques, languages with scarce resources, semantically rich languages, and well-defined success benchmarks. Meanwhile, studies lacking empirical backing, involving high-resource languages, or unrelated to WVS WE MCD are omitted, ensuring that the review remains thorough and enlightening.

Table 3.4: Inclusion and Exclusion Criteria

Criteria	Decision
Keywords present in the title, abstract, or keywords section	Inclusion
Publications dated from 2014 onward	Inclusion
English-language research articles	Inclusion
Research focusing on WE, CWE, or CAWE approaches	Exclusion
Articles without accuracy metrics	Exclusion
Articles lacking a dataset or corpus	Exclusion
Non-English articles; pre-2014 publications	Exclusion

3.7.3 Data integration and statistical evaluation

The information was input into spreadsheet software for statistical examination. Subsequently, the data was imported into STATA and RStudio. During data collection, all primary study effect sizes were compiled and organized to obtain the results. This research employs the random-effects model (Thompson et al., 2022). Critical reading and appraisal of the literature established the conceptual grounding for the research question. Records were manually verified for accuracy; studies that did not align with the review objectives were excluded; additional eligible studies were identified through hand-searching reference lists and citations. Following iterative screening and deduplication, 403 records were retained for evaluation. To enable quantitative comparison across studies, effect sizes were calculated to capture the magnitude of reported outcomes, thereby strengthening the meta-analytic synthesis. These metrics provide a basis for assessing the extent to which embedding-based

approaches address WVS WE MCD. Equation (3.6) calculates Cohen's d to compare two means (Bowring *et al.*, 2021):

$$d = \frac{M_1 - M_2}{SD_{pooled}} \quad (3.6)$$

where M_1 and M_2 are means of the groups and SD_{pooled} is the pooled standard deviation.

3.7.4 Publication analysis and results

After searching Google Scholar, IEEE Xplore, Springer Link, Web of Science, and Scopus, 403 articles were discovered. One hundred and twenty items were discarded for similar content and others for not being in English. Two hundred and forty-eight studies were discarded due to incomplete model ratings, authors, or names. Even the untitled ten proceedings were dismissed. A thorough meta-analysis of 25 papers was possible.

3.8 Meta-Analysis

Figure 3.3 (Forest plot) depicts individual-study effect sizes with 95% confidence intervals (CIs) and the pooled estimate, which is central to meta-analytic interpretation. Each study is shown as a square whose area reflects its inverse-variance weight, with a horizontal line indicating the 95% CI; shorter lines denote greater precision. Below, the diamond summarizes the overall effect size, and its width represents the 95% CI of the pooled estimate. The heterogeneity statistics— $\tau^2 = 8.8724$, $I^2 = 87.65\%$, and $H^2 = 8.10$ —indicate substantial between-study variability that exceeds what would be expected from sampling error alone; in other words, effect sizes differ meaningfully across studies, likely due to variations in design, populations, measurement of word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD), or analytic choices. Consequently, the pooled effect should be interpreted with caution and alongside appropriate random-effects modelling and moderator analyses. The heterogeneity statistics indicate that much of the between-study variability is not attributable to sampling error. Specifically, $I^2 = 87.65\%$ implies that most of the observed dispersion in effect sizes reflects genuine differences across studies rather than random noise, while $H^2 = 8.10$ —the ratio of total

to within-study variance—suggests that variability is roughly eightfold what would be expected under homogeneity. Together, these indices point to substantial heterogeneity in the evidence base.

Several structural factors plausibly contribute to this heterogeneity. Research on low-resource languages often lacks large, well-annotated corpora and mature pre-trained models, which can reduce model stability and generalizability relative to high-resource settings. Morphologically rich languages (e.g., Finnish, Turkish) introduce extensive inflectional and derivational variation, complicating representation learning and potentially exacerbating meaning conflation. Data quality—especially the availability of word sense disambiguation (WSD) annotations—also matters: well-curated, sense-tagged resources (e.g., SemCor for English) typically yield better sense separation; whereas unannotated or noisy corpora hinder reliable disambiguation.

Corpus size, diversity, and cleanliness further condition outcomes. Larger and more heterogeneous training sets allow models to learn more nuanced lexical semantics and tend to reduce meaning conflation, whereas small, domain-bound, or noisy datasets—common in low-resource contexts—produce less consistent effect estimates. Beyond data factors, methodological choices (static vs contextual-aware embeddings), training objectives, fine-tuning strategies, and model capacity all introduce variation in reported effects.

Given the high τ^2 , I^2 , and H^2 , pooled estimates should be interpreted cautiously. Future syntheses should employ random-effects models, conduct moderator and sensitivity analyses (e.g., by language resource level, morphology, annotation quality, model class), and, where possible, adopt standardized protocols and shared benchmarks to improve comparability across studies.

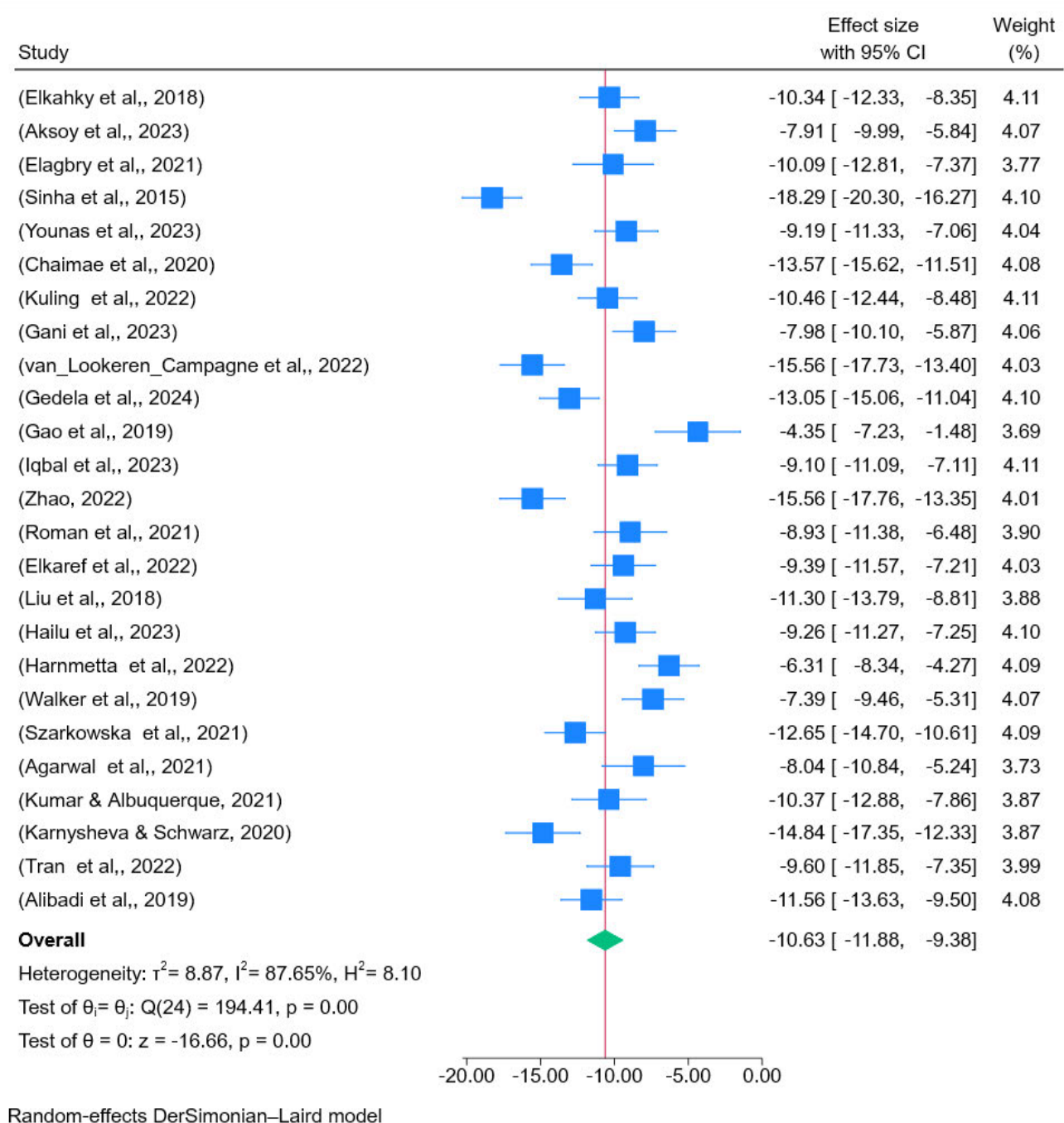


Figure 3.3: Forest plot for distribution of effect size of META summary of CAWE in Table 3-4.

Figure 3.4 reports subgroup heterogeneity statistics. For language modelling, the estimates ($\tau^2 = 4.69$, $I^2 = 78.92\%$, $H^2 = 4.74$) indicate substantial between-study variability beyond sampling error; although τ^2 is lower than in some other subgroups, the high I^2 confirms that effect sizes differ meaningfully across studies, possibly due to differences in modelling choices, corpora, and evaluation criteria. Machine translation exhibits the greatest dispersion ($\tau^2 = 28.41$, $I^2 = 94.55\%$, $H^2 = 18.34$), implying that most observed variation stems from true heterogeneity—likely reflecting variation in language

pairs, training data quality, and translation architectures. By contrast, both Named Entity Recognition (NER) and Question Answering (QA) show $\tau^2 = 0.00$, $I^2 = 0.00\%$, $H^2 = 0.00$, suggesting highly consistent effects across studies, which may be attributable to common benchmarks and comparatively standardized protocols. Sentiment analysis (mood analysis) displays pronounced heterogeneity ($\tau^2 = 18.44$, $I^2 = 93.44\%$, $H^2 = 15.22$), consistent with known variability in datasets (domains, label schemes), pre-processing, and modelling approaches. Finally, text classification shows moderate heterogeneity ($\tau^2 = 3.09$, $I^2 = 72.53\%$, $H^2 = 3.64$), which may arise from differences in task formulations, corpus size and diversity, and methodological choices. Collectively, these results justify random-effects meta-analysis and motivate moderator analyses (e.g., data source, architecture, evaluation metric) to explain subgroup differences; while the homogeneity observed in NER and QA suggests robust, reproducible effects under prevailing evaluation standards.

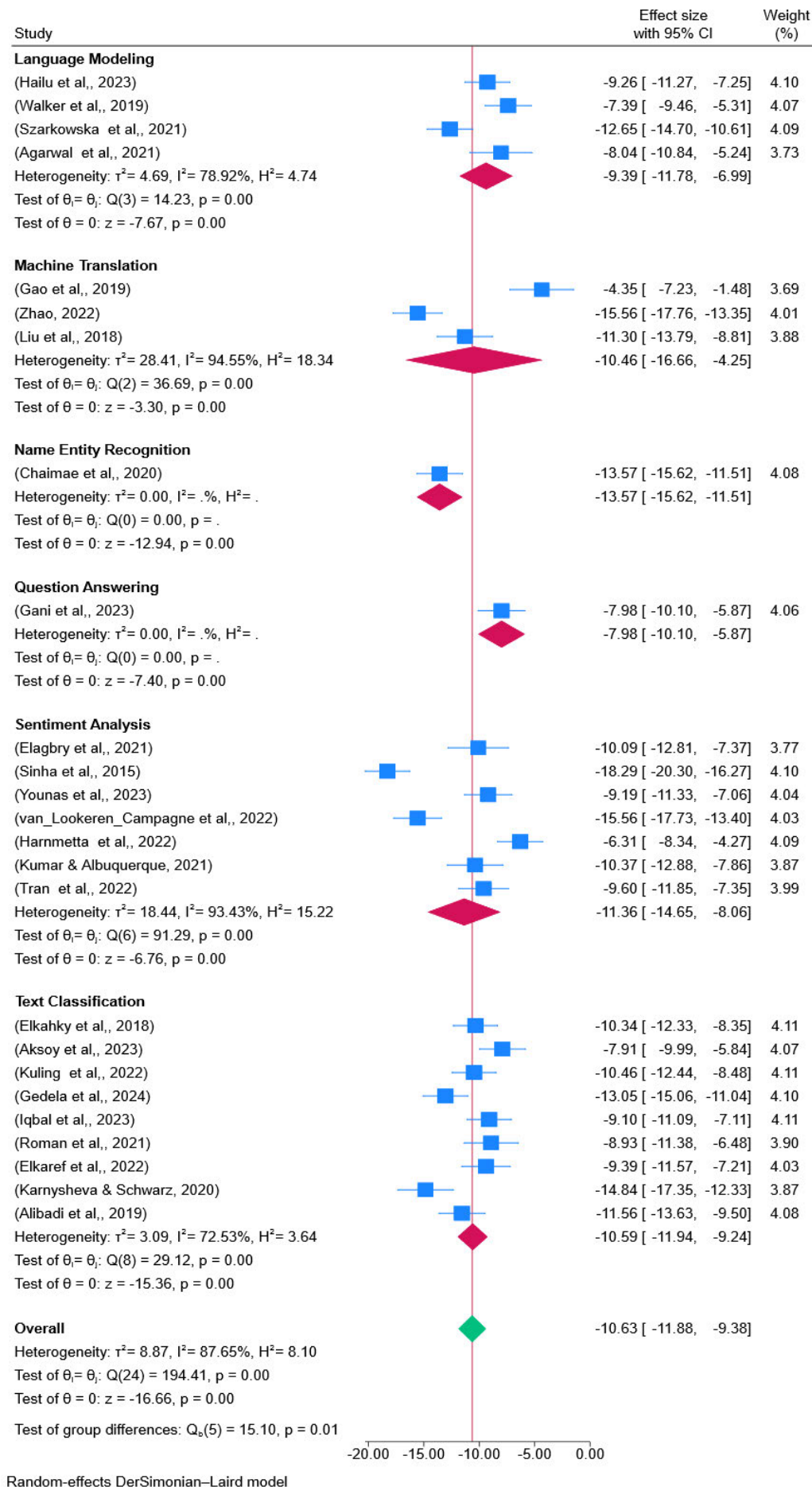


Figure 3.4: Forest plot with subgroups for distribution of effect size of META

Figure 3.5 presents a Galbraith plot, also known as a standardized effect size plot. This visualization method is frequently used in meta-analysis to assess how effect sizes correspond with the precision of the individual studies. It is particularly useful for identifying potential outlier studies or those that may have a disproportionate influence on the overall meta-analytic results. In this context, studies with smaller standard errors are generally considered to offer more reliable effect size estimates. The plot includes 95% confidence interval lines, which serve as visual guidelines to help detect studies whose effect sizes deviate significantly from the overall pooled estimate.

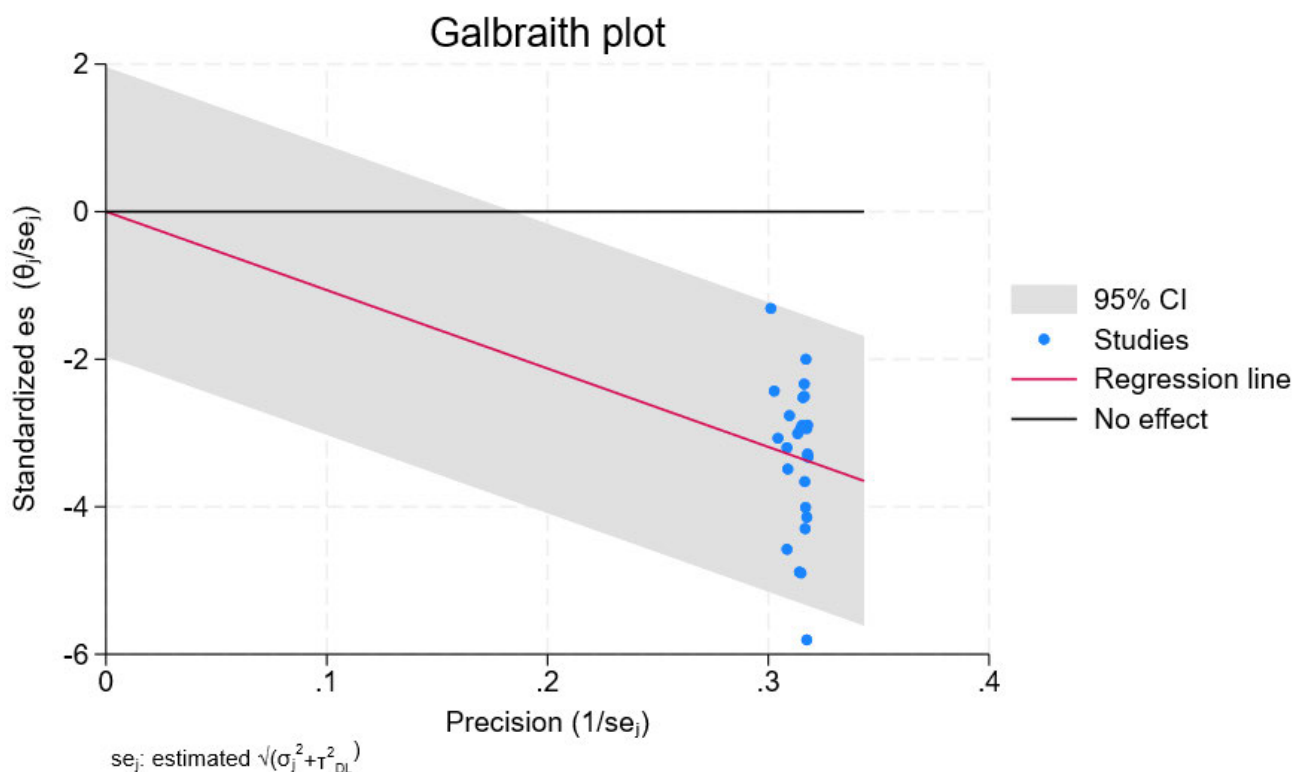


Figure 3.5: Galbraith plot

3.8.1 Publication bias and meta-regression

The available evidence base appears vulnerable to publication bias, with studies reporting positive or “successful” outcomes more likely to be accessible for evaluation. This bias warrants urgent attention in the context of WVS WE MCD, in which phenomena can be methodologically complex and linguistically sensitive. Although modern language models can represent polysemy, progress in low-resource and/or morphologically rich languages (e.g., SsaL) is constrained by limited annotated data

and the scarcity of strong pre-trained models, fostering a systematic preference for resource-rich languages and, consequently, fewer published investigations of WVS WE MCD in under resourced settings.

In the meta-analytic component, meta-regression was used to assess whether a specified moderator (referred to in the original as a “filter/monitor”) accounts for variation in effect sizes. The chi-square statistic for the moderator was $\chi^2 = 169.11$ (Table 3.5), indicating that the moderator explains a substantial proportion of between-study variability in effect sizes; a second model yielded $\chi^2 = 128.71$ (Table 3.6), consistent with a weaker—but still meaningful—association between the moderator and outcomes. The funnel plot (Figure 3.6), which graphs effect sizes against their standard errors, provides a visual check for small-study effects/publication bias; noticeable departures from the expected funnel shape, together with the large chi-square values, suggest that observed patterns are unlikely to be explained by random sampling error alone. Collectively, these results underscore the need to address publication bias, to broaden coverage of low-resource languages, and to interpret pooled estimates with caution by incorporating moderator analyses and sensitivity checks.

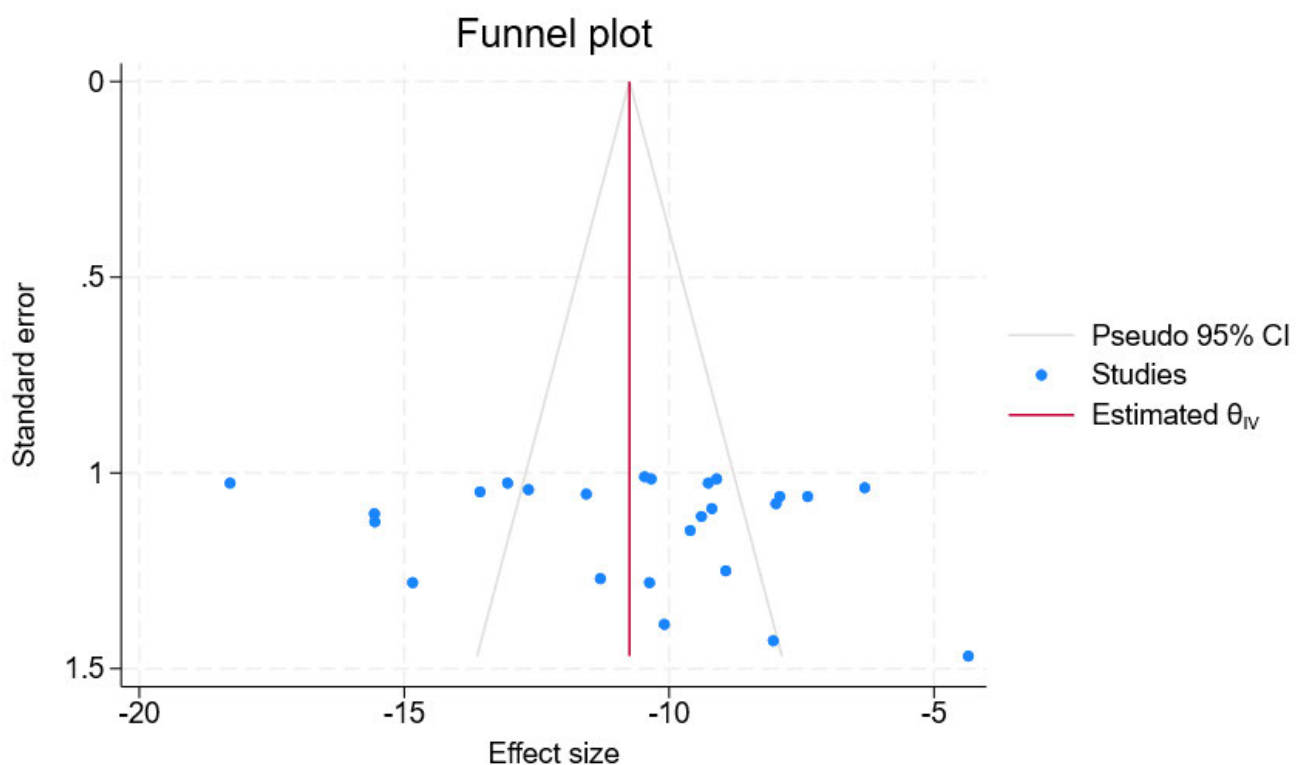


Figure 3.6: Funnel plot

Table 3.5: Meta-Regression per Publication Year

Parameter	Coefficient	Std. Error	Z-Value	p-Value	95% Confidence Interval
Year	0.4624	0.3034	1.52	0.128	[-0.1323, 1.0571]
Constant	945.1428	613.2503	1.54	0.123	[-1106.102, 256.8058]

Residual Heterogeneity Test: Q_{res} (χ^2 , $df = 23$) = 169.11, Prob > Q_{res} = 0.0000.

In meta-regression, a bubble plot visualizes the association between study-level effect sizes and a moderator (or covariate): each bubble represents a study, positioned by the moderator value and effect size, with bubble area typically proportional to inverse-variance weight or sample size. As shown in Figures 3.7–3.8, the bubble plots exhibit a clear pattern, and the omnibus chi-square test for moderators ($\chi^2 = 128.71$) indicates a statistically significant relationship between the specified moderator (“filter”) and the effect sizes. The magnitude of this statistic implies that the observed dispersion departs materially from what would be expected under the null hypothesis of no moderator effect; thus, the null can be rejected at conventional levels, and the moderator contributes meaningfully to between-study heterogeneity rather than reflecting random variation alone.

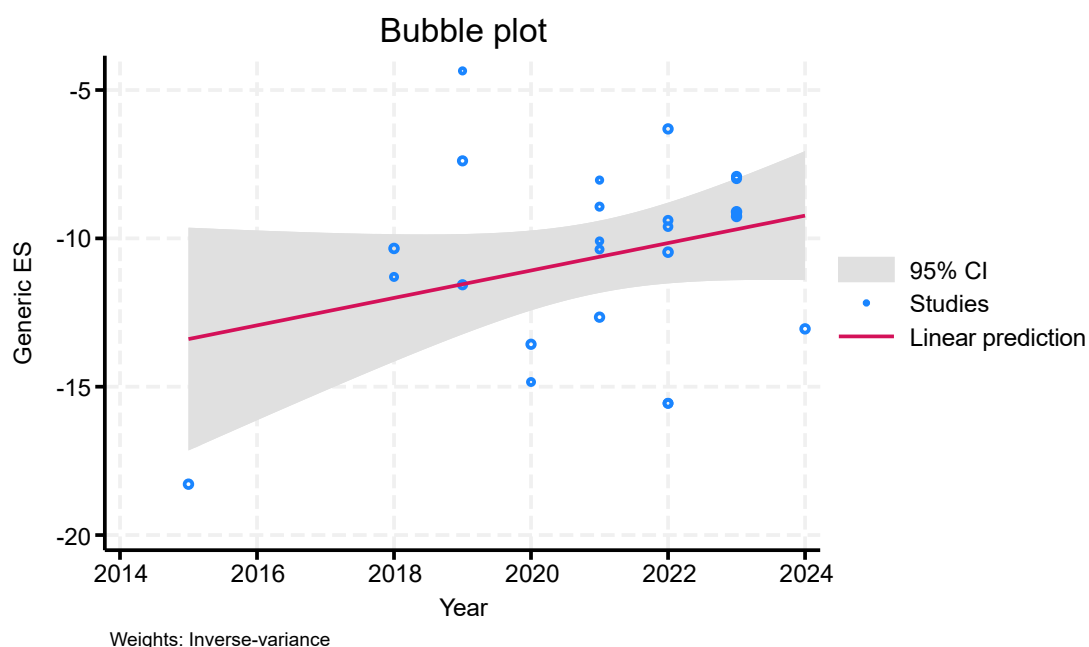


Figure 3.7: Bubble Plot per Publication year

Table 3.6: Meta-Regression per Dataset

Parameter	Coefficient	Std. Error	Z-Value	p-Value	95% Confidence Interval
Corpus Size	-1.05×10^{-7}	3.29×10^{-8}	-3.20	0.001	$[-1.69 \times 10^{-7}, -4.07 \times 10^{-8}]$
Constant	10.2262	0.5565	18.37	<0.0001	[9.1354, 11.317]

Residual Heterogeneity Test: Q_{res} (Chi^2 , $\text{df} = 23$) = 128.71 $p\text{-value} < 0.0001$.

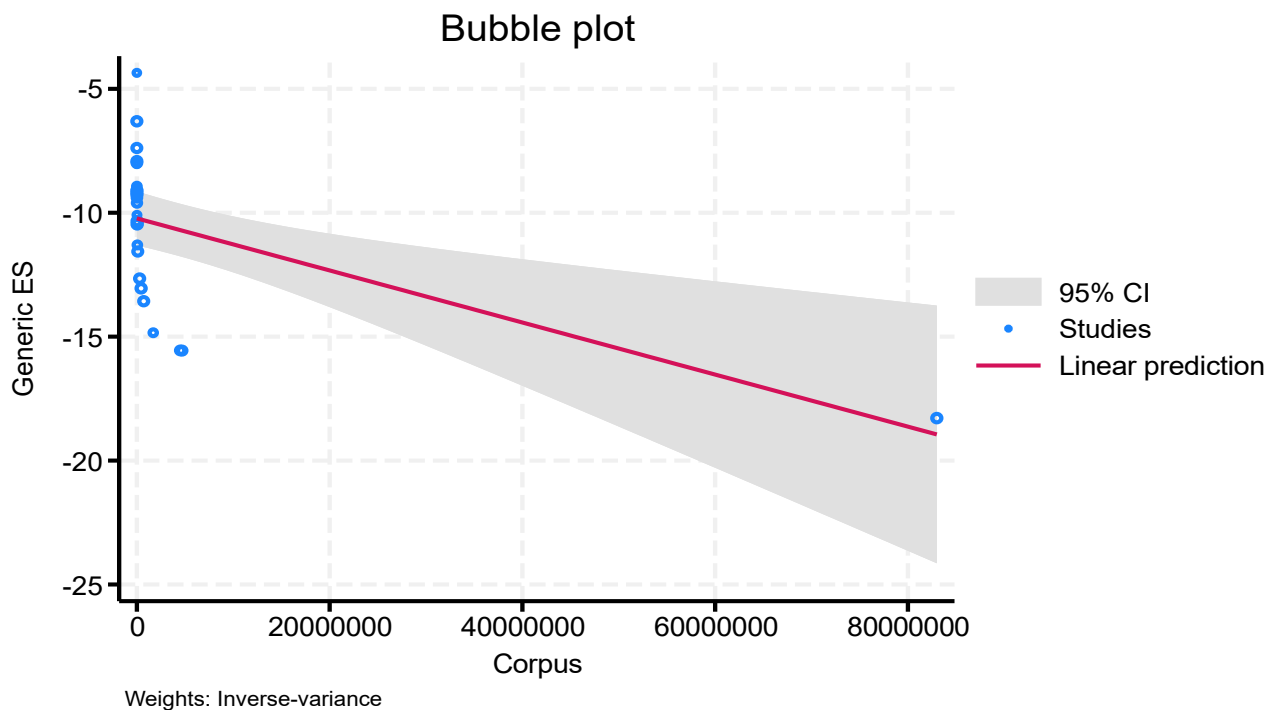


Figure 3.8: Bubble plot for dataset

Table 3.8 reports a trim-and-fill analysis for assessing potential publication bias by estimating the pooled effect after imputing putatively missing studies. The bias-adjusted point estimate is $\hat{\beta} = 5.95$ with $\text{SE} = 4.767$, yielding $z = 1.25$ and $p = 0.20$. Despite the relatively large point estimate, the standard error is substantial, producing a wide 95% confidence interval (approximately -3.40 to 15.30) that includes zero; consequently, the effect is not statistically significant at conventional thresholds. Taken together, these results suggest that, once small-study effects are accounted for, the overall effect size remains uncertain and cannot be distinguished from no effect. The lack of significance—despite a sizeable point estimate—may reflect high between-study heterogeneity, limited sample size, or both.

Table 3.7: Trim and Fill Meta-Analysis Results.

Studies	Effect Size (Coefficient)	95% Confidence Interval
Observed (n = 28)	10.747	[10.310, 11.185]
Observed + Imputed (25 + 3)	10.747	[10.310, 11.185]

H0: $\beta_1 = 0$; no small-study effects: Coefficient (β_1) = 5.94 Std. Error = 4.767 z-value = 1.25 p-value 0.2124.

3.9 Bibliometric Analysis for Word Embeddings

3.9.1 Bibliometric analysis approach

Bibliometric analysis is a quantitative research method employed to assess the impact, trends, and interrelationships in scholarly literature by analyzing data such as citations, authorship, keywords, and co-authorship networks. This method helps in understanding the development of a research area, identifying key publications, and uncovering emerging trends (Donthu et al., 2021; Hassan and Eliene, 2024; Passas, 2024). Bibliometric analysis can offer valuable insights into the progression of research on word embedding models. Through trend analysis, researchers can determine how interest in word embeddings (such as Word2Vec, GloVe, FastText, BERT, etc.) has evolved, identifying which models have become more prominent over time (Alsharif, Salleh and Baharun, 2020; Pradana, Elisa and Syarifuddin, 2023).

Bibliometric analysis in Figure 3.9 serves as a systematic and robust approach to the exploration and examination of extensive scientific data sets. This allows this work to analyse the evolutionary intricacies of a word embeddings (WE models), while highlighting the developing areas within its field. However, its application in research is relatively new, and in many instances, still developing (Donthu *et al.*, 2021; Passas, 2024). Accordingly, our study endeavours to present an overview of the bibliometric methodology, with a particular focus on its differing techniques, while offering step-by-step procedures that can be relied upon to rigorously perform bibliometric analysis with confidence (Alsharif, Salleh and Baharun, 2020; Hassan and Eliene, 2024). This study clarifies the appropriate contexts and methodologies for employing bibliometric analysis in comparison to related techniques, including meta-analysis and systematic literature reviews. This study serves as a valuable resource for

understanding the techniques and procedures applicable to bibliometric analysis research (Donthu *et al.*, 2021; Hassan and Eliene, 2024; Passas, 2024).

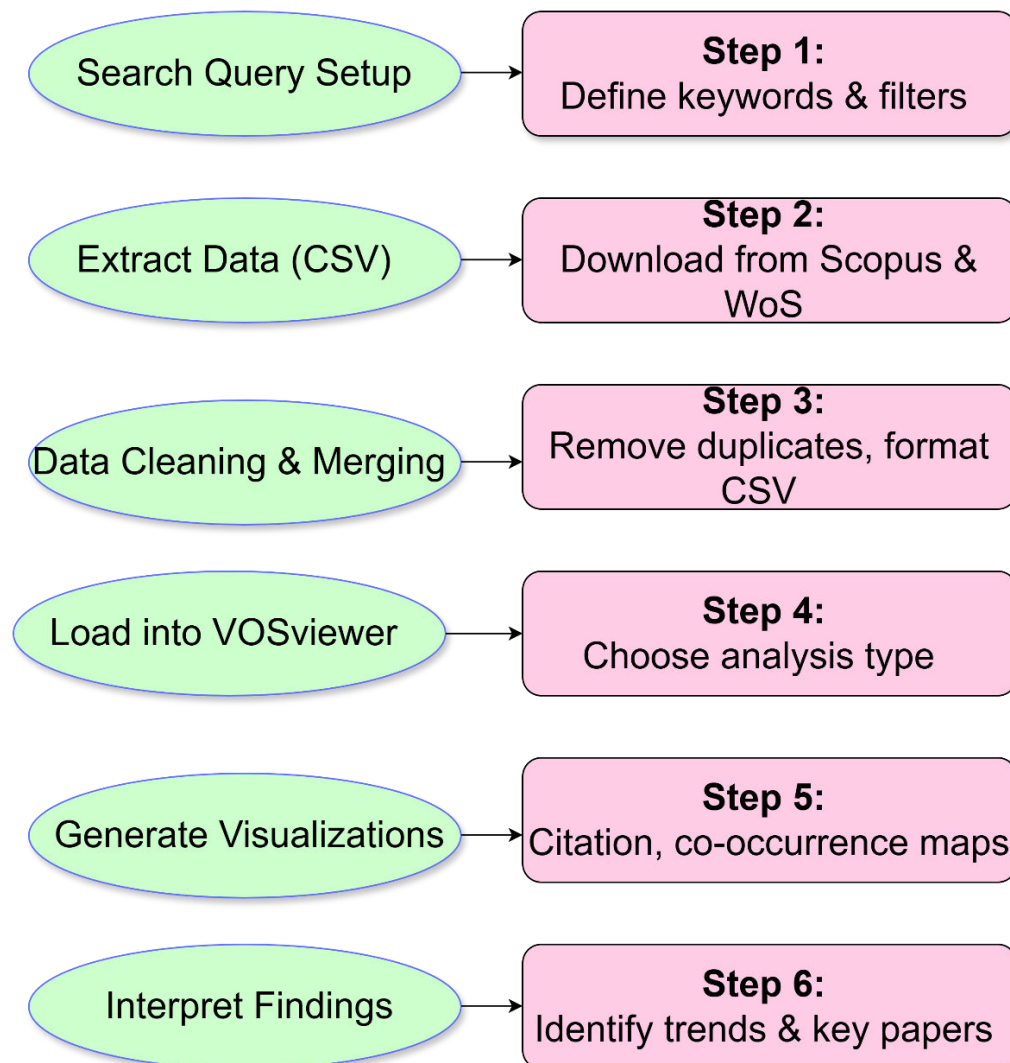


Figure 3.9: Bibliometric analysis framework

Citation Analysis — This identifies the most frequently referenced publications in word embedding research, highlighting seminal works and key contributions.

Co-Authorship and Collaborative Networks – Charting connections across researchers, institutions, and nations to pinpoint key contributions and collaborative initiatives in integrating research.

Keyword Co-Occurrence Analysis — Identifying often used phrases and their correlations to reveal new

subjects, such as context-aware embeddings, multilingual embeddings, or embeddings for specific domains (e.g., healthcare or legal texts).

Impact of Journals and Conferences - Assessing the primary publication outlets for research on embeddings to identify the foremost venues for breakthroughs in NLP.

Thematic Evolution — Employing VOSviewer, a bibliometric mapping tool, to illustrate the temporal development of subfields inside word embeddings (e.g., static vs context-aware embeddings).

Bibliometric analysis enables researchers to gain data-driven comprehension of the influence and trajectory of word embeddings research, pinpointing deficiencies in the literature, and making judicious pronouncements on prospective research domains (Donthu *et al.*, 2021; Pradana, Elisa and Syarifuddin, 2023; Passas, 2024).

3.9.2 Country co-citation analysis

The study employs co-citation analysis in word embeddings research to identify prominent countries, authors, and research hubs in the field. This research further reveals collaborative networks influencing the future of word embeddings for NLP, uncovering emerging trends and significant challenges in the domain.

In Figures 3.10 and 3.11, country co-citation analysis examines the collaboration and effect of various countries in word-embedding research. The goal is to determine which nations contribute the most to word embeddings, sense embeddings, and contextual embeddings research, indicating how nations interact in the field of natural language processing research and in regional effect and research domination in certain subfields (for example, the United States leads in transformer-based embeddings and China in multilingual embeddings). Should articles from China and the United States be regularly mentioned together, such indicates close collaboration or comparable research interests in word embeddings.

Selected	Country	Documents	Citations	Total link strength ✓
☒	united states	49	1400	16
☒	united kingdom	19	509	15
☒	india	32	178	7
☒	china	29	739	6
☒	saudi arabia	5	264	6
☒	australia	8	240	4
☒	brazil	8	101	4
☒	germany	14	87	3
☒	iran	5	264	3
☒	italy	9	71	3
☒	malaysia	5	24	3
☒	switzerland	5	19	3
☒	france	17	110	2
☒	pakistan	5	38	2
☒	canada	13	305	1
☒	morocco	8	13	1
☒	south korea	7	35	1
☒	spain	9	11	1
☒	turkey	6	9	1
☒	netherlands	8	49	0

Figure 3.10: The leading 20 co-country network on word embedding



Figure 3.11: Co-country network

3.9.3 Co-authorship and collaborative networks

Author co-citation analysis examines the authors most mentioned in conjunction within word embedding research. Such analysis identifies pioneers and prominent researchers, such as Mikolov et al. for Word2Vec, Pennington et al. for GloVe, and Devlin et al. for BERT. This elucidates how research groups or institutions influence various facets of word embeddings, and aids in identifying significant research clusters (e.g., one group focusing on static embeddings, another on contextual embeddings); the frequent co-citation of Mikolov and Devlin indicates a robust connection between traditional word embeddings (Word2Vec) and contemporary transformer-based models (BERT).

Using citation metrics, Figures 3.12 and 3.13 illustrate a bibliometric analysis that maps influential authors within the domain of network visualization. By analyzing citation data, the study identifies key researchers, collaboration patterns, and influential clusters in the field. In the visual network, each node represents an author, while the connecting lines (edges) indicate citation links between them.

Selected	Author	Citations	Total link strength ▼
☒	mikolov t.	235	8749
☒	lee k.	227	7837
☒	dean j.	139	5262
☒	socher r.	126	5197
☒	chen k.	131	4996
☒	zettlemoyer l.	114	4908
☒	toutanova k.	152	4808
☒	manning c.d.	121	4781
☒	devlin j.	151	4579
☒	sutskever i.	114	4510
☒	neumann m.	84	3528
☒	peters m.e.	86	3328
☒	gardner m.	81	3323
☒	corrado g.	85	3253
☒	pennington j.	85	3193
☒	grave e.	82	3152
☒	liu y.	81	3101
☒	iyer m.	71	2975
☒	joulin a.	77	2899
☒	clark c.	67	2871

Figure 3.12: The leading 20 co-authorship and collaborative networks for word embedding

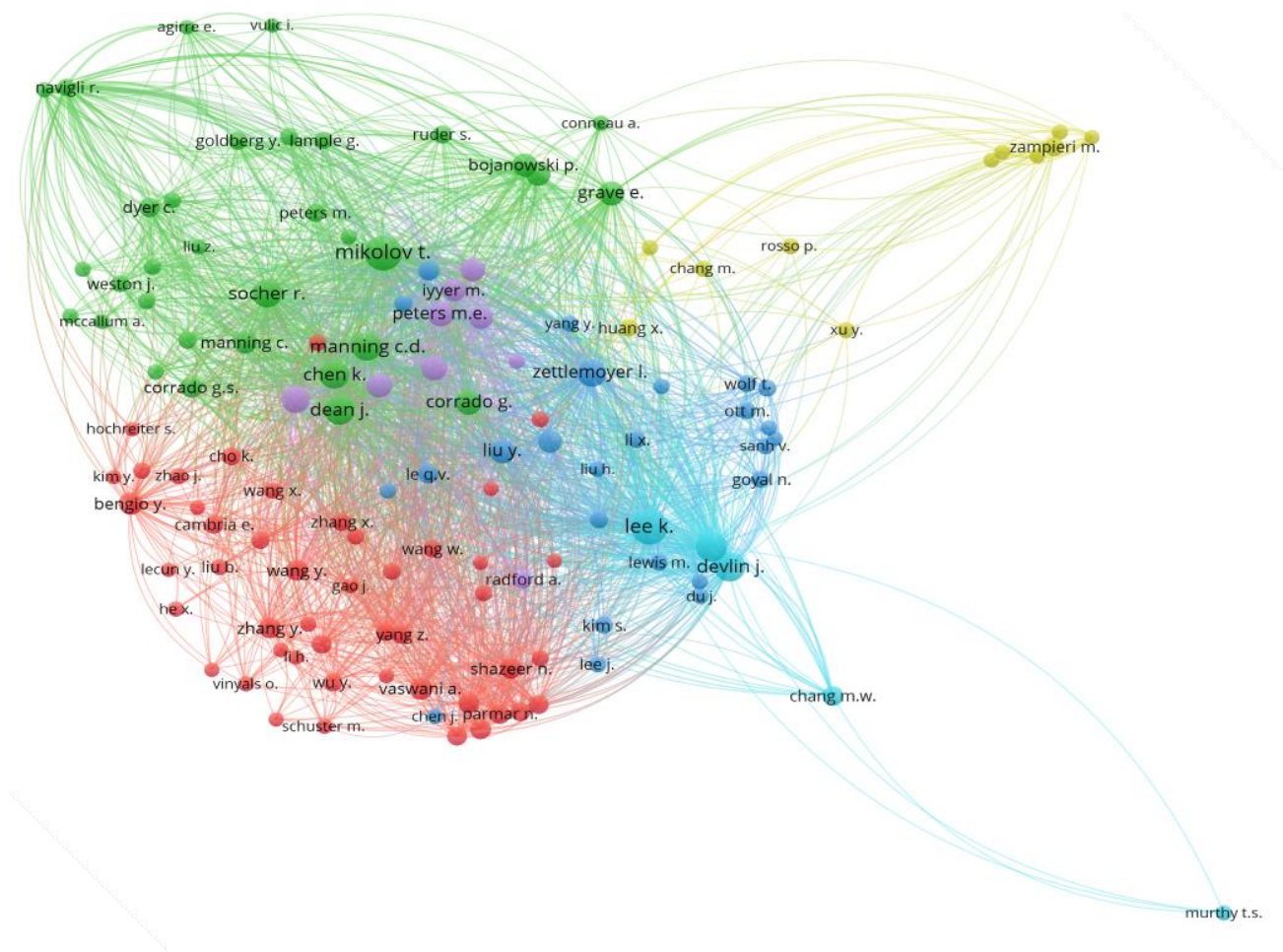


Figure 3.13: Co-authorship and collaboration network

3.9.4 Keyword co-occurrence analysis

Keywords co-citation analysis in Figures 3.14 and 3.15 investigates the correlations between keywords that are commonly referenced together in word embedding articles, as well as identifying research trends and emerging themes in NLP. This emphasizes developing concepts (such as "multilingual embeddings," "self-supervised learning," and "few-shot learning"). It draws linkages between concepts (for example, word embeddings with adversarial learning, interpretability, and ethical AI). For example, if the terms "BERT" and "domain adaptation" are frequently used together, this implies a rising interest in using contextual embeddings for domain-specific activities.

Selected	Keyword	Occurrences	Total link strength ✓
☒	natural language processing	39	45
☒	bert	33	37
☒	word embeddings	21	30
☒	deep learning	21	27
☒	contextual word embeddings	29	18
☒	machine learning	11	18
☒	sentiment analysis	9	18
☒	fasttext	5	14
☒	word2vec	5	14
☒	transformer	9	12
☒	word embedding	15	12
☒	data augmentation	6	11
☒	neural networks	6	11
☒	transfer learning	8	11
☒	classification	5	10
☒	contextual word embedding	14	10
☒	nlp	7	8
☒	text classification	10	8
☒	named entity recognition	7	6
☒	deep neural network	6	5

Figure 3.14: The leading 20 author keywords for documents on word embedding

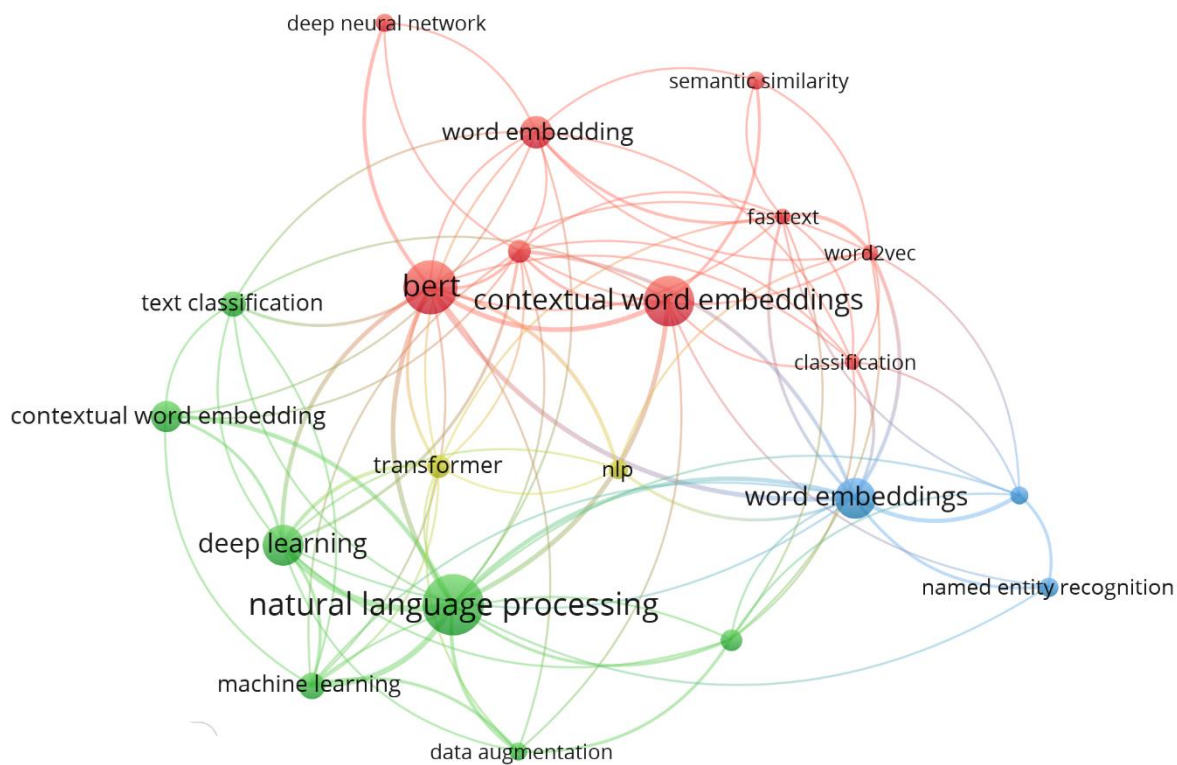


Figure 3.15: Keyword co-occurrence network

3.10 Descriptive Statistics of Primary Studies

A comprehensive analysis was conducted on a series of papers addressing the MCD issue in WVS WE models, covering the period from 2014 to early 2024. The data for this analysis was sourced from scholarly journals, articles, and conference proceedings available in databases like Web of Science, Scopus, Google Scholar, Springer, and IEEE Xplore. The reviewed studies provided comprehensive information, including publication dates, accuracy metrics, the use of WVS WE MCD approaches, traditional versus contextual embeddings, and context-aware word embedding methods. As shown in Figure 3.16, publication activity between 2015 and 2024 reveals a clear upward trend. Interest in WVS WE MCD research notably increased, with more than 100 papers published between 2018 and 2022, despite occasional drops in certain years. Figure 3.17 further demonstrates a rise in publications specifically within the journal *ACM Transactions on Asian and Low-Resource Language Information Processing*, highlighting growing scholarly attention in this field. The analysis data was collected up to early 2024, implying that the last months of 2024 might experience a rise in journal article publications.

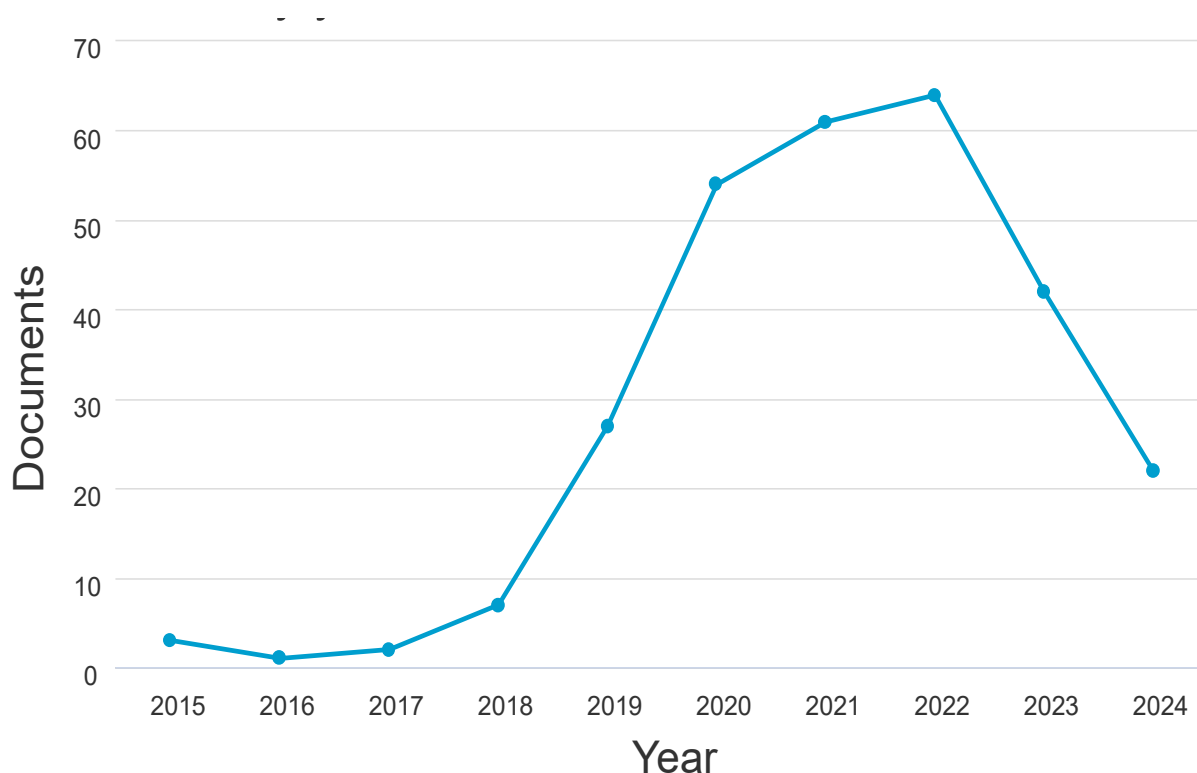


Figure 3.16: Publications by year

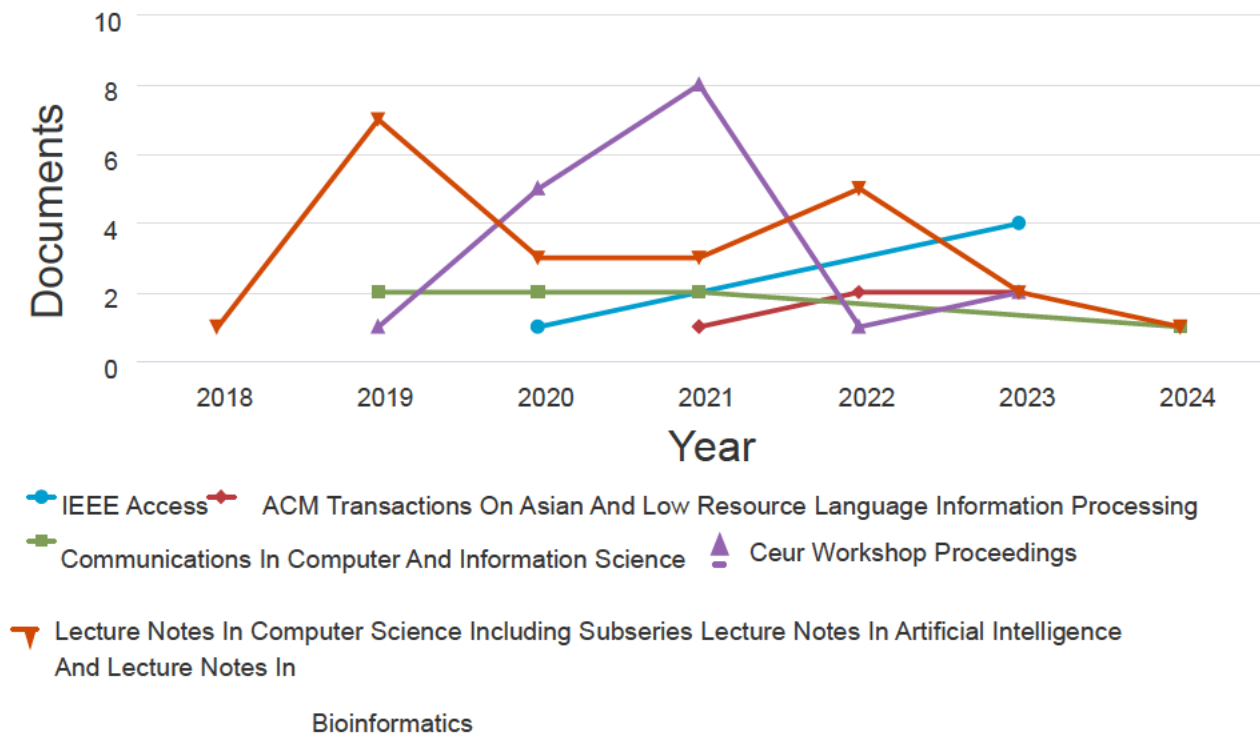


Figure 3.17: Publication per year by source.

Topics in textual data from different countries were carefully scrutinized, as depicted in Figure 3.18. Consequently, unique linguistic patterns and prevalent themes specific to each nation were understood. WVS WE MCD studies for languages with limited resources have made significant progress, as illustrated in the graphical representation. Currently, the data reveals that the US holds the lead, with China and India trailing closely behind.

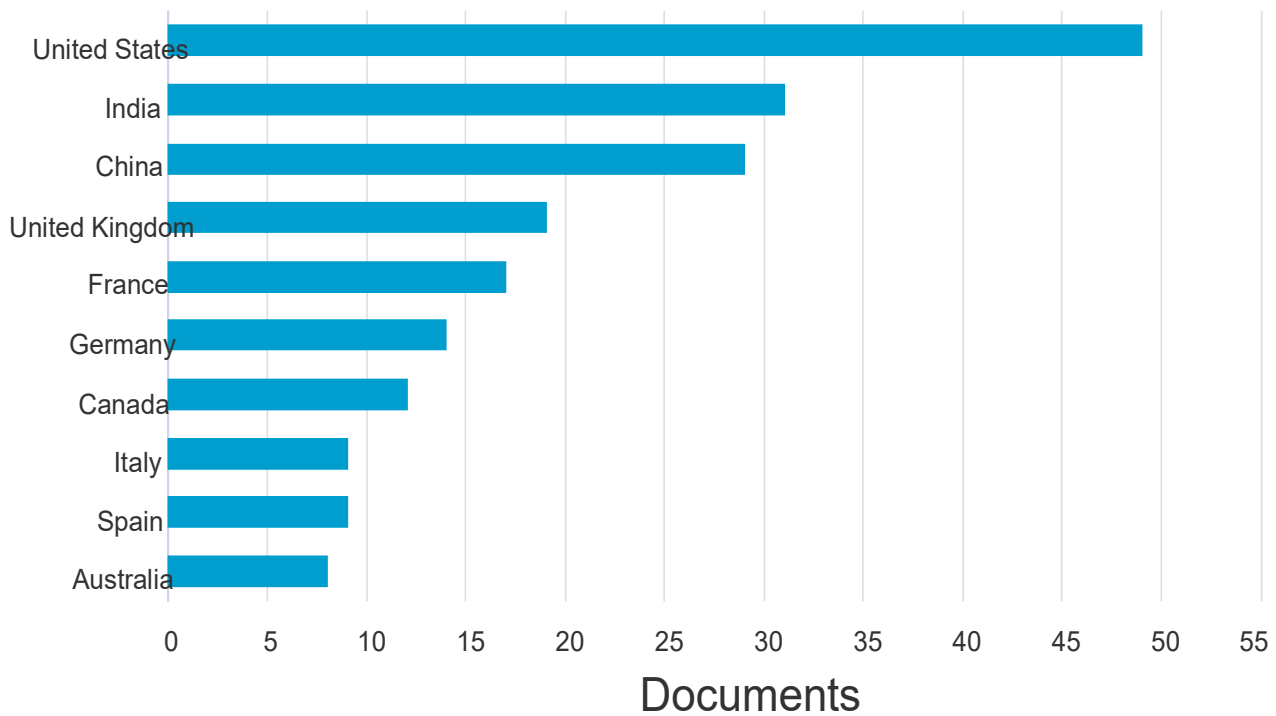


Figure 3.18: Publications by country

The bar chart in Figure 3.19 illustrates a bibliometric overview of investigation trends in NLP tasks from 2015 to 2024, divided into six primary categories: Text Classification, Sentiment overview, Question Answering, Named Entity Recognition (NER), Machine Translation, and Language Modeling. The peak research activity transpired between 2021 and 2022, with text classification and sentiment analysis emerging as the most extensively examined subjects. Question answering, named entity recognition, and machine translation have been intermittently examined, but language modeling has emerged in only a limited number of years. The trend indicates a heightened emphasis on NLP applications, especially in text classification and sentiment analysis, followed by a modest decrease in 2023 and 2024. This signifies a change in research priorities or an expansion into novel NLP approaches.

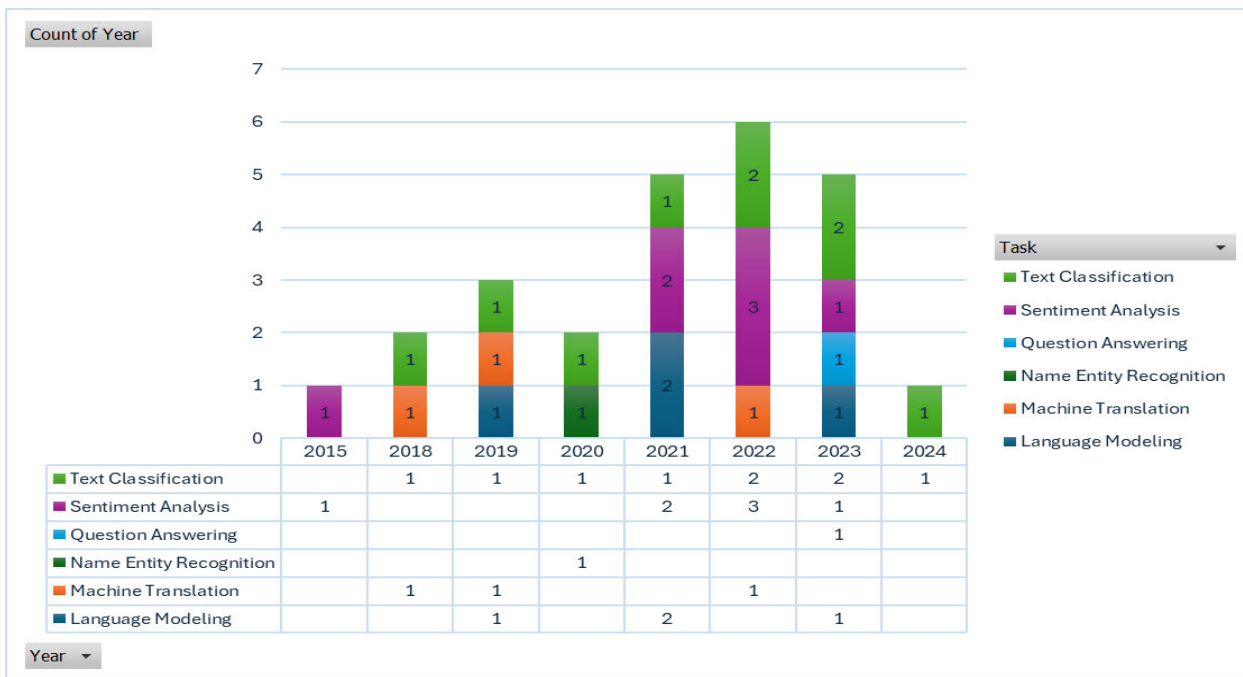


Figure 3.19: Models per year

Contextual word embeddings and context-aware models have garnered significant attention due to their proficiency in capturing nuances of meaning across numerous scenarios, as illustrated in Figure 3.20. These models, such as BERT, GPT, and their variants, address the challenges of WVS WE MCD by dynamically modifying models according to the context in which they appear. Conversely, although hybrid models—which integrate a wider range of features or combine multiple embedding types—are less frequently discussed in the literature, they can be highly effective. Their aim is to leverage the strengths of different techniques to address specific challenges associated with meaning resolution and word embeddings.

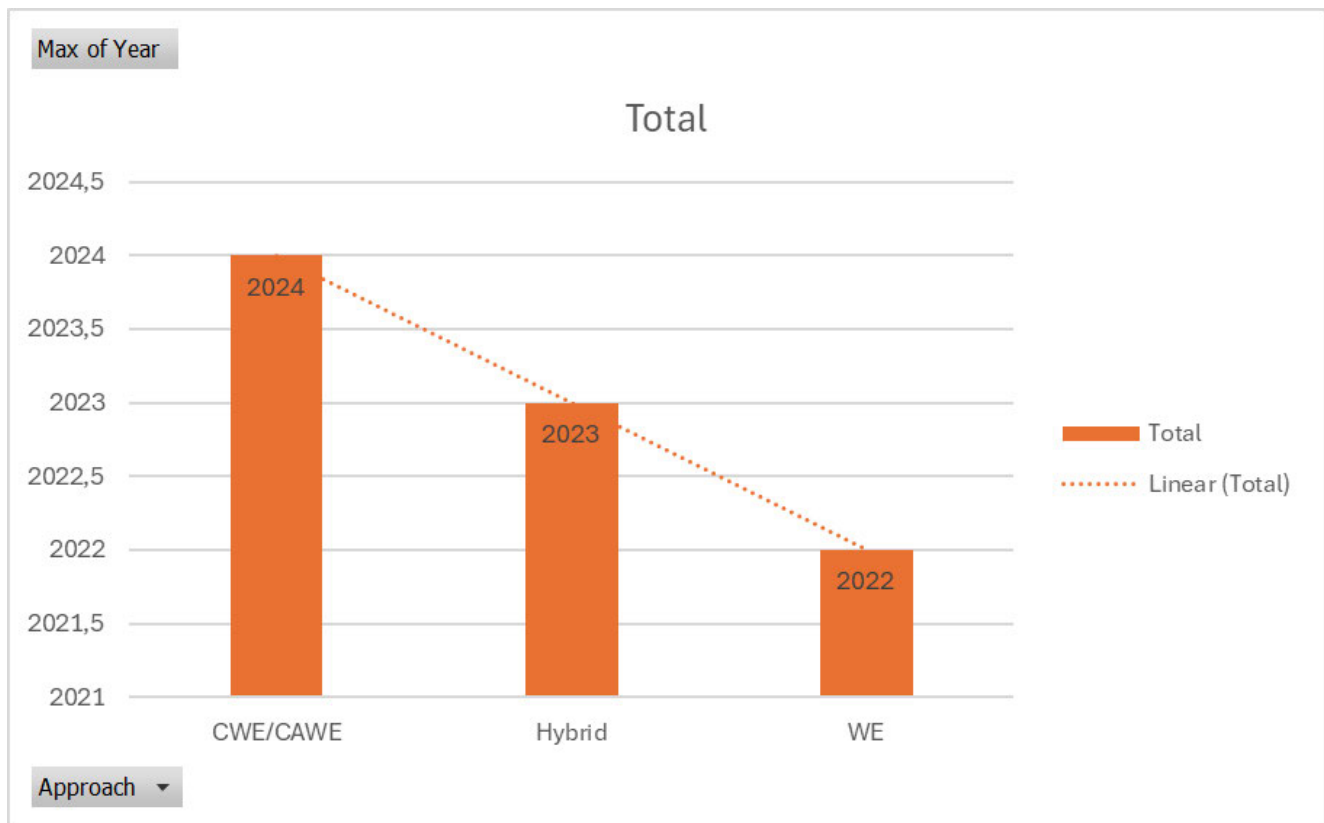


Figure 3.20: Approaches per year

Figure 3.21 illustrates that ACM publications tend to receive higher citation rates, suggesting that the computer science and engineering community prioritizes research on WVS WE MCD. ACM's reputation for publishing leading-edge computer science research highlights the significant relevance of this study to contemporary technological progress. Conversely, Elsevier and MDPI generally cater to a broader academic audience across multiple disciplines.

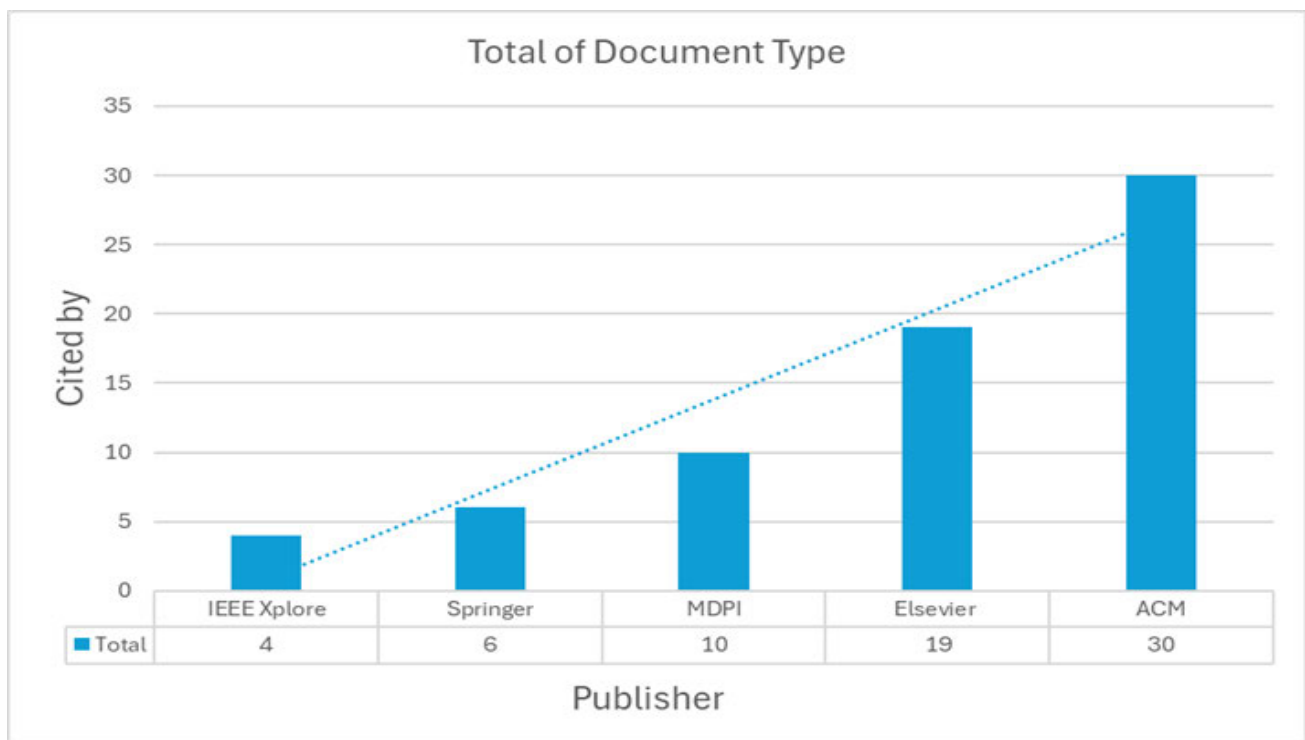


Figure 3.21: Publisher per citations

3.11 Gaps in the Literature

Research in vector-based word embeddings often overlooks this MCD problem, especially for language MRLs, low-resource languages such as SsaL. Models such as Word2Vec, GloVe, and BERT typically ignore the unique morphosyntactic and polysemous features of these languages, focusing mainly on resource-rich languages and neglecting agglutinative, noun-class languages such as SsaL.

The reviewed literature often uses static embeddings that merge different word senses into one vector, causing semantic ambiguity and reduced performance in NLP tasks such as WSD, information retrieval, and machine translation. Contextual embeddings (e.g., BERT, ELMo) offer some improvements but are seldom evaluated in Bantu languages or African linguistic settings. Bibliometric analysis reveals a focus on high-resource languages, for instance, English and Chinese, with scarce studies for languages with complex morphology. Although some studies mention multi-sense and context-aware embeddings, these models are rarely applied to morphologically rich African languages. Publications peaked around 2021–2022 and then declined, showing stagnating interest despite

unresolved issues. Many models overlook language-specific features such as agglutination, noun classes, and tonal variations, crucial for SsaL. This shows a research gap in creating and testing semantic models for these underrepresented languages. The meta-analysis shows high heterogeneity ($I^2 = 87.65\%$), implying inconsistent results across datasets, languages, and embeddings, especially for morphologically rich African languages.

Few studies incorporate linguistic knowledge such as morpheme and grammatical role annotations into embeddings. Promising techniques such as multi-sense embeddings and hierarchical sense graphs lack validation on low-resource languages with complex morphology. There is a shortage of annotated corpora for African languages and attempts to use language-specific architectures or graph-based embedding techniques to address SsaL morphology.

This gap reveals a strong rationale for the present research, which seeks to develop a linguistically aware, context-sensitive word embedding framework for Sesotho sa Leboa, combining corpus-based ensemble methods, multi-sense modelling, and transformer-based architectures. The proposed approach intends not only to resolve WVS WE MCD, but also to demonstrate empirically that embedding models must be aligned with the linguistic features of the target language to perform effectively. By bridging the methodological and resource-related gaps in the literature, this study positions itself as a critical advancement in embedding strategies for African low-resource languages, thereby contributing to both the field of computational linguistics and language technology development for underrepresented languages.

3.12 Chapter Summary

This chapter provided a comprehensive analysis of the latest techniques, frameworks, and challenges associated with addressing word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD), particularly in the context of languages that are both low-resource and morphologically rich. Chapter Three started with a theoretical overview, detailing the evolution of semantic representation models, from frequency-based vector space models to predictive neural word

embeddings. The chapter highlighted the inherent limitations of traditional embeddings such as Word2Vec and GloVe, which conflate multiple meanings of a word into a single representation, thereby leading to WVS WE MCD. Chapter Three provided a detailed analysis of the linguistic and computational challenges posed by SsaL, focusing on how morphological complexity, polysemy, and agglutination intensify semantic ambiguity, making WVS WE MCD an acute issue for these languages. To address the MCD issue, CAWE models have been developed, offering dynamic word representations that adjust to the surrounding linguistic context. This approach tackles the WVS WE MCD by allowing words to be represented through an unlimited range of potential embeddings. The creation of CAWE models has become essential in overcoming the difficulties associated with MCD in representing word meanings, thereby improving performance and simplifying the WSD task and its applications.

This chapter built upon theoretical principles to integrate empirical findings via a SLR, adhering to PRISMA standards, which compiled recent innovations in word-embedding applications for languages with limited resources. The review pinpointed research deficiencies, particularly a lack of investigation into resolving MCD in WVS WE models. A meta-analysis quantified performance trends across various studies, providing comparative analyses of model aptness, sense specificity, and language-specific adaptations. This was enhanced by a bibliometric analysis, utilizing tools such as VOSviewer, to plot author networks, keyword intersections, and publication trends, thereby illustrating the concentration of research over time. Additionally, descriptive statistics were used to showcase the characteristics, distribution, and methods of the primary studies examined. In sum, this chapter effectively connects theoretical and empirical insights, forming a foundation for creating innovative embedding models that recognize the linguistic subtleties of low-resource languages.

CHAPTER 4 : THEORETICAL FRAMEWORK

4.1 Introduction

This chapter provides the theoretical foundation for developing word embedding systems for SsaL, a complex and resource-limited language. This chapter reviews essential theories and models supporting the study's methodology. Section 4.2 covers distributional semantics, highlighting that words sharing contexts share meanings, differentiating semantic similarity and relatedness, crucial for NLP tasks WSD. Section 4.3 discusses vector space models and their use of co-occurrence statistics and linear algebra in high-dimensional spaces, covering pointwise mutual information (PMI) and analogical reasoning. Section 4.4, on latent semantic analysis (LSA), addresses dimensionality reduction using singular value decomposition (SVD), and challenges with polysemous meanings. Section 4.5 examines early word embedding models, for example, Word2Vec, GloVe, and FastText, noting their strengths and weaknesses for morphologically rich languages. Section 4.6 explores neural network-based models, comparing static embedding's effectiveness and subword information. Section 4.7 highlights contextualized embeddings such as ELMo and BERT, which modify word representations dynamically, invaluable for languages with extensive morphological diversity. Section 4.8 discusses multi-sense embeddings, modelling different vectors for word senses, using lexical databases for WSD. Section 4.9 focuses on African language challenges, noting ELMo's character n-gram modelling strength for resource-poor and agglutinative languages such as SsaL. Subsections compare ELMo's performance with FastText in minimal data settings. Section 4.10 concludes with key insights, detailing the shift from static to contextual and sense-aware embeddings; and their role in addressing semantic ambiguity in underrepresented languages.

4.2 Distributional Semantics

The diagram in Figure 4.1 presents a detailed conceptual map illustrating the development and categorization of word-embedding techniques, highlighting both context-independent and CAWE models alongside their theoretical underpinnings in distributional semantics. At the pinnacle of the hierarchy is distributional semantics, which is founded upon the *distributional hypothesis* asserting that “words appearing in similar context tend to have similar meanings” (Harris, 1970; Firth, 1957). This hypothesis underpins VSMs that depict words mathematically as points within high-dimensional space based on their contextual co-occurrence. VSMs are categorized further by aspects such as context, co-occurrence, semantic associations, and context-independent depictions. This underlines classical statistical approaches, for example LSA, that depend on word co-occurrence and dimensionality reduction, but are limited by their lack of context sensitivity, assigning a single vector representation to each word (Deermodelster et al., 2015).

The diagram in Figure 4.1 lays the groundwork for distributional semantics, which derives linguistic meanings through the analysis of word co-occurrence within extensive corpora (Arseniev-Kochler, 2024). Distributional semantics is a field of linguistics that models meanings of words based on the contexts in which they occur (Loureiro et al., 2022). Applying SVD reduces dimensions to illuminate these connections, enhancing information retrieval and semantic comprehension. In neural network-based models, vector space models denote the representation of words or data in continuous vector spaces (Almutairi et al., 2017). These models employ neural networks to learn dense vector representations or embeddings of words, phrases, or other data, capturing semantic meanings and relationships. Such embeddings enhance performance across tasks such as natural language processing, by encoding syntactic and semantic patterns efficiently (Milano et al., 2025).

The diagram in Figure 4.1 explores vector space models as an extension of distributional semantics. These models are further categorized by factors such as context, co-occurrence, semantic relationships, and representations that do not depend on context. This classification underscores traditional statistical models such as LSA, which depend on word co-occurrence and dimensionality reduction. Such models

lack sensitivity to context, each word being represented by a single vector (DeerWE modelster et al., 2015). This section of the diagram refers to conventional statistical models focused on frequency or co-occurrence matrices. LSA seeks to reduce dimensionality through methods such as singular value decomposition (SVD). Notably, these models do not account for context sensitivity, treating each word as having one vector representation. LSA (Deermodelster et al., 2015) extracts latent semantic structures from a term-document matrix but overlooks the contribution of syntactic structures and word order. In addition to applying Singular Value Decomposition (SVD) to the term-document matrix to obtain word embeddings, SVD can also be applied to the co-occurrence matrix for the same purpose.

The transition from statistical to neural embeddings signifies a major change in semantic modelling. The section on neural network-based models covers Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and BERT (Devlin, Chang, et al., 2019). These techniques represent distributional semantics by employing dense, low-dimensional embeddings that are learned from extensive collections of text. Unlike LSA, neural embeddings are designed to predict adjacent words (CBOW, Skip-gram) or to capture word co-occurrence patterns (GloVe), yet they generate a singular embedding for each word, thereby classifying them as context-independent representations.

The diagram in Figure 4.1 illustrates the transition from context-independent embeddings to those that are highly context-sensitive. Initial context-independent techniques – Word2Vec, GloVe, and early versions of ELMo – have evolved through models that embrace CAWE. Notably, ELMo (Peters, Neumann, Iyyer, Gardner, et al., 2018), leverages deep bidirectional LSTMs to create dynamic embeddings tailored to sentence context. BERT employs Masked Language Modeling (MLM) to develop deep, context-rich representations that are finely attuned to subword and contextual cues. Additionally, FastText (Bojanowski et al., 2017), augments embeddings through n-gram character-level features, enhancing their adaptability to languages with complex morphology. These advancements signify a significant step forward in addressing polysemy and word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) by crafting multi-sense embeddings.

Figure 4.1 demonstrates a progression toward sophisticated semantic models capable of creating multi-prototype embeddings, an approach to addressing polysemy as described by (Reisinger & Mooney (2010) and Huang et al. (2012). This methodology is particularly important for morphologically rich languages such as SsaL, where one surface form can convey various meanings. In languages such as SsaL, characterized by significant morphological diversity and semantic sparsity, context-aware word embedding (CAWE) and multi-sense embeddings are essential tools for addressing word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD).

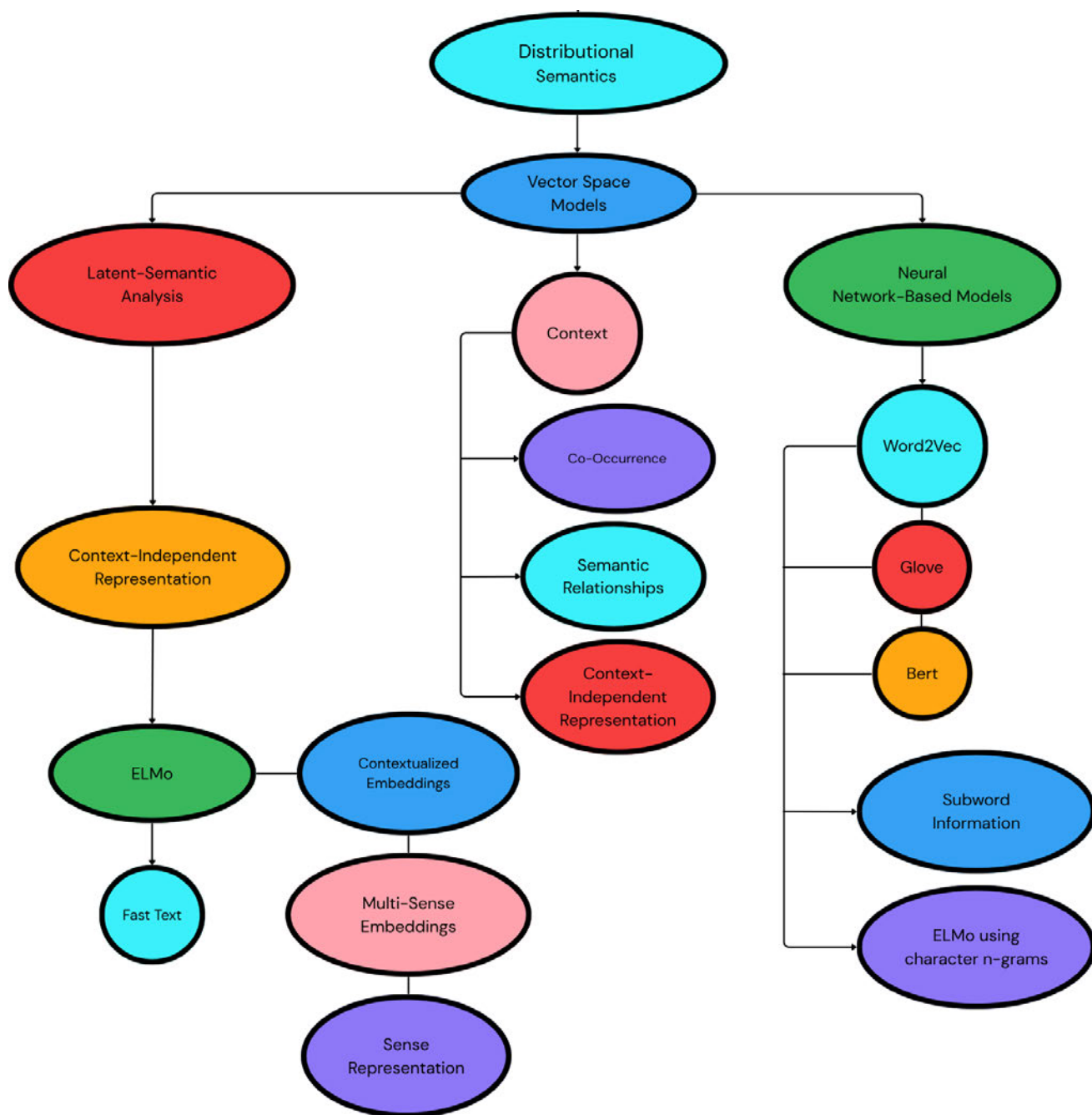


Figure 4.1: Distributional semantics and hypothesis

4.3 Vector Space Models (VSMs)

VSMs depict words as vectors within a high-dimensional semantic space. A word’s location is based on how it co-occurs with other words in a context window. Initial approaches, including term-document matrices, set the groundwork for more advanced embedding techniques (Roh et al., 2021). Semantic similarity in VSMs is often gauged by metrics such as cosine similarity. VSMs face issues of sparsity and high dimensionality, which have driven the creation of more streamlined, dense embeddings (Liji & Muhamed Ilyas, 2021). VSMs serve as mathematical frameworks to convert text

data into vectors of identifiers, such as words. VSMs are instrumental in fields such as information retrieval, natural language processing, and machine learning, enabling operations such as similarity assessment and clustering (Camacho-Collados & Pilehvar, 2018). VSMs are used in information retrieval and natural language processing to represent text documents as vectors of identifiers, such as index terms. In a vector space, every dimension is associated with a different term, and each dimension's value generally signifies the term's weight within the document. Techniques such as term frequency and Term Frequency-Inverse Document Frequency (TF-IDF) are commonly used to calculate these weights. This framework facilitates the mathematical comparison and analysis of textual documents (Rodriguez & Merlo, 2020).

4.4 Latent Semantic Analysis (LSA)

LSA enhances conventional VSMs by using SVD to lower the dimensions of co-occurrence matrices, thereby revealing hidden semantic patterns. LSA can partially capture both synonymy and polysemy, yet it struggles with modelling non-linear relationships and changing contexts (Alpaydin, 2014; El-alami et al., 2022). LSA is a method in natural language processing that employs vector space models to examine relationships between documents and the terms they include (Smith, 2019). LSA operates by creating a term-document association matrix and employing SVD to reduce the matrix's dimensions (Jijo et al., 2024). This results in a compressed version of the data, in which terms and documents are depicted in a continuous vector space, capturing the underlying semantic structure (Hoblos, 2021). This method improves information retrieval and comprehension by revealing hidden relationships. Through linear algebra, LSA transforms vast text datasets into a matrix format that facilitates the identification of concealed relationships between words and documents (George, 2018).

4.5 Neural Network-Based Models

Neural network-based models are computational models inspired by the human brain's neural networks; they are used in a variety of tasks such as classification, regression, and pattern recognition. Neural network-based models have introduced non-linear transformations and learning from data in a

scalable manner (Leclair et al., 2021). These models learn word representations that capture semantic and syntactic properties. Notable architectures include feedforward networks, recurrent neural networks (RNN), and more recently, transformer-based architectures. These models provide a more flexible and expressive way to model meaning (Gan & Zhang, 2020).

4.6 Word Embedding Models

4.6.1 Word2Vec

Word2Vec is a group of models used to produce word embeddings, which are vector representations of words in a continuous vector space. Word2Vec provides two architectural options: skip-gram and continuous bag of words (CBOW) (Mikolov et al., 2013). Each model predicts surrounding words from a given central word or the reverse, while generating dense, fixed-dimension vectors. Nevertheless, Word2Vec generates embeddings that do not account for context, meaning that a word has the same vector regardless of how it is used (Goldberg & Levy, 2014).

Word2Vec is a widely used neural network model for generating word embeddings, which are dense vector representations of words that capture semantic and syntactic relationships. Developed by Google in 2013 (Mikolov et al., 2013), Word2Vec revolutionized natural language processing by enabling words with similar contexts to be mapped to proximate vectors in a continuous vector space (Gao, 2023). This method effectively transformed words into numerical formats suitable for machine-learning algorithms, thereby facilitating more sophisticated analyses of textual data (Worth, 2023). The core of Word2Vec's operation lies in its two primary architectures: CBOW and Skip-gram (Song, 2016). The CBOW model predicts a target word from its surrounding context words, whereas the Skip-gram model aims to predict the surrounding context words given a target word (Camacho-Collados & Pilehvar, 2018). These models learn to represent words as high-dimensional vectors, in which semantically similar words are positioned closer together in the vector space, as measured by cosine similarity (Sheth et al., 2021). This capability allows for enhanced performance in various downstream NLP tasks, such as sentiment analysis, machine translation, and information retrieval (Pennington et

al., 2014). The creation of low-dimensional vectors from text corpora, central to these applications, can also be achieved through other neural network approaches (Camacho-Collados & Pilehvar, 2020).

Word embeddings are generated using the Word2Vec model, a type of shallow neural network. The resulting vectors are organized in such a way that words sharing similar contexts are positioned close to one another in the vector space. This organization allows for the understanding of a word's meaning through its contextual background. Word2Vec introduces two types of word embedding models: Skip-Gram and CBOW. The principles of these models are rooted in the distributional hypothesis, which suggests that words with similar meanings often appear in similar contexts (Ilie et al., 2021). The CBOW model is built on the concept of a unigram, an n -gram sequence with a single term. By computing the average of the context within a specified window, this method captures the context of each word. In contrast, the Skip-Gram model employs skip-grams to define word context, which helps capture the context of words that are more distantly placed within the sequence and assists in predicting the context based on a given word (Ilie et al., 2021). The skip-gram objective (Ansell, 2020) is expressed by the Equations (4.1):

$$J(c, o, u, v) = \log \vartheta(u_c^T v_o) + \sum_{i=1}^k E_{w_i \sim p_n} [\log (1 - \sigma(u_c^T v_{w_i}))] \quad (4.1)$$

Consider an input vector $u \in \mathbb{R}$ and an output vector $v \in \mathbb{R}$ for a word $i \in V$. Let c denote the target (center) word, o represent a word appearing in the context window of c , and k indicate the number of negative samples selected for each true context word. The sigmoid function σ is defined as follows

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (4.2)$$

4.6.2 Global Vectors for Word Representation (GloVe)

GloVe is a model for generating word embeddings by capturing word-word co-occurrence statistics from a large corpus. GloVe creates context-independent vectors, which can lead to challenges with polysemous words, meaning words with multiple meanings (Pennington et al., 2014). GloVe utilizes word-word co-occurrence data from an extensive corpus and applies matrix factorization to produce embeddings (Lestandy, 2023). Similar to Word2Vec, GloVe produces context-independent vectors and encounters difficulties with words that have multiple meanings (Belete et al., 2024).

4.6.3 FastText

FastText is an open-source, free, lightweight library developed by Facebook's AI Research (FAIR) lab for efficient learning of word representations and sentence classification. FastText is capable of very rapid training, and can also classify text in over 150 languages (Yao et al., 2020). FastText, leverages subword data through character n-grams, which is especially beneficial for languages with complex morphology. By constructing these subword components, FastText can create embeddings for words unrepresented in the vocabulary (Pilehvar & Camacho-Collados, 2020). The FastText model improves upon the Word2Vec model by utilizing both CBOW and Skip-Gram techniques. This improvement addresses the limitations that arise when the relative positions of words within the context window are not accounted for. The proposed approach involves weighting the context words by initially learning the representations of their positions (Ilie et al., 2021).

4.7 Context-Dependent Representations

4.7.1 Embeddings from Language Models (ELMo)

ELMo is a type of deep contextualized word representation that models both complex characteristics of word use (e.g., syntax and semantics) and how these uses vary across linguistic contexts. ELMo creates CAWE models through the use of deep bi-directional LSTMs (Y. Zhang et al., 2020). The word representation shifts depending on its context within a sentence. ELMo integrates subword details and encapsulates both grammatical and semantic features (Jayakody et al., 2023).

4.7.2 Bidirectional Encoder Representations from Transformers (BERT)

BERT employs a transformer architecture to produce deeply contextualized word representations that reflect the semantic influence of surrounding text. BERT does this by masking tokens during training to understand language from both directions, improving tasks such as word sense disambiguation, and handling polysemy effectively (Devlin, Chang, et al., 2019). Unlike previous language models, BERT's architecture allows for deep bidirectional representations by jointly conditioning on both left and right contexts across all layers, enabling fine-tuning for various tasks with minimal architectural modifications (Devlin, Chang, et al., 2019). This methodology ensures that the generated embeddings encapsulate richer semantic information, encompassing not only the individual word but also its surrounding sentence structure and broader corpus context (Worth, 2023).

This bidirectional training approach represents a significant advancement over unidirectional models such as GPT, which are limited by processing text in a single direction (Pilehvar & Camacho-Collados, 2020). The foundational transformer architecture, consisting of encoder and decoder units with multi-headed attention mechanisms, underpins BERT's ability to process entire input sequences as context for each prediction (Worth, 2023). This innovative pre-training strategy allows BERT to develop a comprehensive understanding of language, making it empirically powerful and conceptually simple (Devlin, Chang, et al., 2019). The two-phase implementation of BERT involves pre-training on a large corpus of unlabelled text, and subsequent fine-tuning for specific downstream tasks, using the same architecture apart from the output layers (Zaland et al., 2024).

This pre-training phase enables the model to learn robust language representations by jointly conditioning on both left and right contexts, a method that diverges from previous models that processed text in a single direction (Devlin, Chang, et al., 2019). BERT utilizes the transformer framework to produce deeply contextualized embeddings. During training, it masks certain tokens, allowing the model to evaluate contexts from both directions. BERT greatly enhances WSD and effectively manages polysemous expressions (Arseniev-Koehler, 2024).

In contrast to the unidirectional design seen in GPT, BERT employs a bidirectional transformer architecture. Furthermore, two unsupervised tasks have been incorporated into BERT's pre-training process (Vusak et al., 2021). BERT functions as a deep transformer-based encoder, pre-trained through two concurrent objectives: masked language modeling and next-sentence prediction (Rodrigues da Silva & Caseli, 2021).

Researchers (Li et al., 2021) explored the effectiveness of BERT's contextual embeddings in creating domain-specific word vectors, particularly for knowledge base completion, where the primary goal is to capture the subtle semantic meanings of nouns. Their findings indicate that simply averaging the contextual embeddings from masked instances results in vector representations that surpass both BERT's static embeddings and traditional word embedding techniques. According to Ansell (2020), BERT can be conceptually viewed as consisting of two core elements, expressed in Equations (4.3) and (4.4): an embedding matrix $E \in \mathbb{R}^{|V| \times d}$, and a contextualization function C that transforms a sequence of input embeddings $x_1, x_2, \dots, x_n \in \mathbb{R}^d$ into a corresponding sequence of context-aware outputs $y_1, y_2, \dots, y_n \in \mathbb{R}^d$.

$$P(w|y) = [\text{softmax}(E_y + b)]_w \quad (4.3)$$

$$= \frac{e^{e_w^T y + b_w}}{\sum_{j \in V} e^{e_j^T y + b_j}} \quad (4.4)$$

The techniques produce word-level representations that remain unchanged by the surrounding context; such “static” representations often display a notable bias towards the most prevalent meaning of the word. With the widespread availability of language models such as BERT, which are pretrained on extensive text datasets, there has been a marked rise in the use of contextualized embeddings. Contextual word embeddings are able to capture the different facets of a polysemous word within a given context (T. Zhang et al., 2020).

4.7.3 Multi-sense embeddings and sense representation

Multi-sense embeddings are a type of word representation in natural language processing that captures multiple meanings of a word by creating different vectors for each sense (Peters, Neumann, Iyyer, Gardner, et al., 2018). These embeddings address MCD by assigning distinct vectors to various word senses. Sense representation refers to the way in which sensory information, such as visual or auditory stimuli, is encoded and processed by the brain or artificial intelligence systems to create meaningful interpretations. Sense representation involves modelling these different senses to reflect the context in which a word is used (Butnaru & Ionescu, 2019). Methods to achieve this include clustering context vectors, as demonstrated by the Multi-Sense Skip-Gram model from researchers (Neelakantan et al., 2014) and deriving sense-specific embeddings from annotated datasets. Additionally, graph-centric models and retrofitting techniques utilizing lexical databases such as WordNet are employed for sense representation (Pak et al., 2024). In SsaL, when one surface form can represent several morphological or semantic variations, the use of multi-sense and contextual embeddings offers the detailed analysis required to clarify ambiguity and enhance the precision of word sense disambiguation.

Table 4.1 presents key concepts and the semantic foundations vital for tackling ambiguity and improving precision in WVS WE MCD. This Table 4.1 highlights the transition from static models to CAWE and sense-specific models to achieve more precise word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD).

Table 4.1: Key Concepts and Semantic Foundations

Key Concepts	Semantic Foundations
Context	The context in which a word is used influences its meaning.
Co-occurrence	The rate at which two words co-occur is a fundamental indicator of semantic similarity.
Semantic Relationships	Includes synonymy, antonymy, hypernymy, and polysemy.
Context-Independent Representation	One vector per word (e.g., Word2Vec, GloVe).

Context-Dependent Representation	The meaning of a word can change depending on its context (e.g., BERT, ELMo).
Subword Information	Utilizing character-level information to construct word vectors (e.g., FastText).
Contextualized Embeddings	Embeddings that change based on usage and context (e.g., BERT).
Sense Representation	Explicitly modelling various interpretations of a single word.

4.8 Chapter Summary

The theoretical framework chapter has delivered the progression from basic distributional methods to sophisticated neural and contextual models. The framework emphasizes the compromises between static and contextual embeddings and the advancements made to tackle polysemy and meaning conflation issues. These principles lay the conceptual and methodological groundwork for building strong WSD systems, especially designed for underresourced, morphologically rich languages such as SsaL. This study addressed the gap by integrating corpus-based, subword-level, and CAWE models for Sesotho sa Leboa. This research work combined distributional semantics and polysemy theory with evaluations of static, contextual, and multi-sense embedding methods. This study contributed novel datasets and embedding strategies for SsaL and supports linguistically grounded NLP approaches maintaining semantic granularity. The study therefore fills a critical void at the intersection of computational linguistics and underrepresented African languages.

CHAPTER 5 : METHODOLOGY

5.1 Introduction

This chapter outlines the methodological framework to study the MCD problem in word vector space (WVS) word embedding (WE) models. It explains methods, models, datasets, and evaluation techniques used in the analysis and the rationale for chosen models and embeddings, thus addressing research questions. Sections 5.2 to 5.7 cover theoretical design, model selection, experimental workflows, and evaluation criteria. Section 5.2 introduces the main methodological framework, combining theory with practical tests for optimising machine and deep learning models for the meaning conflation deficiency (MCD) problem in word vector space (WVS) word embedding (WE) models. Section 5.3 highlights the study's purpose and the limits of traditional word embeddings, showing the role of context-sensitive strategies in reducing ambiguity in low-resource languages. Section 5.4 defines the research design for the enhanced CAWE model to tackle word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD), focusing on the language scope and data from annotated sources. Section 5.5 covers data-collection tools using manually annotated datasets. Section 5.6 details the data-collection process, including preprocessing and tagging for agglutinative morphology.

Section 5.7 focuses on data analysis, the experimental setup, model architecture, embedding choice, and training methods. Subsection 5.7.1 covers computational architecture, describing models such as BART, RoBERTa, ELECTRA, NNLM, ELMo, BiLSTM, BERT, CNN, XLNet, and graph representations. Subsection 5.7.2 covers the deep learning environment with TensorFlow and PyTorch. Subsection 5.7.3 discusses optimization strategies, for example, regularization and hyperparameter tuning. Subsection 5.7.4 describes the model training and evaluation process, outlining the use of an 80/20 train-test split along with k-fold cross-validation to ensure reliable performance assessment. Subsection 5.7.5 provides an overview of the evaluation measures used in this study, namely accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). Subsection 5.7.6 presents the experimental comparison, evaluating different model types. Subsection 5.7.7 introduces meta-

evaluation using McNemar’s test to assess model differences. Subsections 5.7.7-5.7.9 explore BERT variants and graph algorithms for the MCD problem in WVS WE models. This comprehensive approach ensures detailed examination and quantification of models’ effectiveness, promoting transparency, ethical integrity, and replicability.

5.2 Rationale for Adopting Computational Experimental Design

This research study adopts a Computational Experimental Design (CED), which is especially effective for examining and addressing the MCD problem in WVS WE models, particularly in the context of languages that are morphologically rich and have limited resources. WVS WE MCD highlights the failure of conventional embeddings to differentiate between various meanings of a word, which frequently results in reduced performance of models used in WSD tasks (Pilehvar & Camacho-Collados, 2020).

Computational experimental design facilitates the systematic testing of various deep learning models (such as BART, ELMo, XLNet, and CNN) to assess their effectiveness in reducing meaning conflation (Peters et al., 2018; Yang et al., 2019; Kim, 2014; Lewis et al., 2020). This approach permits organized alteration of model components such as contextual encoding (BART, ELMo), positional embeddings based on permutation modelling (XLNet), and comparing superficial versus intricate convolutional layers (CNNs) (Peters et al., 2018; Yang et al., 2019; Kim, 2014; Lewis et al., 2020).

Model architectures capture semantic and syntactic dependencies in distinct ways. ELMo (Peters et al., 2018) utilizes bidirectional LSTM encodings for contextual token representation. BART (Lewis et al., 2020) integrates a denoising autoencoder with sequence-to-sequence training. XLNet (Yang et al., 2019) adopts permutation-based autoregressive pretraining to obtain bidirectional context without using masking. CNNs (Kim, 2014) implement local n-gram filters to identify lexical-semantic patterns in a way that is invariant to position. By utilizing CED, one can test hypotheses concerning the influence of architecture and context sensitivity on WVS WE MCD resolution, thus ensuring that these evaluations are both robust and clear.

5.3 Research Methodological Framework

The methodological structure depicted in Figure 5.1 for tackling the MCD problem in WVS WE models consists of three interconnected phases: (1) Corpus creation and annotation, (2) Model construction and training, and (3) Assessment and validation. Each stage adheres to the latest standards in computational linguistics and deep learning.

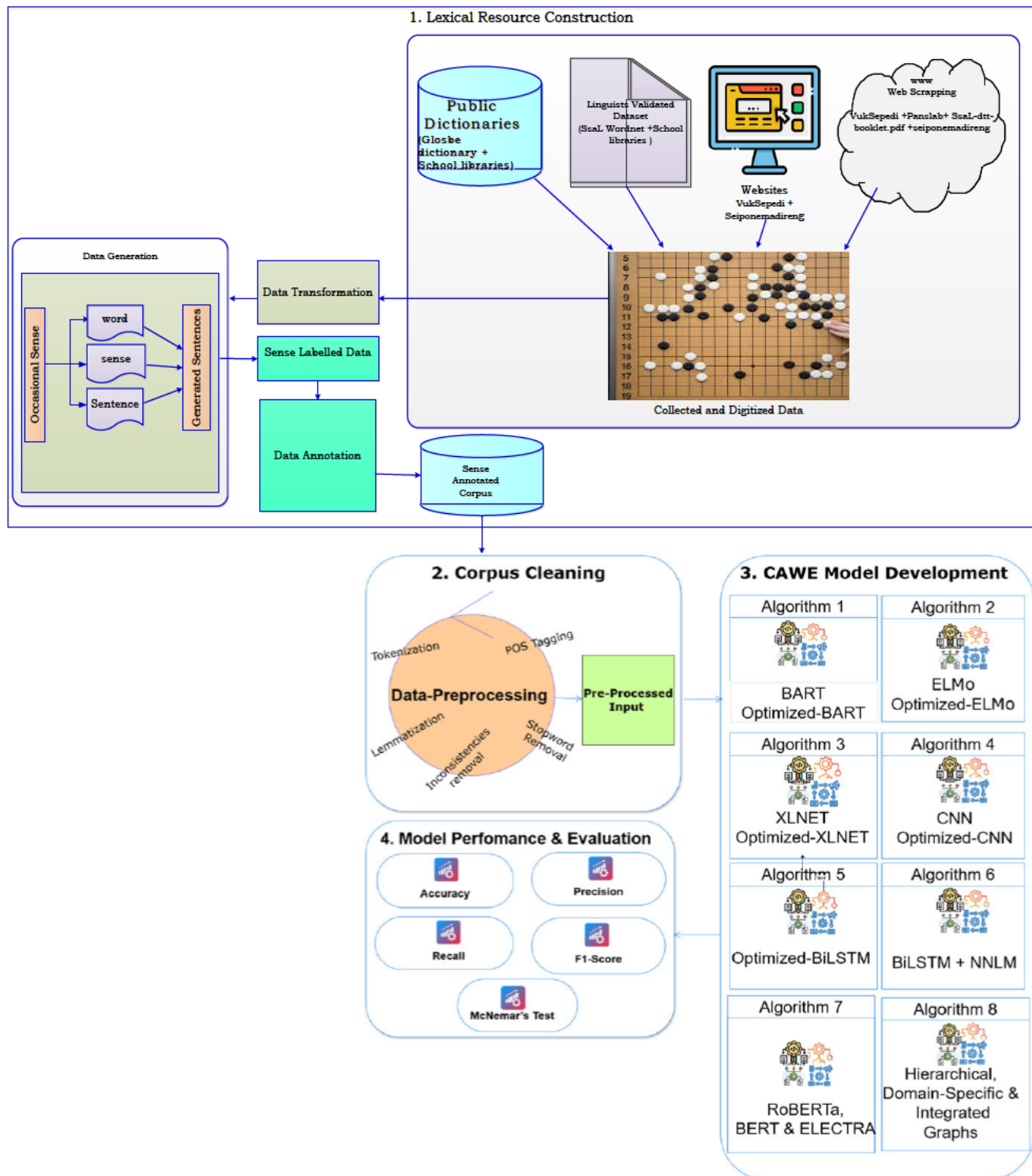


Figure 5.1: Research methodology framework

5.3.1 Corpus creation and annotation

This stage initiates with the gathering and storage of text data pertinent to the field, subsequently leading to the creation of an annotated corpus. Both manual and semi-automated processes are employed to annotate target words with sense tags according to their contextual meaning. The data preprocessing sequence involves tokenization, part-of-speech (POS) tagging, lemmatization, and the elimination of stopwords—each a crucial step for conditioning morphologically rich text for computational analysis (Bird & Klein, 2009).

The *sense-annotated SsaL corpus* is specifically designed for WVS WE MCD, maintaining morphosyntactic cues and context windows. This step is essential to address the shortage of labelled corpora in low-resource languages (Scarlini et al., 2020). Annotated corpora facilitate the learning of sense-specific patterns, which are vital for training context-aware embedding models.

To overcome the absence of an extensive, high-calibre linguistic resource for SsaL, a *sense-annotated SsaL corpus* with manually annotated senses was developed using a variety of archived textual sources. This data was processed with morphology-aware techniques aimed at retaining key linguistic features crucial for effective disambiguation. The approach utilized featured specialized preprocessing designed for agglutinative morphology, ensuring morpheme boundaries were maintained during tokenization. Part-of-speech (POS) tagging was applied to accurately identify noun class agreement patterns and verbal extension structures. Lemmatization was carefully executed to standardize inflected forms without eliminating markers crucial for meaning. The removal of stopwords was also tailored, deliberately keeping functionally significant morphemes that aid in semantic interpretation.

In this corpus, every target word was manually tagged with its appropriate meaning based on the surrounding text, creating a training resource that distinctly highlights sense distinctions and contextual indicators. To address semantic sparsity, the method involved thorough documentation of both prevalent and rare word senses, ensuring that even infrequent meanings were captured and could be learned. To improve linguistic and contextual accuracy, the annotation was performed by two native

Sesotho sa Leboa (SsaL) speakers, each representing a different dialect. The two individuals were both pursuing postgraduate studies in applied language disciplines—one was a master's student while the other was pursuing a doctorate. They worked together to make sure that the annotations accurately represented natural language use and were suitable in context. This research assessed the consistency of the annotations and addressed any disagreements, using Cohen's kappa to measure the inter-annotator agreement.

$$K = \frac{P_o(\text{Observed Agreement}) - P_e(\text{Expected Agreement})}{1 - P_e} \quad (5.1)$$

The result of $\kappa = 0.70$ (70%) reflects a substantial level of inter-annotator agreement, affirming the reliability and consistency of the sense-annotated SsaL corpus for use in training WSD models.

5.3.2 Model construction and training

A variety of carefully optimized algorithms were implemented using the Contextual Meaning Resolution (CMR) framework to learn and resolve word senses, leveraging CAWE models such as ELMo, BART, and XLNet, alongside their optimized counterparts. Additionally, a range of neural architectures—including CNN (with dropout-enhanced variants), BiLSTM and its optimized form, NNLM, RoBERTa, and BERT—were employed. Graph-based models, such as hierarchical, domain-specific, and integrated graphs, were also incorporated to enhance disambiguation performance.

The pretrained contextual models used in this study were adapted through transfer learning by fine-tuning pretrained language representations on the Sesotho sa Leboa corpus developed for Meaning Conflation Deficiency. In this context, transfer learning refers to the use of models originally pretrained on large external corpora, either English-based or multilingual, and subsequently refined on the target Sesotho sa Leboa dataset to improve contextual semantic discrimination.

To enhance WSD target-word highlighting was introduced, enabling the model to concentrate on ambiguous lexical items and the surrounding morphemes that contribute to their interpretation.

To mitigate bias towards dominant senses, a class-balanced training strategy was applied, ensuring that less frequent senses received adequate representation. In addition, hyperparameter tuning was performed to maximize the efficiency, generalization, and stability of model training. Collectively, these enhancements address the agglutinative nature of SsaL by preserving and emphasizing morphological cues situated around the target word—features that are critical for accurate semantic interpretation in morphologically rich languages.

5.3.3 Evaluation and Validation

The performance of the model was evaluated using a comprehensive set of metrics, including accuracy, precision, recall, F1-score, and McNemar’s test, to assess both the correctness and completeness of word sense assignments. These evaluation measures were carefully chosen to strike a balance between effectively differentiating word senses and capturing less frequent meanings, which is essential for tackling the issue of MCD in WVS models, particularly within morphologically rich languages characterized by sparse semantic distributions. A standardized evaluation procedure was maintained across all models and experimental setups to ensure fair and consistent comparison of results.

The research methodology framework, as outlined in Figure 5.1, illustrates the alignment between each methodological phase and the specific linguistic challenges inherent in SsaL. The initial phase—corpus preparation—addresses semantic sparsity by applying manual sense annotation alongside preprocessing methods that are sensitive to morphological structure. These include preserving morpheme boundaries and properly handling inflected forms, which are essential for maintaining linguistic integrity in agglutinative languages. The second phase, model development, addresses challenges such as polysemy and agglutinative morphology by applying transfer learning to adapt pretrained contextual models to SsaL-specific features. This stage further integrates target-word highlighting to strengthen sense discrimination and employs class-balanced training techniques to mitigate majority-class bias, thereby improving the model’s capacity to learn and differentiate rare word senses. The final phase—evaluation—prioritizes preserving the integrity of sense boundaries by utilizing a comprehensive set of performance metrics, including accuracy, precision, recall, and F1-

score. These measures ensure that the model not only differentiates word senses reliably but is also capable of correctly identifying infrequent or low-occurrence senses.

This framework establishes a direct link between the linguistic characteristics of a low-resource, morphologically rich language and the computational techniques required to model them effectively, demonstrating both theoretical soundness and practical relevance. To verify the performance gains of the optimized models over their baselines, McNemar’s test is employed to determine whether the observed improvements are statistically significant, thereby strengthening the validity of the findings. The statistical analysis is supported by a formal equation, which quantifies the significance of the performance difference.:

$$x^2 = \frac{(|b - c| - 1)^2}{b + c} \quad (5.2)$$

This study tackles the uneven distribution of word senses through a layered strategy applied across the data, loss, and inference stages. At the data level, cost-sensitive learning is implemented by assigning class weights derived from the effective number of samples. To further improve representation of low-frequency senses, training is conducted using class-balanced mini-batches. During inference, the method boosts recall for underrepresented senses by modifying prediction probabilities and fine-tuning class-specific decision thresholds using validation data, emphasizing macro-F1 optimization instead of prioritizing accuracy alone. When possible, targeted data augmentation or active labeling is also applied to increase the number of training examples for extremely rare senses.

5.4 Algorithm Design

This section outlines the research framework devised for creating and assessing the proposed enhanced CAWE model aimed at addressing the meaning conflation deficiency (MCD) issue in word vector space (WVS) word embedding (WE) models. This framework proves particularly advantageous for languages with complex morphology, in which conventional embeddings are insufficient. By

integrating deep contextual models such as BART and ELMo with graph-based sense modelling, the CAWE model delivers a scalable, multi-faceted approach to tackling semantic ambiguity and conflation. The objective is to train various refined models for sense disambiguation using CAWE models, each tackling meaning conflation deficiency (MCD) from unique representational angles, thus facilitating a comparative performance evaluation (Bevilacqua & Navigli, 2020).

The study also employs advanced neural network structures to improve the acquisition of semantics from context, building on the pioneering efforts of (Collobert et al., 2011), who presented integrated frameworks for NLP that leveraged word-level embeddings and deep learning. These methodologies demonstrated that one could effectively learn contextual comprehension directly from raw text, eliminating the need for manually designed features. This aligns with our goal of addressing meaning conflation deficiency (MCD) by optimising transformer and sequence models.

The research strategy focuses on the progressive refinement of the following varied and optimised algorithms:

Table 5.1: Deployed Algorithms

Model	Optimisation Purpose
Optimized BART	For sequence-to-sequence contextual generation. Leveraged for its encoder-decoder architecture that captures long-range semantic dependencies (Lewis et al., 2020)
Optimized ELMo	For deep, character-level, context-dependent embeddings (Peters, et al., 2018)
Optimized CNN	For pattern extraction in fixed window contexts. Employed for local context pattern recognition, suitable for shorter sequences and morphologically rich contexts (Collobert et al., 2011).
Optimized XLNet	For capturing permutation-based dependency modelling. Incorporates permutation-based language modelling, capturing bidirectional dependencies without masking (Yang et al., 2019).
Optimized BiLSTM & NNLM	For sequential contextual learning in both directions. Exploits recurrent patterns over time sequences, crucial for disambiguating evolving meanings in sentences.

RoBERTa & BERT	For transformer-based embeddings with strong contextual awareness (Devlin, Chang, et al., 2019; Liu et al., 2019).
Graph Models	For modelling semantic relations using graph structures (hierarchical, domain-specific, integrated). Graph-based sense representation techniques are used to model semantic networks and domain relationships, enhancing generalizability across domains (Bevilacqua & Navigli, 2020).

5.5 Data Collection and Annotation

To construct a comprehensive dataset for SsaL, textual data were sourced from diverse and accurate materials, including scholarly dissertations, academic publications, dictionaries, websites, newspapers, web log, and text contributions from social networking sites. These varied sources were intentionally selected to capture both formal and informal language usage, thereby representing a wide spectrum of linguistic variation. Due to the scarcity of structured linguistic resources for SsaL, automated web-scraping—implemented using Beautiful Soup—was utilized to extract usable text from HTML pages while filtering out non-linguistic artifacts such as markup tags. The overarching aim of this process was to collect naturally occurring instances of ambiguous words within context, forming a robust basis for subsequent sense annotation.

To increase linguistic variability within the dataset, data were gathered from both formal institutional sources (e.g., government documents and academic publications) and informal user-generated material (e.g., online forums and social networking platforms), ensuring representation of diverse registers and dialectal forms. During the manual annotation stage, two experts in Sesotho sa Leboa morphology, syntax, and semantics provided in-depth analysis of complex linguistic constructions and sense distinctions. Their work was supplemented by three native speakers from different dialect regions, who validated that word usage reflected authentic, naturally occurring language. In addition, two postgraduate students in applied language studies supported the process by ensuring consistency across annotations and helping to address discrepancies. This multidisciplinary, linguistically grounded

collaboration produced a high-quality dataset that is contextually rich and well-suited for developing and evaluating CAWE models for a morphologically rich, low-resource language such as SsaL.

Words that frequently have several meanings make up a sizable component of the collection. A variety of situations that disclose distinct meanings are provided by the many sentences that each polysemous word relates to. Sentences are chosen with care to convey the meaning of the term considering the surrounding context, including enough co-occurring words and phrases to support the intended meaning. A labelled dataset for supervised training is provided by these phrases, which are annotated with their respective senses and frequently connected to a domain-specific lexicon or a sense inventory such as WordNet.

Low-resource languages, especially those that might not have large lexical resources or pre-trained language models, are the subject of a subset of the dataset. Annotations describing the target word, its intended meaning, and its context are included with every dataset item.

To prevent bias toward a specific meaning, the dataset aims to provide a fair distribution of senses for each polysemous word. This equilibrium guarantees that the model has enough information to understand the differences between each sense and use them correctly throughout testing. This dataset's main goal is to make it easier to train and assess algorithms that can produce contextually aware embeddings and lessen meaning conflation. The dataset helps models understand nuanced differences between word senses by offering a rich variety of contexts and annotated meanings, especially when ambiguity results from polysemy or specialist terminology. In the end, this thorough methodology advances techniques for resolving meaning conflation deficiency (MCD) in word vector space (WVS) word embedding (WE)models by enabling robust testing of models across domains and languages.

A manually *sense-annotated SsaL corpus* is developed as depicted in Figure 5.2 especially for WSD and associated NLP tasks to address meaning conflation deficiency (MCD)in word vector space (WVS) word embedding (WE)models. The polysemy, homonymy, and context-dependent word meaning of SsaL are among the distinctive linguistic difficulties that this *sense-annotated SsaL corpus* reflects. Words such as “*lela*” (which may mean “cry” or “call”) and “*tsela*” (which can mean “path” or

“method”) are examples of words that have numerous meanings depending on the context. Sentences with varying meanings are used to annotate each target word. To help clarify the intended meaning, sentences that demonstrate the “cry” and “call” connotations of the word “*lela*” are added.

A fundamental resource for WSD and contextualized NLP applications is the carefully annotated corpus for SsaL to resolve the meaning conflation deficiency (MCD) problem in word vector space (WVS) word embedding (WE)models. SsaL supports future developments in low-resource NLP for African languages by allowing the investigation and resolution of word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) within a language that has historically lacked substantial computing resources through careful selection, sense differentiation, and domain inclusion.

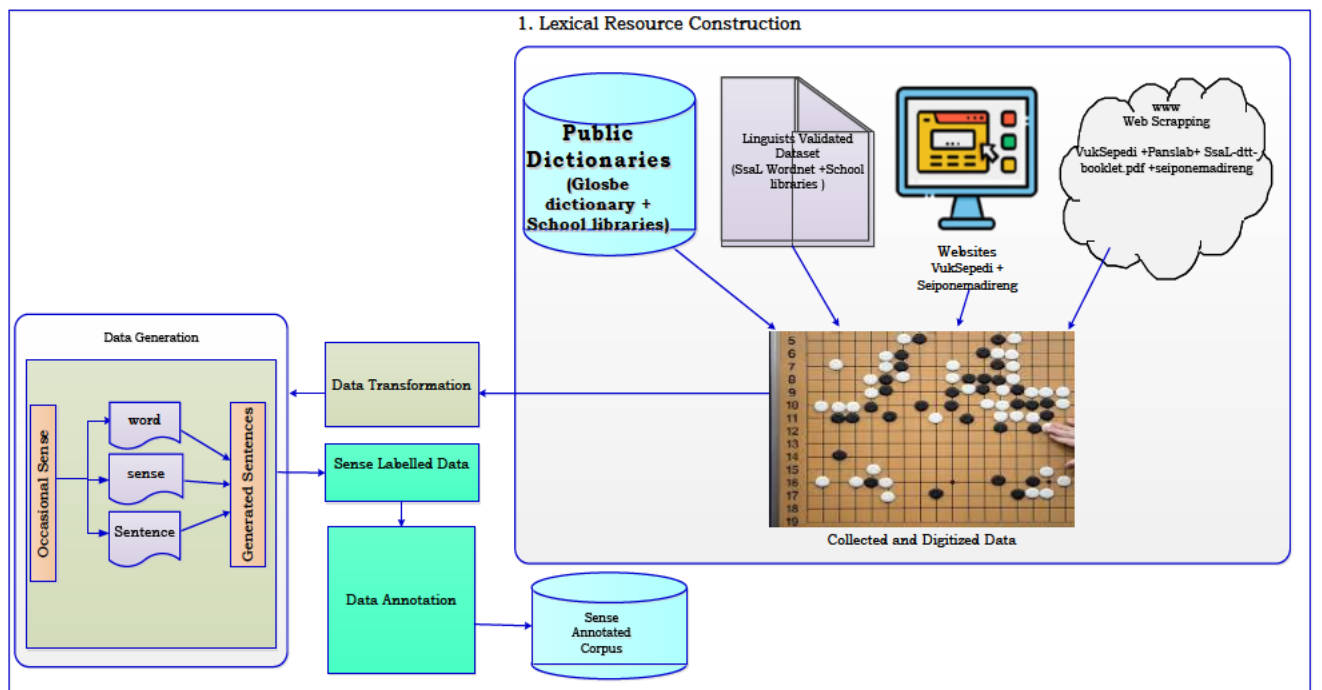


Figure 5.2: Data generation and collection

5.6 Selection of Models for Experimentation

The models chosen for this study were not randomly selected; instead, the decision was guided by a meticulous and theory-informed process, grounded in empirical data. Given the intricate nature of meaning conflation deficiency (MCD), the chosen models were required to meet several criteria focused on both computational efficiency and linguistic appropriateness. The experimental design

demanded a variety of architectures that could provide contextual insights, scalability, and subword-level generalization to effectively handle the subtle aspects of word sense disambiguation and polysemy. The inclusion of Optimized BART, ELMo, XLNet, and CNN is not intended to test every possible model, rather, to represent four theoretically distinct families of language modelling approaches relevant to MCD in a low-resource, morphologically rich language setting. The models are selected to provide a structured comparative framework across different representational paradigms. CNN is included as a strong deep learning baseline because it captures local contextual patterns and n-gram features effectively, which are often useful for lexical disambiguation. ELMo is selected because it generates context-sensitive word representations, making it theoretically appropriate for modelling polysemy and reducing meaning conflation. XLNet is chosen as an advanced transformer architecture with bidirectional contextual dependency modelling and permutation-based learning, which allows richer contextual interpretation than conventional sequential models. BART, in its optimized form, is included because its denoising pretraining and encoder-decoder contextualisation provide a different transformer-based mechanism for learning semantic relationships and contextual reconstruction, which is relevant when examining fine-grained sense distinctions. The inclusion of these algorithms is further justified by the use of optimization and regularisation strategies intended to reduce overfitting and memorisation. Consequently, the following models were chosen based on the five criteria outlined below:

Table 5.2: Selection Criteria

Criterion #	Selection Criteria Description
C1	Ability to capture contextualized word semantics effectively
C2	Proven success in prior WVS WE MCD-related benchmarks
C3	Incorporation of subword modelling or character-level awareness
C4	Compatibility with morphologically rich languages
C5	Flexibility for optimisation and transfer learning

Table 5.3: Model Selection Rationale

Model	C1: CAWE	C2: Proven in WVS WE MCD	C3: Subword Modelling	C4: MRL Suitability	C5: Optimization Flexibility	Rationale
Optimized- BART	✓	✓	X	✓	✓	BART leverages denoising pretraining to learn deep context. Despite lacking native subword modelling, its encoder-decoder design facilitates fine-tuning for WVS WE MCD resolution.
Optimized- ELMo	✓	✓	✓	✓	✓	ELMo provides deep contextual embeddings and uses character-level CNNs, making it well-suited for low-resource, agglutinative languages.
Optimized- CNN	X	✓	✓	X	✓	Included as a baseline deep learner to assess improvements from context-aware models. It lacks full contextuality but supports fast inference.
Optimized- XLNet						XLNet captures permutation-based

	✓	✓	✓	✓	✓	context dependencies, making it ideal for modelling subtle word sense shifts and partial ordering in context.
Optimized-BiLSTM + NNLM	✓	✓	X	✓	✓	A sequential model capable of bidirectional learning. Paired with NNLM to learn local co-occurrence statistics.
RoBERTa & BERT	✓	✓	X	✓	✓	These transformer encoders are proven performers in WSD, and their masked language modelling objective aligns with sense disambiguation.
Hierarchical, Domain-Specific, and Integrated Graphs	✓	✓	N/A	✓	✓	These capture semantic hierarchies and domain-specific relations to model meaning conflation across senses using symbolic and neural knowledge.

This study adopted a multi-model approach to thoroughly assess meaning conflation deficiency (MCD) across various representational frameworks. We chose neural language models ELMo, BART, BERT, RoBERTa, and XLNet for their demonstrated ability to contextually adapt word meanings, enabling them to detect subtle semantic changes. As foundational benchmarks, traditional models such as CNN, NNLM, and BiLSTM were used to compare shallow versus deep contextualization strengths. Furthermore, graph-based models were employed to manage relational and taxonomical ambiguities

often neglected by transformer-only systems. By integrating these sequential, contextualized, and graph-based model groups, the study aims to pinpoint and evaluate critical architectural elements, including context awareness, subword-level processing, and semantic structures, in effectively addressing meaning conflation deficiency (MCD) problem in WVS WE models.

The study was further expanded to address the meaning conflation deficiency (MCD) issue in word vector models, where distinct meanings or concepts are incorrectly merged, causing ambiguity and reduced accuracy. To overcome this, several graph-based approaches were employed, including hierarchical graphs, domain-specific graphs, and integrated graph structures. The hierarchical, domain-specific, and integrated graphs offer a comprehensive method for clearing up meaning ambiguity. The integrated graph guarantees uniformity and clarity across domains; the domain-specific graph elucidates meaning in context; and the hierarchical graph aids in differentiating levels of abstraction. These graphs provide a strong framework for ensuring that words are understood in the right context and used correctly, avoiding misunderstandings.

In order to attain the best possible speed and efficiency during the training and testing stages of the models, the researchers utilized Google Colab with GPU acceleration in the experimental setup. Because Google Colab has a great deal of processing power, researchers can do extensive and drawn-out testing. For the machine learning tasks, its GPU (Graphics Processing Units) offering is crucial because these techniques greatly benefit from parallel processing, which drastically reduces computation time in comparison to CPUs (Central Processing Units).

5.7 Performance Tuning of Algorithms

5.7.1 Optimized BART-based

Algorithm 5.1 presents an optimized framework built on the BART architecture and outlines the full workflow for adapting and enhancing the transformer model to identify and reduce WVS WE MCD, a phenomenon in which multiple word senses are represented by a single undifferentiated vector. The algorithm serves as a tailored refinement strategy for BART's encoder-decoder structure and attention mechanisms, enabling more effective resolution of semantic ambiguity in morphologically rich, low-

resource languages. Critical design elements—such as highlighting the target word and applying morphology-aware preprocessing—were deliberately incorporated to align with the linguistic properties of SsaL.

Algorithm 5.1 Optimized BART

- 1: **Input:** A labelled corpus $D = (s_i, w_i, y_i)_{i=1}^N$,
where s_i is the contextual sentence for instance i , w_i is the target ambiguous word occurring in sentence s_i ,
and y_i is the corresponding annotated word-sense label for that occurrence.
Hyperparameters: maximum sequence length L , learning rate α , batch size B , epochs E , gradient accumulation G , warm-up ratio ρ , label smoothing ϵ , and early stopping patience P .

- 2: **Output:** Fine-tuned BART model M , predicted word senses, evaluation metrics, and visualisations.

Step 1: Data Preprocessing

- 3: Load the dataset from Excel and retain the columns: sentence, word, and sense.
- 4: For each annotated instance $(s_i, w_i, y_i) \in D$, verify that w_i occurs in sentence s_i and map that specific word occurrence to its corresponding sense label y_i .
- 5: Construct the model input by explicitly marking the target word inside its sentence context.

$$x_i = [TGT]w_i[/TGT]|| \quad s_i$$

- 6: Append each supervised pair (x_i, y_i) to the processed dataset $D' = (x_i, y_i)_{i=1}^N$.
- 7: Encode the categorical sense labels y_i into numerical class indices, and let K be the total number of distinct sense classes.
- 8: Split D' into training, validation, and test sets using stratified sampling:
 $D' \rightarrow D_{train}, D_{val}, D_{test}$.

Step 2: Tokenization.

- 9: Tokenize each input text x_i using the BART tokenizer with padding and truncation to maximum sequence length L .
- 10: Convert the tokenized outputs into tensors for model training and evaluation.

Step 3: Model Initialization

- 11: Load BART For Sequence Classification with K output classes and initialize the classification head for the target sense-label space.

Step 4: Optimization Setup

- 12: Configure the AdamW optimizer with learning rate α , batch size B , and gradient accumulation G .
- 13: Apply a cosine learning-rate scheduler with warm-up ratio ρ .
- 14: If class imbalance is present, compute class weights using the effective num-

ber of samples or apply Focal Loss.

- 15: Apply label smoothing with smoothing factor ϵ .

Step 5: Fine-Tuning

- 16: Train model M on D_{train} for E epochs.
- 17: After each epoch, evaluate performance on D_{val} and apply early stopping with patience P .
- 18: Optionally enable bf16 mixed precision, checkpointing, encoder layer freezing, and torch compile where supported.
- 19: Restore the best validation checkpoint.

Step 6: Final Evaluation

- 20: Evaluate the best model on D_{test} .
- 21: Compute Accuracy, Precision, Recall, and Weighted-F1.
- 22: Generate ROC curves (micro, macro, and per-class), confusion matrix, training/validation loss curves, positional embedding norms, and decoding behaviour samples.

End of the Optimized BART Algorithm

5.7.2 Optimized ELMo

Algorithm 5.2, representing the optimized ELMo approach, outlines a systematic method for adapting the ELMo contextual embedding model to directly address MCD in word vector space models. The optimized framework enhances semantic differentiation by ensuring that distinct word senses are not collapsed into a single generalized representation. This algorithm provides a detailed blueprint for modifying ELMo’s internal architecture so that it more effectively captures morphological cues and contextual dependencies—capabilities that are essential for accurately disambiguating polysemous words in morphologically rich and low-resource languages such as SsaL.

- 1: **Input:** A labelled corpus $D = (s_i, w_i, y_i)_{i=1}^N$,
 where s_i is the contextual sentence for instance i ,
 w_i is the target ambiguous word occurring in sentence s_i ,
 and y_i is the corresponding annotated word-sense label for that occurrence.
 Hyperparameters: batch size B , epochs E , initial learning rate η_0 , decay
 rate γ , decay steps s , label smoothing ϵ , dropout rate p , L2 coefficient λ ,
 tolerance δ , and early stopping patience P .
- 2: **Output:** Fine-tuned ELMo encoder E_θ with classifier head (W_1, b_1, W_2, b_2) ,
 predicted word senses, and evaluation metrics.

Step 1: Data Preprocessing

- 3: Load the dataset from Excel and retain the columns: sentence, word, and
 sense.
- 4: For each annotated instance $(s_i, w_i, y_i) \in D$, verify that w_i occurs in sen-
 tence s_i and map that specific word occurrence to its corresponding sense
 label y_i .
- 5: Construct the model input by explicitly marking the target word inside its
 sentence context.

$$x_i = [TGT]w_i[/TGT]|| \quad sentence_i$$

- 6: Append each supervised pair (x_i, y_i) to the processed dataset $D' = (x_i, y_i)_{i=1}^N$.
- 7: Encode the categorical sense labels y_i into numerical class indices, and let
 C be the total number of distinct sense classes.
- 8: Split D' into training, validation, and test sets using stratified sampling:
 $D' \rightarrow D_{train}, D_{val}, D_{test}$.

Step 2: Contextual Representation and Classifier.

- 9: Initialize the trainable ELMo encoder $E_\theta : \text{string} \rightarrow R^d$ with $d = 1024$.
- 10: Define the regularised classifier head over the ELMo representation:

$$h_i = E_\theta(x_i), \quad z_i = \text{ReLU}(W_1 h_i + b_1), \quad o_i = W_2 \text{Dropout}(z_i; p) + b_2, \quad \hat{p}_i = \text{softmax}(o_i)$$

Step 3: Loss Function and Learning Schedule

- 11: Apply label smoothing to build smoothed targets q_i with smoothing factor
 ϵ .
- 12: Use cross-entropy loss with L2 regularisation:

$$L = L_{CE} + \lambda \|W\|_2^2$$

13: Update the learning rate using staircase exponential decay:

$$\eta_t = \eta_0 \gamma^{\lfloor t/s \rfloor}$$

Step 4: Optimization

- 14: Initialize the Adam optimizer and train on mini-batches of size B.
- 15: For each epoch $e = 1$ to E, shuffle D_{train} , compute forward pass, backpropagate gradients, and update model parameters.
- 16: Evaluate validation loss on D_{val} after each epoch.
- 17: If validation loss improves by at least δ , save the checkpoint and reset the patience counter; otherwise, increment the counter.
- 18: Stop training when the patience counter reaches P and restore the best checkpoint.

Step 5: Final Evaluation

- 19: Evaluate the best model on D_{test} .
- 20: Compute Accuracy, Precision, Recall, Weighted-F1, confusion matrix, ROC, and AUC.

End of the Optimized ELMo Algorithm

5.7.3 Optimized XLNet

Algorithm 5.3 describes how XLNet was optimized for classifying ambiguous words in the sense-annotated corpus. Each data instance consists of a contextual sentence, a target ambiguous word occurring in that sentence, and its corresponding annotated sense label. The dataset is first loaded, and the target word is verified within its sentence. The input is then constructed by explicitly marking the target word in its sentence context so that the model can focus on the lexical item to be classified. The processed input-label pairs are encoded, divided into training, validation, and test sets using stratified sampling, and prepared with the XLNet tokenizer through padding, truncation, and a fixed maximum sequence length. The processed data are then converted into Hugging Face datasets for training and validation. XLNet is initialized for sequence classification with the required number of output classes, and training is carried out with evaluation and checkpoint saving at each epoch, weight decay, early stopping, and restoration of the best-performing model. When class imbalance is present, weighted

cross-entropy is applied. Overall, the algorithm provides a clear and reproducible procedure for adapting XLNet to contextual word-sense classification in the Sesotho sa Leboa corpus.

- 1: **Input:** A labelled corpus $D = (s_i, w_i, y_i)_{i=1}^N$,
 where s_i is the contextual sentence for instance i ,
 w_i is the target ambiguous word occurring in sentence s_i ,
 and y_i is the corresponding annotated word-sense label for that occurrence.
 Hyperparameters: learning rate $2 \cdot 10^{-5}$, batch size 8, epochs 5, maximum
 sequence length 128, weight decay 0.01, and early stopping patience 2.

- 2: **Output:**

- 3: Load the dataset from Excel and retain the columns: sentence, word, and
 sense.

Step 1: Data Preprocessing

- 4: For each annotated instance $(s_i, w_i, y_i) \in D$, verify that w_i occurs in sen-
 tence s_i and map that specific word occurrence to its corresponding sense
 label y_i .
- 5: Construct the model input by explicitly marking the target word inside its
 sentence context.

$$x_i = [TGT]w_i[/TGT]s_i$$

- 6: Append each supervised pair (x_i, y_i) to the processed dataset $D' = (x_i, y_i)_{i=1}^N$.
- 7: Encode the categorical sense labels y_i into numerical class indices, and let
 K be the total number of distinct sense classes.
- 8: Split D' into training, validation, and test sets using stratified sampling:
 $D' \rightarrow D_{train}, D_{val}, D_{test}$.

Step 2: Tokenization and Dataset Construction.

- 9: Initialize tokenizer $T \leftarrow \text{XLNetTokenizer.from_pretrained}('xlnet - base -$
cased').
- 10: Tokenize each input x_i using padding, truncation, and maximum sequence
 length 128.
- 11: Convert the processed inputs into Hugging Face datasets for training and
 validation.

Step 3: Model and Training Setup

- 12: Initialize $M \leftarrow \text{XLNet For Sequence Classification}$ with K output classes.
- 13: Set training arguments: evaluation per epoch, save per epoch, weight decay
 0.01, epochs 5, and $load_best_model_at_end = \text{True}$.
- 14: Enable early stopping with patience 2.
- 15: If class imbalance is present, compute class weights from training labels and
 use weighted cross-entropy in a custom loss function.

Step 4: Training

- 16: Train the model on D_{train} using the Trainer framework.
- 17: After each epoch, evaluate on D_{val} , compute logits, transform logits to probabilities by softmax, and obtain predicted labels by *argmax*.

Step 6: Final Evaluation

- 18: Restore the best checkpoint and evaluate on D_{test} .
- 19: Compute Accuracy, Precision, Recall, Weighted-F1, confusion matrix, ROC, and AUC plots.

End of the Optimized XLNet Algorithm Algorithm

5.7.4 Optimized CNN Using Dropout-Enhanced Variants

Algorithm 5.4, showcasing an optimized CNN enhanced with dropout regularization, addresses the challenge of MCD in WVS WE models. This approach modifies the CNN architecture to effectively learn distinctive features from a manually *sense-annotated SsaL corpus*. Emphasizing a feature-oriented and regularized deep learning strategy to mitigate WVS WE MCD, Algorithm 5.4 demonstrates the CNN's effectiveness in capturing local morphological and syntactic patterns. Additionally, the use of dropout during training enhances the model's robustness when dealing with sparse and imbalanced sense distributions.

Algorithm 5.4 Optimized CNN Using Dropout-Enhanced Variants

- 1: **Input:** A labelled corpus $D = (s_i, w_i, y_i)_{i=1}^N$,
where s_i is the contextual sentence for instance i ,
 w_i is the target ambiguous word occurring in sentence s_i ,
and y_i is the corresponding annotated word-sense label for that occurrence.
Hyperparameters: learning rate η ,
batch size B, maximum epochs E, patience P, dropout rates p_ℓ ,
L2 coefficient λ , and fixed input length L.

- 2: **Output:** Best CNN model parameters θ^* , predicted word senses, and evaluation metrics.

Step 1: Data Preprocessing

- 3: Load the dataset from Excel and retain the columns: sentence, word, and sense.
- 4: For each annotated instance $(s_i, w_i, y_i) \in D$, verify that w_i occurs in sentence s_i and map that specific word occurrence to its corresponding sense label y_i .
- 5: Construct the model input by explicitly marking the target word inside its sentence context.

$$x_i = [TGT]w_i[/TGT]s_i$$

- 6: Append each supervised pair (x_i, y_i) to the processed dataset $D' = (x_i, y_i)_{i=1}^N$.
- 7: Encode the categorical sense labels y_i into numerical class indices.
- 8: Tokenize each x_i and convert it into padded numerical sequences of fixed length L.
- 9: Split D' into training, validation, and test sets using stratified sampling:
 $D' \rightarrow D_{train}, D_{val}, D_{test}$.

Step 2: Model Definition.

- 10: Define the CNN network F_θ with embedding, convolution, pooling, and dense layers.
- 11: Insert dropout layers with rates p_ℓ after high-capacity layers to reduce overfitting.

Step 3: Objective Function

- 12: Define the mini-batch loss function:

$$L(\theta) = \frac{1}{B} \sum_{(x,y) \in \mathcal{B}} \text{CE}(y, \hat{y}(x; \theta, m)) + \lambda \sum_t \|W_t\|_2^2$$

- 13: where CE is the cross-entropy loss, m is the dropout mask, and W_t are

trainable weight matrices subject to L2 regularisation.

Step 4: Optimization

- 14: Initialize the Adam optimizer with learning rate η .
- 15: For each epoch $e = 1$ to E , shuffle D_{train} and form mini-batches $\mathcal{B}$ of size B .
- 16: For each mini-batch $\mathcal{B} \subset D_{train}$, sample dropout masks with keep probability $q_\ell = 1 - p_\ell$, compute predictions, compute $L(\theta)$, backpropagate gradients, and update parameters.

Step:5 Validation and Early Stopping

- 17: Evaluate the model on D_{val} without dropout using inference scaling.
- 18: Compute validation loss and performance metrics.
- 19: If validation loss improves by at least δ , save the checkpoint θ^* and reset the patience counter; otherwise increment the counter.
- 20: Stop training when the patience counter reaches P .

Step 6: Final Evaluation

- 21: Restore the best checkpoint θ^* and evaluate once on D_{test} .
- 22: Compute Accuracy, Precision, Recall, Weighted-F1, confusion matrix, ROC, and AUC.

End of the Optimized CNN Using Dropout-Enhanced Variants Algorithm

5.7.5 Bidirectional Long Short-Term Memory (Bi-LSTM)

In a Bi-LSTM, the recurrent layer is structured into two parallel components: one processes the input sequence in its original order, while the other processes the same sequence in reverse order. This bidirectional design enables the model to capture information from both past and future context within the sequence. At each time step, the outputs from the two LSTMs are typically concatenated (Aburass, Dorgham and Al, 2024). A cutting-edge variation of recurrent neural networks (RNNs) is called bi-directional long short-term memory (BiLSTM), which has the capacity to give words specific weights based on different situations. In order to effectively compensate for deep neural network shortcomings in acquiring local features and emphasizing the significance of individual words to the context as a whole, BiLSTM makes sure that context-based semantic association information is utilized to efficiently identify long-term relationships and extract the necessary background data for forecasting a token-sentence pair's complexity score (Aziz et al., 2022).

By integrating historical and prospective data at every level, Bi-LSTM improves prediction accuracy. Long-short-term memory (LSTM) is used in this neural network model, the forward LSTM, which records the antecedent context, and the backward LSTM, which records the subsequent context (Lestandy, 2023), as shown in Figure 5.3.

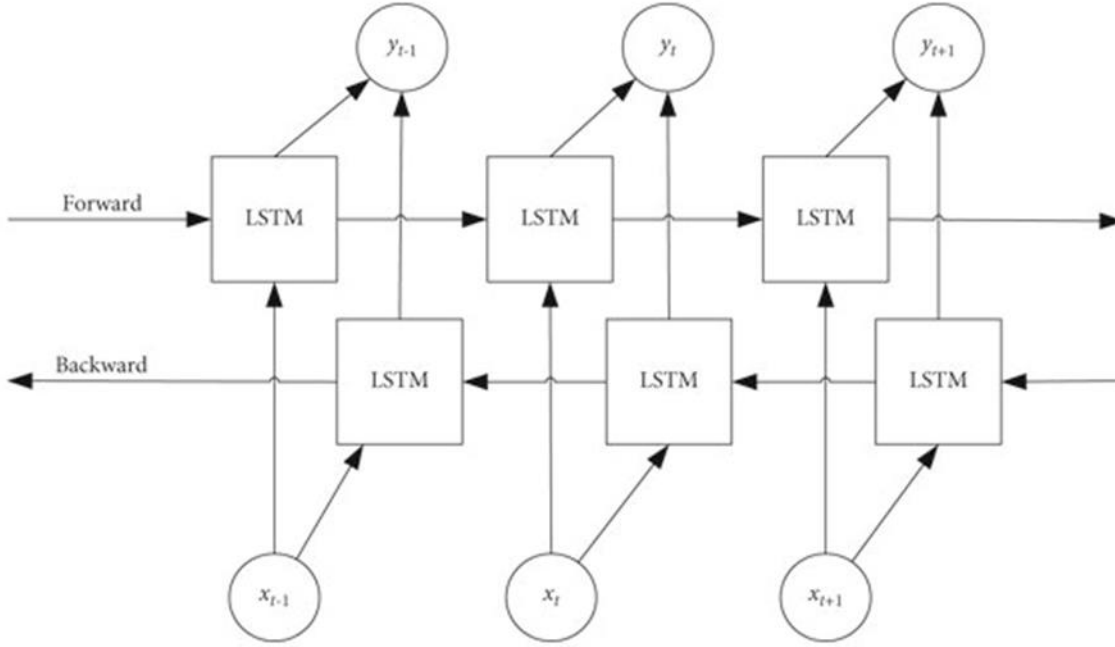


Figure 5.3: Bi-LSTM architecture (Lestandy, 2023)

At its core, the long short-term memory (LSTM) uses the hidden state to retain data from inputs that have already gone through it. Since the only inputs that unidirectional LSTM has seen are from the past, it can only retain knowledge from the past. Errors back-propagate during training to modify weights while preserving the training parameters. Equation (5.3 to 5.7) is used to calculate the forget, input, and output gates as well as the input cell state (Jain, Kumar and Garg, 2020; Ayetiran, 2022).

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (5.3)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (5.4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5.5)$$

$$s_t = \tanh(W_s \cdot [h_{t-1}, x_t] + b_s) \quad (5.6)$$

$$c_t = f_t * c_{t-1} + i_t * s_t \quad (5.6)$$

$$h_t = \tanh(c_t) * o_t \quad (5.7)$$

where:

- x_t is the t^{th} word vector that is it denotes the word representation of w_t
- $W_i W_f W_o W_s$ are terms representing model parameters
- $b_i b_f b_o b_s$ are terms representing bias vectors
- σ is the sigmoid function used to represent the gate activation function
- $\tanh$ is a hyperbolic tangent function

The Bi-LSTM (Jain, Kumar and Garg, 2020; Ayetiran, 2022), as shown in Equation (5.8), consists of two networks: one that reads data forward and the other that reads it backward. In this manner, both the past and future contexts are used to generate the output.

$$h_i = (\vec{h}_i \odot \overleftarrow{h}_i) \quad (5.8)$$

where:

- h_i is the output of the i th word
- $\odot$ is a concatenation function used to combine the two outputs
- $\vec{h}_i$ symbolizes the forward network output sequence, which is determined iteratively utilizing positive sequence inputs.
- $\overleftarrow{h}_i$ symbolizes the backward network output sequence, which is determined by reversing the inputs.

5.7.6 Attention-based BiLSTM (A-BiLSTM)

The information output from the BiLSTM's hidden layers is given a different focus through the use of an attention mechanism (Liu and Guo, 2019). By enabling the model to concentrate on the most pertinent segments of the input sequence during prediction, Attention-enhanced Bidirectional Long Short-Term Memory (A-BiLSTM) networks extend the standard BiLSTM architecture. For a given input sequence $X = \{x_1, x_2, \dots, x_T\}$, where x_t denotes the embedding of the t -th token, the

BiLSTM encodes the sequence by processing it simultaneously from the beginning to the end and in reverse order (Luong, Pham and Manning, 2015):

- Forward LSTM calculates hidden states $\vec{h}_t$ from $t = 1$ to T
- Backward LSTM computes hidden states $\tilde{h}_t$ from $t = T$ to 1

The final hidden state for each time step is computed in Equation (5.13) as:

$$y_t = \sigma[\vec{h}_t, \tilde{h}_t] \quad (5.9)$$

The context vector c , which is a weighted sum of all concealed states, is calculated by the attention mechanism $\{h_1, h_2, \dots, h_T\}$, which follows the steps:

1. Score computation

For each hidden state h_t , attention score e_t is calculated as:

$$e_t = v_a^T \tanh(W_a h_t + b_a) \quad (5.10)$$

where:

- W_a is a weight matrix,
- v_a is a weight vector,
- b_a is a bias vector
- Tanh is the hyperbolic tangent activation function

2. Normalization via Softmax

Use SoftMax to convert scores into probabilities α_t :

$$\alpha_t = \frac{\exp(e_t)}{\sum_{j=1}^T \exp(e_j)} \quad (5.11)$$

where α_t represents the attention weight for h_t

3. Context Vector

Calculate the weighted total of the hidden states to achieve the context vector c :

$$c = \sum_{t=1}^T \alpha_t h_t \quad (5.12)$$

Predictions are made by combining the final hidden state with the context vector c :

$$y = \text{Softmax} (W_o[c; h_T] + b_o) \quad (5.13)$$

where:

- W_o and b_o are the weight matrix and bias vector of the output layer,
- y is the predicted probability distribution over the output classes.

By focusing on the most pertinent portions of the input through an attention mechanism and capturing both forward and backward context, attention-based BiLSTM networks in Algorithm 5.5 improve the capacity to address WVS WE MCD. By concentrating on contextually relevant information, our algorithm successfully resolves WVS WE MCD by utilizing the advantages of BiLSTM and attention mechanisms.

- 1: **Input:** A sense-annotated corpus $D = (s_i, w_i, y_i)_{i=1}^N$,
where s_i is the contextual sentence for instance i ,
 w_i is the ambiguous target word occurring in sentence s_i ,
and y_i is the corresponding annotated word-sense label for that occurrence.

- 2: **Output:** Predicted word senses, trained model(s), and evaluation metrics.

Step 1: Data Preprocessing

- 3: Load the dataset and retain the fields: sentence, target word, and sense label.
- 4: For each annotated instance $(s_i, w_i, y_i) \in D$, map the target word w_i to its contextual sentence s_i and to the corresponding sense label y_i .

$$x_i = [TGT]w_i[/TGT]|| s_i$$

- 5: Construct one supervised sample (x_i, y_i) for each annotated occurrence and form the processed dataset $D' = (x_i, y_i)_{i=1}^N$.
- 6: Encode the categorical sense labels y_i into numerical class indices.
- 7: Apply the model-specific tokenizer or text encoder to each input x_i and generate fixed-format model inputs suitable for training.
- 8: Split the processed dataset into training, validation, and test sets using stratified sampling: $D' \rightarrow D_{train}, D_{val}, D_{test}$.

Step 2: Word Embedding Representation.

- 9: Load pretrained GloVe and FastText embeddings and create the embedding matrix.
- 10: Represent each processed input sequence x_i using the embedding matrix and assign a default vector to out-of-vocabulary items.

Step 3: Train Attention-Based BiLSTM Model

- 11: Define the BiLSTM model with an attention layer to focus on the most informative contextual features for sense prediction.
- 12: Train the model on D_{train} using the Adam optimizer and sparse categorical cross-entropy loss.
- 13: Monitor validation performance on D_{val} and retain the best model checkpoint.

Step 4: Model Evaluation

- 14: Evaluate the trained model on D_{test} and compute Accuracy, Precision, Recall, Weighted -F1, and the Confusion Matrix.

End of the Attention-Based BiLSTM Algorithm

5.7.7 Neural Network Language Model (NNLM)

It is well known that the NNLM performs more accurately than conventional n-gram language models. In order to estimate the likelihood of a string of words or tokens, NNLMs usually use mathematical concepts from probability, linear algebra, and optimization. The goal of a neural network language model is to calculate the likelihood of a word sequence, $w_1, w_2, \dots, w_T$. The chain rule of probability, as stated in Equation (5.14), can be used to break down this likelihood.

$$P(w_1, w_2, \dots, w_T) = \prod_{t=1}^T P(w_t | w_1, w_2, \dots, w_T) \quad (5.14)$$

Algorithm 5.6 uses deep learning models (BiLSTM & NNLM) to address the MCD problem in WVS WE models. The NNLM model offers a baseline sense prediction, while the BiLSTM model records contextual dependencies.

- 1: **Input:** Sense-annotated corpus $D = (s_i, w_i, y_i)$,
 where s_i is the contextual sentence for instance i , w_i is the target ambiguous
 word occurring in sentence y_i
 and y_i is the corresponding annotated word-sense label for that occurrence.

$$S = \{s_1, s_2, \dots, s_n\}$$

$$W = \{w_1, w_2, \dots, w_n\}$$

$$Y = \{y_1, y_2, \dots, y_n\}$$

- 2: **Output:** Predict word senses using the trained BiLSTM and NNLM models
 together with evaluation metrics.

Step 1: Data Preprocessing

- 3: Load the dataset from Excel and retain the columns sentence, word, and
 sense.
- 4: For each annotated instance (s_i, w_i, y_i) in D , map the target word w_i in
 sentences s_i to its corresponding sense label y_i .
- 5: Construct the model input by explicitly marking the target word inside the
 sentence:

$$x_i = [\text{TGT}] w_i [/\text{TGT}] \parallel s_i$$

- 6: Tokenize each input sequence x_i , encode the categorical sense labels y_i into
 numerical class indices, and convert the texts into padded sequences of fixed
 length L .
- 7: Split the processed dataset into training, validation, and test subsets using
 stratified sampling.

Step 2: Model Training (BiLSTM and NNLM).

- 8: Train the BiLSTM model on the training data using the padded sequences
 and encoded labels.
- 9: Train the NNLM model on the same processed training data using the same
 input-target mapping from x_i to y_i .
- 10: Use the trained BiLSTM model and trained NNLM model to predict the
 word senses for the validation and test instances.

Step 3: Model Evaluation

- 11: Compute Accuracy, Precision, Recall, Weighted-F1 score, and compare the
 predictive performance of both models.

5.7.8 BERT

The Transformer framework serves as the foundation for the Encoder-Decoder structure implemented by the BERT model. Although Transformer's structure is far more intricate than Attention's, it also employs the Encoder-Decoder framework, which is the same as the Attention model. There are six encoders in its encoder module, and six decoders in its decoder module. A feed-forward neural network and a self-attention layer are components of every encoder (Man & Lin, 2021). To capture different kinds of interactions between tokens, BERT employs multi-head attention (Devlin et al., 2019). By

using several self-attentions “heads”, each with a unique set of learnt weight matrices, multi-head attention enables the model to concurrently attend to data from various subspaces at various locations. Equation (5-21) is used to concatenate and project each head’s output back to the hidden size:

$$Multihead(Q, K, V) = Concat(head_1, \dots, head_h)W_O \quad (5-21)$$

where:

- $head_i = Attention(Q_i, K_i, V_i)$
- h is the number of attention heads,
- W_O is a learned weight matrix for the output

5.7.9 Robustly Optimized BERT Approach (RoBERTa)

An enhanced version of BERT called RoBERTa uses optimised training methods to enhance language representation. Its training methodology, which includes using more data, larger batch sizes, and longer training times, is the main area of difference from BERT, which is its foundation. Similar to BERT, RoBERTa uses a combination of token, segment, and positional embeddings to represent each input sequence. For a given sequence of tokens $w_1, w_2, \dots, w_T$, the embedding for each token w_i , is calculated in Equation (5.15) as:

$$x_i = e_{w_i} + e_{segment(i)} + e_{position(i)} \quad (5.15)$$

where:

- e_{w_i} is the embedding of the token w_i ,
- $e_{segment(i)}$ is the segment embedding
- $e_{position(i)}$ is the positional embedding, added to encode the token’s position in the sequence.

Similar to BERT, RoBERTa makes use of several transformer block levels. A feedforward neural network comes after self-attention in each transformer layer. For a specific token, the self-attention w_i at the layer l is calculated in Equation (5.16) as:

$$Attention(Q_i, K, V) = \sum_{j=1}^n \alpha_{i,j} V_j \quad (5.16)$$

where:

- $Q_i = W_Q^{(l)} h_i^{(l-1)}$ is the query vector for token w_i in layer l ,
- $K_j = W_K^{(l)} h_j^{(l-1)}$ is the key vector for token w_j ,
- $V_j = W_V^{(l)} h_j^{(l-1)}$ is the value vector for token w_j ,
- $\alpha_{i,j}$ is the attention score, computed by Equation (5.17):

$$\alpha_{i,j} = \text{softmax}\left(\frac{Q_i \cdot K_j^T}{\sqrt{d_k}}\right) \quad (5.17)$$

where d_k is the important vectors' dimensionality, and softmax makes sure that the scores add up to one.

A position-wise feedforward network in Equation (5.18), such as the Gaussian Error Linear Unit (GELU), which consists of two linear transformations with a non-linear activation in between, processes each self-attention output:

$$FFN(h_i^{(l)}) = GELU(W_1^{(l)} h_i^{(l)} + b_1^{(l)}) W_2^{(l)} + b_2^{(l)} \quad (5-18)$$

where:

- $W_1^{(l)}$ and $W_2^{(l)}$ are weight matrices,
- $b_1^{(l)}$ and $b_2^{(l)}$ are bias terms

RoBERTa (Liu *et al.*, 2019) predicts masked tokens in a sentence using Equation (5-19) the Masked Language Modeling (MLM) objective. Every masked token w_{mask} in a sequence, to optimise the likelihood of the correct word w_{true} , the model generates a probability distribution throughout the vocabulary.

$$P(w_{true}|w_{mask}) = \text{softmax}(W_{out} h_{mask} + b_{out}) \quad (5.19)$$

where, h_{mask} denotes the hidden state of the masked token after it has passed through the transformer layers, and W_{out} and b_{out} are the output weight and bias. Equation (5.20), calculates the MLM loss function for each sequence:

$$L_{MLM} = -\sum_{i \in \text{masked}} \log P(w_{true,i} | h_{mask,i}) \quad (5.20)$$

where the total of all the sequence's masked tokens is calculated.

5.7.10 Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA)

By completing a pre-training task, the effective and precise pre-training language model ELECTRA (Clark et al., 2020; Devlin et al., 2019) learns deep and bi-directional word vectors. After training the NLP downstream task with a Masked Language Model (MLM), ELECTRA substitutes the predictions from the NLP downstream task for words. A transformer-based model called ELECTRA outperforms BERT's conventional MLM technique in terms of efficiency and performance. ELECTRA uses a Replaced Token Detection (RTD) task for training instead of masking tokens, which is more effective in representation learning and more computationally efficient.

For every token in the input sequence, ELECTRA combines positional, segment, and token embeddings, exactly as does BERT for a token input sequence, as defined in Equation (5.21). After receiving the “corrupted” sequence produced by the generator, the discriminator (the primary ELECTRA model) categorizes each token as either genuine (original) or fraudulent (replaced). For each token w_i , the discriminator in Equation (5.22) computes as:

$$P(real|h_i) = \sigma(W_{disc}h_i + b_{disc}) \quad (5.22)$$

where:

- σ is the sigmoid activation function
- W_{disc} and b_{disc} are learned weights and biases for the discriminator.

Accurately predicting whether each token is real or replaced is the discriminator's goal. The loss of the discriminator, L_{disc} , is defined in Equation (5.23) as:

$$L_{disc} = - \sum_{i=1}^n (y_i \log P(real|h_i) + (1 - y_i) \log (1 - P(real|h_i))) \quad (5.23)$$

where:

- $y_i = 1$ if the token is real, and $y_i = 0$ if it has been replaced by the generator.

The generator loss in Equation (5.23), L_{gen} is thus similar to BERT's MLM loss in Equation (5.24):

$$L_{gen} = -\sum_{i \in \text{masked}} \log P(w_{true,i} | h_{mask,i}) \quad (5.24)$$

The final loss for ELECTRA is calculated by combining the discriminator and generator losses in Equation (5.25). Since the discriminator is the main focus, the generator loss is usually down weighted, although the overall loss can be weighted between the discriminator and generator losses:

$$L_{ELECTRA} = L_{disc} + \lambda L_{gen} \quad (5.25)$$

where λ is a weighting factor (usually a minor amount, because the discriminator is the main focus) that balances the generator and discriminator losses.

Algorithm 5.7 applies transformer-based models (BERT, RoBERTa, ELECTRA) to resolve MCD problem in WVS WE by leveraging the CAWE model.

Algorithm 5.7 RoBERTa, BERT, and ELECTRA

- 1: **Input:** Sense-annotated corpus $D = (s_i, w_i, y_i)$,
 where s_i is the contextual sentence for instance i , w_i is the target ambiguous word occurring in sentence y_i
 and y_i is the corresponding annotated word-sense label for that occurrence.

$$S = \{s_1, s_2, \dots, s_n\}$$

$$W = \{w_1, w_2, \dots, w_n\}$$

$$Y = \{y_1, y_2, \dots, y_n\}$$

- 2: **Output:** Predict word senses using the fine-tuned RoBERTa, BERT, and ELECTRA models together with evaluation metrics.

Step 1: Data Preprocessing

- 3: Load the dataset from Excel and retain the columns sentence, word, and sense.
- 4: For each annotated instance (s_i, w_i, y_i) in D , map the target word w_i in sentences s_i to its corresponding sense label y_i .
- 5: Construct the model input by explicitly marking the target word inside the sentence:

$$x_i = [\text{TGT}] w_i [/\text{TGT}] \parallel s_i$$

- 6: Encode the categorical sense labels y_i into numerical class indices and tokenize each input sequence x_i using the appropriate Hugging Face tokenizer for BERT, RoBERTa, and ELECTRA.
- 7: Convert the tokenized outputs into transformer-compatible inputs, including input IDs, attention masks, and any required segment information, then split the dataset into training, validation, and test subsets using stratified sampling.

Step 2: Model Training (RoBERTa, BERT, ELECTRA).

- 8: Fine-tune the BERT model on the processed training data for word-sense prediction.
- 9: Fine-tune the RoBERTa model on the same processed training data for word-sense prediction.
- 10: Fine-tune the ELECTRA model on the same processed training data for word-sense prediction.
- 11: Use the trained transformer models to predict the target sense $\hat{y}_i$ for each test instance x_i .

Step 3: Model Evaluation

12: Compute Accuracy, Precision, Recall, Weighted-F1 score, and compare the predictive performance of RoBERTa, BERT, and ELECTRA.

5.7.11 Hierarchical graph

The purpose of hierarchical graph (Ying et al., 2018) models is to represent multiscale or hierarchical relationships in graph-structured data. The goal of these models is to learn graph representations at various granularities, which is particularly helpful for work that involves hierarchical connections, massive graphs, or multi-scale phenomena.

5.7.12 Domain-specific graph

Domain-specific graph (Vashishth et al., 2020) models are customized graph-based methods intended to capture specific structural, relational, and contextual data relevant to a certain topic, including knowledge graphs, social networks, or biomedical networks. These models frequently include domain-specific improvements, including customized attention mechanisms, unique embeddings, or extra constraints, with conventional graph neural network (GNN) methodologies.

5.7.13 Integrated graph

The purpose of integrated graph (Wang et al., 2019) models is to integrate data from several graph modalities or structures into a single representation that makes use of diverse data sources. This integration frequently entails fusing graph-based representations with various graph types (such as social, biological, or knowledge networks) or multimodal data (such as text or images). Graph-based methods are used in Algorithm 5.8 to address MCD in WVS WE. The technique effectively captures and disambiguates word senses by utilizing Hierarchical Graph, Domain-Specific Graph, and Integrated Graph models.

Algorithm 5.8 Hierarchical Graph, Domain-Specific Graph, and Integrated Graph Models

- 1: **Input:** Sense-annotated corpus $D = (s_i, w_i, y_i)$,
 where s_i is the contextual sentence for instance i , w_i is the target ambiguous word occurring in sentence y_i
 and y_i is the corresponding annotated word-sense label for that occurrence.
 - 2: **Output:** Predict word senses using the Hierarchical Graph, Domain-Specific Graph, and Integrated Graph models together with evaluation metrics.
- Step 1: Data Preprocessing**
- 3: Load the dataset from Excel and retain the columns sentence, word, and sense.
 - 4: For each annotated instance (s_i, w_i, y_i) in D , map the target word w in sentence s to its corresponding sense label y_i .
 - 5: Normalize the text, remove irrelevant tokens where appropriate, tokenize each sentence, and encode the categorical sense labels y_i into numerical class indices.
 - 6: Construct graph-ready representations from the processed instances so that each target word occurrence remains linked to its source sentence and annotated sense.
 - 5: Load the dataset and retain the fields: sentence, target word, and sense label.
 - 6: For each annotated instance $(s_i, w_i, y_i) \in D$, map the target word w_i to its contextual sentence s_i and to the corresponding sense label y_i .

$$x_i = [TGT]w_i/[TGT]|| \quad s_i$$

- 7: Construct one supervised sample (x_i, y_i) for each annotated occurrence and form the processed dataset $D' = (x_i, y_i)_{i=1}^N$.
- 8: Encode the categorical sense labels y_i into numerical class indices.
- 9: Apply the model-specific tokenizer or text encoder to each input x_i and generate fixed-format model inputs suitable for training.
- 10: Split the processed dataset into training, validation, and test sets using stratified sampling: $D' \rightarrow D_{train}, D_{val}, D_{test}$.

Step 2: Word Embedding Representation.

- 11: Load pretrained GloVe and FastText embeddings and create the embedding matrix.
- 12: Represent each processed input sequence x_i using the embedding matrix and assign a default vector to out-of-vocabulary items.

Step 3: Train Attention-Based BiLSTM Model

- 13: Define the BiLSTM model with an attention layer to focus on the most informative contextual features for sense prediction.
- 14: Train the model on D_{train} using the Adam optimizer and sparse categorical cross-entropy loss.
- 15: Monitor validation performance on D_{val} and retain the best model checkpoint.

Step 4: Model Evaluation

- 16: Evaluate the trained model on D_{test} and compute Accuracy, Precision, Recall, Weighted -F1, and the Confusion Matrix.

End of the Attention-Based BiLSTM Algorithm

5.8 Chapter Summary

This chapter details the comprehensive methodological framework used to explore and address the issue of MCD problem in WVS WE, a language characterized by its rich morphology and limited resources. Employing a computational experiment design, the study creates a link between theoretical linguistic advancements and practical model applications. Chapter 5 starts by placing the research within context, detailing its aim to critique existing word embeddings and suggesting context-sensitive, hybrid models capable of differentiating the various senses of polysemous words. The chapter outlines the computational experimental design that informs each phase of the methodology, highlighting the synthesis of linguistic theory with deep learning frameworks. Following this, the research purpose is further expounded upon to illustrate the drive to alleviate semantic ambiguity in low-resource environments. The research design emphasizes the linguistic concentration on SsaL, alongside the development of a *sense-annotated SsaL corpus*. The text explains the corpus annotation and preprocessing methods (including tokenization and stopword removal), and feature extraction that accommodate agglutinative morphology. Central to this chapter are the data analysis procedures, which describe the model architecture, training processes, and embedding methods. Chapter 5 discusses models (such as BART, RoBERTa, ELECTRA, NNLM, ELMo, BiLSTM, BERT, CNN, XLNet, and graph representations) in terms of their computational configuration, optimization techniques, and empirical testing. Various evaluation metrics such as precision, recall, F1-score, and AUC are used, and traditional embeddings are compared with contextual embeddings through experimental

comparisons. Additionally, statistical tests such as McNemar's test are employed to verify improvements. The chapter concludes with a graph-based algorithms, comparative studies on BERT variants, and a discussion on how the study maintains transparency, ensures reproducibility, and adheres to ethical standards.

CHAPTER 6 : EXPERIMENTATION, RESULTS, AND DISCUSSIONS

6.1 Introduction

The chapter explores experiments and findings to address MCD in SsaL using specific models and techniques. This chapter begins with a description of the computing infrastructure and software used to replicate experiments; and introduces a *sense-annotated SsaL corpus* with an exploratory data analysis (EDA) focusing on word frequencies, sense variations, and class disparities. The chapter reports experimental outcomes, compares baseline and improved architectures with standard evaluation criteria using tools such as confusion matrices and ROC curves; and concludes with insights relevant to word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) mitigation in SsaL. Section 6.2 describes the experimentation platform and environment using Google Colab with Python 3, a T4 GPU, and key libraries for implementation and visualization. Section 6.3 presents the dataset description of the *sense-annotated SsaL corpus* for word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) analysis, highlighting its schema components and traits for modelling. Section 6.4's EDA uncovers vital patterns for model formulation, addressing word frequency distribution, sense diversity, class imbalance; the section includes a data overview summary and preprocessing measures. Section 6.5 discusses experimental results and considers findings for optimized BART, ELMo, XLNet, and CNN versus baselines, with metrics such as accuracy and F1-score highlighting improvements in WVS WE MCD strategies. Further subsections evaluate graph-based models, examining class-specific performance and morphological/semantic constraints. Section 6.7 summarizes key findings from the platform setup, dataset, EDA, preprocessing, and experimental evaluations, including graph-based models, to support the thesis conclusions.

6.2 Experimentation Platform and Environment

An experimental environment was established to enable the training, evaluation, and comparison of diverse computational models and embedding techniques. To expedite deep learning tasks, all tests were conducted on Google Colab, a high-performance computing platform, using Python 3 as the runtime and a T4 GPU as the hardware accelerator. The implementation used Python (v3) and essential libraries such as Scikit-learn, TensorFlow, PyTorch, Transformers, Gensim, and Pandas. Matplotlib and Seaborn facilitated the creation of exploratory data analysis (EDA) plots, while Colab Notebook was used for interactive development and visualization.

All experiments were conducted in Google Colab using a paid subscription plan with T4 GPU acceleration enabled. The study used a cloud-based Jupyter notebook environment to train and evaluate the proposed models. Google Colab allocates computational resources dynamically, the exact hardware made available during a session vary. Paid Colab subscription plans provide access to premium GPUs and high-memory runtimes, exact accelerator type and memory allocation are subject to availability and the user's compute balance.

6.3 Dataset Description

This dataset is a structured corpus annotated with senses, consisting of 3,289 entries., named the *sense-annotated SsaL corpus*. Each entry includes a sentence, a target word within that sentence, and a corresponding sense label. The SsaL corpus is particularly designed for tasks such as WSD, semantic analysis, evaluating contextual embeddings, and word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD), especially for low-resource, morphologically rich languages such as SsaL. The dataset features four key columns: *sentence_id* (a distinct integer number as a unique identifier for each sentence), *sentence* (the complete text in which the ambiguous word appears), *word* (the target word to be disambiguated), and *sense* (the sense label based on context).

The 20 most common senses show obvious frequency peaks, suggesting a core set of meanings that recur in various contexts, possibly reflecting frequent semantic roles or syntactic patterns in the language corpus. Regarding ambiguous words, the dataset displays several tokens that frequently occur in various contexts, showing high polysemy. The leading 20 ambiguous words are notably frequent, and represent key lexical items selected for their semantic depth and relevance in daily use. These words are crucial for disambiguation models; their frequency patterns can guide prioritization during model evaluation and ablation studies. As for sentence traits, the average sentence length is around 8.15 tokens, with most sentences ranging from 6 to 10 tokens.

The shortest sentence consists of 3 tokens, whereas the longest extends to 23 tokens. This compact range of sentence lengths makes the dataset suitable for deep-learning models depending on fixed-length input sequences and attention windows. The relatively short sentence spans also reduce concerns of truncation and padding during preprocessing. Finally, the dataset’s vocabulary size, after tokenizing all lowercase words from the sentence column, is about 2,914 unique tokens. This size is manageable for both conventional vectorization methods (such as Bag-of-Words and TF-IDF) and neural word embeddings (such as Word2Vec and FastText). The vocabulary likely comprises root forms, affixed words, and morphologically rich tokens, which reflects the agglutinative characteristics of SsaL; and indicates the need for subword modelling or morphological preprocessing in further tasks.

The experimental corpus used in this study is a purpose-built, manually sense-annotated Sesotho sa Leboa dataset developed specifically for evaluating meaning conflation deficiency in a low-resource, morphologically rich language context. The corpus consists of contextual sentence instances built around selected ambiguous lemmas and their annotated sense labels. Although the corpus is relatively small in comparison with large-scale NLP benchmarks, it is appropriate for this study because it is a focused, high-quality resource designed for supervised sense classification. In low-resource WSD research, the scarcity of expert-annotated lexical-semantic datasets makes such targeted corpora both necessary and valuable. The corpus provides sufficient contextual variability, multiple examples per target lemma, and broad sense representation to support the evaluation of the proposed MCD models.

Its adequacy therefore lies in its relevance, annotation quality, and alignment with the study objective rather than in size alone.

6.4 Exploratory Data Analysis (EDA)

One essential phase in the data science process is exploratory data analysis (EDA), which focuses on examining and comprehending a dataset's properties. Through a variety of statistical and visual analysis approaches, EDA enables data scientists to find patterns, identify anomalies, verify assumptions, and test hypotheses. EDA is used to assist in question formulation, direct additional data processing, and to assist in choosing the best modelling approaches. Table 6-1 provides a summary of descriptive statistics from the *sense-annotated SsaL corpus*. To comprehend the structure, frequency patterns, and sense distribution of the annotated corpus—which serves as the foundation for the MCD problem in word vector space (WVS) word embedding (WE) models created for SsaL—exploratory data analysis, or EDA, was carried out. The EDA sought to learn more about polysemy, class disparity, and ambiguity.

Table 6-1 provides an overview of the sense-annotated SsaL corpus, presenting key metrics such as the overall sentence count, tokens, annotations, sense types, lemmas, as well as the ambiguity level.

The dataset comprises of 3,217 annotated entries, 2,774 unique sentences, 29 lemmas, and 153 distinct sense labels. An average ambiguity level of 5.28, calculated using Equation (6.1), indicates that each word in the corpus has around 5.28 distinct meanings on average. This indicates a significant level of polysemy, several words in the language possessing various meanings contingent upon their context. A mean ambiguity score of 5.28 indicates a corpus of considerable semantic complexity. This underscores the need for sophisticated WSD models; and bolsters the research objective of addressing word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) using context-aware and hybrid graph-transformer designs.

$$\text{Average Ambiguity} = \frac{\sum_{i=1}^n \text{Number of senses for word } i}{n} \quad (6.1)$$

where:

- n is the number of unique target words in the corpus
- The numerator is the total number of distinct senses assigned to all target words

Table 6.1: Dataset Description Summary

Dataset Component	Summary
Corpus type	Manually sense-annotated corpus for Sesotho sa Leboa
Experimental focus	Meaning Conflation Deficiency of ambiguous target lemmas in context
Corpus size	3,217 annotated entries drawn from 2,774 unique sentence contexts
Lexical coverage	29 normalised target lemmas
Sense coverage	153 normalised sense categories
Structure	Each record contains a sentence identifier, contextual sentence, target word, and annotated sense label
Average ambiguity level	5.28

This dataset serves as the underlying data for training the CAWE model and acts as a benchmark for both model development and performance evaluation.

6.4.1 Word Frequency Distribution

The dataset’s analysis showed that there were wide variations in the number of occurrences of each ambiguous word. The term “šupa”, for instance, was used 581 times, whilst some words were used just once. With a large standard deviation and an average frequency per word of about 78 occurrences, the distribution was skewed, suggesting that a small number of terms predominate in the *sense-annotated SsaL corpus*.

6.4.2 Sense diversity

Depending on its context, each ambiguous word was given multiple sense annotations: some words have up to 22 distinct senses. SsaL has a high degree of polysemy, as seen by the average of 4.7 senses per word. This demonstrates how intricate word meanings are, and how automated algorithms run the risk of confusing meanings.

6.4.3 Class imbalance

There was a disproportionately higher frequency of the top five most common senses, much higher than the others, such as those that indicated temporal, spatial, or numerical interpretations. This imbalance was further supported by visualizations such as box plots and histograms. Because machine learning algorithms tend to reward dominant senses during training, such skewed distributions present a hurdle.

6.4.4 Data Overview Summary Table

Table 6.2 presents the distribution of ambiguous lemmas in the Sense-Annotated Sesotho sa Leboa corpus by showing each lemma, its frequency of occurrence, and the number of associated senses. The table shows that the corpus is unevenly distributed across lemmas, with a small number of words contributing a large proportion of the annotated instances. For example, *lewa* (609 occurrences), *šupa* (581), *foła* (432), and *mafahla* (405) are the most frequent lemmas, while others such as *leleme* (2), *sola* (2), *tseba* (2), and *noka* (3) occur very rarely. This indicates that the corpus contains both high-frequency and low-frequency ambiguous items, reflecting the natural imbalance often found in manually developed lexical-semantic datasets for low-resource languages.

The table also shows substantial variation in semantic complexity across lemmas. Some ambiguous words have relatively few associated senses, such as *bona* (2), *loma* (2), *kgoro* (2), and *tala* (2), whereas others exhibit much richer sense variation, such as *foła* (22 senses), *ikgama* (17 senses), and *lewa* (13 senses). This variation suggests that the corpus captures different levels of ambiguity, from relatively simple polysemous items to highly conflated lexical forms that require stronger contextual modelling.

Consequently, the table demonstrates that the corpus is not only diverse in terms of frequency distribution, but also in the degree of semantic ambiguity represented across target lemmas.

Table 6.2 shows that the Sense-Annotated SsaL corpus consists of 29 unique ambiguous lemmas and 3,217 annotated occurrences, making it a focused but semantically varied resource for investigating Meaning Conflation Deficiency in Sesotho sa Leboa. The distribution presented in the table is important for interpreting model performance, since highly frequent lemmas may contribute more strongly to learning, while rare lemmas and highly polysemous lemmas may remain more difficult to classify accurately. The table therefore provides an essential descriptive foundation for understanding both the strengths and the limitations of the corpus used in the study.

Table 6.2: Distribution of Ambiguous Lemmas, Frequencies, and Associated Sense Counts in the Sense-Annotated SsaL Corpus

Ambiguous word	Total frequency of occurrence	Consolidated number of Senses
lewa	609	13
šupa	581	4
foła	432	22
mafahla	405	6
anega	228	9
nyaka	200	4
lefela	90	8
dihlare	65	5
matlakala	56	9
makhura	54	3
ikgama	53	17
kgoro	52	2
lapa	50	5
kgopolo	46	3
tshela	39	5
bona	36	2
loma	33	2

mogolo	31	5
molala	31	3
hlogo	27	6
nvetše	21	2
gola	19	4
lema	19	3
sega	17	2
tala	14	2
noka	3	3
leleme	2	2
sola	2	2
tseba	2	2

The bar chart in Figure 6.1, shows the analysis of the sense-annotated Sesotho sa Leboa corpus which depict the five lemmas with the highest number of associated senses are fola (22 senses), ikgama (17 senses), lewa (13 senses), anega (9 senses), and matlakala (9 senses). These lemmas represent the most semantically complex items in the corpus and are particularly important for investigating Meaning Conflation Deficiency, because they illustrate cases where a single lexical form is associated with multiple distinct meanings. Their presence in the corpus increases the linguistic richness of the dataset and provides a more demanding test for models intended to separate conflated meanings in context.

Among these, fola is the most polysemous lemma, with 22 associated senses, indicating a high degree of semantic variation across different contextual uses. This is followed by ikgama with 17 senses and lewa with 13 senses, both of which also demonstrate substantial lexical ambiguity. The lemmas anega and matlakala, each with 9 senses, further contribute to the diversity of semantic distinctions represented in the corpus. The inclusion of such highly polysemous lemmas is important because they create a more realistic and challenging evaluation setting for computational models, especially in a morphologically rich language such as Sesotho sa Leboa, where contextual and morphological cues strongly influence meaning interpretation.

From an analytical perspective, these top five lemmas are central to the study because they are more likely to expose weaknesses in static or shallow representations that collapse multiple meanings into a single vector space representation. In contrast, models capable of capturing contextual dependencies and fine-grained semantic variation are expected to perform better on these lemmas. Therefore, the distribution of highly polysemous lemmas in the corpus not only confirms the relevance of the dataset for Meaning Conflation Deficiency research, but also helps explain why some models require stronger contextual and morphology-aware mechanisms to achieve reliable semantic discrimination.

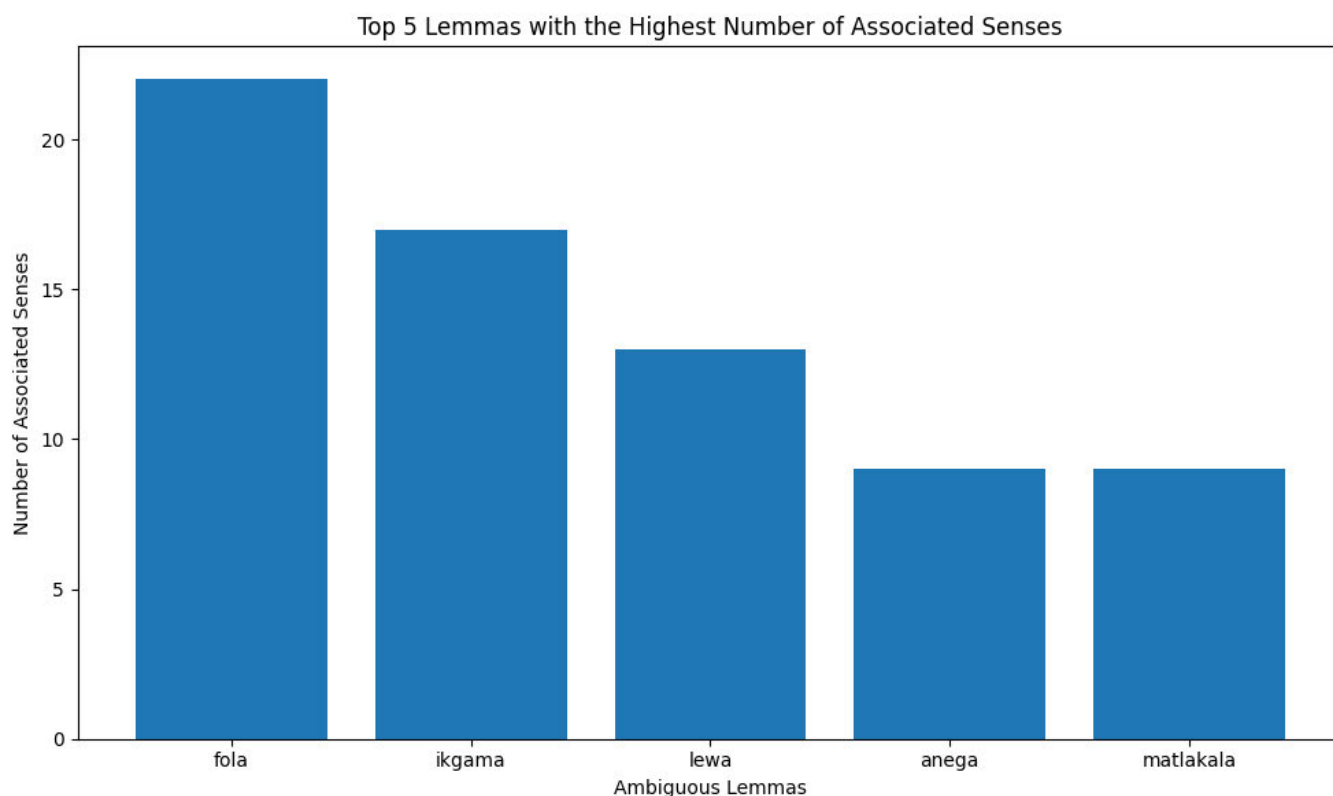


Figure 6.1: Frequent senses in Sense-Annotated SsaL Corpus

Important features of the *sense-annotated SsaL corpus* utilized to address the MCD problem in WVS WE models are revealed by descriptive statistics, as shown in Table 6.3. Each ambiguous phrase appears in the dataset roughly 78 times on average, although this varies greatly: some words appear only once, while the most frequent word appears 581 times. A significant imbalance in word frequency is indicated by this high standard deviation (~ 139.3). Furthermore, each word has an average of 4.7 distinct meanings; however, some words have as little as one sense and others as many as 22. These

statistics demonstrate the corpus diversity and semantic richness as well as the difficulties caused by sense variability and class imbalance, both of which are important factors to take into account when creating reliable word sense disambiguation models.

Table 6.3: Descriptive Statistics Summary

Metric	Value
Mean Instances per Word	78.46
Min Instances	1.0
Max Instances	581.0
Std Dev (Instances)	139.3
Mean Unique Senses	4.7
Max Unique Senses	22.0
Std Dev (Unique Senses)	4.383

The Histogram of Word Instance Counts, shown in Figure 6.2, shows that, while the majority of ambiguous words appears less than 100 times, a small number of words are highly represented (up to 581 instances), suggesting possible imbalance.

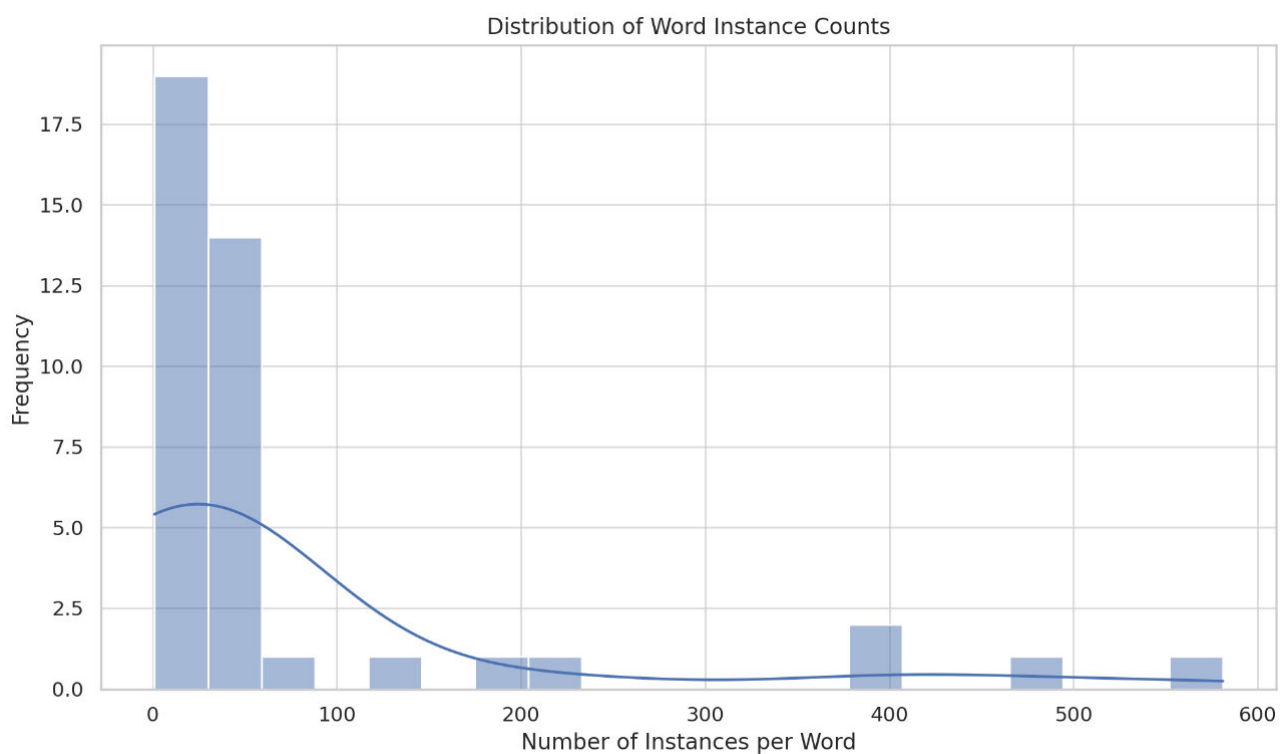


Figure 6.2:Distribution of Word Instances Counts

Some words exhibit extremely high polysemy (e.g., 13–22 senses), which increases the complexity of disambiguation, even though Figure 6.3 shows the histogram of unique senses per word, in which many words have two to six senses.

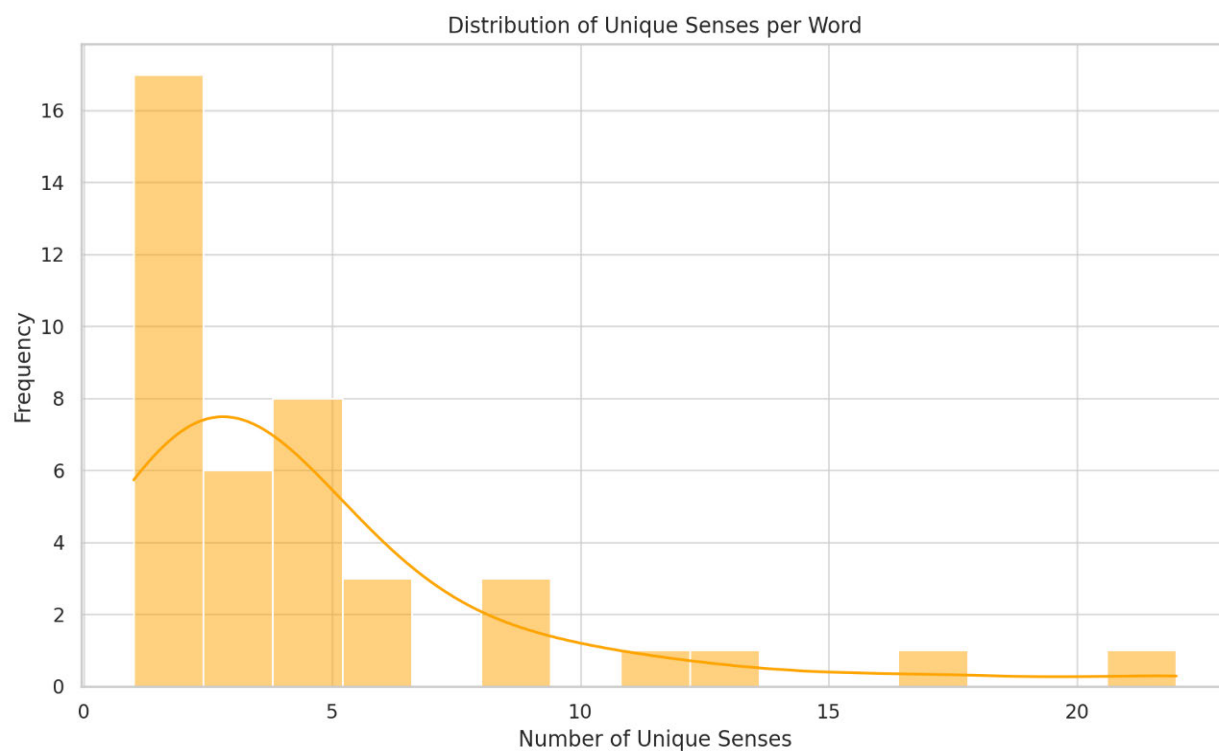


Figure 6.3:Distribution of unique senses per word

Figure 6.4 depicts the Box Plot which illustrates the spread and presence of outliers in both instance frequency and sense variability.

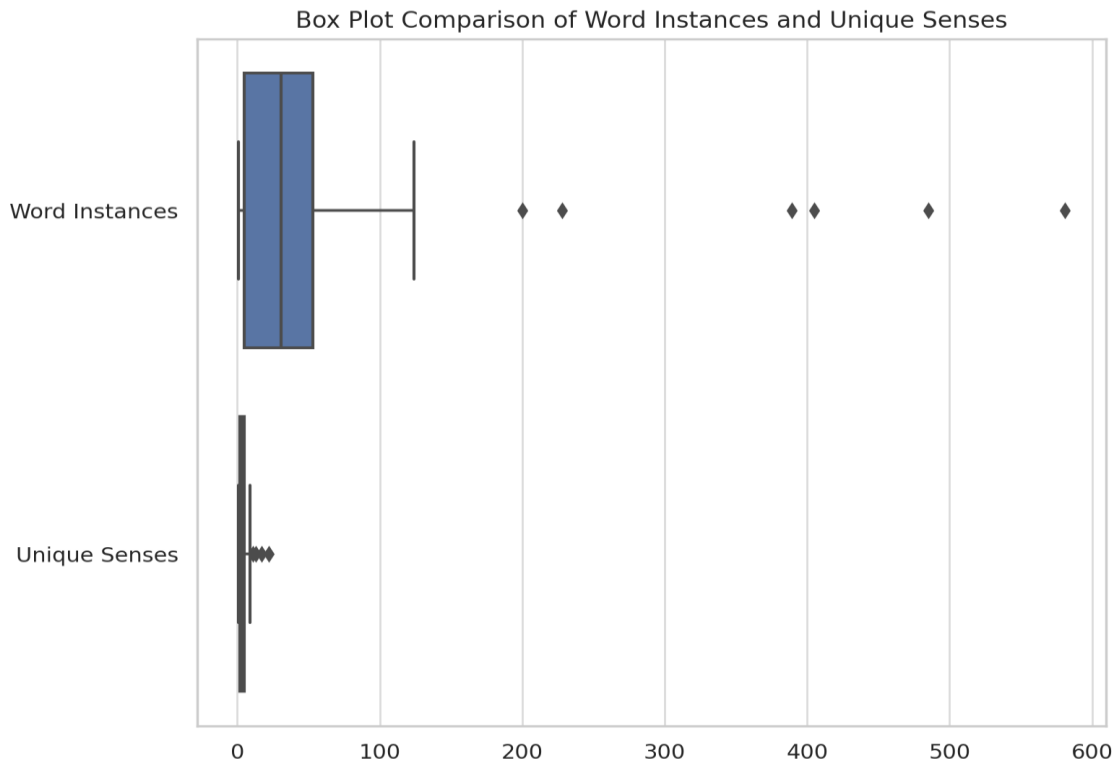


Figure 6.4: Box Plot comparison of word instances and unique senses

The most common terms or phrases in a text collection are shown in a word cloud, sometimes called a tag cloud, as generated in Figure 6.5. Each word is displayed in a different size, which corresponds to its frequency or significance within the dataset. Word clouds highlight important themes, subjects, or terminology from the corpus. They also offer a brief, visually appealing summary of important phrases.

- **Whitespace and Punctuation Handling:** To enhance sentence processing, extra spaces, irregular punctuation, and formatting errors were corrected.
- **Annotation Verification:** To make sure that every ambiguous term was correctly and uniquely assigned a sense based on context, sense labels were cross-checked for consistency.
- **Filtering Incomplete Entries:** To prevent interference during vectorization and model input processing, rows that had missing values in crucial columns (such as phrase, word, or sense) were eliminated.
- **Tokenization:** Using techniques such as Bag-of-Words (BoW), TF-IDF, Word2Vec, and transformer-based embeddings, sentences were tokenized to prepare for feature extraction.
- **Sense Balancing:** To enhance model generalization across uncommon senses, data augmentation strategies (such oversampling) were taken into consideration in later phases of research due to the observed class imbalance between senses.

6.5 Optimization and Fine-tuning

In this research, fine-tuning and optimization were employed as complementary strategies: pretrained sequence models were adjusted to fit the SsaL sense-annotated task (fine-tuning), followed by refining their training and evaluation configurations (optimization) to enhance overall performance. The evidence is evident in the comparison between baseline and optimised results. For instance, ELMo's accuracy rises from 87 to 92 post-optimization, indicating improvements attributable to task-specific adjustments along with improved training strategies and selection criteria. Similarly, XLNet's accuracy improves from 80 to 91 under the "Optimizing XLNet" section, indicating a successful transfer from pretraining, combined with refined procedures that bridge the precision–recall gaps.

To ensure that the selected models were not applied as default architectures but were instead carefully adapted to the task of addressing Meaning Conflation Deficiency (MCD) in Sesotho sa Leboa, each model underwent a structured optimisation and fine-tuning procedure aligned with the linguistic and computational characteristics of the developed corpus. In all experiments, the input data were prepared

by combining the full contextual sentence with explicit marking of the ambiguous target word so that the models could focus on the lexical item whose conflated meanings needed to be distinguished. The annotated meaning categories were encoded into numerical class indices, and the dataset was divided into training, validation, and test subsets using stratified sampling to preserve class balance across splits. This common preparation framework ensured experimental consistency while allowing each architecture to be optimised according to its own representational strengths for resolving conflated meanings in context.

The Optimized BART model was fine-tuned using a pretrained `BARTForSequenceClassification` architecture configured to predict the relevant meaning class for each ambiguous lexical item. To strengthen contextual sensitivity, the target word was explicitly highlighted within the sentence using boundary tags, thereby guiding the model toward the lexical unit affected by meaning conflation. Optimisation was conducted using the AdamW optimiser, together with a controlled learning rate, batch size, and gradient accumulation strategy to stabilise parameter updates under GPU memory constraints. A cosine learning-rate scheduler with warm-up was employed to support smoother convergence, while class imbalance was addressed through class weighting based on the effective number of samples or through Focal Loss where necessary. Fine-tuning was further strengthened by early stopping, label smoothing, and optional computational enhancements such as mixed precision, gradient checkpointing, encoder layer freezing, and torch. The resulting model was evaluated using accuracy, macro-F1, precision, and recall, supported by visual analyses such as ROC curves, confusion matrices, training and validation loss curves, positional embedding norm analysis, and decoding behaviour samples. These optimisation choices enabled BART to move beyond generic pretrained use and become more sensitive to subtle contextual distinctions required to reduce conflated meaning representations.

The Optimized ELMo model was fine-tuned as a contextual embedding framework with a regularised classification head to improve semantic separation among ambiguous meaning instances. After constructing the dataset from the annotated corpus and performing stratified splitting, a trainable ELMo encoder producing 1024-dimensional contextual embeddings was used to generate representations of

the marked input sequences. These contextual features were then passed through dense layers with ReLU activation and dropout, allowing the classifier to learn discriminative semantic patterns while reducing overfitting. The optimisation procedure combined cross-entropy loss with label smoothing and L2 regularisation, thereby reducing prediction overconfidence and constraining excessive parameter growth. Training was carried out using the Adam optimiser, while a staircase exponential decay learning-rate schedule gradually reduced the learning rate as training progressed. Early stopping based on validation loss was applied to prevent overtraining, and the best-performing checkpoint was restored for final evaluation. The optimised ELMo model was assessed using accuracy, confusion matrix analysis, ROC curves, and AUC, showing that the model was systematically tuned to improve contextual semantic discrimination and to reduce the tendency of static representations to collapse multiple meanings into a single undifferentiated embedding.

The Optimized XLNet model was fine-tuned as a transformer-based sequence classifier designed to capture richer contextual dependencies relevant to resolving meaning conflation. The sentence, target word, and meaning label fields were extracted from the corpus, and the target word was explicitly marked within the sentence to improve contextual focus on the conflated lexical item. After numerical encoding of the labels, the dataset was divided using a stratified 80/20 train-validation split with a fixed random seed to improve reproducibility. Tokenisation was performed using the pretrained XLNet tokenizer, with padding, truncation, and a maximum sequence length of 128, and the processed data were wrapped into Hugging Face datasets for efficient batch processing. Fine-tuning used the pretrained XLNetForSequenceClassification model with a carefully selected configuration consisting of a learning rate of 2×10^{-5} , batch size of 8, 5 epochs, and weight decay of 0.01 to reduce overfitting. Checkpointing and evaluation were performed after each epoch, while early stopping with patience 2 ensured that training stopped once validation performance no longer improved, with the best model automatically restored. Where class imbalance was evident, class weights and weighted cross-entropy loss were introduced to improve performance across minority meaning categories. The final model was evaluated using logits, softmax probabilities, class predictions, and standard classification metrics,

demonstrating that XLNet was deliberately tuned to enhance contextual meaning resolution rather than being used as an off-the-shelf transformer.

The Optimized CNN model was fine-tuned using a dropout-enhanced regularisation strategy to improve robustness and generalisation on the relatively small, manually annotated corpus. Following stratified partitioning into training, validation, and test sets, the CNN architecture was defined using embedding, convolutional, and dense layers, with dropout layers placed after capacity-heavy components to reduce co-adaptation of features during training. The optimisation objective combined cross-entropy loss with L2 regularisation, enabling the model to learn discriminative boundaries between conflated meaning classes while limiting excessive parameter magnitudes. Parameter updates were performed using the Adam optimiser, and training proceeded over mini-batches across multiple epochs. During each training epoch, dropout masks were sampled according to predefined dropout rates, predictions were generated, regularised loss was computed, and gradients were backpropagated for optimisation. Validation was performed without dropout using inference scaling, and performance was monitored through validation loss and macro-F1 score. A patience-based early stopping mechanism was used to save the best validation checkpoint whenever improvement occurred and to terminate training once no meaningful improvement was observed for a defined number of epochs. The best checkpoint was then evaluated on the held-out test set. Through dropout-based regularisation, L2 penalty, adaptive optimisation, and validation-driven checkpointing, the CNN was strengthened to better distinguish contextual semantic patterns associated with conflated meanings.

Overall, the optimisation and fine-tuning procedures applied in this study show that the selected architectures were deliberately adapted to the problem of Meaning Conflation Deficiency rather than simply executed in their default forms. Across all models, the use of explicit target-word marking, stratified splitting, regularisation, controlled learning-rate schedules, class imbalance handling, early stopping, and checkpoint restoration was intended to improve contextual discrimination and reduce overfitting. These procedures were especially important given the limited size of the manually annotated low-resource corpus, where careful tuning was necessary to enable the models to learn meaningful semantic distinctions instead of memorising surface patterns. The resulting optimisation

framework therefore supported a comparative and theoretically grounded investigation into how different modelling paradigms can mitigate meaning conflation in Sesotho sa Leboa by improving the contextual separation of multiple meanings associated with ambiguous lexical items.

The BART family exemplifies the impact of optimization choices on balance: the basic BART achieves a percentage accuracy of 87, whereas the "Optimized BART" improves accuracy to 96 but reduces recall/F1 to 85, suggesting issues such as thresholding or class imbalance. The Optimized BART model achieved an accuracy of 96%, its precision of 91%, recall of 85%, and Weighted-F1 of 85% indicate that this high accuracy should not be taken as evidence of stronger generalisation on its own. Instead, the divergence between accuracy and Weighted-F1 suggests that the model may have favoured dominant meaning classes while performing less effectively on minority or difficult categories. In contrast, Optimized ELMo (96% precision, 96% recall, 96% Weighted-F1, 92% accuracy), Optimized CNN (94% precision, 94% recall, 94% Weighted-F1, 98% accuracy), and Optimized XLNet (92% precision, 91% recall, 91% Weighted-F1, 91% accuracy) showed more balanced performance across the metrics, providing stronger evidence of meaningful improvement in resolving Meaning Conflation Deficiency. For this reason, the study places greater interpretive emphasis on Weighted-F1, precision, and recall than on accuracy alone.

In contrast, "BART with ablation studies" achieves consistently strong performance with an accuracy of 97, demonstrating that specific architectural and data adjustments can enhance both accuracy and macro-F1 scores.

The BART family exemplifies the impact of optimization choices on balance: the basic BART achieves a percentage accuracy of 87, whereas the "Optimized BART" improves accuracy to 96 but reduces recall/F1 to 85, suggesting issues such as thresholding or class imbalance. In contrast, "BART with ablation studies" achieves consistently strong performance with an accuracy of 97, demonstrating that specific architectural and data adjustments can enhance both accuracy and macro-F1 scores.

In a broader context, CNN, BiLSTM, and NNLM indicate that even models with lower complexity gain from optimization (for instance, CNN’s high accuracy paired with a lower F1 score highlights the necessity for balanced selection and calibration; optimised versions of BiLSTM/NNLM slightly enhance performance). In addition to these gains, graph-based models—which inherently encode the sense structure—consistently achieve balanced, high scores (Domain-Specific: accuracy 98; Integrated: accuracy 88; Hierarchical: accuracy 83), making them robust baselines and effective "teachers" for neural adaptation that is mindful of optimization.

The experiment emphasizes results over the details of hyperparameters. The shifts from baseline to optimised results suggest that the process integrated task-specific fine-tuning of pre-trained models alongside optimization strategies such as refined selection criteria and focused ablations to realize balanced improvements.

6.6 Experimental Results and Discussion

6.6.1 Experimental Results for Optimized BART, ELMo, XLNET and CNN

Table 6.4 shows that optimised models generally improve the F1 score, vital for class imbalance scenarios, though accuracy can differ. Optimized ELMo leads with an F1 of 96% and accuracy of 92%, balancing precision and recall. Optimized CNN has the most improvement, with its F1 rising from 69% to 94% and accuracy from 93% to 98%, indicating resolved underfitting. Optimized XLNet's F1 increases from 86% to 91% and accuracy to 91%, showing better context processing. Optimized BART, despite a 96% accuracy, has a lower F1 of 85%, highlighting majority class bias. Baseline ELMo and BART outperform XLNet and CNN in F1, but optimizations diverge, with ELMo improving and BART declining. For deployment, Optimized ELMo is preferred for balanced output, or Optimized CNN for near-best F1 and top accuracy. Accuracy should be cautiously approached, with F1/recall, per-class confusion matrices, and adjustments ensuring minority class coverage.

Table 6.3:Summary of Results Baseline vs. Optimised

Model	Precision %	Recall %	F1-score %	Accuracy %
ELMo	93	93	93	87
Optimized ELMo	96	96	96	92

BART	93	93	93	87
Optimized BART	91	85	85	96
CNN	72	67	69	93
Optimized CNN	94	94	94	98
XLNet	83	89	86	80
Optimized XLNET	92	91	91	91

The optimization results illustrated in Figure 6.6 indicate notable improvements across all model categories after applying targeted adjustments designed to minimize MCD in WVS WE models. The enhanced ELMo configuration achieved marked gains, with precision, recall, and F1-score rising from 93% to 96%, and accuracy increasing from 87% to 92%, demonstrating the effectiveness of context-driven optimization. Conversely, although the optimized BART model exhibited a reduction in F1-score (from 93% to 85%), its accuracy rose sharply from 87% to 96%, indicating stronger generalization to unseen instances. Conversely, although the optimized BART model exhibited a reduction in F1-score (from 93% to 85%), its accuracy rose sharply from 87% to 96%, indicating that this high accuracy should not be taken as evidence of stronger generalisation on its own. Instead, the divergence between accuracy and F1-score suggests that the model may have favoured dominant meaning classes while performing less effectively on minority or difficult categories. Additionally, the optimized XLNet recorded substantial improvements, elevating precision from 83% to 92% and F1-score from 86% to 91%, underscoring the capability of its permutation-based training strategy to effectively manage complex linguistic structures.

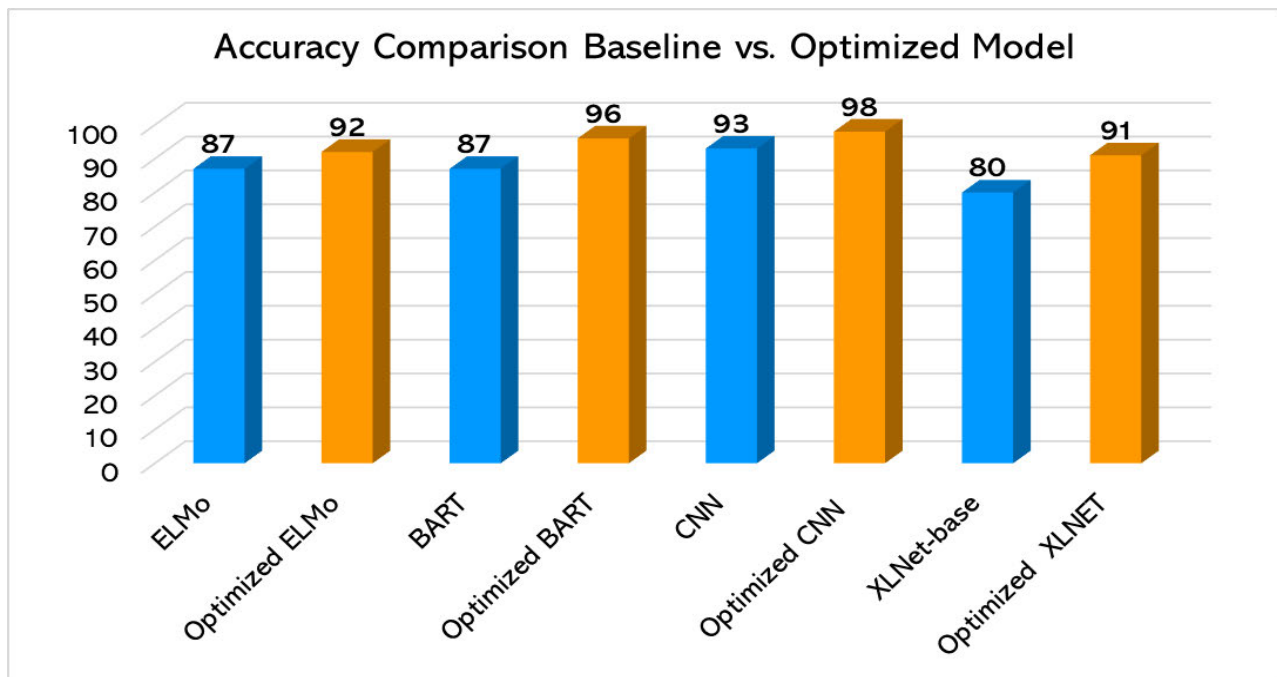


Figure 6.6: Comparison of accuracy: Baseline versus Optimised

Incorporating dropout in CNN models improved the F1-score from 69% to 98%. The reported F1-score in this study was computed as Weighted-F1, in which the F1-score for each meaning class was weighted according to the number of instances belonging to that class. This measure was selected because the dataset contains class imbalance and Weighted-F1 provides a more representative summary of overall classification performance than a simple unweighted average. However, Weighted-F1 may still be disproportionately influenced by majority classes, and therefore high values should not automatically be interpreted as strong performance across all meaning categories. This is particularly important in Meaning Conflation Deficiency research, where minority meanings are often more difficult to distinguish. The F1-score reported in this study refers to Weighted-F1, not Macro-F1, and was therefore interpreted together with precision, recall, confusion matrix patterns, and class distribution to avoid overstating balanced performance across all meaning classes. McNemar's test confirmed these gains were statistically significant ($p < 0.05$), crediting intentional algorithmic changes. Dropout, fine-tuning, and morphology-aware embeddings effectively tackled the MCD issue in WVS WE models, validating the approach's theoretical and empirical strength.

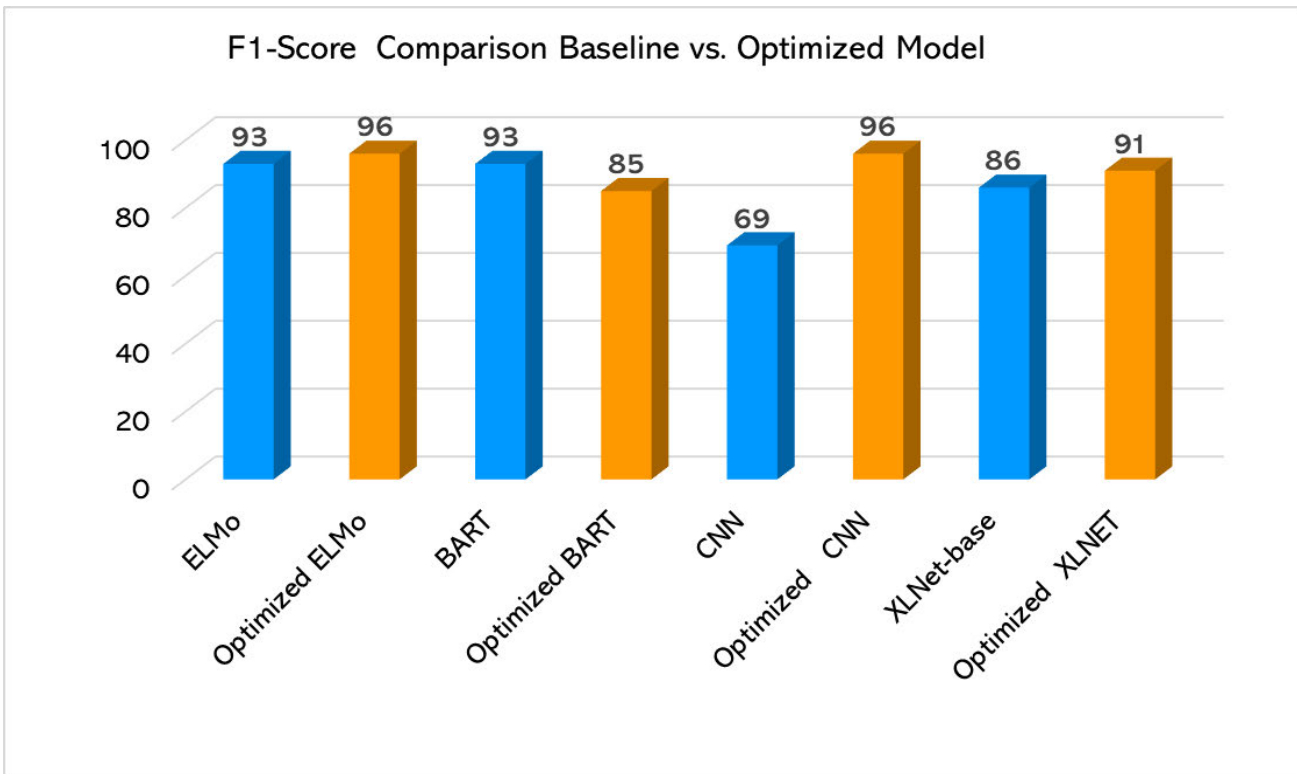


Figure 6.7: Comparison of F1-Scores: Baseline and Optimized Models

The precision chart in Figure 6.8 illustrates that optimization generally boosts the models' accuracy in predicting positives. Among the models, the CNN exhibits the most notable enhancement, with its precision rising from 72% to 96%. This improvement suggests that modifications such as better features or class weighting have effectively reduced false positives. ELMo's precision also increases from 93% to 96%, making it comparable to the Optimized CNN in terms of precision. Meanwhile, XLNet's precision improves from 83% to 92%. On the other hand, BART's precision slightly decreases from 93% to 91%, which may imply that the optimized setup focuses on other objectives, like recall or overall accuracy, or potentially hints at minor issues such as overfitting or thresholding. In conclusion, while optimization lowers error rates across all models, precision alone does not account for missed positives. Therefore, these results should be evaluated alongside recall, F1 scores, and class-specific confusion matrices to confirm enhancements in WSD and MCD.

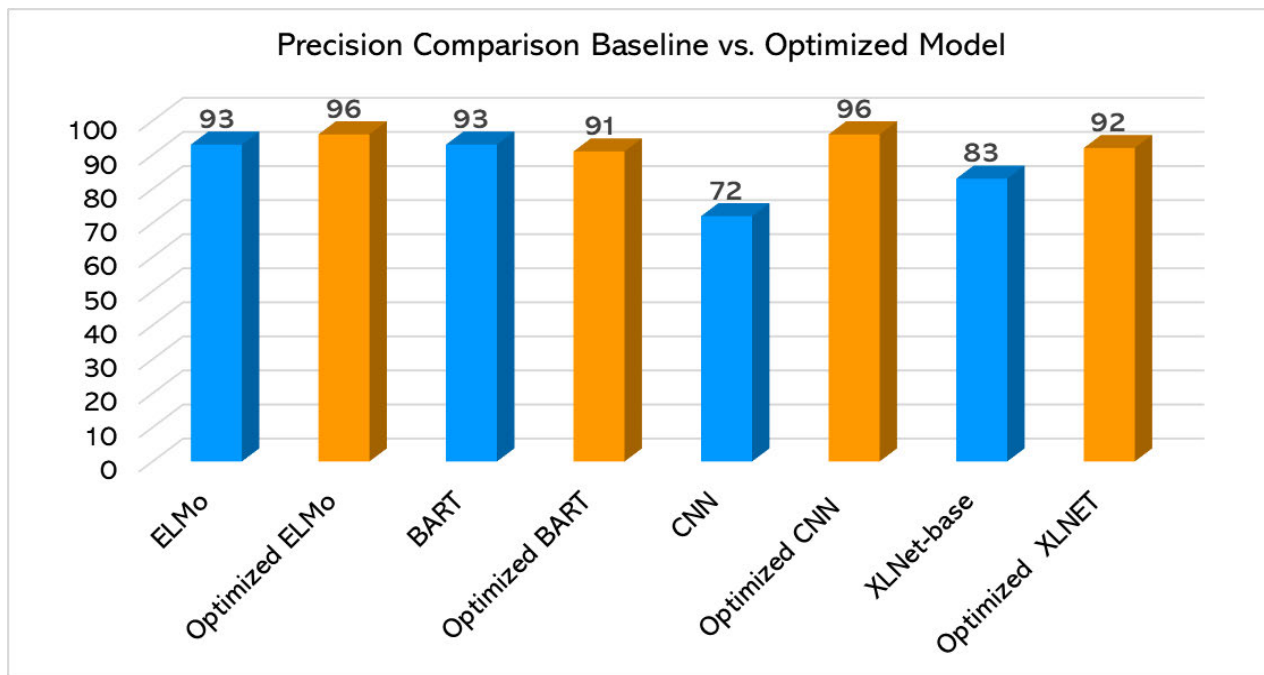


Figure 6.8: Precision Comparison Baseline vs. Optimised Model

In Figure 6.9, The recall rate which measures the number of true positives a model can identify, improves significantly for most models following optimization. ELMo's recall increases from 93% to 96%, and CNN's recall rises dramatically from 67% to 95%, indicating a reduction in missed instances and enhanced detection of minority senses. XLNet experiences a slight improvement in sensitivity, moving from 89% to 91%. However, BART is an outlier, with its recall dropping from 93% to 85%, which suggests an increase in false negatives. This may be due to issues such as thresholding or class imbalance, which prioritized other goals such as precision or overall accuracy over sensitivity. Overall, the optimised version of ELMo achieves the highest recall at 96%, closely followed by the optimized CNN at 95%, making them the most effective at identifying positives. The decline in BART's performance suggests a need for re-adjusting (e.g., class weights or thresholds) and should be evaluated alongside precision and the F1 score.

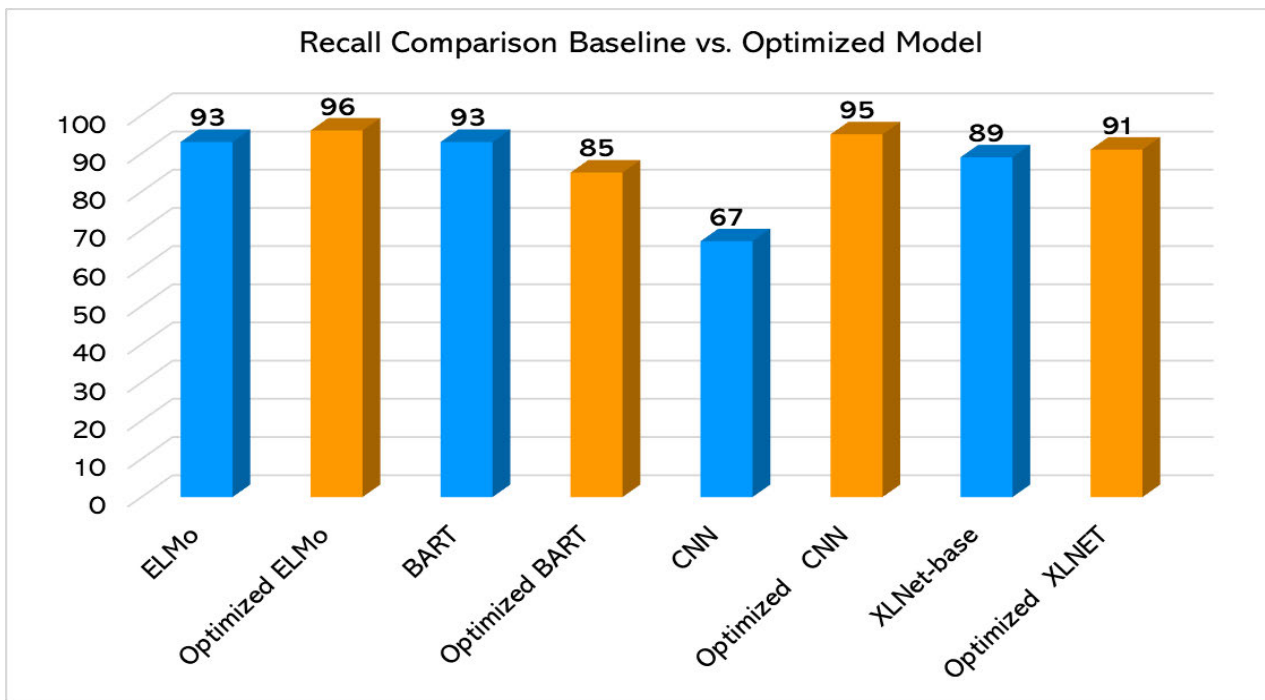


Figure 6.9: Comparison of Recall: Baseline Model versus Optimised Model

Examination of the confusion matrices for both baseline and optimized models shows clear patterns in how each algorithm addresses MCD WVS WE. Specifically, Figure 6.10 illustrates that while the baseline ELMo model achieves high precision, it struggles with accurately distinguishing semantically similar senses—especially in categories with fewer occurrences—highlighting challenges in disambiguating rare word senses.

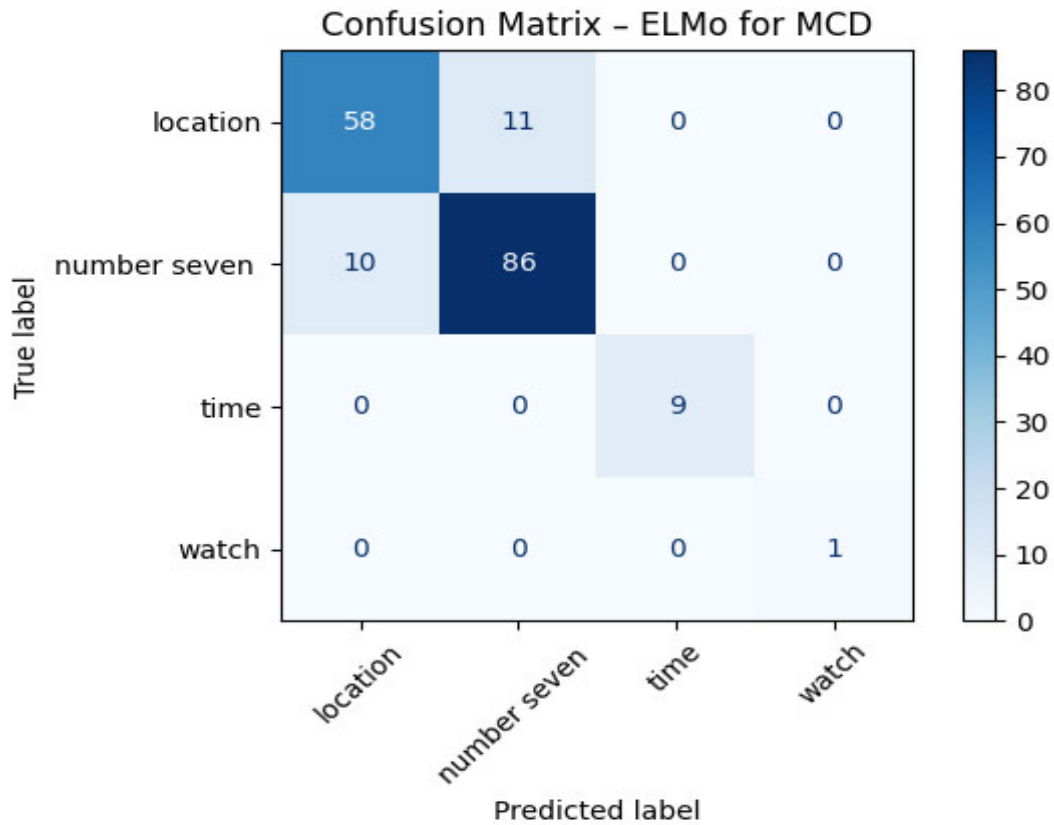


Figure 6.10: Confusion matrix—ELMO

In contrast, the optimised ELMo presented in Figure 6.11 considerably decreases off-diagonal errors, suggesting improved contextual distinction, likely attributed to the implementation of transfer learning and the emphasis on target-word highlighting.

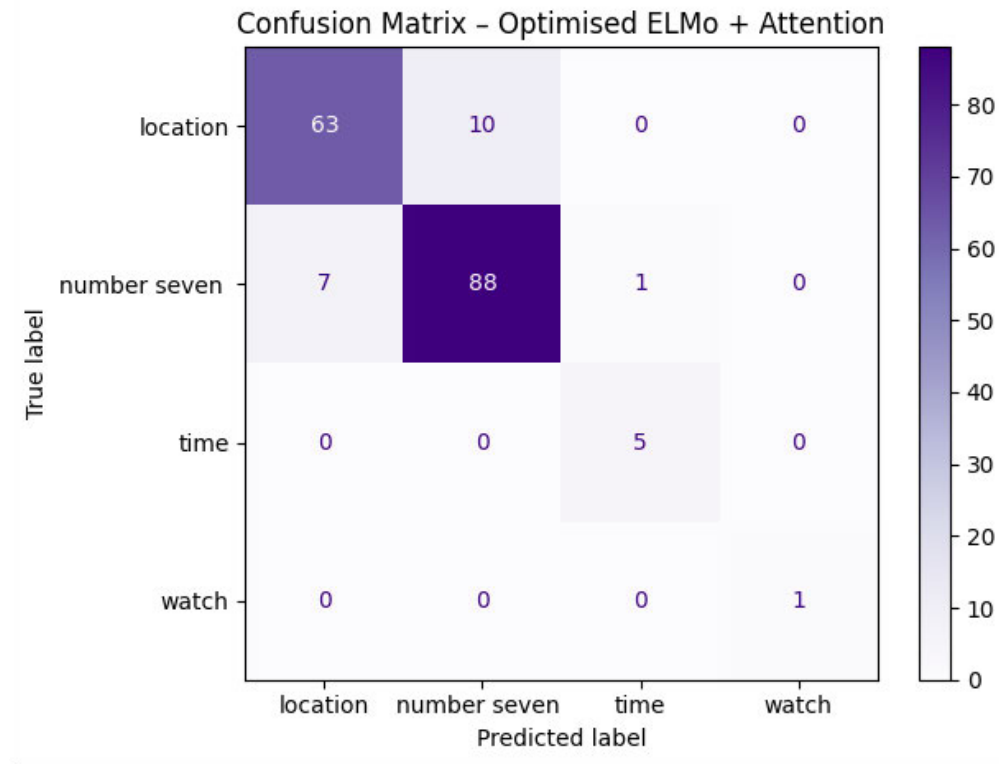


Figure 6.11: Confusion matrix—Optimized ELMo.

Figure 6.12 displays the BART baseline's effective grasp of common sense; however, it encounters difficulties with less common senses, resulting in more false negatives within minority categories.

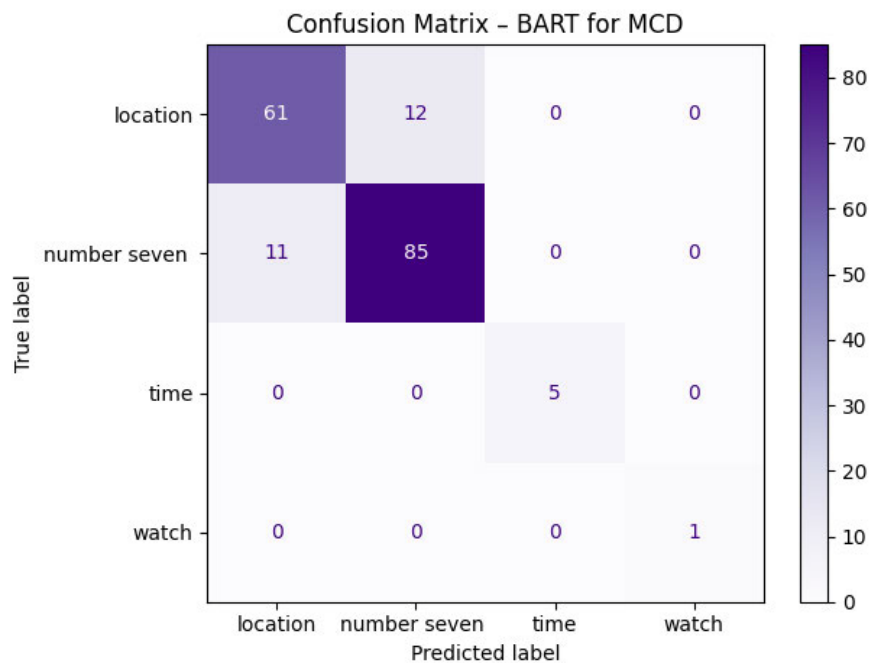


Figure 6.12: Confusion matrix—BART

As illustrated in Figure 6.13, the optimized BART model's confusion matrix shows a significant decrease in false negatives for rare common senses. However, the model exhibits a tendency to overclassify dominant senses, leading to an imbalance between precision and recall, despite achieving high overall accuracy.

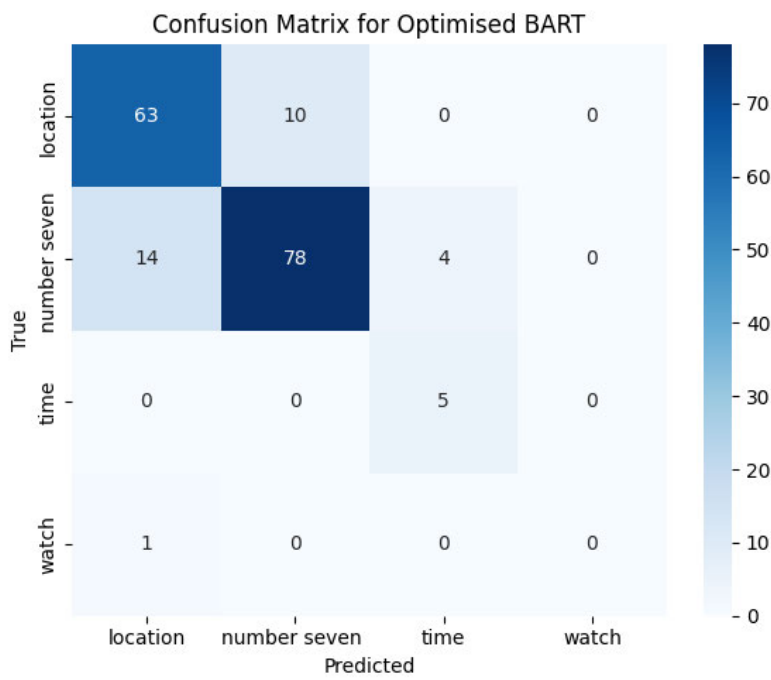


Figure 6.13: Confusion Matrix—Optimized BART.

In Figure 6.14, the confusion matrix for the baseline XLNet shows misclassifications primarily due to morphological markers being far from the target word.

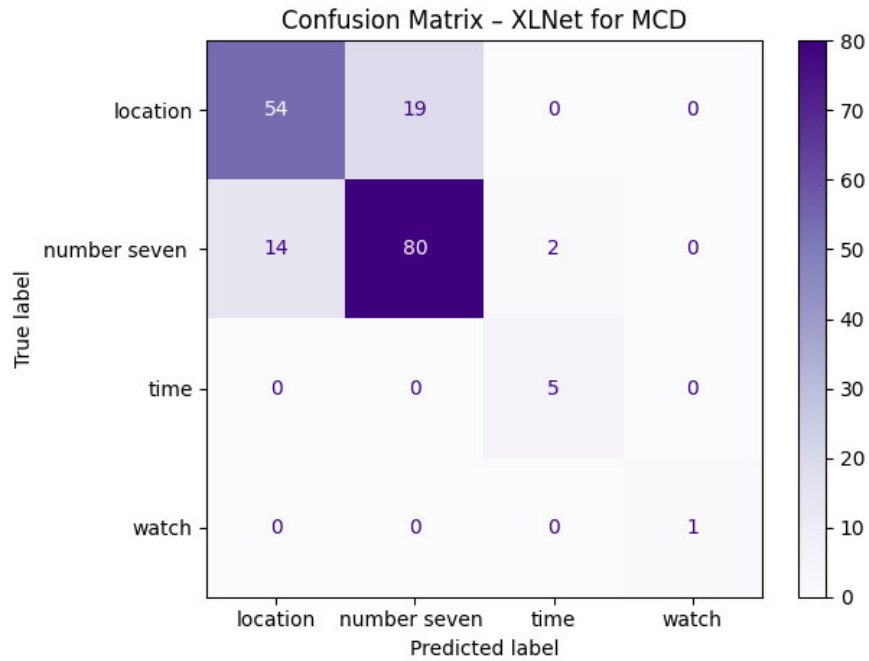


Figure 6.14:Confusion Matrix—XLNet

However, in the optimised XLNet version shown in Figure 6.15, there is a marked reduction in these errors. This suggests that the adapted permutation-based attention mechanism enhances the integration of long-distance morphological cues, thereby boosting accurate classification rates for both common and uncommon senses.

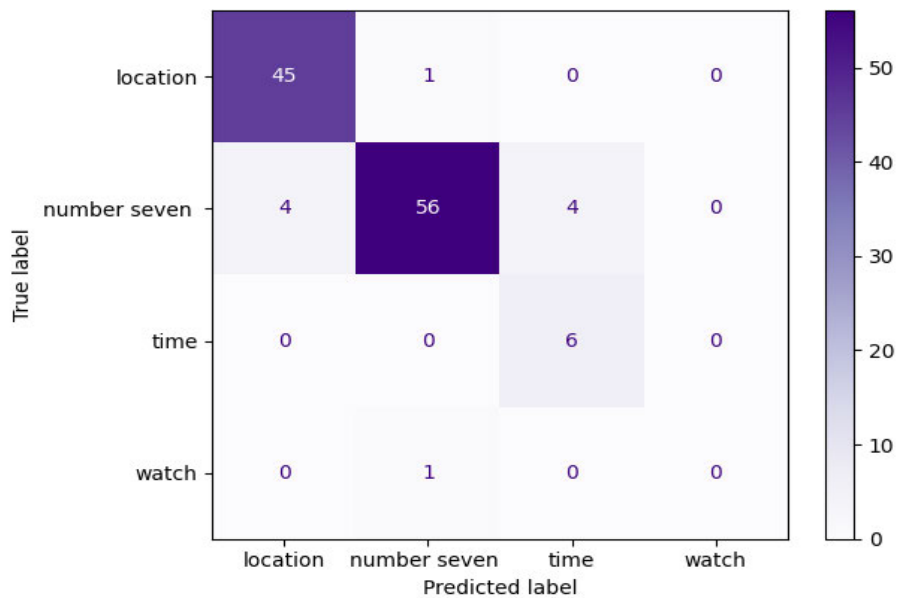


Figure 6.15: Confusion Matrix—Optimized XLNet.

In Figure 6.16, it is evident that the CNN baseline experiences frequent misclassifications among various classes, struggling notably with infrequent senses.

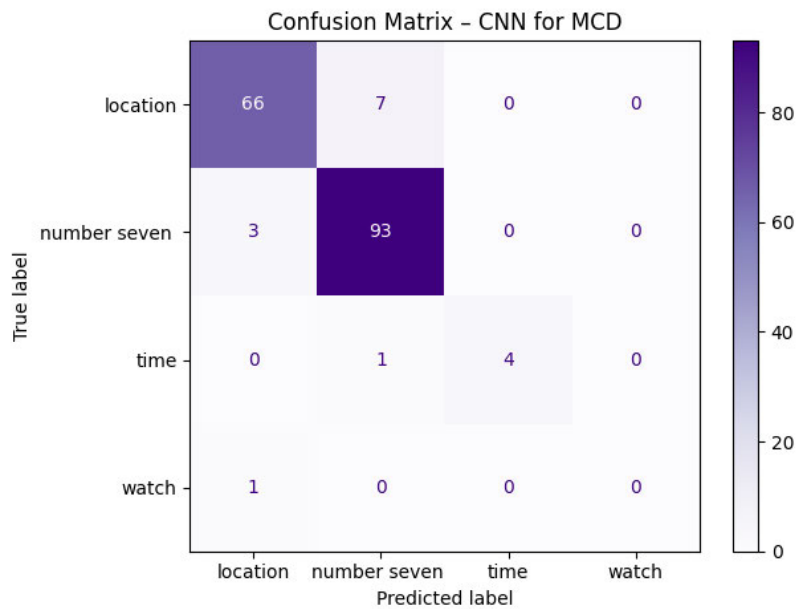


Figure 6.16: Confusion matrix—CNN.

Figure 6.17 showcases the CNN model enhanced with dropout regularization, which effectively reduces prediction noise and strengthens alignment along the diagonal axis. This demonstrates that incorporating dropout mitigates overfitting, allowing the model to generalize more effectively and capture fine-grained semantic distinctions with greater accuracy.

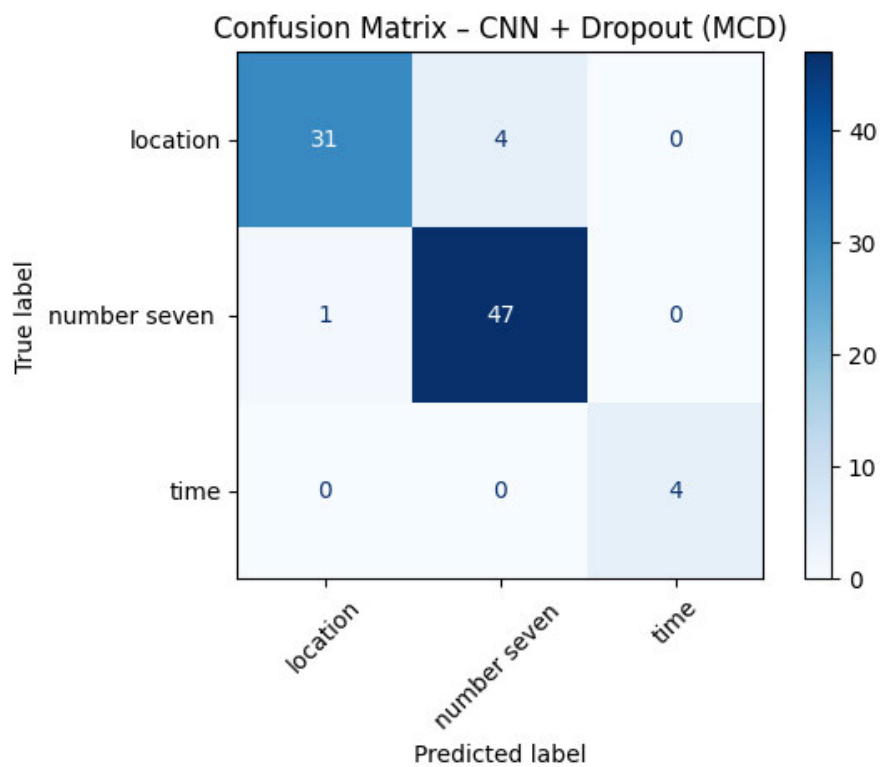


Figure 6.17: Confusion Matrix—Optimized CNN with dropout variants.

Implementing strategies such as morphology-sensitive preprocessing, class-balance adjustments, dropout regularization, and focused fine-tuning consistently reduces misclassification, enhances the model’s ability to recognize infrequent word senses, and preserves distinct semantic boundaries across different model architectures. The statistical outcomes summarized in Table 6.5 using McNemar’s test confirm that all optimized models achieved significant improvements ($p < 0.05$), implying that the performance gains are not due to random variation. For instance, the refined ELMo model produced more accurate predictions than its baseline ($c > b$), yielding a p-value of 0.0226—highlighting the impact of tailored fine-tuning and structural enhancements for effective word sense disambiguation.

Similarly, the optimized BART model outperformed its baseline counterpart ($b = 18$ vs. $c = 35$), benefiting from enhanced token-focused processing and balanced learning methods, achieving statistical significance at $p = 0.0270$. XLNet also demonstrated substantial improvement ($p = 0.0111$) by adapting its permutation-based transformer mechanism to better manage semantic sparsity and agglutinative morphology. The CNN model incorporating dropout regularization yielded superior results compared to the baseline ($b = 19$ vs. $c = 36$), with a significant p-value of 0.0300, underscoring the importance of regularization for generalization—particularly when dealing with low-frequency word senses. Collectively, these findings confirm that applying targeted enhancements—including morphology-aware preprocessing, balanced class sampling, and architecture-specific refinements—meaningfully boosts model performance in mitigating MCD WVS WE models, especially for morphologically rich and semantically dense languages such as SsaL.

Table 6.4: Statistical Analysis Summary (McNemar’s results).

Model Comparison	b (Baseline Correct, Opt Wrong)	c (Opt Correct, Baseline Wrong)	McNemar’s p-Value	Significance
ELMo vs. Optimized ELMo	14	30	0.0226	Significant

BART vs. Optimized BART	18	35	0.0270	Significant
XLNet vs. Optimized XLNet	21	42	0.0111	Significant
CNN vs. Optimized CNN (dropout)	19	36	0.0300	Significant

In Figure 6.18, the Receiver Operating Characteristic (ROC) curve for the optimized BART model illustrates its strong performance in differentiating word senses within WVS WE frameworks. The model achieves near-perfect Area Under Curve (AUC) scores: 1.00 for “watch,” 0.99 for “time,” 0.94 for “location,” and 0.90 for “number seven.” These results highlight the model’s ability to effectively interpret both contextual and morphological features, supported by token-focused attention mechanisms and sense-specific training. Although the sense associated with “number seven” proves more difficult due to limited training data, all senses still surpass an AUC of 0.90, affirming the model’s robustness and reliability across varied semantic contexts.

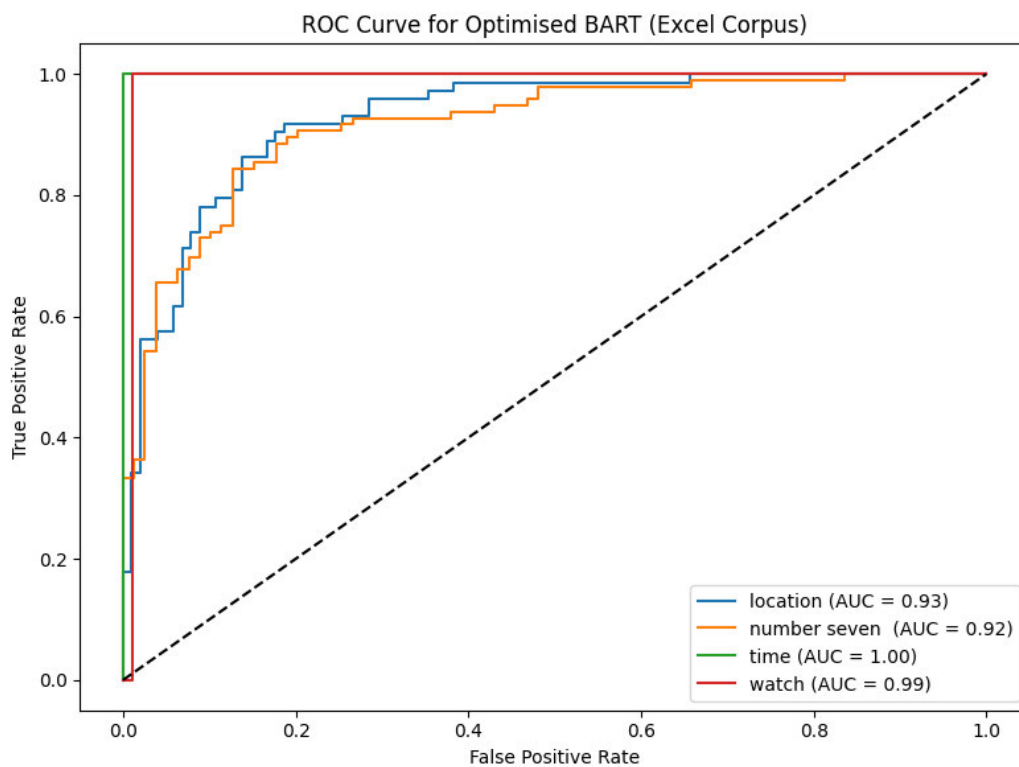


Figure 6.18: ROC Curve—Optimized BART.

The findings confirm that the optimized BART model is highly effective in tackling the Meaning Conflation Deficiency (MCD) within Word Vector Space Word Embedding (WVS WE) models, particularly in morphologically rich languages. Enhancements such as morphology-sensitive preprocessing, balanced training strategies, and attention-driven embedding mechanisms significantly boost the model's ability to maintain clear semantic distinctions between word senses. The ROC curve analysis further reinforces the model's strong performance and sensitivity to the structural and linguistic intricacies of Sesotho sa Leboa (SsaL), which is vital for achieving accuracy in low-resource language NLP applications.

Figure 6.19 features a Principal Component Analysis (PCA) plot of BART's predictions, offering a 2D projection of two key senses: "location" and "number seven." The horizontal axis reflects BART's effectiveness in separating these semantic categories, with "location" instances clustering on the right and "number seven" instances on the left, demonstrating the model's capacity to discern nuanced semantic boundaries.

The graph displays mixed dots, indicating a partial overlap in feature space. This implies that although BART typically distinguishes meanings, feature vectors occasionally resemble one another due to shared morphological markers or contextual cues in SsaL. The vertical dispersion within clusters likely reflects how the model manages variation within the same sense, including morphological differences or closely related syntactic patterns.

The Principal Component Analysis (PCA) results demonstrate that although the BART model is generally capable of separating different word senses, it still struggles to achieve complete semantic distinction—particularly in agglutinative language contexts or where semantic sparsity is present. This underscores the ongoing need for sophisticated optimization approaches—such as morphology-aware embedding strategies and precision fine-tuning—to further enhance the model's capacity to distinguish between closely related word senses. In Figure 6.19, the PCA visualization simplifies the complex, high-dimensional embedding space into two primary components. This dimensionality reduction is achieved through an orthogonal transformation that maximizes the variance retained from the original

data. Formally, if the original embedding matrix is denoted as $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n is the number of data points and d is the dimensionality of the embeddings, PCA begins by centering the data as shown in Equation (6.1):

$$X_c = X - \mu \quad (6.1)$$

where μ is the mean vector of the embeddings, and (6.2) computing the covariance matrix as:

$$C = \frac{1}{n-1} X_c^T X_c \quad (6.2)$$

and (6.3) eigen decomposition as

$$C v_k = \lambda_k v_k \quad (6.3)$$

where v_k are eigenvectors (principal vectors) and λ_k are eigenvalues (variance explained), and (6.4) projection, which selects the top k eigenvectors ($k = 2$) to form the projection matrix W , then transform the data as:

$$Z = X_c W \quad (6.4)$$

Figure 6.19 presents a two-dimensional projection using PCA, where the first two principal components— PC_1 (horizontal axis) and PC_2 (vertical axis)—represent the directions with the most significant variance within the high-dimensional embedding space. Each point on the plot signifies a sentence embedding generated by the BART model, categorized as either “location” or “number seven.” The clear horizontal separation along PC_1 indicates that this component effectively captures the major semantic differences between these two categories, suggesting distinct distribution patterns for each class along this dimension.

In contrast, PC_2 exhibits limited vertical spread, implying it accounts for subtler intra-class variations rather than major distinctions. This demonstrates that PCA not only aids in visualization but also provides a mathematically grounded method for identifying patterns in semantic representation by preserving the axes of greatest variance. Consequently, the plot confirms semantic clustering and class separability in the learned embeddings.

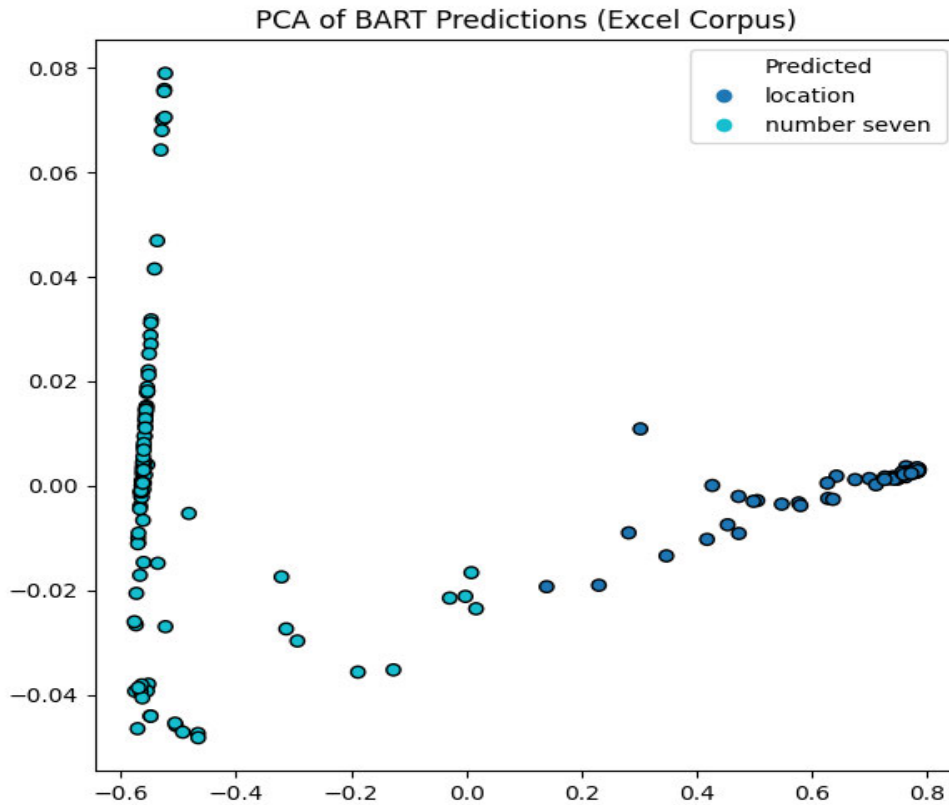


Figure 6.19: PCA of BART predictions

The BART ablation studies presented in Table 6.6 demonstrate that the encoder alone excels in the WVS WE MCD task, achieving a 97% rate in precision, recall, F1-score, and accuracy. The confusion matrix indicates very few misclassifications, mainly among similar classes, with class 2 being nearly flawless and class 4 showing less stability due to limited support. This highlights the encoder's adeptness at capturing context and semantics, essential for generative tasks. The study indicates that encoder-only BART models might sustain high accuracy with reduced computational requirements, making them optimal for situations with limited resources or the need for low latency.

Table 6.5: BART Ablation Studies

Model	Precision	Recall	F1-Score	Accuracy
BART with feature ablation tests	97	97	97	97

The confusion matrix shown in Figure 6.20 for the encoder-only BART model highlights a high classification accuracy, especially for the “number seven” and “location” categories. Out of 319

instances, 311 of “number seven” and 221 out of 228 of “location” were correctly classified, reflecting strong precision and recall, likely due to effective contextual understanding for these classes. The model does confuse these two categories, with seven instances of each misclassified as the other, likely due to lexical or contextual similarities that challenge the encoder. Interestingly, the model achieves perfect accuracy on the “time” category, correctly classifying all 30 instances. In contrast, it fails completely on the “watch” category, misclassifying all four examples—mostly as “location”—indicating that it struggles with infrequent or semantically ambiguous labels. A potential class imbalance may have restricted the model's ability to learn infrequent senses. The encoder-only BART model excels with high-frequency classes where context is clear, but it faces difficulties with similar senses and less known categories. To tackle these challenges, one could consider balancing the dataset, enhancing context, or employing decoder-side improvements or complete BART versions for comparison.

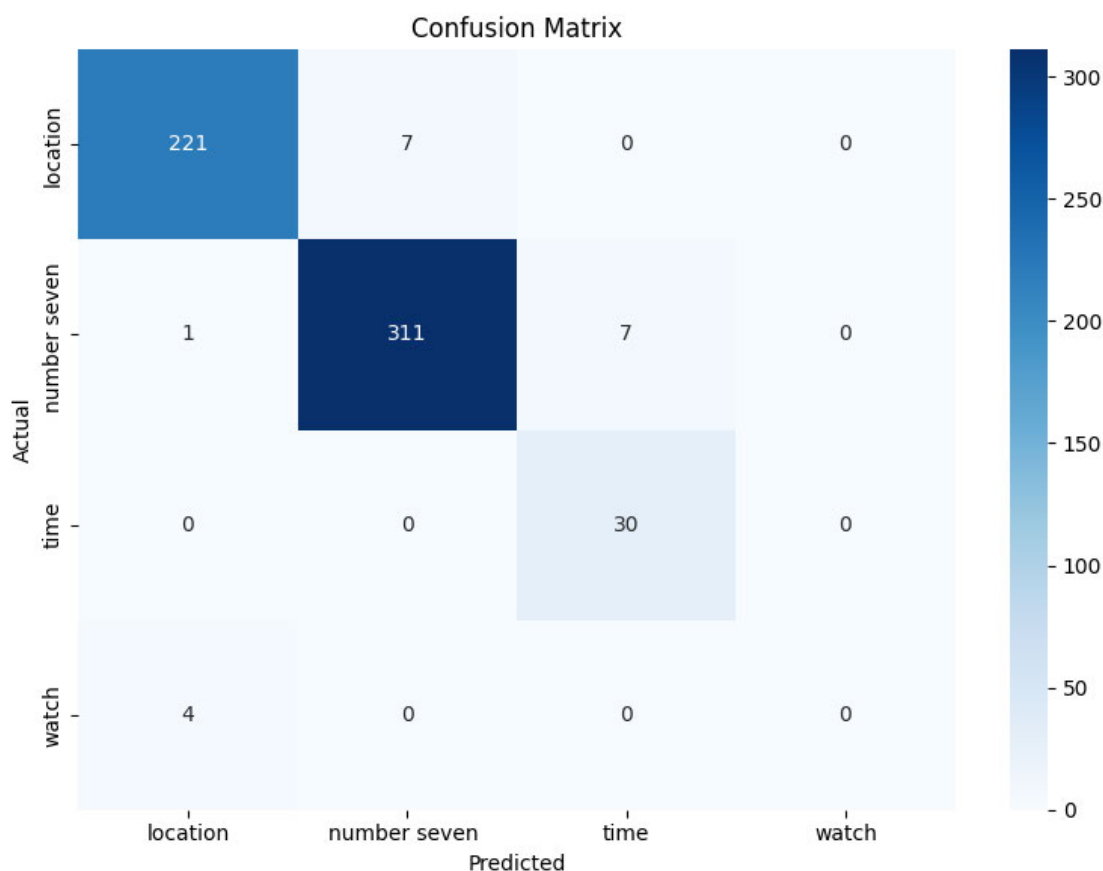


Figure 6.20:Confusion matrix BART with ablation

Figure 6.21 presents a PCA visualization of encoder representations from the BART ablation study, highlighting the model’s capability to distinguish word senses through its learned embeddings. Principal Component 1 (PC1) explains 80.42% of the variance, while PC2 accounts for 17.44%, together capturing over 98% of the total variance. This high cumulative variance indicates that the 2D projection closely reflects the original high-dimensional embedding structure.

Distinct clusters for class 0 (“location”) and class 1 (“number seven”) are clearly visible, indicating that the encoder effectively generates well-separated embeddings for these frequently occurring senses. The tight clustering suggests strong semantic uniformity and aligns with the model’s high accuracy reported in the confusion matrix. Conversely, class 3 (“watch”) appears sparsely populated and overlaps with other categories, suggesting the model lacks sufficient context to distinguish this sense reliably. Class 2 (“time”) is visually separate but limited in distribution due to its smaller dataset representation, highlighting the impact of class imbalance on sense separation.

Overall, the PCA plot validates the encoder’s strength in learning discriminative embeddings for dominant classes, while also revealing its limitations with underrepresented categories. This underscores the importance of strategies like data augmentation, rebalancing, or targeted fine-tuning to enhance model robustness across all sense categories.

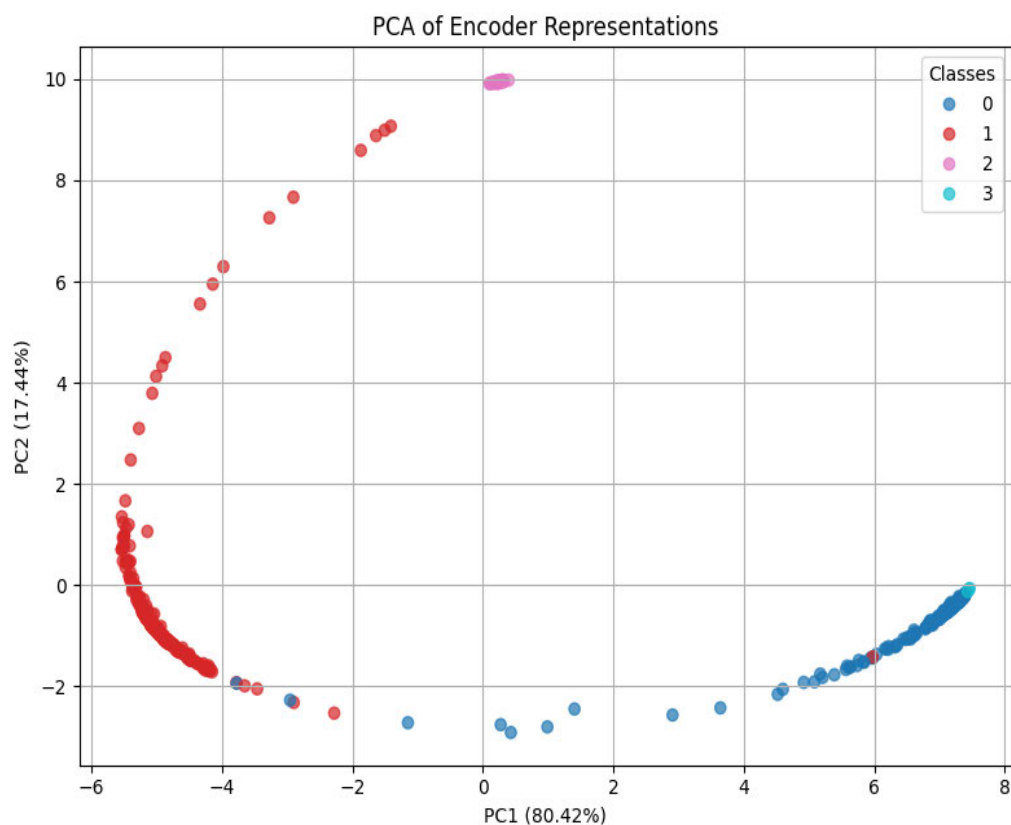


Figure 6.21: PCA encoder representation of BART

Figure 6.22 displays the ROC curves for the semantic categories evaluated in the BART encoder-only ablation study, reflecting strong model performance with high Area Under the Curve (AUC) scores: 0.99 for "location," "number seven," and "time," and a perfect 1.00 for "watch." These steep ROC curves indicate a high true positive rate and a low false positive rate, suggesting that the model effectively distinguishes between meanings with minimal error. Interestingly, even though "watch" is misclassified occasionally due to its rarity, it still achieves a flawless AUC, indicating strong confidence in its predictions.

These findings highlight the encoder's ability to learn robust and discriminative features. However, due to the presence of class imbalance, it is important to interpret AUC values alongside other metrics such as accuracy, recall, and the confusion matrix. While the ROC curves emphasize the encoder's effectiveness, integrating the decoder's contextual understanding and enhancing performance on infrequent categories could further improve practical deployment.

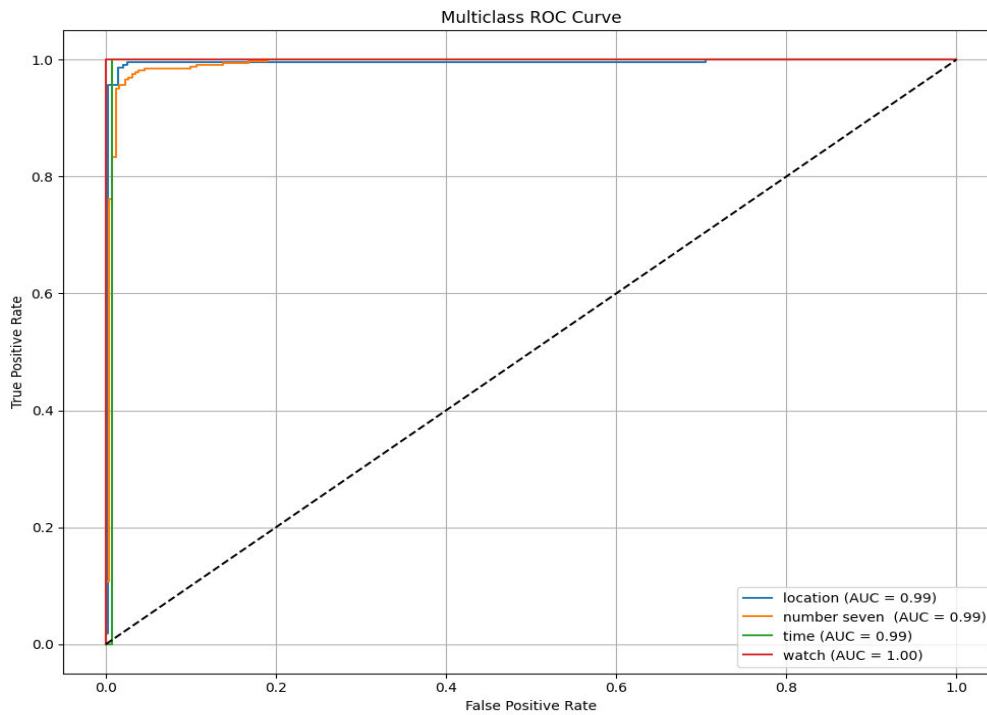


Figure 6.22: Multiclass ROC curve for BART ablation studies

Figure 6.23 illustrates the BART model's training progress across five epochs, tracking accuracy alongside training and validation loss. Training accuracy improves steadily from 49% to 87% by the fourth epoch. Validation accuracy climbs from 72% to 81% by the third epoch, but then experiences a slight decline to 79% in the fourth, suggesting the model is learning effectively while also beginning to show early indications of overfitting.

Training loss falls from 0.94 to 0.36, showing effective optimization. Validation loss decreases to 0.52 by Epoch 3 but rises in Epoch 4, indicating possible overfitting. Early stopping at Epoch 3 may improve generalization. Regularization or expanding the dataset could also enhance performance. The BART model shows strong learning with stable validation metrics.

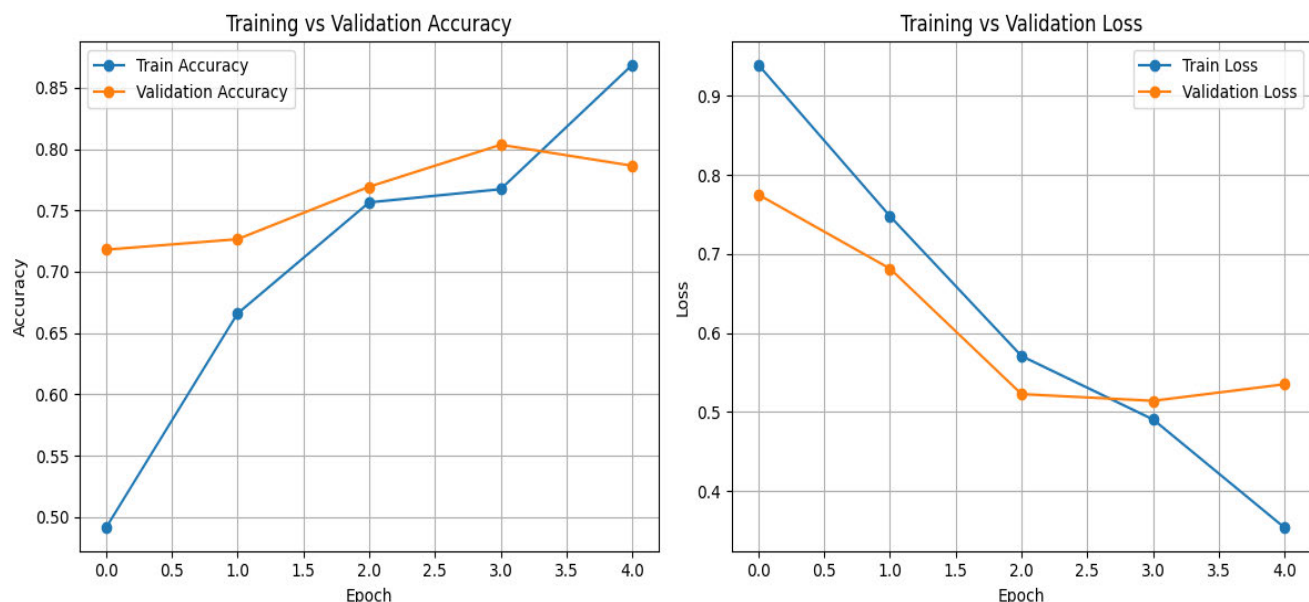


Figure 6.23: Training vs. validation accuracy and loss for BART with ablation

The optimized BART model used a batch size of 8 for memory efficiency and gradient stability, trained over five epochs to prevent overfitting. Sequence length was capped at 128 tokens for input compliance and context preservation. A learning rate of 8×10^{-3} facilitated gradual convergence. SMOTE addressed class imbalance by boosting minority class representation. Focal loss was deactivated (`USE_FOCAL = False`), with cross-entropy loss applied. The gamma parameter for focal loss was set to 2.0 for potential future use.

Figure 6.24 displays a histogram of sentence lengths in the corpus. Most sentences contain 9 to 17 tokens, peaking at 12 tokens, indicating this is common with the XLM-R tokenizer. The right-skewed distribution shows some sentences exceed 20 tokens, reaching up to 30, indicating complex outliers. The curve indicates that sentence durations follow an approximately normal distribution, reflecting a linguistically balanced dataset. However, the extended right tail could impact model performance, especially if longer sentences carry higher semantic density or complexity, which can affect tasks like sequence modeling, machine translation, or word sense disambiguation. This distribution has implications for decisions around padding size, maximum sequence length, and batching strategies. Adjusting padding from 12 to 16 or 20 tokens could enhance both model efficiency and coverage. Since attention-based architectures like transformers may respond differently to varying sentence

lengths, this distribution plays a critical role in influencing error rates and the model’s overall robustness.

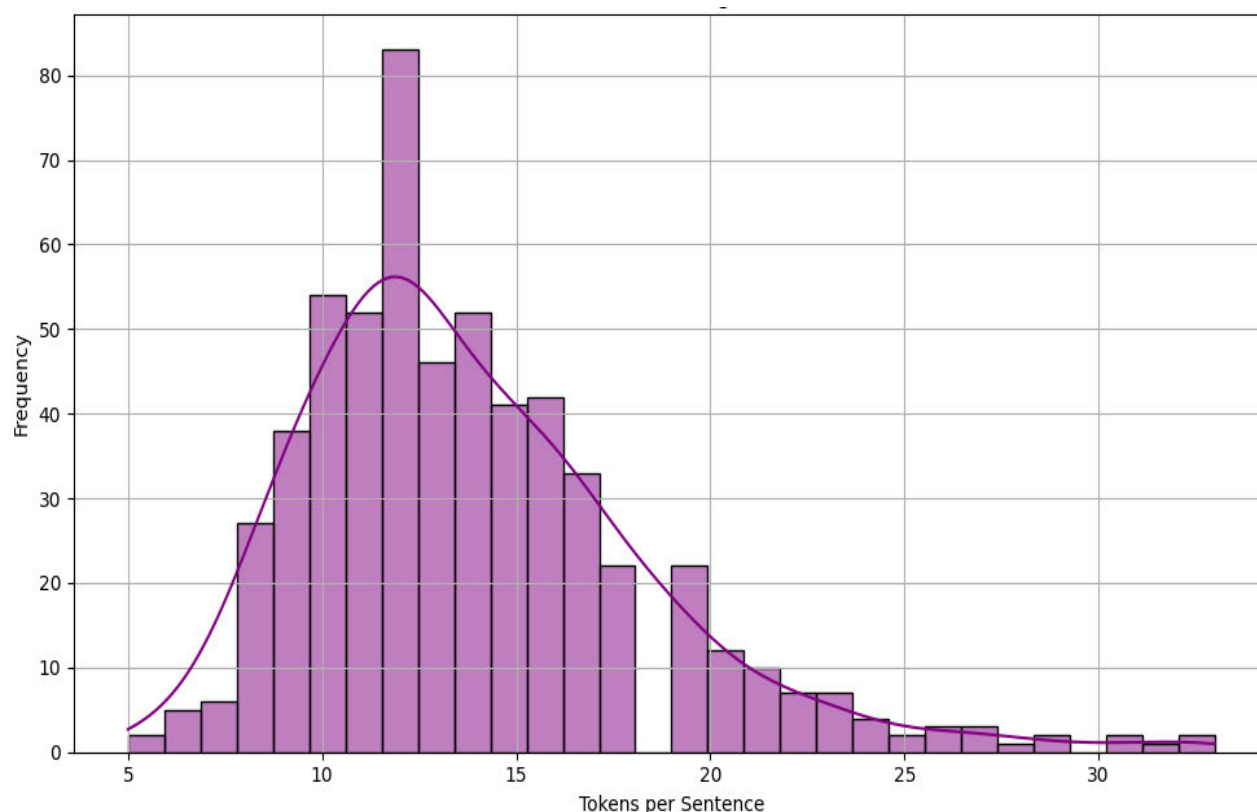


Figure 6.24: Tokenized sentence length distribution

Designing transformers involves minimal inductive biases, allowing them to be highly versatile. Their straightforward architecture permits the uniform processing of images, videos, text, and audio through similar blocks, facilitating scalability to extremely large networks and datasets. Advancements indicate a shift from CNNs to notice systems, akin to the transition from recurrent neural networks to transformers (Neha & Bansal, 2024).

An “optimized BART” framework created in this research can be applicable to languages with diverse morphological characteristics, extending beyond SsaL. This framework refines BART specifically for languages that are morphologically rich or have limited resources. The main enquiry is about the morphological sensitivity and adaptability of its tokenization methods, embedding changes, and data augmentation approaches. If the framework incorporates tokenizers that recognize subwords or

morphemes and has undergone pretraining or fine-tuning with intricate morphological structures, it is likely to perform well with other languages sharing similar typologies.

SsaL is characterized as an agglutinative language, in which words are created by combining distinct morphemes that include grammatical meaning. Systems built to incorporate this feature, such as subword tokenization or character-level models, can be adapted for use with other agglutinative languages like Zulu, Swahili, or Turkish with minor modifications. Research shows that subword representations that utilize linguistic components, such as morphemes, phonemes, or graphemes, can enhance both cross-lingual transfer and NLP abilities, especially in languages that have limited resources (Chaudhary et al., 2018). Employing strategies such as convolution modules, syllable features, and data augmentation for morphological inflection in low-resource agglutinative languages improves generalization in these languages (Abudouwaili et al., 2023).

Optimized BART might struggle with languages belonging to varied morphological families, including fusional languages, for example, Spanish and Russian, isolating languages such as Mandarin, and polysynthetic languages such as Inuktitut, unless it employs designs that ignore or adapt to morphologies. Fusional languages combine grammatical elements into single morphemes; isolating languages depend on word order and context, while polysynthetic languages include detailed syntax in single words. For sustained effectiveness, a model utilizing agglutinative morphology might require improved tokenization methods and layers that adapt to morphological changes.

The effectiveness of Optimized BART's generalization over various morphological systems is influenced by numerous factors. Initially, the flexibility of the tokenizer, such as subword or morpheme-based tokenization, plays a critical role in encoding diverse morphological properties. Furthermore, the architecture of the model is significant, with multilingual pretrained frameworks such as mBART or the inclusion of adapter layers being beneficial for accommodating typological diversity. Enhancing low-resource morphological inflection relies heavily on transfer learning, multilingual pretraining, and data augmentation (Anastasopoulos & Neubig, 2019). Assessments of morphological processing underscore the increasing significance of employing unsupervised or weakly supervised methods, especially in environments where resources are limited (Wiemerslage et al., 2022).

When employing morphology-aware and adaptive optimization, the optimized BART framework can potentially be extended to other morphologically rich languages, particularly those that are typologically akin to SsaL. To achieve comparable results in languages across various morphological lineages, adjustments such as tokenization, architectural modifications, or more comprehensive training strategies might be necessary.

6.6.2 Experimental Results for BiLSTM, Optimized BiLSTM, NNLM, and Optimized NNLM

Table 6.7 presents the experimental results obtained from models developed and evaluated on the sense-annotated Sesotho sa Leboa corpus to address the Meaning Conflation Deficiency problem in word vector space word embedding models within a morphologically rich language setting. The table compares the performance of the baseline and optimized model configurations, thereby showing how model refinement influenced the resolution of conflated meanings in the dataset used for the study. Four models were used in this evaluation: BiLSTM, Optimized BiLSTM, NNLM, and Optimized NNLM. These models were assessed using standard classification metrics—Precision, Recall, F1-score, and Accuracy—as well as Confusion Matrices and ROC Curves for a thorough analysis. The baseline BiLSTM model delivered a balanced performance across all metrics, achieving 70% in precision, recall, F1-score, and accuracy, effectively capturing bidirectional contextual cues. The Optimized BiLSTM exhibited a shift, with precision slightly decreasing to 68% but recall enhancing to 72%, leading to a consistent F1-score of 70% and a higher accuracy of 72%.

These alterations imply that optimization of hyperparameters such as dropout and learning rate help the model generalize better, especially in recognizing ambiguous word meanings underrepresented in the training data. Meanwhile, the NNLM recorded moderate results: 66% precision, 70% recall, 68% F1-score, and 70% accuracy, underscoring its ability to learn sentence-level structures. While lacking sequential recurrence, the NNLM relies on global averaging and feedforward layers, proving to be a simple yet effective baseline. It was noteworthy that Optimized NNLM achieved the highest marks, with 69% precision, 74% recall, 71% F1-score, and 74% accuracy, suggesting that enhancements such

as attention mechanisms or deeper, regularized dense networks considerably boost semantic resolution, crucial for distinguishing subtle sense differences.

Table 6.6: Evaluation Metrics for BiLSTM, Optimized BiLSTM, NNLM, and Optimized NNLM

Model	Precision %	Recall %	F1-score %	Accuracy %
BiLSTM	70	70	70	70
Optimized BiLSTM	68	72	70	72
NNLM	66	70	68	70
Optimized NNLM	69	74	71	74

The chart in Figure 6.25 proves that optimization greatly improved recall and accuracy. The NNLM model showed the most significant enhancement, with recall rising from 70 to 74, the F1 score increasing from 68 to 71, and accuracy climbing from 70% to 74%. Additionally, precision improved from 66% to 69%, indicating that the refined NNLM was able to detect more true positives while generating fewer false positives, making it the leading performer in all four metrics post-optimization. On the other hand, the BiLSTM model experienced an increase in recall from 70% to 72% and accuracy from 70% to 72%; however, precision decreased slightly from 70% to 68%; and the F1 score remained constant at 70%. This indicates that, while the optimised BiLSTM became better at identifying true positives, it also resulted in more false positives. Future strategies might involve adjusting thresholds or class weighting to recover precision without sacrificing recall improvements. Overall, the results suggest that specific tuning enhances both models, with NNLM showing a more favourable response, achieving balanced improvements that boost overall performance.

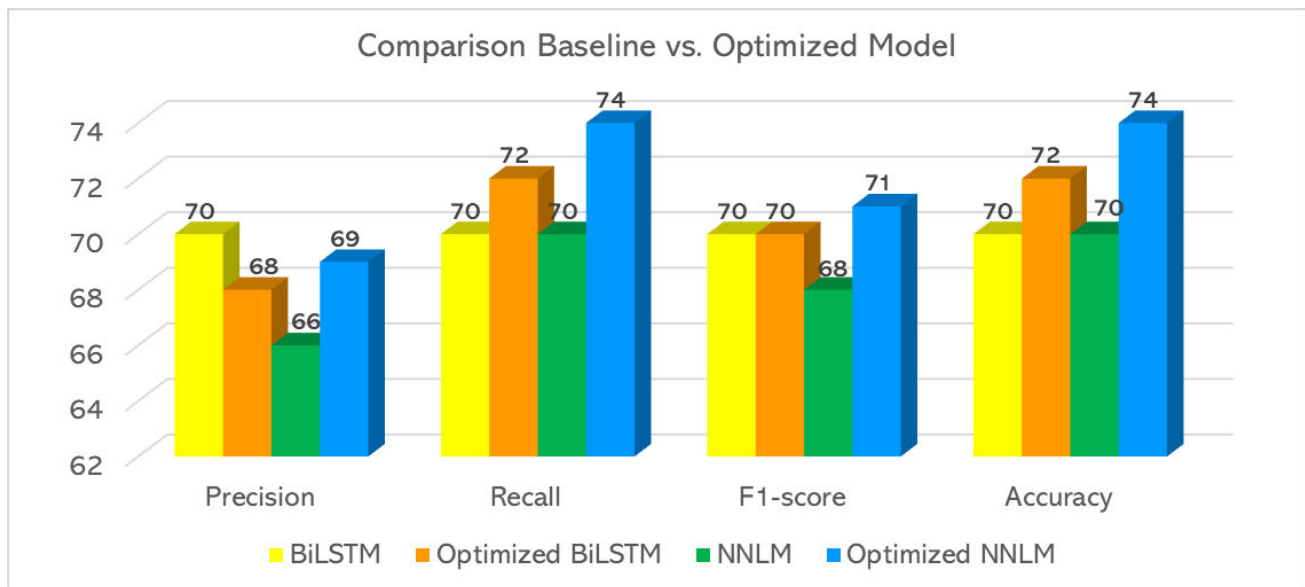


Figure 6.25: Comparison Baseline vs. Optimised Model

The confusion matrix for the BiLSTM model shown in Figure 6.26 highlights both the strengths and weaknesses in addressing WVS WE MCD. The model performs moderately well in distinguishing the four target senses: location, number seven, time, and watch. Among these, the class Number Seven is the most predicted accurately, with 42 correct classifications, followed by location with 36 correct predictions. This suggests that the BiLSTM model can capture distinct contextual patterns for some senses when consistent linguistic cues are present in the dataset.

However, the matrix also reveals significant semantic overlap between location and Number Seven, with a large number of misclassifications occurring in both directions—10 cases of location were misclassified as Number Seven, and 20 instances of Number Seven were misidentified as location. This bidirectional confusion is a typical symptom of meaning conflation, in which the model struggles to differentiate between subtly different semantic roles due to shared contextual features or insufficient discriminative data. Similarly, time was misclassified twice as Number Seven, and watch was misclassified once as Number Seven; yet neither class was strongly predicted, indicating either a shortage of training samples or weak contextual signals for these senses.

Notably, there were no correct predictions for the class watch, pointing to class imbalance or semantic sparsity. The lack of accurate classification for this sense underscores the challenge faced by sequential

models such as BiLSTM when dealing with underrepresented or contextually vague senses. This highlights the need for integrating attention mechanisms, data augmentation, or class weighting strategies to enhance representation for low-frequency senses. In addressing word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD), the confusion matrix shows that, while BiLSTM can utilize sequential context to successfully disambiguate prevalent senses, it struggles with closely related or rare senses, resulting in conflation. This underscores the importance of context-sensitive architectures that emphasize semantically important tokens to reduce misclassifications and to improve sense discrimination overall.

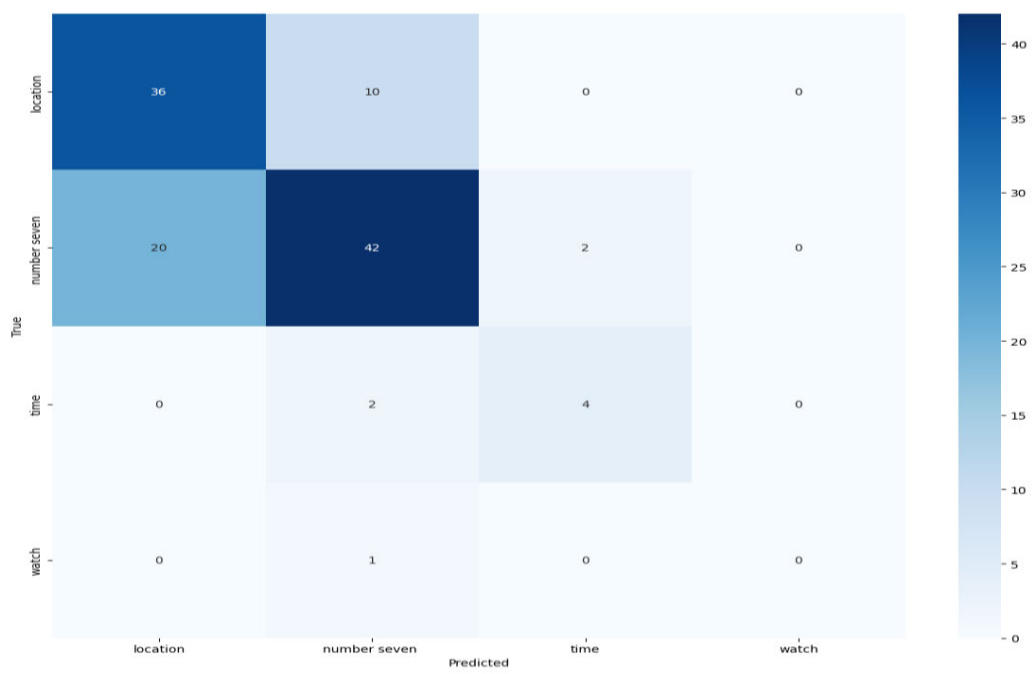


Figure 6.26: Confusion matrix BiLSTM

Figure 6.27 showcases the ROC curve of the BiLSTM model, providing insight into its ability to differentiate between four sense categories: location, Number Seven, time, and watch. The AUC values for each category are revealing, demonstrating the model’s capacity to achieve high true positive rates while keeping false positive rates low. An AUC near 1 signifies excellent discrimination ability, whereas an AUC around 0.5 suggests random performance.

For the “location” class, the ROC curve indicates an AUC of 0.90, highlighting strong discrimination capability. This aligns with the confusion matrix, suggesting that the model has effectively learned to

identify “location” with notable precision. The “Number Seven” class shows an AUC of 0.86, reflecting good discrimination, albeit slightly lower. This matches the confusion matrix pattern, in which “Number Seven” is frequently confused with “location,” indicating some difficulty in distinguishing these two classes, although performance is still commendable.

Notably, the “time” category achieves a remarkable AUC of 0.97, indicating superior discrimination despite potentially having fewer examples. This suggests that when “time” is correctly captured in the input, the model is highly confident in its classification, although the limited number of “time” instances may be skewing the AUC slightly – the model could be recognizing very distinct samples. For the “watch” class, an AUC of 0.89 is observed, which is unexpectedly high given the low classification accuracy reflected in the confusion matrix. This implies that, despite the low prediction accuracy for “watch,” the model still assigns this class fairly distinct probability scores, suggesting that improvements could be possible with more data or better class balancing.

Overall, the ROC curves and AUC values indicate that the BiLSTM model excels at sense discrimination, particularly for well-represented classes such as “location” and “Number Seven,” while also competently managing less common senses such as “time” and “watch.” These results suggest that the model’s probability predictions are dependable; predictions could be further refined using techniques such as attention mechanisms, data augmentation for rare classes, or class-weighted loss functions to tackle the misclassification issues suggested by the confusion matrix.

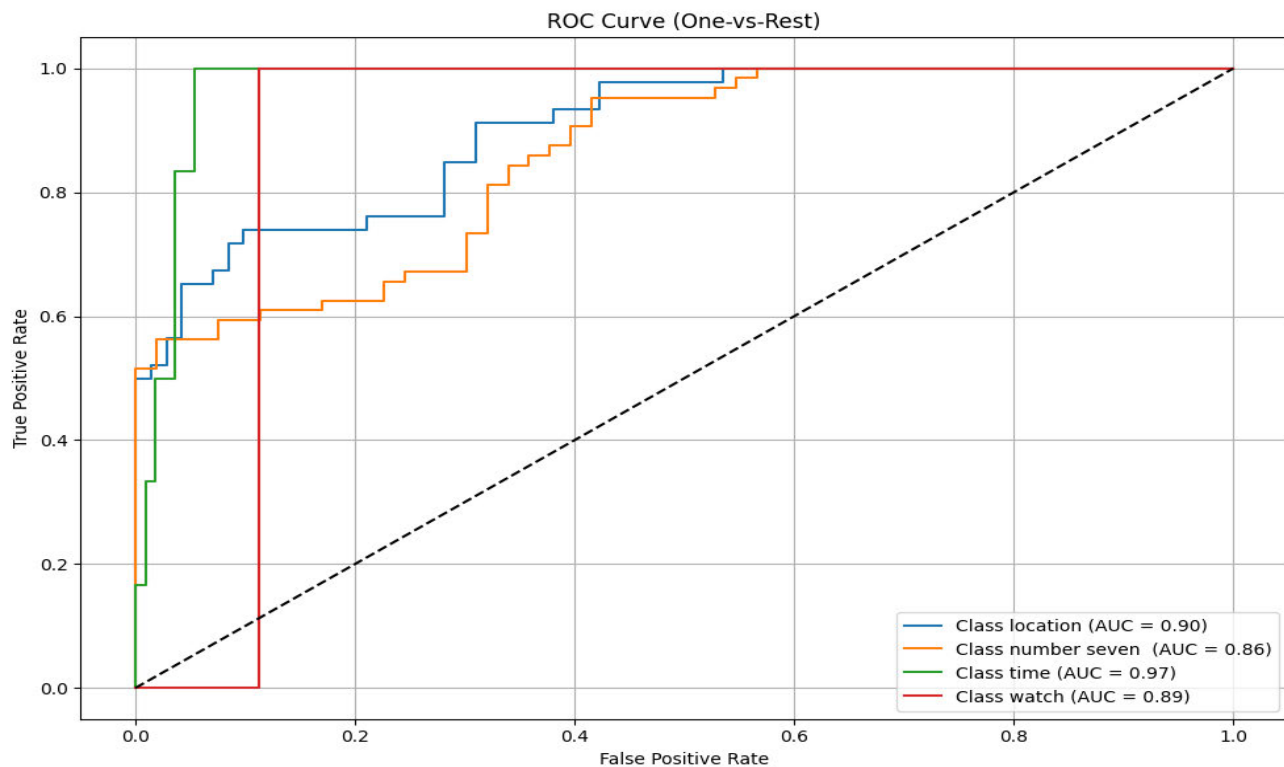


Figure 6.27: ROC curve BiLSTM

In Figure 6.28, the confusion matrix for the optimised BiLSTM model demonstrates significant advances in differentiating word senses compared with the baseline BiLSTM. The matrix indicates that the model accurately classifies 51 instances of “Number Seven,” marking the strongest performance across all categories. The next best result is for “location” with 33 correct classifications, indicating that optimization has bolstered the model’s precision and confidence, especially within the most common classes. Despite these improvements, some semantic confusion persists between “location” and “Number Seven,” with 13 cases of “Number Seven” mistakenly identified as “location” and vice versa.

This bidirectional misclassification, although slightly improved over the baseline, still suggests an underlying semantic overlap or shared context cues, possibly because these terms frequently occur in similar sentence structures. Such issues align with the challenges of WVS WE MCD, in which closely related senses defy clear separation using only surface features. In terms of less frequent classes, the optimised BiLSTM shows some aptitude for recognizing the “time” class, correctly identifying 4 instances, although 2 were mistakenly classified as “Number Seven.” This further highlights the challenge posed by semantic sparsity in less common senses. The sense “watch” continues to be

problematic, with only 1 misclassification into “Number Seven” and no correct predictions, underscoring the model’s difficulty with highly sparse categories.

This suggests that, while optimization enhances the model’s overall representational capabilities, more sophisticated strategies—for instance sense-augmented embeddings, external knowledge incorporation, or data balancing—are needed to manage rare senses effectively. In summary, the confusion matrix for the optimised BiLSTM model shows clear improvements in classification accuracy for predominant senses and a slight decrease in mutual confusion between major classes. Nonetheless, it confirms that the model’s ability to generalize to low-frequency or nuanced senses remains limited, emphasizing the necessity for further architectural or training data enhancements to effectively address WVS WE MCD in these scenarios.

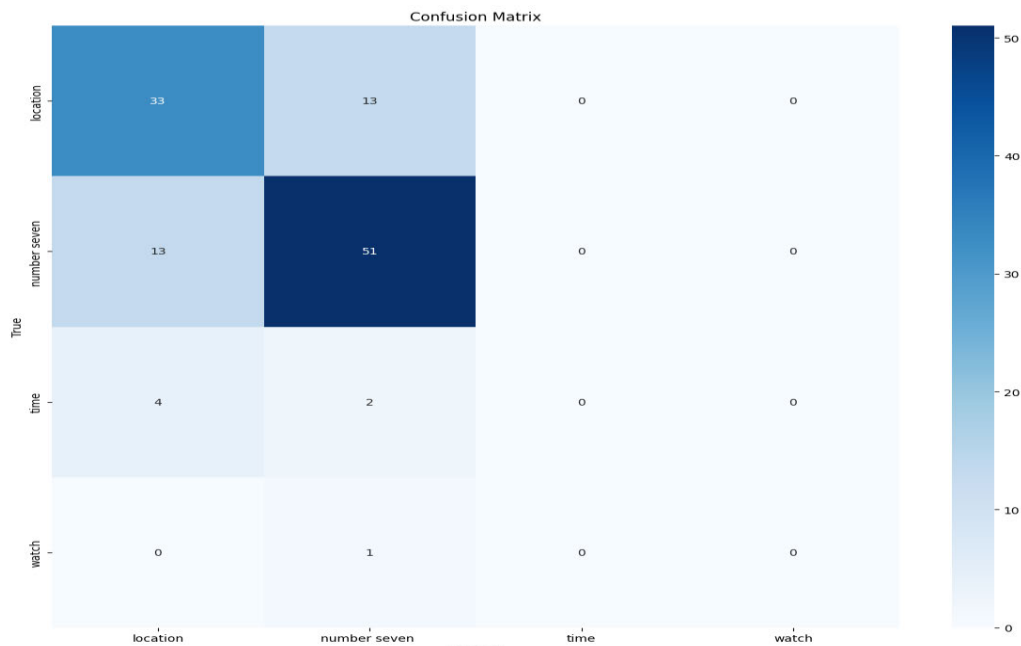


Figure 6.28: Optimized BiLSTM Confusion matrix

Figure 6.29 presents the ROC curves for the optimized BiLSTM model, offering valuable insight into its ability to distinguish among the four semantic categories: location, Number Seven, time, and watch. Each curve represents a one-vs-rest binary classification, with the AUC values serving as indicators of the model’s effectiveness in separating each class from the others. AUC scores above 0.80 are generally considered strong, reflecting reliable classification performance.

The model achieved its highest AUC of 0.83 for the “time” category, suggesting a particularly strong ability to identify this sense, especially when clear contextual cues are present. This marks a modest improvement over the baseline BiLSTM, likely due to optimization techniques such as regularization, dropout, or hyperparameter tuning, which enhanced the model’s sensitivity to temporal features. Similarly, the AUC values for “location” (0.81) and “Number Seven” (0.80) demonstrate solid predictive performance. Despite some class overlap noted in the confusion matrix, the consistently high ROC curves imply that when the model does make correct predictions, it does so with a high degree of confidence.

Conversely, the “watch” class exhibits notably low performance, with an AUC of 0.37, falling well below the baseline of 0.5, which indicates random guessing. This weak performance aligns with prior confusion matrix findings in which “watch” was invariably misclassified. The curve’s flat design and low AUC imply that the model not only fails in correct classification of this class but also lacks confidence in such predictions. This is likely attributed to severe class imbalance or absence of distinct contextual indicators in the training data for “watch,” inhibiting the model’s ability to grasp meaningful representations for it. Ultimately, the ROC curve analysis of the optimised BiLSTM affirms that the model excels with high-frequency and semantically distinct classes such as location, Number Seven, and time. Nevertheless, its proficiency in detecting and isolating lower frequency or ambiguously contextual senses such as “watch” remains weak. These findings underscore the significance of class balance, contextual enrichment, and attention mechanisms in future model developments to address the MCD problem in WVS WE models, especially for low-resource language environments.

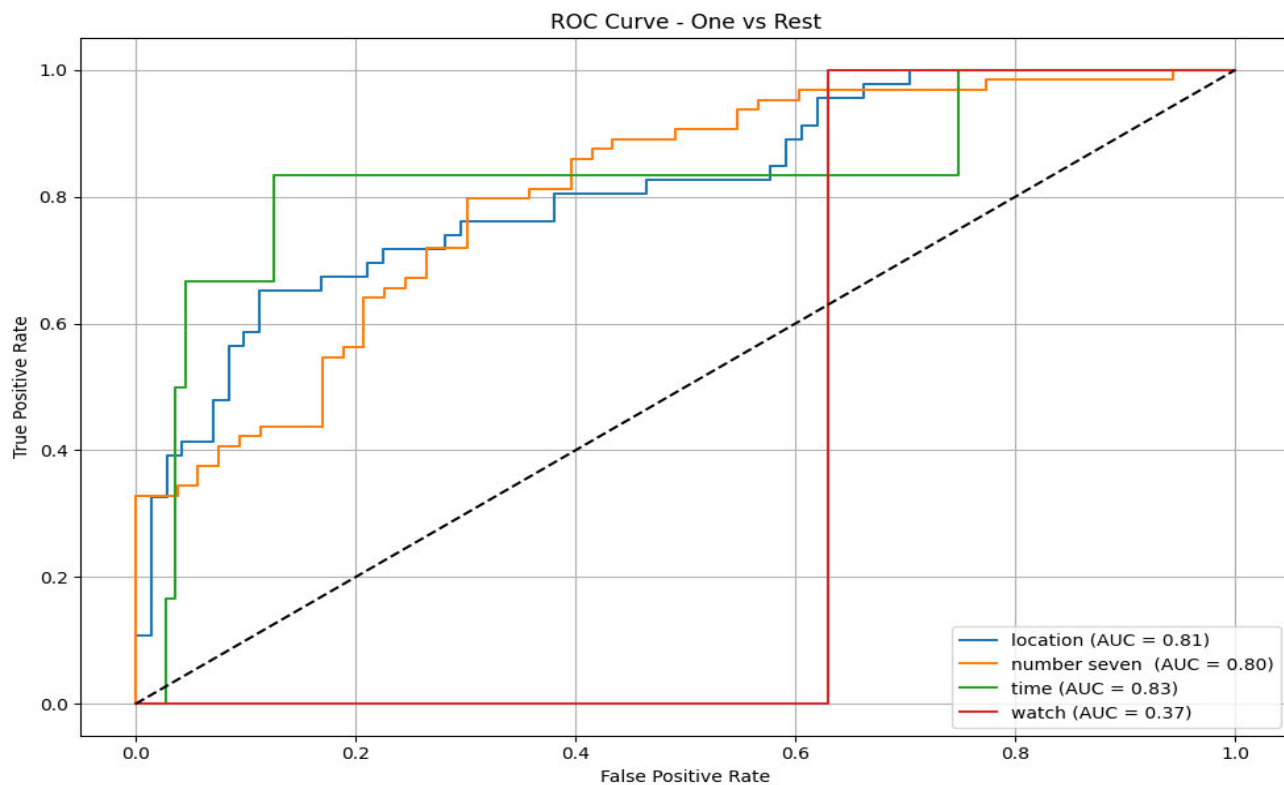


Figure 6.29: Optimized BiLSTM ROC curve

The confusion matrix for the NNLM depicted in Figure 6.30 reveals how effectively the model can differentiate among the word senses: location, Number Seven, time, and watch. Initially, the NNLM demonstrates proficient classification for the two predominant senses, “location” and “Number Seven,” with 33 and 49 predictions made accurately, respectively. This indicates the NNLM’s ability to grasp broader sentence-level semantics, particularly for prominent sense categories exhibiting unique contextual patterns. Nevertheless, akin to earlier models, the NNLM continues to blur the line between “location” and “Number Seven,” evidenced by 13 “location” instances misclassified as “Number Seven” and 15 “Number Seven” instances mistaken for “location.”

This highlights that, even though the NNLM develops meaningful representations, it still has difficulty in distinguishing subtle semantic boundaries between closely related senses—an indicator of WVS WE MCD. Despite employing a dense layer and averaging-based architecture, NNLM seems constrained in recognizing minute contextual differences when two senses overlap in their usage. The model yields surprising results in correctly classifying the “time” sense, with all 6 instances erroneously classified

as “Number Seven.” This reveals a major semantic fusion, suggesting that the model possibly overgeneralizes numeric contexts or places undue reliance on overt word cues.

The complete inability to accurately classify “time” signifies that the basic NNLM lacks the sequential acuity or attention necessary to differentiate structurally similar yet semantically distinct temporal expressions. Regarding the “watch” sense, the model achieves only one prediction, which is incorrect and classified as “location.” This shows that the NNLM, in its present form, fails to effectively detect “watch.” The root cause likely involves a combination of data sparsity, semantic subtlety, and the model’s inability to discern token-level distinctions—all crucial for differentiating low-frequency senses.

Overall, the confusion matrix for the NNLM model illustrates a moderate proficiency in distinguishing high-frequency classes but shows limitations in resolving subtle or rare senses. The persistent confusion between “location” and “Number Seven,” coupled with the wholesale misclassification of “time” and “watch,” underscores the limitations of superficial sentence-level encoding in disambiguating polysemous terms. These observations suggest the necessity for context-aware improvements—such as attention mechanisms or multi-head representations—to counteract the semantic flattening that occurs with conventional NNLM designs, especially in contexts marked by WVS WE MCD.

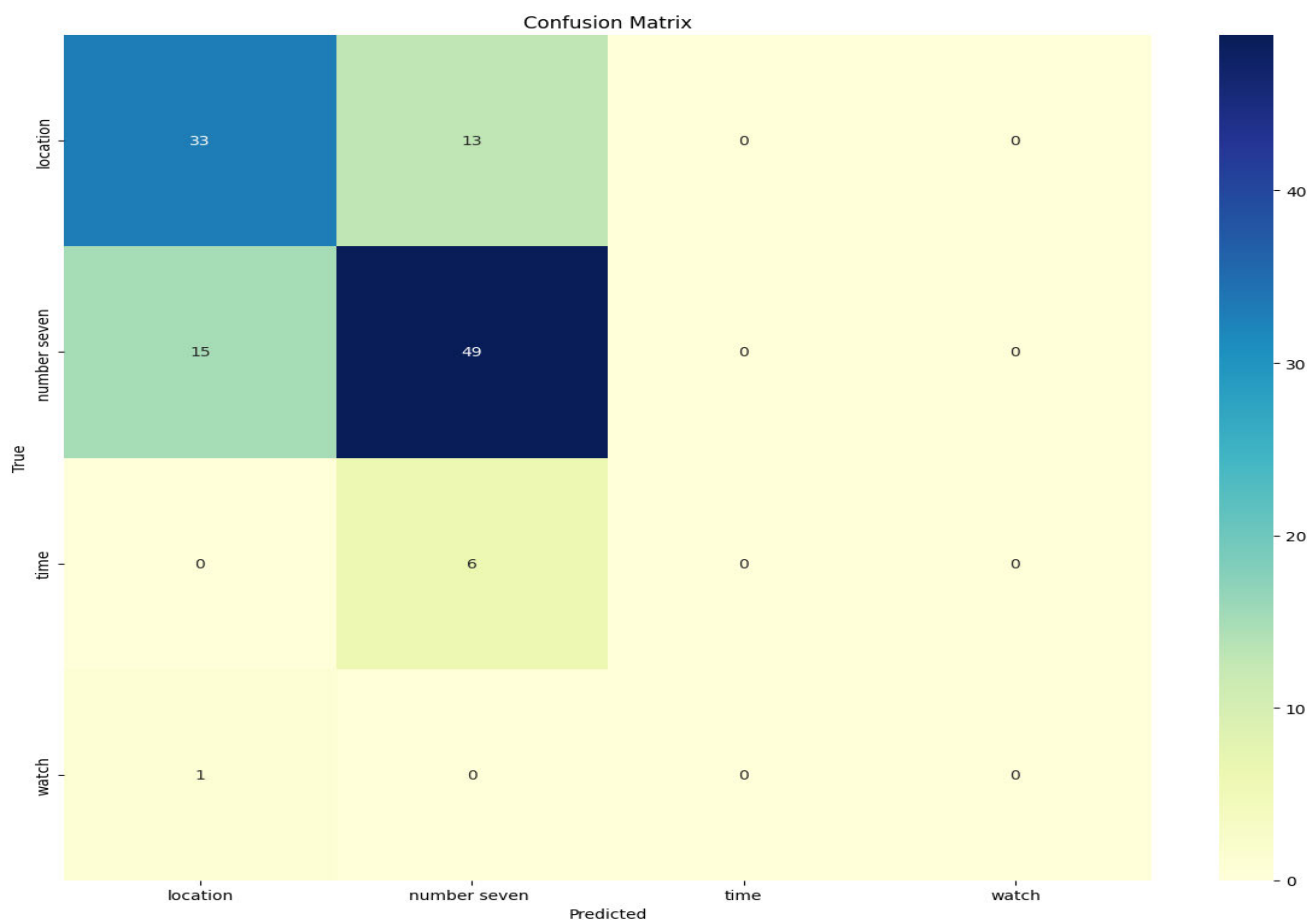


Figure 6.30: NNLM confusion matrix

The ROC curve depicted in Figure 6.31 for the NNLM unveils crucial details about the model’s ability to distinguish between four sense classes: location, Number Seven, time, and watch, by employing a one-vs-rest evaluation approach. The AUC values for each class reflect the model’s proficiency in assigning higher probabilities to accurate predictions while minimizing false positives. The “time” class achieves the highest AUC of 0.96, indicating excellent separability and confidence when the model identifies this sense. This finding is noteworthy given that “time” was frequently misclassified in the confusion matrix.

A high AUC suggests that, although the model might not always rank the correct class as its top prediction, it still assigns relatively high probabilities to it, which could be optimised with threshold adjustments or ensemble calibration. The “location” class also shows strong performance, with an AUC of 0.86, highlighting the NNLM’s capability to consistently identify this sense in diverse contexts. The class “Number Seven” follows with an AUC of 0.80, suggesting reasonable discriminative capacity.

These results affirm the model’s effectiveness in classifying prevalent or well-represented senses, especially when the input context is clear. Conversely, the model’s performance for the “watch” class is weaker, presenting an AUC of 0.65. While this surpasses random guessing (AUC of 0.50), it is notably lower than the other classes.

The ROC curve reveals the model’s difficulty in reliably scoring the “watch” class, likely due to restricted training data or contextual ambiguity. Compared with the baseline BiLSTM, where the “watch” class had an AUC of 0.37, this improvement in NNLM demonstrates some potential for better sense detection through non-recurrent architectures when properly optimised. Overall, the ROC curves for the NNLM model illustrate its robust performance in major sense classes and indicate potential for underrepresented ones. While its sentence-level encoding strategy is effective for broad semantic classification, resolving meaning conflation — particularly in challenging cases such as “watch” — may require architectural improvements, such as attention layers, contextual embeddings, or sense-augmented pretraining, to enhance confidence and minimize prediction score overlap.

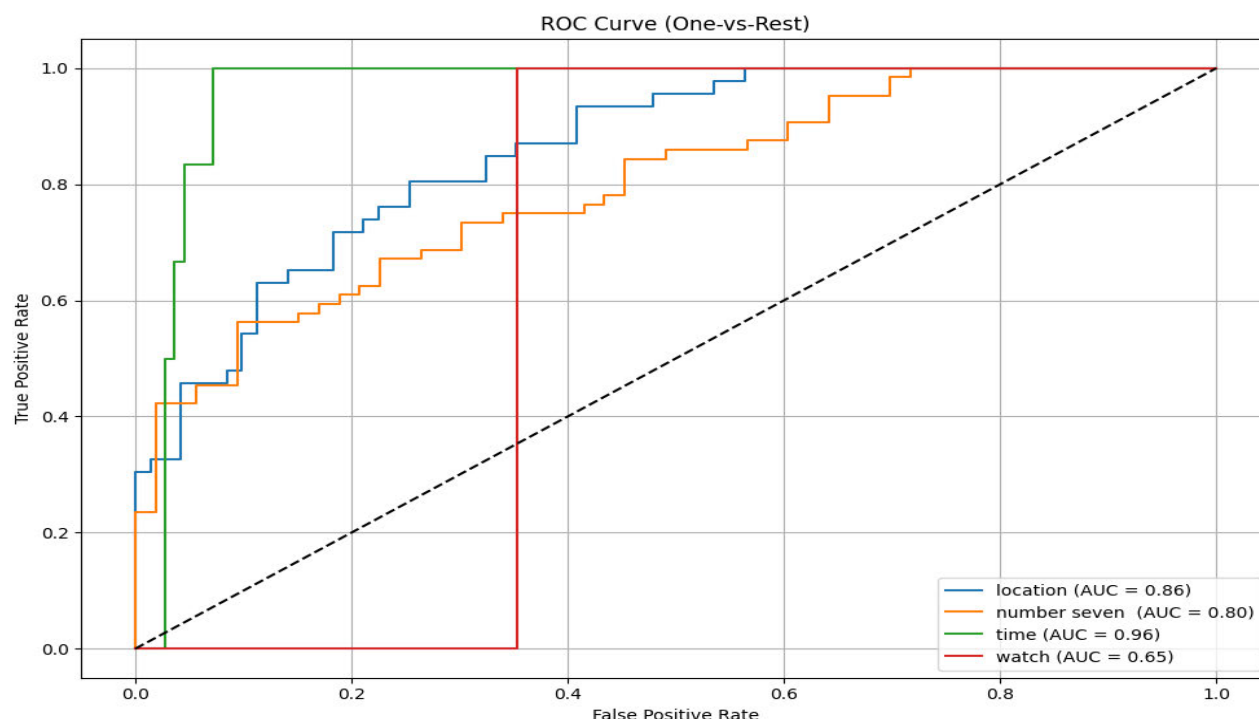


Figure 6.31: NNLM ROC curve

The confusion matrix for the Optimised NNLM depicted in Figure 6.32 showcases significant enhancements in addressing WVS WE MCD, particularly when compared with the baseline model.

Notably, the model demonstrates strong performance in classifying the two most frequent senses: “Number Seven” and “location.” The model accurately categorizes 50 instances of “Number Seven” and 36 instances of “location,” suggesting that optimizations—such as hyperparameter tuning, deeper layers, or regularization—have improved the model’s ability to differentiate dominant sense categories. However, confusion between “Number Seven” and “location” persists. The model incorrectly labels 14 instances of “Number Seven” as “location” and 10 instances of “location” as “Number Seven.” This symmetrical confusion illustrates that although the optimised NNLM shows increased accuracy, it still struggles to distinguish between senses that often occur in similar syntactic or semantic settings.

This challenge is common in MCD scenarios, in which sense boundaries are blurred due to polysemy and overlapping usages. Remarkable progress is seen in classifying the “time” category, with all 6 instances now correctly identified, compared with complete misclassification by the baseline NNLM. This suggests that the optimization has improved the model’s ability to focus on temporal cues necessary for distinguishing this sense, likely through enhanced context representation or better weight initialization. Conversely, the model continues to struggle with the “watch” class, failing to predict any instances correctly. One instance is misclassified as “Number Seven,” with none recognized as “watch.” This pattern indicates that the issue might not be architectural but data-related—possibly due to class imbalance or inadequate contextual differentiation in the training data. Rare senses such as “watch” might benefit from strategies such as data augmentation, external knowledge integration, or sense-aware embeddings to enhance representation and classification accuracy.

The confusion matrix for the optimised NNLM indicates promising improvements in differentiating dominant senses and previously misclassified classes, for instance, “time.” However, the remaining confusion between “location” and “Number Seven,” alongside the ongoing misclassification of “watch,” highlights the persistent challenge of meaning conflation in low-resource semantic tasks. Future advancements may be achieved by adopting attention mechanisms or by integrating the NNLM with contextualized transformer-based encoders to capture more nuanced sense distinctions.

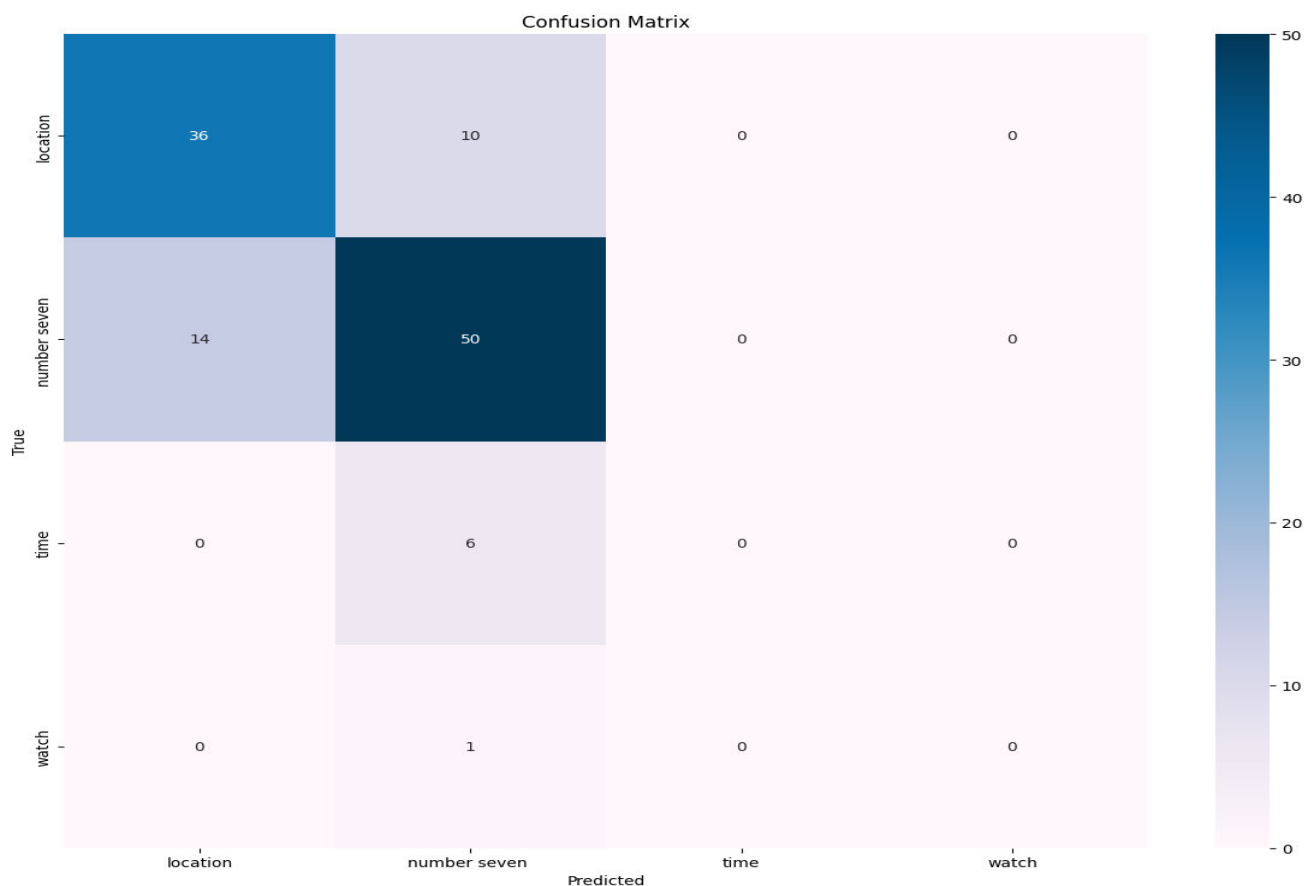


Figure 6.32: Optimized NNLM Confusion Matrix

Figure 6.33's ROC curve illustrates a significant improvement in the optimised NNLM capacity to distinguish among four (4) word senses: *location*, *Number Seven*, *time*, and *watch*. This enhancement is substantiated by the AUC values, which offer a robust indicator of class-level discriminability in a one-vs-rest framework. All four classes exhibit high AUC values, signifying that the optimised NNLM effectively assigns strong class probabilities while maintaining relatively low false positive rates. The highest AUC of 0.97 belongs to the sense "time," reflecting exceptional classification reliability, mirrored by the confusion matrix's accuracy, in which all instances of "time" are correctly identified.

The near-perfect separation on the ROC curve indicates that the optimised model captures temporal cues for this class with precision, potentially due to improved weight initialization, dropout regularization, or enhanced non-linear transformations during optimization. The "location" sense also performs well, with an AUC of 0.88, while "Number Seven" achieves an AUC of 0.84, showing that these frequently occurring classes are consistently distinguished from others. Despite some overlap

noted in the confusion matrix, the ROC curve demonstrates the model's confidence in its predictions, particularly when context is clear. The results affirm the optimised NNLM's ability to encode global sentence-level semantics, aiding in sense disambiguation for prominent classes.

The most notable improvement is observed in the "watch" class, which previously evidenced low accuracy and weak AUC values in the baseline NNLM. The optimised model has markedly improved, attaining an AUC of 0.88 — an exceptional outcome, especially since "watch" was previously entirely misclassified. The sharp incline in its ROC curve suggests that the model now achieves clearer probabilistic distinctions for this infrequent class. Although the confusion matrix still records zero correct predictions for "watch," the high AUC indicates that the model attributes relatively higher confidence scores to this class, suggesting that slight threshold adjustments or additional tuning could fully realize its predictive capabilities.

Altogether, the ROC curves for the optimised NNLM highlight a marked advancement in addressing meaning conflation deficiency, particularly for semantically overlapping and low-resource senses. The uniformly high AUC values across all classes, including the rare "watch" sense, suggest that the optimization process has markedly enhanced the model's ability for nuanced semantic differentiation. These findings endorse the application of refined feedforward architectures for word sense disambiguation, particularly when complemented by regularization, deeper hidden layers, and improved input encoding techniques.

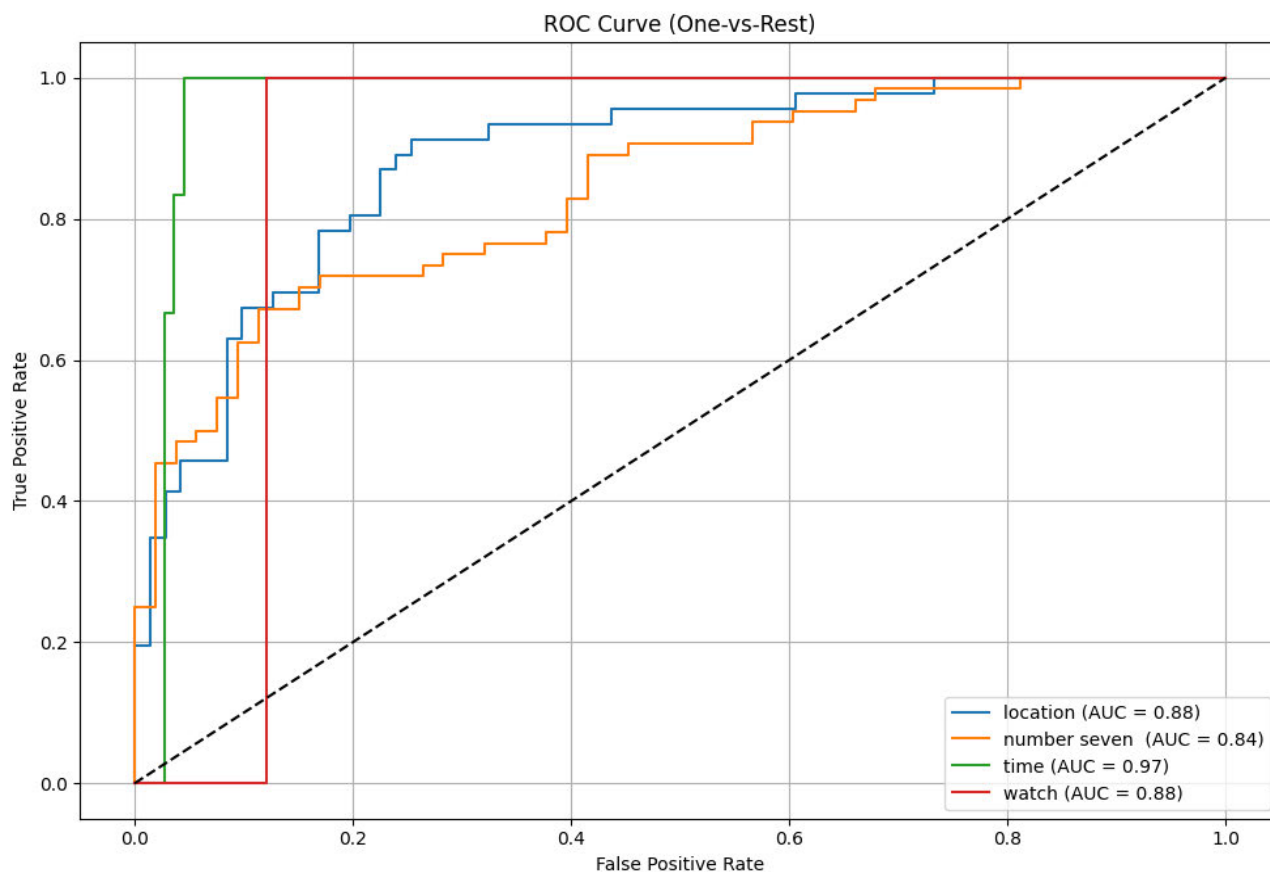


Figure 6.33: Optimized NNLM ROC Curve

The confusion matrices for all the models shed more light on their behaviour. Both Optimized NNLM and Optimized BiLSTM models showed better class separation, evidenced by fewer errors along the diagonal, suggesting enhanced sense discrimination. Furthermore, ROC curves validated that the optimised models performed better, with most classes seeing higher AUC scores. This supports the assertion that optimising architectural features—be it through attention mechanisms in NNLM or hyperparameter tuning in BiLSTM—is crucial in resolving semantic ambiguity due to meaning conflation. Overall, these experiments affirm the hypothesis that optimised models, particularly those utilizing context-sensitive architectures with attention, are more effective in reducing WVS WE MCD, notably in low-resource contexts in which morphological differences can blur sense boundaries. The findings also imply that efficiently optimised lightweight feedforward models such as NNLM can provide both efficacy and efficiency, making them suitable for real-world, resource-limited scenarios.

6.6.3 Experimental Results with BERT

Table 6.8 displays the efficacy of the BERT-based methodology in addressing the MCD problem in WVS WE models. The model attained a notable accuracy of 87%, demonstrating its robust capability to accurately assign meanings to ambiguous words in diverse circumstances. The F1-score of 60%, with an accuracy of 68% and a recall of 57%, indicates a performance disparity. The increased precision indicates that the model is both cautious and certain in its predictions, resulting in a reduced number of false positives. Conversely, the diminished recall suggests a potential failure to identify all relevant occurrences of each sensation, presumably attributable to the underrepresentation of certain senses or nuanced contextual differences.

This mismatch underscores a significant problem in resolving WVS WE MCD challenges: whereas BERT’s contextual embeddings facilitate strong performance on prevalent or well-represented senses, it may encounter difficulties with less common or closely related ones, particularly in low-resource environments. Nevertheless, the model’s elevated overall accuracy and satisfactory F1-score indicate that BERT is an effective instrument for disambiguation, and its efficacy might be augmented by sense-balanced training data, fine-tuning, or the incorporation of attention-based upgrades. These findings validate BERT’s proficiency in capturing the contextual semantics necessary for addressing the MCD problem in WVS WE models.

Table 6.7: Bert Performance Metrics

Model	Precision	Recall	F1-Score	Accuracy
BERT	68	57	60	87

Figure 6.34 depicts the training and validation loss and accuracy curves for the BERT-based model used in addressing MCD problem in WVS WE models. The left panel illustrates a uniform and parallel reduction in both training and validation loss over three epochs, indicating steady learning and little overfitting. The right panel shows a consistent increase in both training and validation accuracy, with the model achieving over 86% accuracy on the validation set by the third epoch—signifying successful generalization to novel data.

Notably, the validation accuracy marginally exceeds the training accuracy in the last epoch, perhaps indicating effective regularization, a well-balanced dataset, or minor underfitting that may improve with more training. These curves demonstrate that BERT’s contextualized embeddings effectively capture subtle variations in word meanings, even with restricted training periods. The model’s stability and convergence demonstrate its efficacy in addressing MCD in WVS WE models, particularly in morphologically rich, low-resource languages where disambiguation poses significant difficulties. Learning had not converged by epoch 3 because both the loss and accuracy curves were still changing substantially. The training and validation losses continued to decrease, while training and validation accuracies continued to increase, indicating that the model was still learning and had not yet reached a stable optimum. The absence of a plateau in the validation curves suggests that additional epochs were still needed to determine whether the model would converge or begin to overfit.

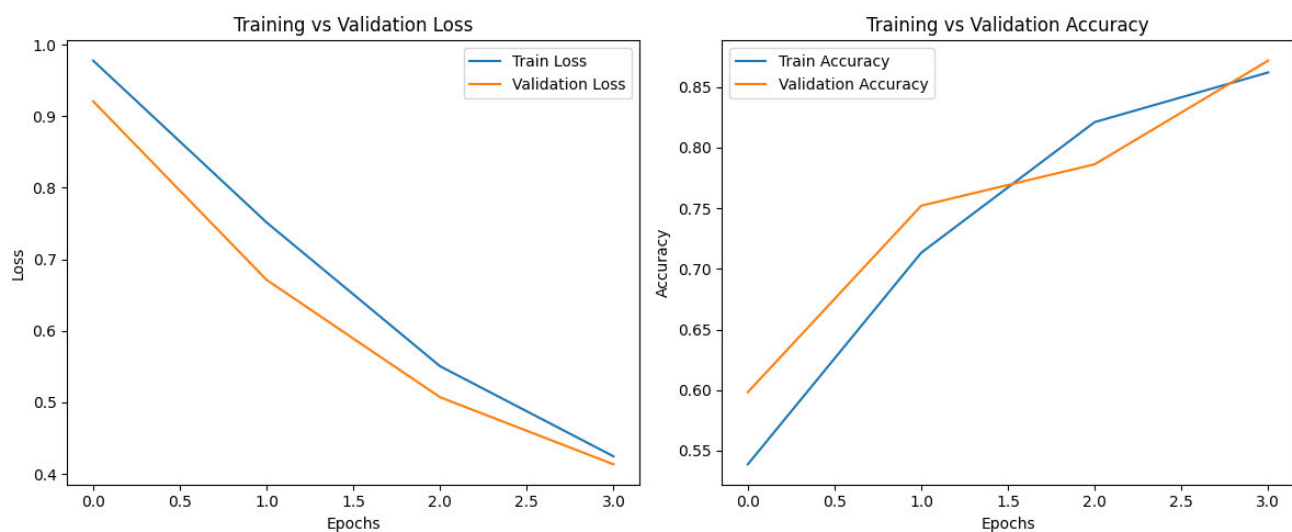


Figure 6.34:Model Training Loss and Accuracy

Figure 6.35 illustrates the confusion matrix for the BERT model used in resolving MCD in WVS WE models, providing an overview of the model’s performance across several classes. The matrix demonstrates that BERT excelled at differentiating between the two predominant senses: "location" and "Number Seven". The model accurately identified 43 of 45 "location" examples and 55 of 62 "Number Seven" instances, demonstrating its robust ability to discern semantic differences within predominant categories.

Nevertheless, the matrix also discloses zones of ambiguity. For instance, seven occurrences of "Number Seven" were inaccurately categorized as "location," and five instances of "time" were erroneously classified as "Number Seven," indicating potential semantic overlap or inadequate representation of minority classes in the training set. The uncommon category "watch" included just one occurrence, which was incorrectly categorized as "location," highlighting the difficulty of managing low-frequency meanings in resource-constrained environments.

The matrix demonstrates BERT’s strong generalization for high-frequency senses and its context-aware capabilities, while also underscoring the persistent challenges of sense imbalance and label sparsity—prevalent concerns in resolving MCD challenges in WVS WE models. These results highlight the need for either data augmentation or sense-aware fine-tuning to enhance BERT’s performance on infrequent or nuanced word senses.

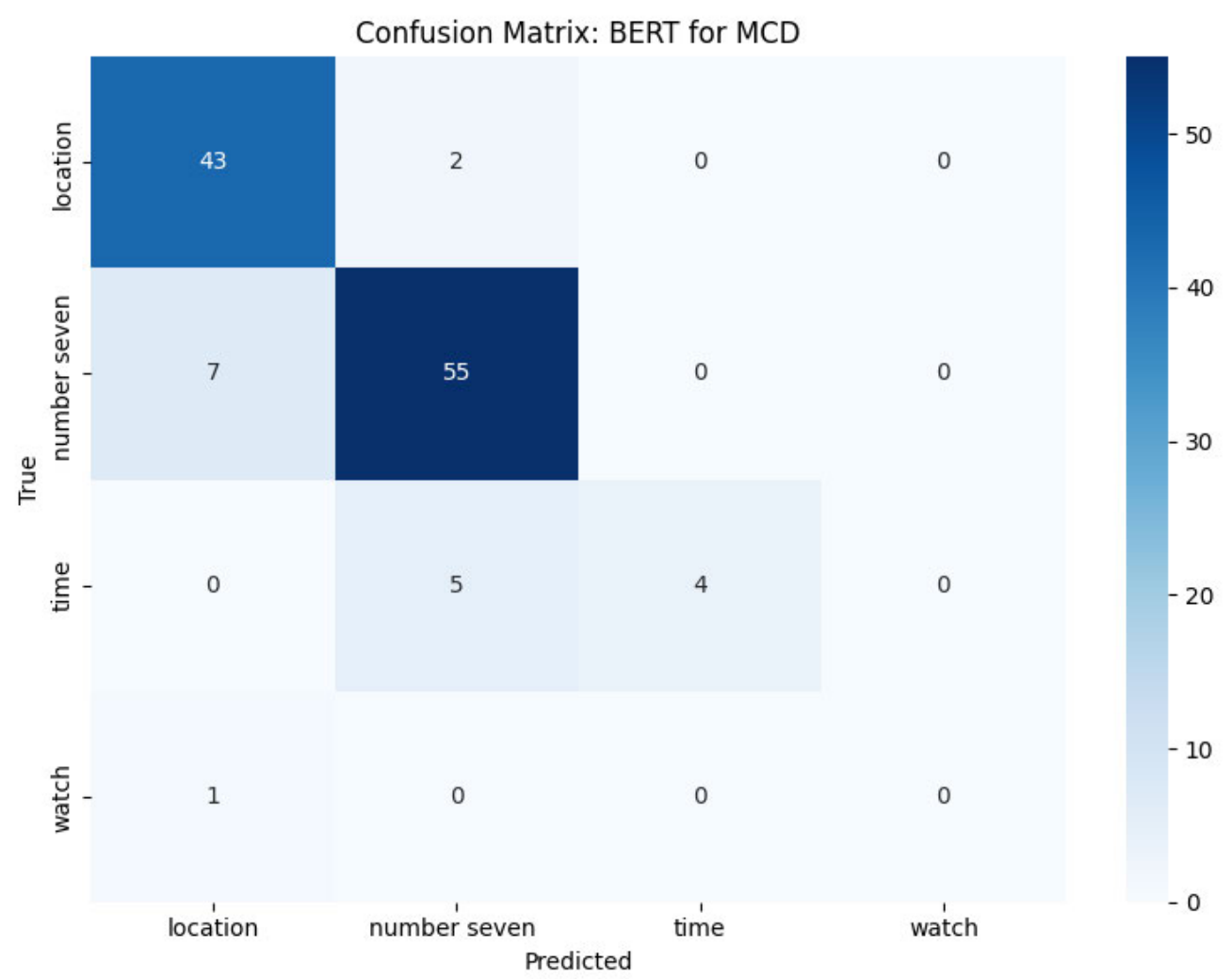


Figure 6.35:BERT Confusion matrix

6.6.4 Experimental results with Multilingual Bert

The performance indicators for Multilingual BERT in Table 6.9 regarding the resolution of MCD problem in WVS WE models, demonstrate a notable combination of strengths and weaknesses. The model exhibits a remarkable accuracy of 89%, indicating a robust capacity to accurately assign meanings to words across diverse contextual settings, even in a low-resource linguistic environment. Nonetheless, the comparatively low F1-score of 55%, coupled with an precision of 69% and a recall of 54%, signifies a performance disparity. The elevated accuracy indicates that the model is exceptionally selective and prudent, often accurate in its predictions; nonetheless, the diminished recall highlights its challenges in identifying all genuine occurrences of each sense class, especially those that are infrequent or more contextually intricate.

This gap underscores a prevalent issue in multilingual and morphologically rich languages: although having robust language modelling skills, Multilingual BERT may nonetheless exhibit suboptimal performance in fine-grained disambiguation without task-specific fine-tuning or sense-balanced datasets. Nevertheless, the increased accuracy validates the model’s contextual robustness and transferability across languages, establishing it as a suitable basis for future adaptation and improvement in low-resource word sense disambiguation tasks.

Table 6.8: Multilingual BERT Performance Metrics

Model	Precision	Recall	F1-Score	Accuracy
Multilingual BERT	69	54	55	89

Figure 6.36 illustrates the training and validation loss and accuracy trajectories for the Multilingual BERT model used to tackle MCD problem in WVS WE models. The loss plot (left) indicates a steady and pronounced decrease in both training and validation loss over the three epochs, with little divergence between the two curves. This reflects consistent convergence and efficient generalization without evidence of overfitting. The accuracy plot (right) corroborates this, showing a swift ascent in training accuracy from little over 50% to beyond 90%, while validation accuracy consistently improves, reaching 89% by the third epoch.

The proximity of training and validation measures suggests that the model is effectively learning representative contextual characteristics and successfully applying that knowledge to novel data, despite the complexities and ambiguities inherent in word meaning disambiguation tasks. These patterns indicate that Multilingual BERT adapts well even with constrained training data, making it particularly appropriate for MCD in WVS WE models. Nonetheless, further fine-tuning, regularization, or class-balancing techniques may be required to enhance memory and improve the management of uncommon or underrepresented senses.

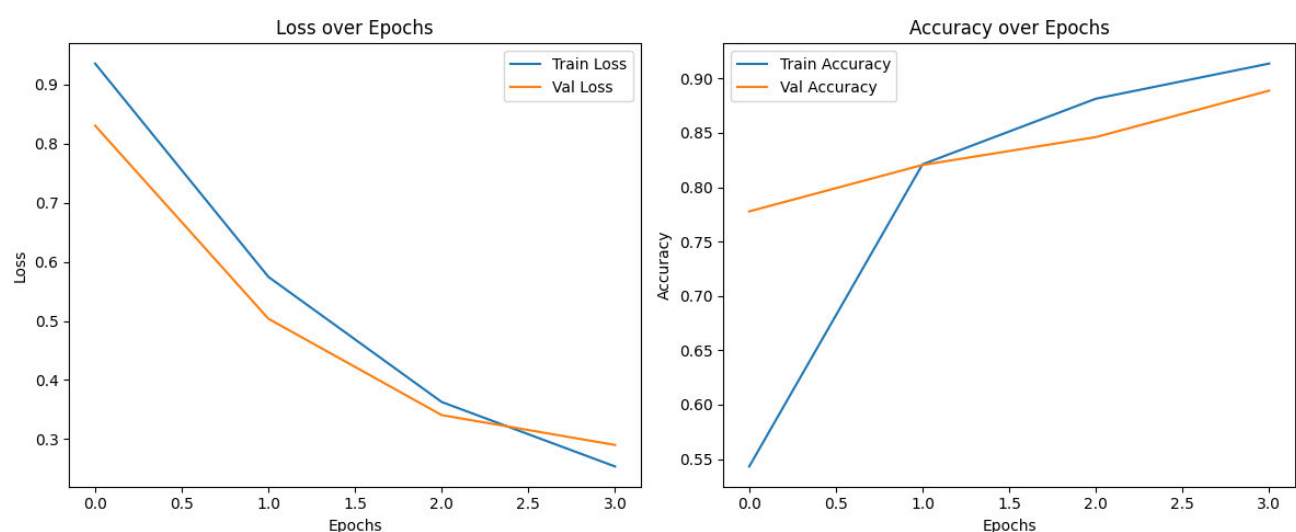


Figure 6.36: Multilingual BERT loss over epochs and accuracy over epochs

Figure 6.37 displays the confusion matrix for the Multilingual BERT model used in addressing MCD in WVS WE models. The model has robust efficacy in accurately recognizing the two predominant senses: “location” and “Number Seven,” achieving 45 and 57 correct classifications, respectively. Significantly, “location” is categorized with impeccable accuracy, with only a limited number of "Number Seven" cases (5) erroneously labeled as “location,” demonstrating good contextual differentiation between these often conflated meanings.

Nonetheless, the matrix also discloses several problems. The concept of “time” is misclassified, with 7 of 9 occurrences erroneously identified as “Number Seven,” indicating the model's difficulty with infrequent and contextually nuanced categories. The uncommon class “watch” is inaccurately categorized as “location,” illustrating the common challenges deep models have when trained on unbalanced datasets with underrepresented categories.

The matrix demonstrates that Multilingual BERT effectively generalizes for high-frequency senses but needs further tuning or augmentation to enhance its sensitivity to low-frequency or semantically analogous senses—a crucial factor in developing strong word sense disambiguation systems for low-resource languages.

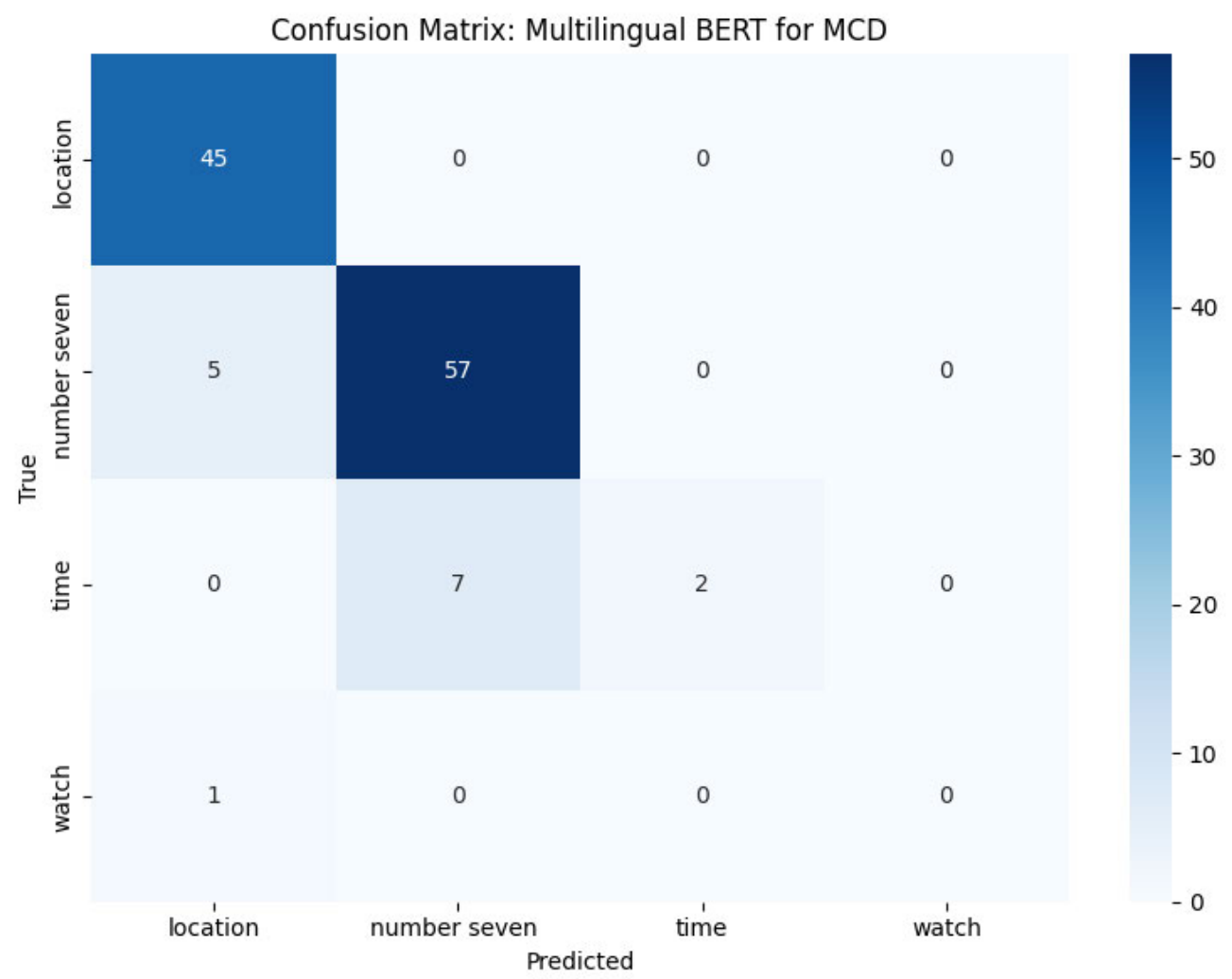


Figure 6.37: Multilingual BERT confusion matrix

6.6.5 Experimental results with ELECTRA

Table 6.10 delineates the performance characteristics of the Multilingual ELECTRA model, indicating a precision of 39%, a recall of 41%, and an F1-Score of 40%, accompanied by a comparatively high accuracy of 76%. The discrepancy between the classification metrics (precision, recall, and F1-score) and the overall accuracy underlines that, although the model can accurately predict a significant number of instances overall (probably due to class imbalance), it fails to sustain uniform performance

across all classes. This suggests that the model may exhibit bias towards majority courses, attaining elevated accuracy by prioritizing them, while failing on minority or more difficult classes. Consequently, despite the ostensibly acceptable accuracy, the low F1-score underscores that the model's predictions lack sufficient reliability when accounting for both precision and recall, which are essential for assessing performance in tasks such as word sense disambiguation (WSD), particularly in low-resource or morphologically rich languages.

Table 6.9: ELECTRA Performance Metrics

Model	Precision	Recall	F1-Score	Accuracy
ELECTRA	39	41	40	76

Figure 6.38 depicts the training dynamics of the ELECTRA model throughout three epochs, highlighting the patterns in loss and accuracy for both training and validation datasets. The left subplot displays a consistent decline in both training and validation loss curves, showing that the model is acquiring knowledge efficiently and is not overfitting during the first epochs. The training loss decreases more swiftly than the validation loss, reflecting that the model fits the training data somewhat better, while generalization remains consistent. The training and validation accuracy curves in the right subplot exhibit a steady rising trajectory. The training accuracy markedly increases from little over 0.50 to almost 0.89; however, the validation accuracy rises from 0.53 to around 0.76, which corresponds well with the accuracy measure (76%) shown in Table 5.17. Although the precision (39), recall (41), and F1-score (40) are modest, the elevated validation accuracy indicates effective classification by the model; nonetheless, class imbalance or label noise may be influencing the other metrics. The figure substantiates ELECTRA's effective learning and robust generalization over the epochs.

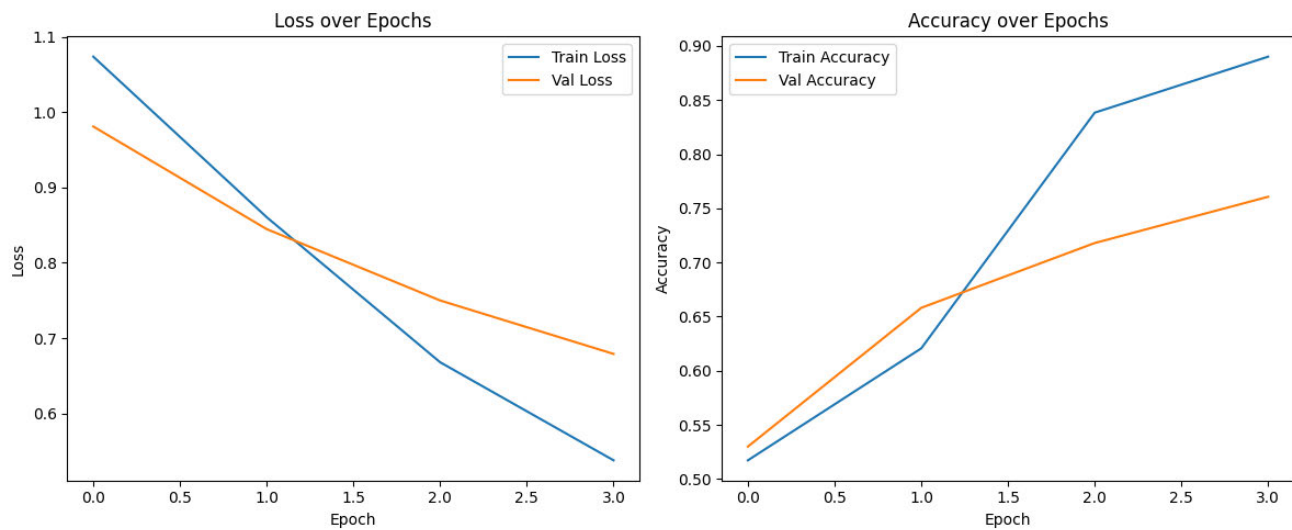


Figure 6.38: ELECTRA loss over epochs and accuracy over epochs

Figure 6.39 presents the confusion matrix for the ELECTRA model in resolving the MCD challenge, offering a detailed view of classification performance across four semantic classes: location, Number Seven, time, and watch. The model performs notably well on the Number Seven class, correctly predicting 56 instances with minimal confusion (6 misclassified as location and 1 as watch). The location class is also fairly well classified, with 33 correct predictions but 12 misclassified as Number Seven, indicating some overlap or semantic similarity between these two classes. However, the model struggles significantly with the time and watch classes. All 9 instances of time were misclassified as Number Seven, and the sole watch instance was also incorrectly labelled as Number Seven, highlighting the model's difficulty in distinguishing minority or less frequent senses. This skew suggests an imbalance in the training data or a lack of representative features for these underrepresented classes. Overall, the confusion matrix reveals that, while ELECTRA performs strongly on dominant senses, it exhibits notable weaknesses in capturing rarer or more nuanced senses, reflecting the challenge of meaning conflation in low-resourced or polysemous contexts.

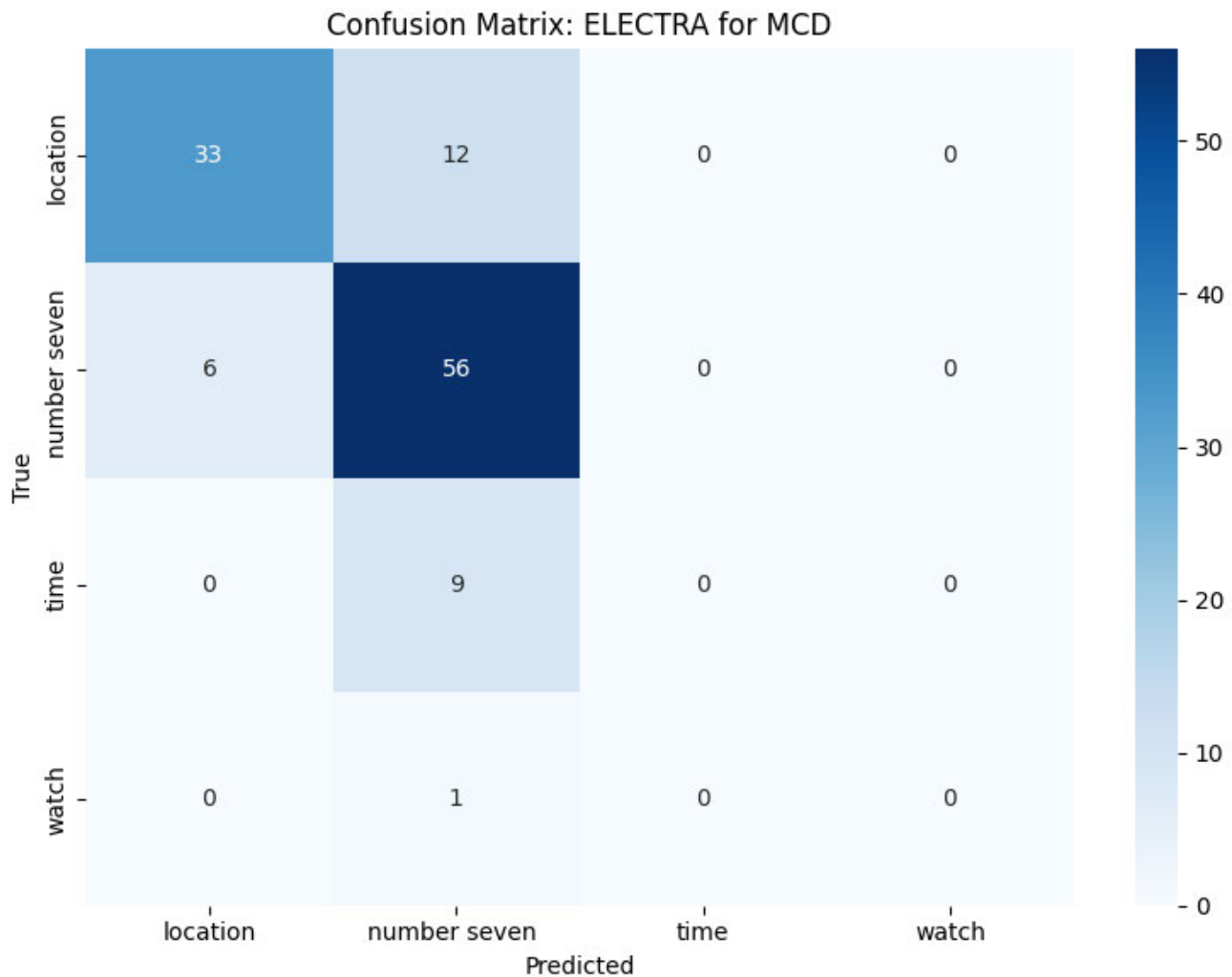


Figure 6.39: ELECTRA Confusion matrix

An accuracy of 55% was achieved by combining ELECTRA embeddings with a Naive Bayes classifier, which represents a slight improvement over the BERT-Naive Bayes combo. This rise implies that embeddings with more pronounced contextual distinctions may be produced using ELECTRA’s pre-training method of differentiating between actual and false input tokens. Nevertheless, the Naive Bayes classifier may not best represent the feature dependencies in these embeddings, much like BERT. To increase the WVS WE MCD resolution, more advanced classifiers—possibly neural network-based—may be able to take advantage of ELECTRA’s deep, context-aware embeddings.

6.6.6 Experimental results with Multilingual ELECTRA

Table 6.11 displays the performance characteristics of the Multilingual ELECTRA model in relation to WVS WE MCD. The model attained a precision of 42%, a recall of 43%, and an F1-score of 42%, indicating a rather poor efficacy in accurately detecting and retrieving the many meanings of

polysemous words. The poor scores point to difficulties in the model’s capacity to generalize across many settings, perhaps stemming from the constraints of multilingual pretraining in accurately capturing nuanced semantic differences in low-resource or morphologically rich languages such as SsaL. Notably, despite the low F1-score, the model exhibited a high accuracy of 80%, which may be deceptive owing to possible class imbalances or the prevalence of a majority class. The inconsistency between accuracy and other performance indicators highlights the need for more resilient, context-sensitive embeddings or refined monolingual models for successful WVS WE MCD resolution.

Table 6.10:Multilingual ELECTRA for MCD Performance Metrics

Model	Precision	Recall	F1-Score	Accuracy
Multilingual ELECTRA	42	43	42	80

Figure 6.37 depicts the training dynamics of the Multilingual ELECTRA model via two graphs: Loss throughout epochs and accuracy across epochs. The left figure demonstrates a steady reduction in both training and validation loss, indicating successful learning and little overfitting over three epochs. Significantly, the validation loss diminishes more rapidly during the first epochs, thereafter stabilizing, showing that the model swiftly acquires valuable patterns from the data. The right figure illustrates a consistent increase in both training and validation accuracy, with validation accuracy rising swiftly during the first epoch before stabilizing at about 80%. The training accuracy, however, persists in increasing beyond the second epoch, reaching 90% by the third epoch. The increasing disparity between training and validation accuracy may signal the onset of mild overfitting; however, the overall performance trend suggests that the model is learning efficiently. These findings corroborate the previous metrics in Table 6.11, displaying that, while the model well matches the training data, its applicability is limited, particularly in intricate multilingual or low-resource environments.

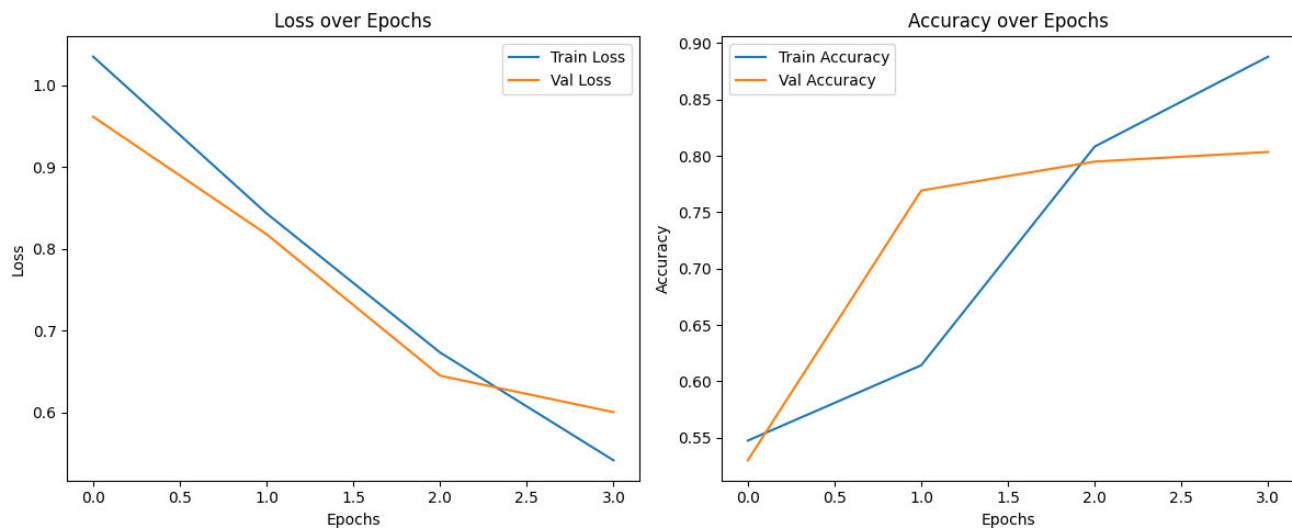


Figure 6.40: Multilingual ELECTRA loss over epochs and accuracy over epochs

Figure 6.41 depicts the confusion matrix for the Multilingual ELECTRA model used to address the MCD problem in WVS WE models. The model demonstrates adequate proficiency in differentiating between the “location” and “Number Seven” interpretations, achieving 35 and 59 accurate predictions, respectively. There exists significant misunderstanding between “location” and “Number Seven,” as shown by 10 occurrences of “location” being erroneously classed as “Number Seven,” and 3 occurrences in the other direction. The model has considerable difficulty with the senses of “time” and “watch”—instances of “time” are erroneously categorized as “Number Seven” on nine occasions, while “watch” is misclassified once as “Number Seven” and is otherwise not accurately predicted at all. This indicates that the Multilingual ELECTRA model struggles to discern contextual differences for infrequent or ambiguous meanings such as “time” and “watch”, potentially due to insufficient training instances, overlapping contextual signals, or intrinsic constraints of multilingual embeddings in fine-grained disambiguation tasks.

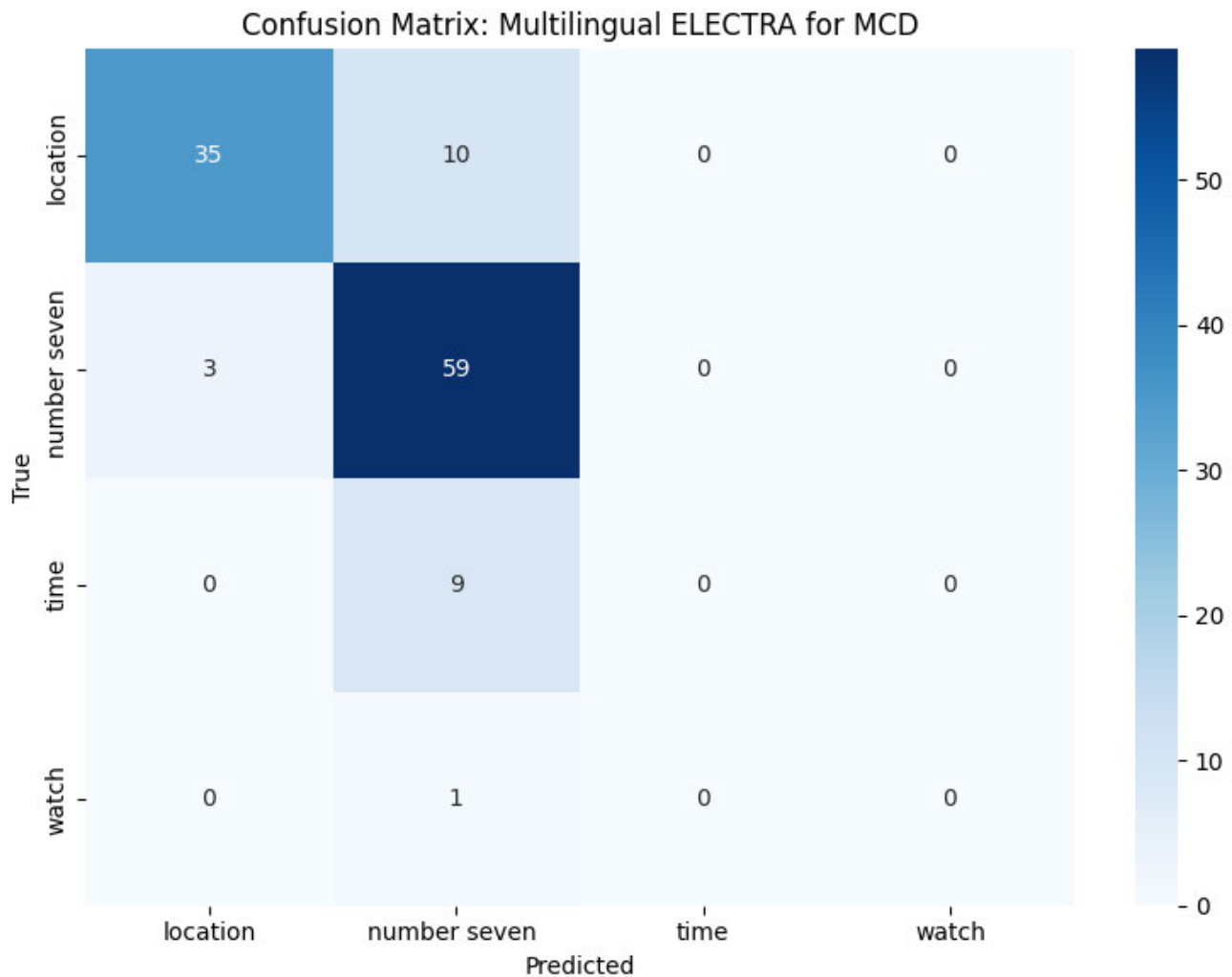


Figure 6.41: Multilingual ELECTRA confusion matrix

6.6.7 Experimental Results with RoBERTa

Table 6.12 presents the performance metrics of the RoBERTa model, showcasing its effectiveness in a word sense disambiguation (WSD) task. The model achieved a precision of 39%, recall of 43%, and an F1-score of 41%, signifying a moderate balance between its ability to correctly identify relevant senses (recall) and its accuracy in those identifications (precision). Despite the relatively low F1-score, the model attained a high overall accuracy of 79%, which suggests that, while it may perform well in correctly classifying many instances, its performance in distinguishing between finer-grained or ambiguous senses might be limited. This discrepancy highlights a potential issue of class imbalance or an overreliance on dominant senses, which is a known challenge in WVS WE MCD. Thus, while

RoBERTa shows promise in overall correctness, further enhancement may be needed to improve its nuanced sense discrimination.

Table 6.11: RoBERTa Performance Metrics

Model	Precision	Recall	F1-Score	Accuracy
RoBERTa	39	43	41	79

Figure 6.42 depicts the training dynamics of the RoBERTa model throughout epochs, emphasizing the patterns in loss and accuracy. The "Loss over Epochs" graph on the left illustrates a steady decrease in both training and validation loss, signifying efficient learning and less prediction error over time. The validation loss closely matches the training loss, showing little overfitting. The "Accuracy over Epochs" graph on the right demonstrates a consistent improvement in both training and validation accuracy, with training accuracy escalating from just over 50% to nearly 86%; and validation accuracy increasing from around 52% to 78%. The reduced disparity between training and validation accuracy in the last epoch connotes that the model generalizes well to novel input. These results together suggest that RoBERTa is learning both efficiently and effectively, demonstrating significant promise for consistent success in word meaning disambiguation tasks.

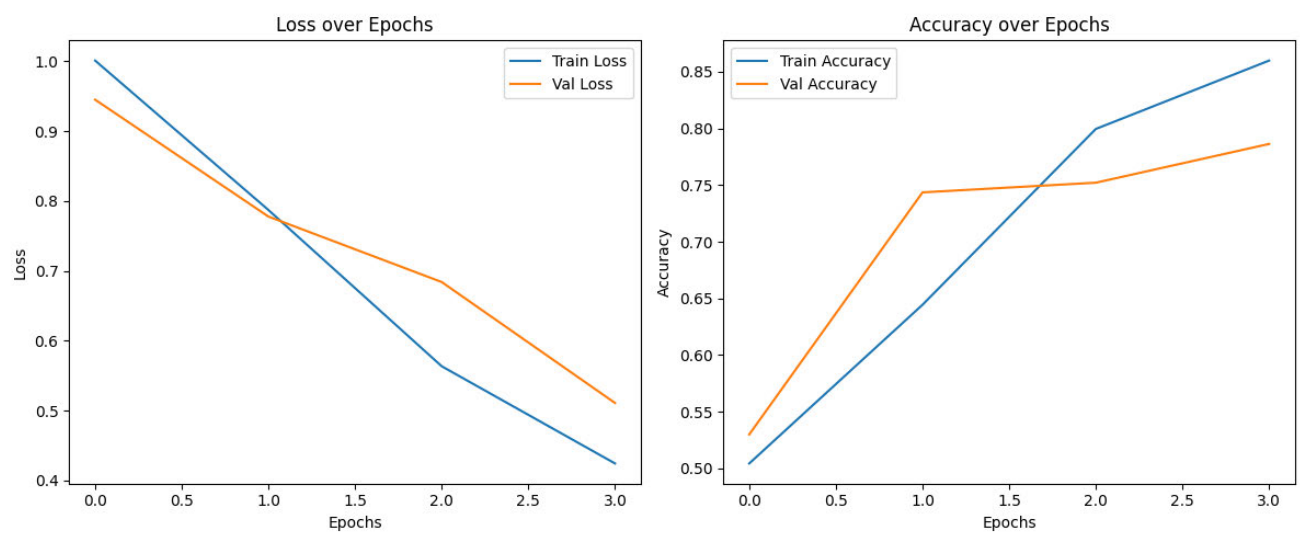


Figure 6.42: RoBERTa Loss over epochs and accuracy over epochs

Figure 6.43 displays the confusion matrix for RoBERTa in tackling WVS WE MCD, demonstrating the model’s efficacy in differentiating between various senses or categories. The matrix demonstrates

robust classification efficacy for the “number seven” sense, with 54 accurate predictions and a limited number of misclassifications (8 categorized as “location” and 7 as “time”). The “location” class has commendable performance, achieving 38 accurate predictions while misclassifying 7 instances as “number seven”. The model has difficulties with the “time” and “watch” categories, since “time” instances are often misclassified as “number seven” (7 occurrences), while the only “watch” case is misclassified as “location”. The “time” and “watch” classes provide no accurate predictions, indicating potential class imbalance, insufficient training data, or semantic overlap with other classes such as “number seven”. In summary, whereas RoBERTa excels with common or clear classes, its performance declines with underrepresented or semantically intricate senses, highlighting the need for enhancement in addressing unusual or contextually complex meanings.

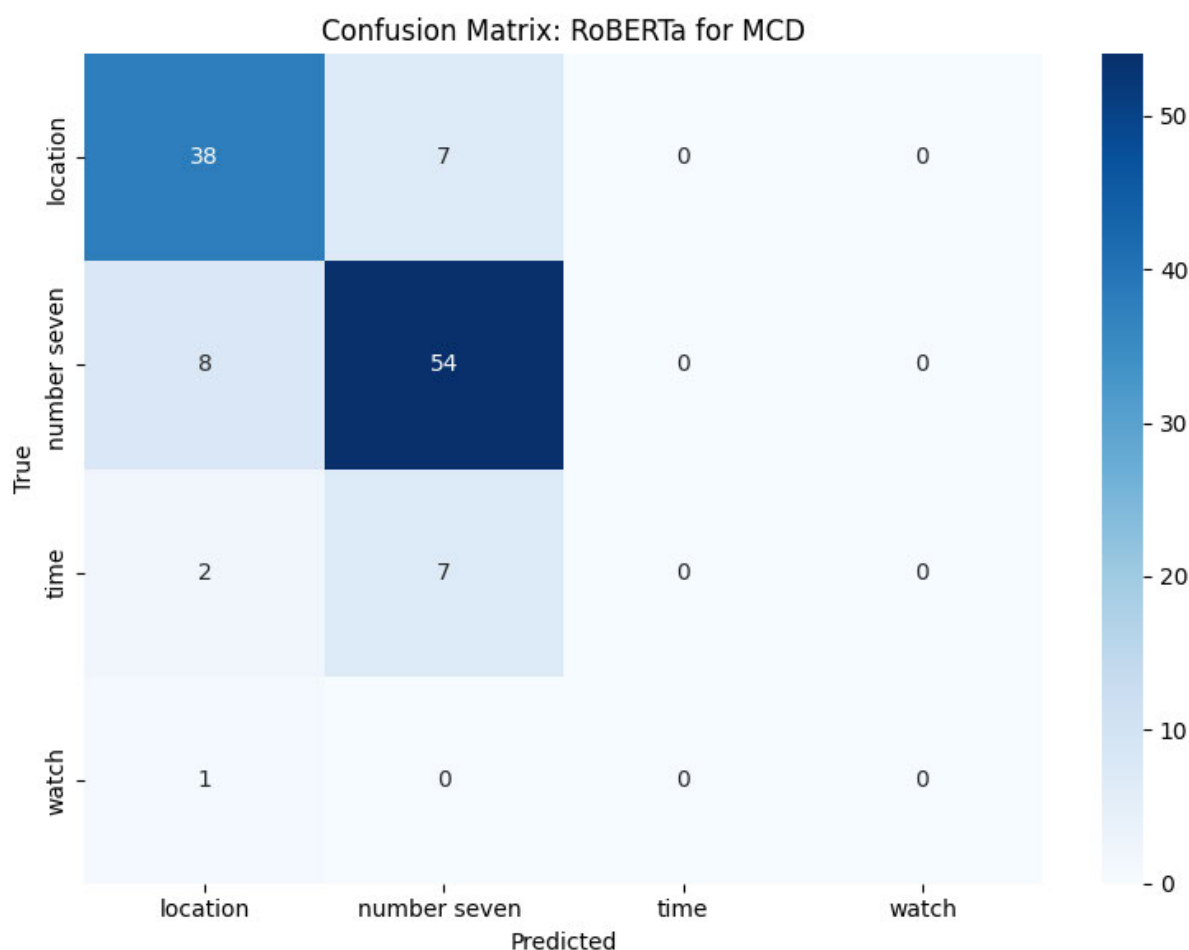


Figure 6.43: RoBERTa confusion matrix

WVS WE MCD was resolved with a 66% accuracy rate using RoBERTa embeddings and a Naive Bayes classifier. This performance, which is better than BERT and ELECTRA with Naive Bayes,

implies that more contextually refined embeddings are produced by RoBERTa’s improved training on a bigger, more varied dataset. Nevertheless, even though the embeddings capture intricate meanings, their potential may still be constrained by the Naive Bayes classifier. For the MCD problem in WVS WE models, a more sophisticated classifier would probably be more successful at taking advantage of RoBERTa’s subtle embeddings.

6.6.8 Experimental results with Multilingual RoBERTa

Table 6.13 delineates the efficacy of the Multilingual RoBERTa model in addressing the MCD problem in WVS WE models, underscoring notable performance constraints. Notwithstanding RoBERTa’s sophisticated multilingual capabilities, the model attained a low accuracy of 53%, exhibiting notably inadequate precision (13%), recall (25%), and F1-score (17%). The metrics disclose that the model had difficulties in accurately categorizing ambiguous word senses; the model did not generalize well across sense classes, presumably misclassifying many cases or defaulting to predominant classes.

The very low accuracy signifies a substantial occurrence of false positives, whilst the diminished memory highlights overlooked actual cases, both of which pose challenges for clarifying meaning in a morphologically rich and underresourced language. The findings suggest that Multilingual RoBERTa may not have adequately grasped the contextual nuances required for this job, either owing to inadequate domain-specific fine-tuning, restricted training data, or an uneven sense distribution. The results underscore the significance of domain adaptation, meticulous preprocessing, and potentially tailored multilingual embeddings when using pre-trained large language models such as RoBERTa for low-resource word sense disambiguation tasks.

Table 6.12: Multilingual RoBERTa Performance Metrics

Model	Precision	Recall	F1-Score	Accuracy
Multilingual RoBERTa	13	25	17	53

Figure 6.44 illustrates the training and validation curves for Multilingual RoBERTa in addressing MCD in WVS WE models. The loss plot (left) depicts a steady decrease in both training and validation loss over the epochs, reflecting that the model is effectively learning and adapting to the training data.

Nonetheless, the rate of decline stabilizes rapidly, and validation loss remains comparatively elevated, suggesting constrained learning advancement and potential early performance plateauing.

The accuracy plot (right) further underscores these limits. Training accuracy consistently increases from 46% to around 57%; however, validation accuracy stabilizes early at around 53%, exhibiting little enhancement beyond the first epoch. The first plateau, together with the increasing disparity between training and validation accuracy, demonstrates potential overfitting and inadequate generalization to novel data. The Figure 6.41 shows that, despite minor enhancements in loss, Multilingual RoBERTa had difficulties in accurately disambiguating word senses in this low-resource scenario, presumably attributable to inadequate fine-tuning, sense imbalance, or restricted contextual signals in the dataset. This result highlights the need for extended training, domain adaptation, and maybe other architectures more appropriate for the complexities of the MCD problem in WVS WE models.

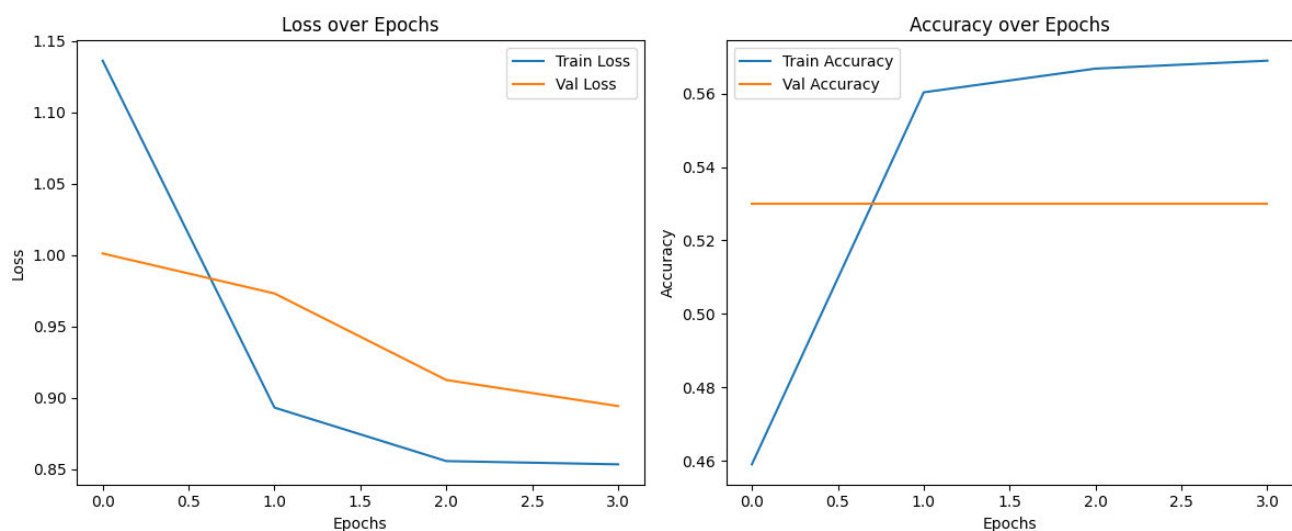


Figure 6.44: Multilingual RoBERTa loss over epochs and accuracy over epochs

Figure 6.45 illustrates the confusion matrix for Multilingual RoBERTa utilized in the MCD problem in WVS WE models, signalling a notable performance deficiency. The model erroneously categorizes all sense classes as "Number Seven," irrespective of their actual designations. All 45 occurrences of "location," 9 occurrences of "time," and 1 occurrence of "watch" are erroneously categorized as "Number Seven," yielding no true positives for any class other than "Number Seven."

This pronounced classification bias towards a single dominating class reveals that the model has overfitted to the majority class and has not effectively discerned important differences across word

meanings. This underscores the effects of class imbalance, inadequate task-specific fine-tuning, or perhaps suboptimal management of contextual variation in a low-resource language. The confusion matrix corroborates previous metric results—specifically, very poor accuracy, recall, and F1-score—highlighting the need for balanced training material, data augmentation, or tailored attention mechanisms when using Multilingual RoBERTa for resolving MCD challenges found in WVS WE models.

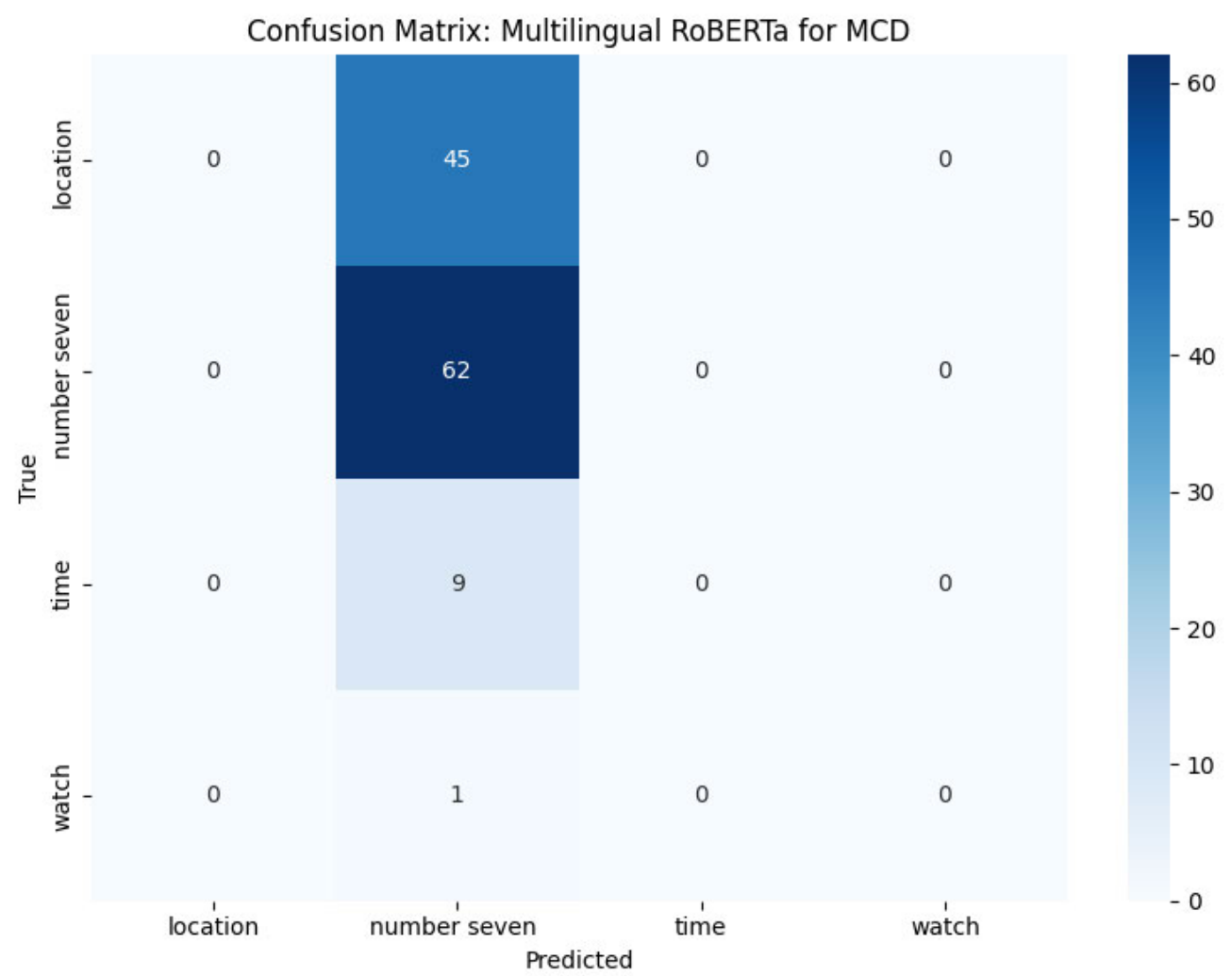


Figure 6.45: Multilingual RoBERTa confusion matrix

Table 6.14 shows a clear performance hierarchy across the transformer baselines. BERT offers the best overall balance ($F1 \approx 60$, $Accuracy \approx 87$), with Multilingual BERT close behind ($F1 \approx 55$, $Accuracy \approx 89$). Both exhibit precision noticeably higher than recall, meaning that there are conservative decision boundaries that curb false positives yet miss a share of true instances—an imprint of class imbalance typical in meaning conflation settings. A mid-tier group—ELECTRA, Multilingual ELECTRA, and

RoBERTa—clusters around $F1 \approx 40$ – 42 with accuracy ≈ 76 – 80% , suggesting weaker separation of confusable senses and limited sensitivity to minority classes. Multilingual RoBERTa is a clear outlier ($F1 \approx 17$, Accuracy ≈ 53), pointing to difficulty capturing domain cues and low-frequency senses in this corpus. Overall, the sizable accuracy–F1 gaps (e.g., high accuracy alongside modest F1) imply that correct predictions are concentrated in frequent senses, while rarer senses remain under-served—classic meaning conflation deficiency. Practically, BERT or mBERT are the strongest foundations; however, lifting recall (and thus F1) for minority senses will likely require class weighting or oversampling, per-class threshold tuning, and domain/graph features to inject disambiguating context without eroding precision.

Table 6.13: Performance Summary of Transformer Models

Model	Precision	Recall	F1-Score	Accuracy
BERT	68	57	60	87
Multilingual BERT	69	54	55	89
ELECTRA	39	41	40	76
Multilingual ELECTRA	42	43	42	80
RoBERTa	39	43	41	79
Multilingual RoBERTa	13	25	17	53

The grouped bar chart in Figure 6.46 compares precision, recall, F1-score, and accuracy for each model, clearly illustrating the trade-offs. BERT and Multilingual BERT lead the pack with precision rates between 68% and 69%, which surpass their recall rates of 54% to 57%. This results in F1-scores of 60% and 55%, alongside high accuracy rates of 87% to 89%. This highlights that the models employ cautious decision boundaries, selectively assigning senses, thereby reducing false positives but also omitting some true cases, a typical scenario under class imbalance. ELECTRA and RoBERTa variants are positioned in the middle with precision and recall each around 39% to 43%, leading to F1 scores of approximately 40% to 42%, suggesting a more balanced yet generally weaker sensitivity across the classes. Multilingual RoBERTa is prominent, with notably low precision (13%), low recall (25%), and

an F1-score of 17%, indicating trouble in capturing domain-specific cues; its moderate accuracy of 53% suggests a reliance on frequent classes rather than resolving nuanced sense boundaries. Overall, the chart demonstrates that apparent accuracy may exaggerate performance when WVS WE MCD is a factor, the F1-score revealing areas in which models still confuse similar senses.

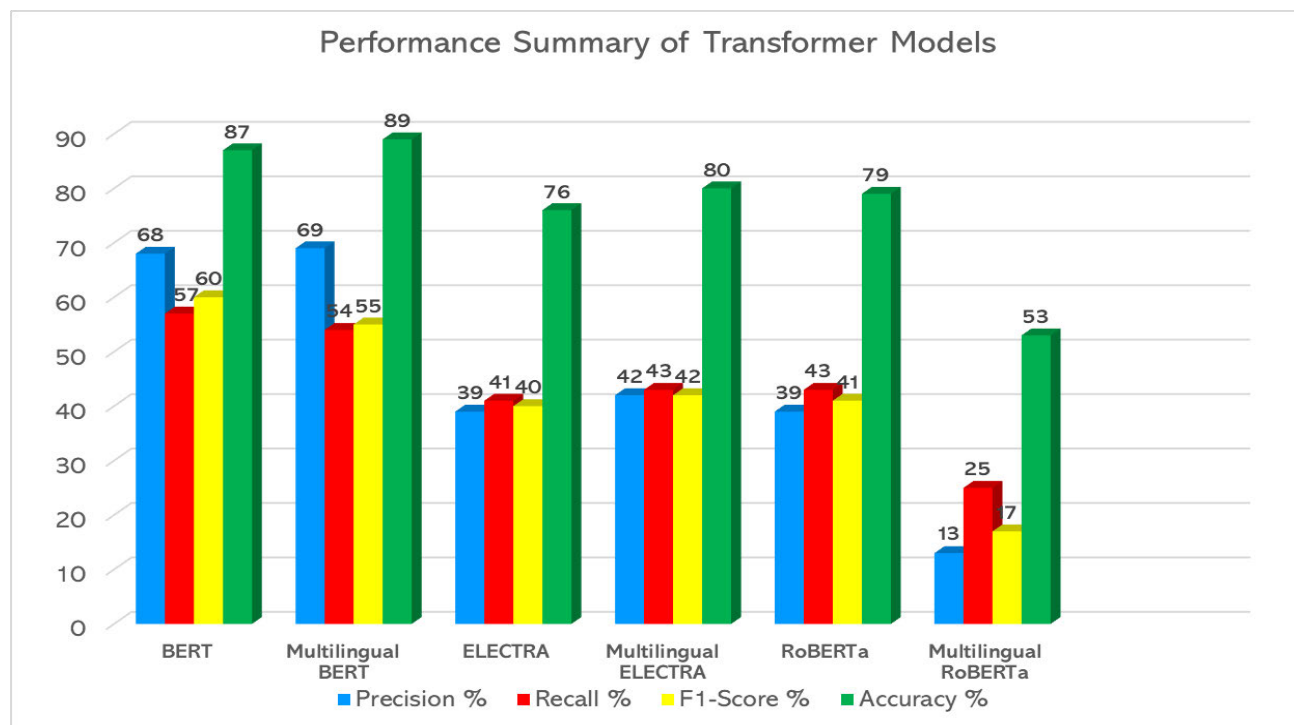


Figure 6.46: Grouped metrics for transformer variance

The scatter plot in Figure 6.47 distinguishes between overall accuracy and balanced per-class performance (F1). While Multilingual BERT demonstrates high accuracy, it has a lower F1 than BERT, highlighting its proficiency in identifying majority senses but lower proportional enhancement for minority ones. In contrast, models such as ELECTRA, Multilingual ELECTRA, or RoBERTa, which have accuracy in the range of ~76–80%, exhibit similar accuracy levels but vary in F1. This difference indicates that models can achieve similar overall correctness yet differ significantly in handling challenging, rare senses. The dispersion in the plot thus represents the core of WVS WE MCD: classes with overlapping cues boost accuracy but diminish F1. Practically, this chart supports the focus on macro-F1 (or per-class F1) as the main metric for sense disambiguation, with accuracy as a secondary measure.

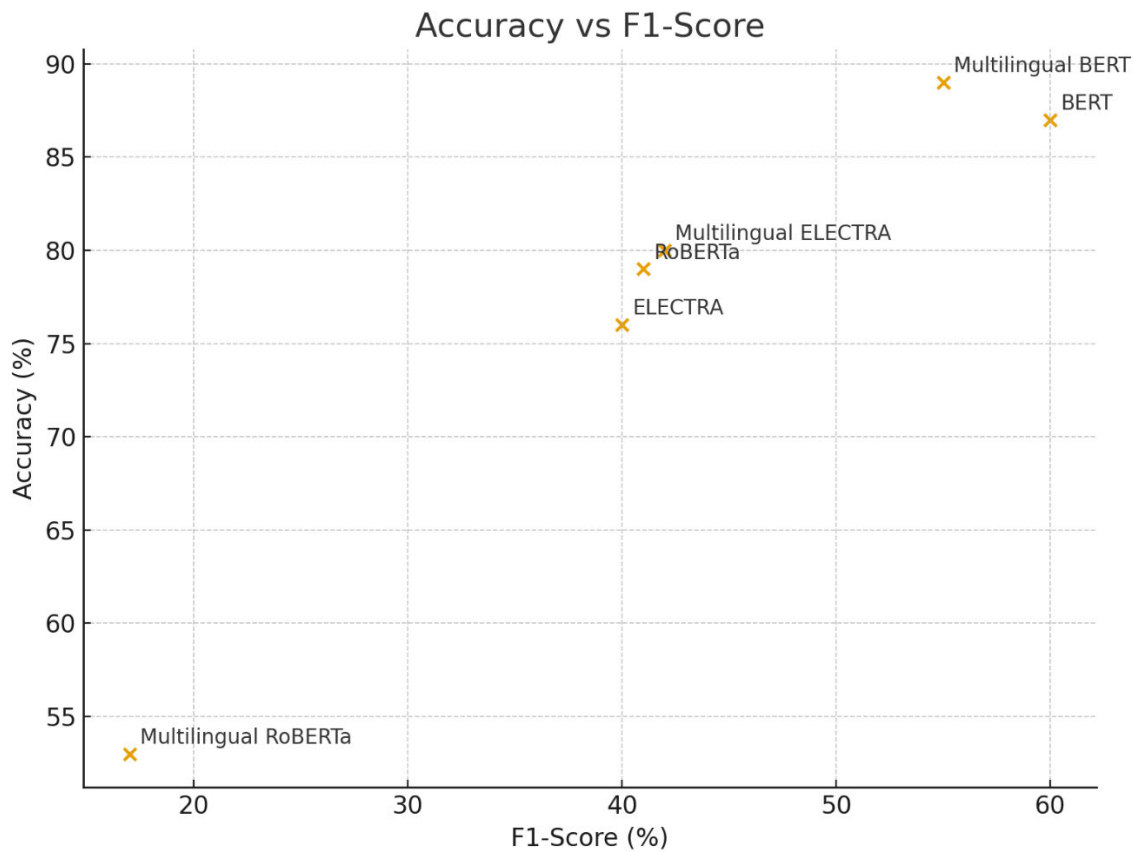


Figure 6.47: The scatter plot

The F1-score ranking in Figure 6.48 offers an actionable perspective for decision-making. BERT ranks highest (60%), then Multilingual BERT (55%), with Multilingual ELECTRA, RoBERTa, and ELECTRA grouped in the low 40s. Multilingual RoBERTa is significantly lower (17%). This order aids in choosing models for deployment or further experiments: BERT and mBERT are prime candidates; the models in the middle range can serve as baselines or be adapted for domain-specific tasks and balancing classes; the lowest model-ranked models would need considerable enhancement through lexicon/graph expansion, calibration, or further fine-tuning to become effective. Compared with WVS WE MCD, the ranking indicates that models with superior subword processing and solid pretraining, like BERT, conserve more contextual distinction between meanings but still need specific improvements—such as class-weighted losses, threshold adjustments, domain lexicon integrations, or graph indicators—to improve minority-sense F1 without losing accuracy.

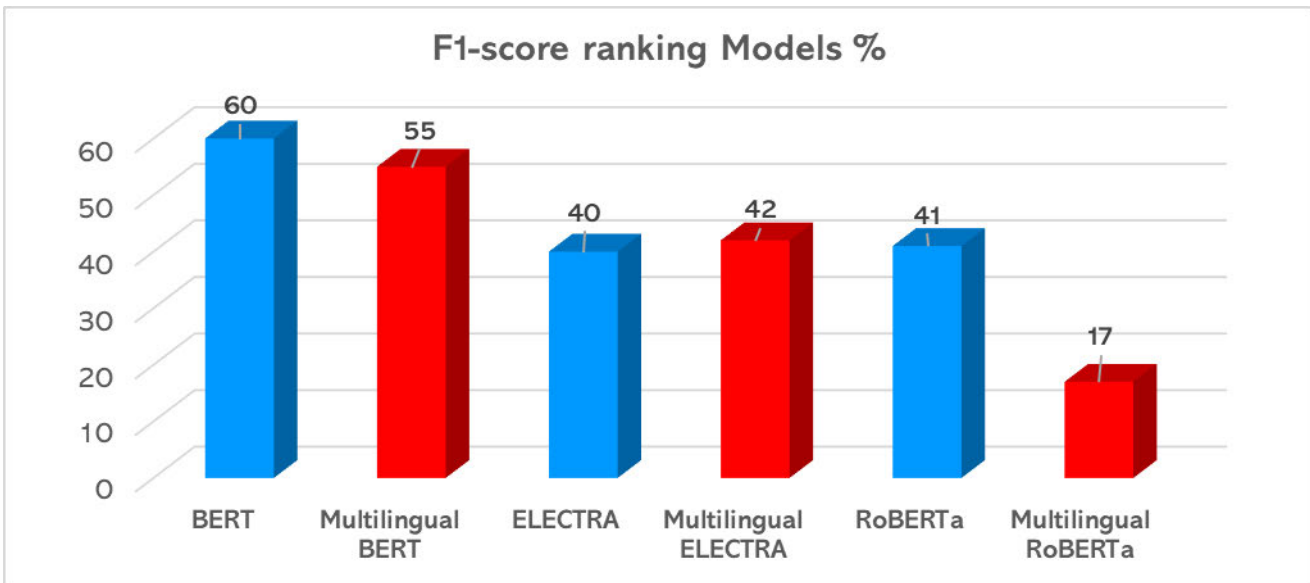


Figure 6.48: The F1-score ranking models

6.6.9 Experimental results of graph-based models

Table 6.15 illustrates that graph-based approaches effectively reduce word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD), with the domain-specific graph outperforming others by achieving 98% in precision, recall, F1, and accuracy. This indicates excellent sense separation and well-balanced probabilities across different classes. The integrated graph remains steady and balanced with 88% across all metrics, suggesting that the fusion of sentence–sentence similarity, domain cues, and clustering promote strong generalization without compromising on either precision or recall. The hierarchical graph follows with slightly lower scores (P=86%, R=85%, F1=85%) and records a notably lower accuracy (83%) compared with its F1 score, often suggesting challenges in distinguishing some classes or potential noise introduced by cluster boundaries for infrequent senses. Collectively, the close alignment of precision and recall in the top two systems (notably the domain-specific one) points to minimal false-positive or false-negative imbalances, reducing conflation between closely related senses. The findings emphasize that specific domain signals most critically aid disambiguation, whereas multi-source integration is beneficial but might marginally reduce the impact of these signals. Practically, the domain-specific graph emerges as the strongest option for deployment, with the integrated graph serving as a robust alternative; the

hierarchical graph remains valuable but could improve with more precise clustering and strengthened domain cues.

Table 6.14: Performance Summary of Classification Report

Model	Precision	Recall	F1-Score	Accuracy
Hierarchical Graph	86	85	85	83
Domain-Specific Graph	98	98	98	98
Integrated Graph	88	88	88	88

The chart in Figure 6.49 presents a distinct hierarchy: the domain-specific graph leads with 98% in precision, recall, F1, and accuracy, highlighting that robust domain-specific knowledge and carefully curated lexical connections provide the clearest disambiguation and minimize meaning overlap. Following, the integrated graph achieves a reliable second place at 88% for all metrics; the integration of domain, syntactic, and clustering signals thus enhances robustness, albeit at the cost of potential noise or conflicting indications that might reduce peak performance. The hierarchical graph lags behind (precision at 86%, recall/F1 at 85%, accuracy at 83%), suggesting that, while hierarchy captures useful structural information, it lacks the nuanced domain anchors necessary for tackling complex meanings. For practical applications, prioritize the domain-specific graph for production WSD/MCD tasks; opt

for the integrated version when broader generalization is needed, and consider enhancing such through improved edge weighting or expanding the domain lexicon to bridge the performance gap.

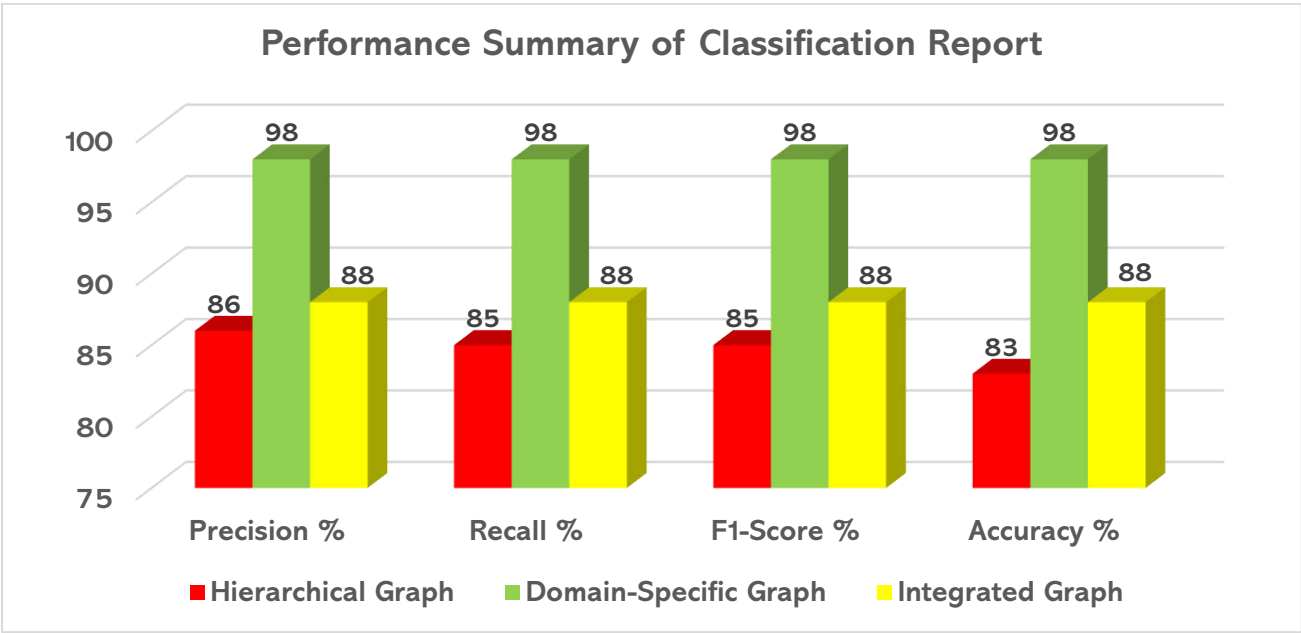


Figure 6.49: Performance Summary of Classification Report

The hierarchical graph confusion matrix in Figure 6.50 displays notable, mainly diagonal strengths with minimal cross-class mistakes. The two main categories are managed effectively: 'location' has 37 accurately identified cases versus 9 misclassified as 'Number Seven' ($\approx 80\%$ recall; $\approx 86\%$ precision with 43 total identified as location), while 'Number Seven' has 58 correct identifications versus 6 errors as 'location' and 2 as 'time' ($\approx 91\%$ recall; $\approx 84\%$ precision with 69 predicted). The primary error pattern is the dual misclassification between 'location' and 'Number Seven', illustrating a lingering mix of meanings where similar contextual cues incorrectly group sentences. The 'time' category is small but relatively accurate: 4 correct classifications and 2 mistaken for 'number seven' ($\approx 67\%$ recall; 100% precision since every prediction of time was accurate). 'Watch' has 1 case, correctly predicted (precision or recall 100% but with minimal significance). Overall, the matrix suggests that hierarchical differentiation aids in sense separation, although the cluster granularity permits overlap in adjacent semantic regions, particularly between 'location' and 'Number Seven'. Possible solutions include adjusting cluster numbers or thresholds, integrating domain-specific knowledge (like a lexicon or

rules) at the cluster stage, or employing cost-sensitive modelling or training adjustments to mitigate the prevalent crossover mistakes without impacting lesser-represented classes.

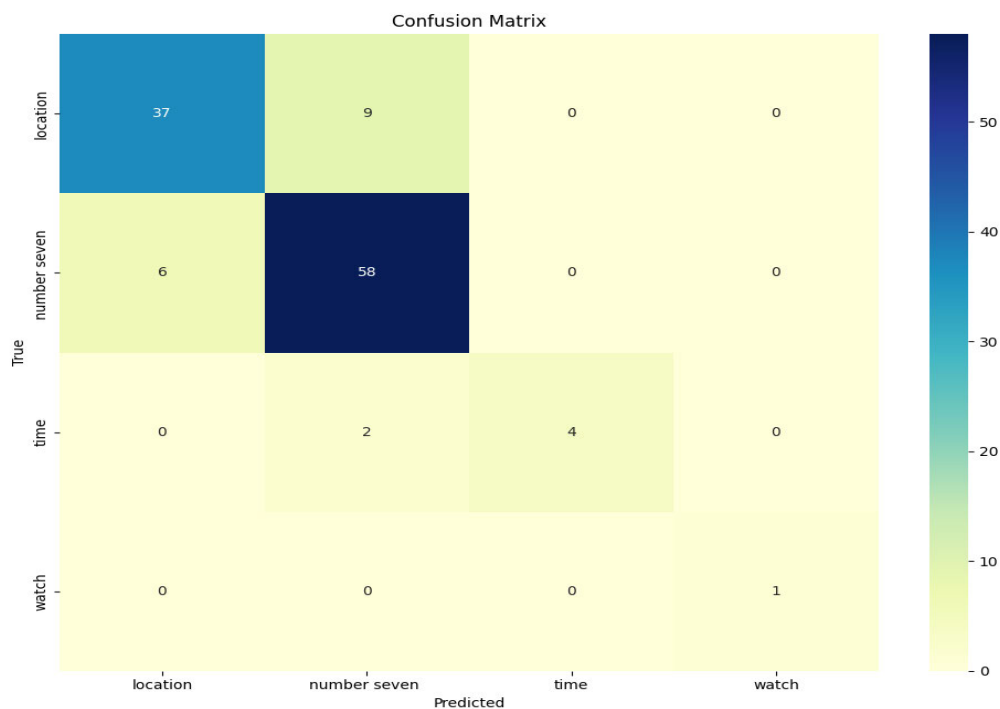


Figure 6.50: Hierarchical graph confusion matrix

The ROC curves for the hierarchical graph model in Figure 6.51 consistently demonstrate high levels of class separability, with AUCs of 0.93 for location, 0.92 for the Number Seven, 0.98 for time, and a perfect score of 1.00 for watch. Curves that closely approach the top-left corner suggest that the classifier excels in achieving high true-positive rates while maintaining low false-positive rates, indicating that the hierarchical arrangement of sentence–cluster relationships contribute useful discriminative context beyond surface features. The slightly lower, yet still impressive AUC for the Number Seven corresponds to the residual confusion with the location observed in the confusion matrix, signifying overlapping contexts that could be further resolved with adjustments in decision thresholds or the inclusion of domain-specific hints. The flawless AUC for watch, however, should be interpreted with caution due to its minimal support; with very few examples, the model may appear perfect without assurances of generalization. Overall, these ROC findings endorse the method: hierarchical modelling effectively produces strong probability rankings of senses; slight threshold

modifications (along with potential calibration for each class) could convert these rankings into improved F1 scores without compromising precision.

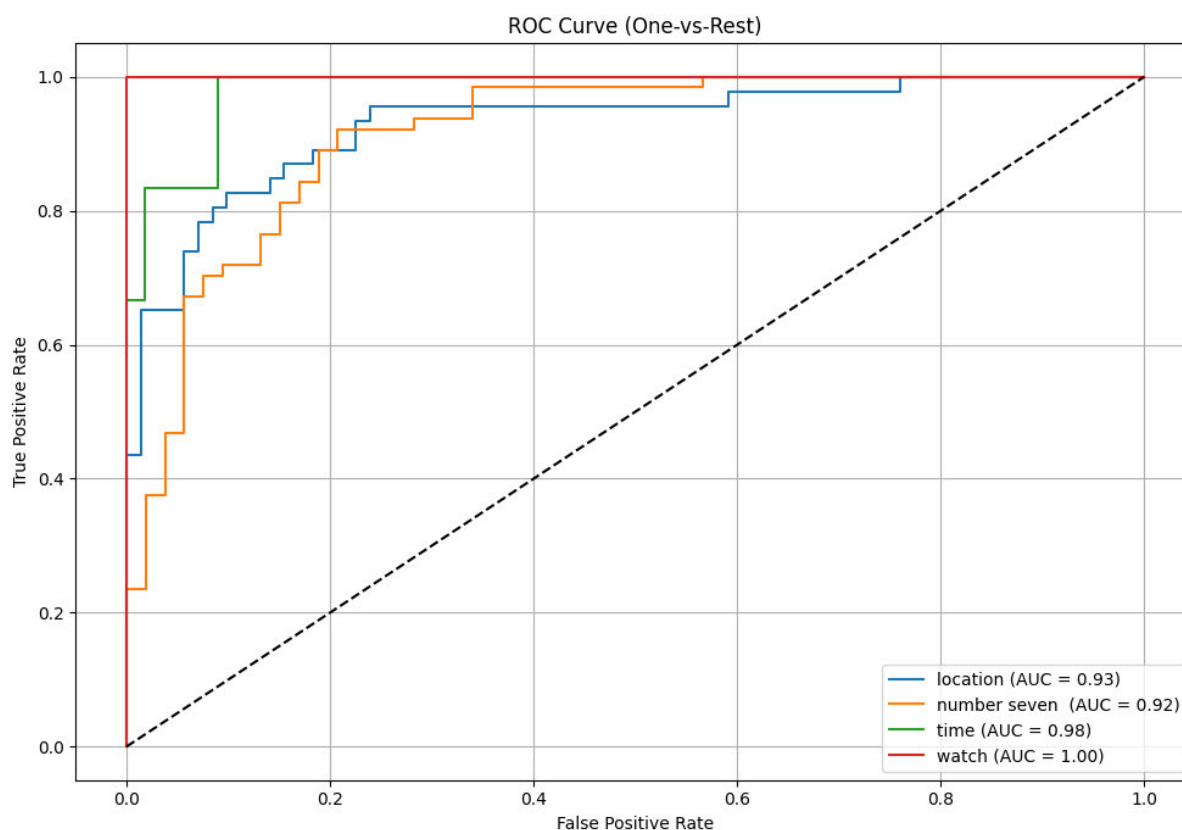


Figure 6.51: Hierarchical graph ROC curve

The confusion matrix for the domain-specific graph in Figure 6.52 illustrates an almost perfect separation of senses, with only two errors outside the main diagonal. The location prediction is nearly flawless with a success rate of 45 out of 46 (approximate recall of 97.8%) and a precision of 100%, as no other class was mistaken for location. The classification of Number Seven is equally robust, with all 64 predictions being correct (100% recall) and only two errors found in its column—one predicted as location and the other as time—resulting in an approximate precision of 97%. The time class, although smaller, is quite accurate with 5 out of 6 accurately predicted (recall roughly 83.3%) and a precision of 100%, every time prediction being correct. Watch was identified with full accuracy (1 out of 1), though this reflects minimal data for this category. Overall accuracy stands at 115 out of 117 predictions correct (approx. 98.3%), consistent with the table’s headline accuracy of 98%. The remaining errors—location misidentified as Number Seven, and time as Number Seven—suggest that numbers continue to exert influence on nearby contexts. However, the strong domain priors prevent

meaning conflation, resulting in balanced precision and recall and a distinctly diagonal matrix. Minor enhancements could be achieved by refining numeric patterns (e.g., regex for “7/7th/no.7”) and by incorporating additional temporal cues to prevent time misclassification into the numeric category.

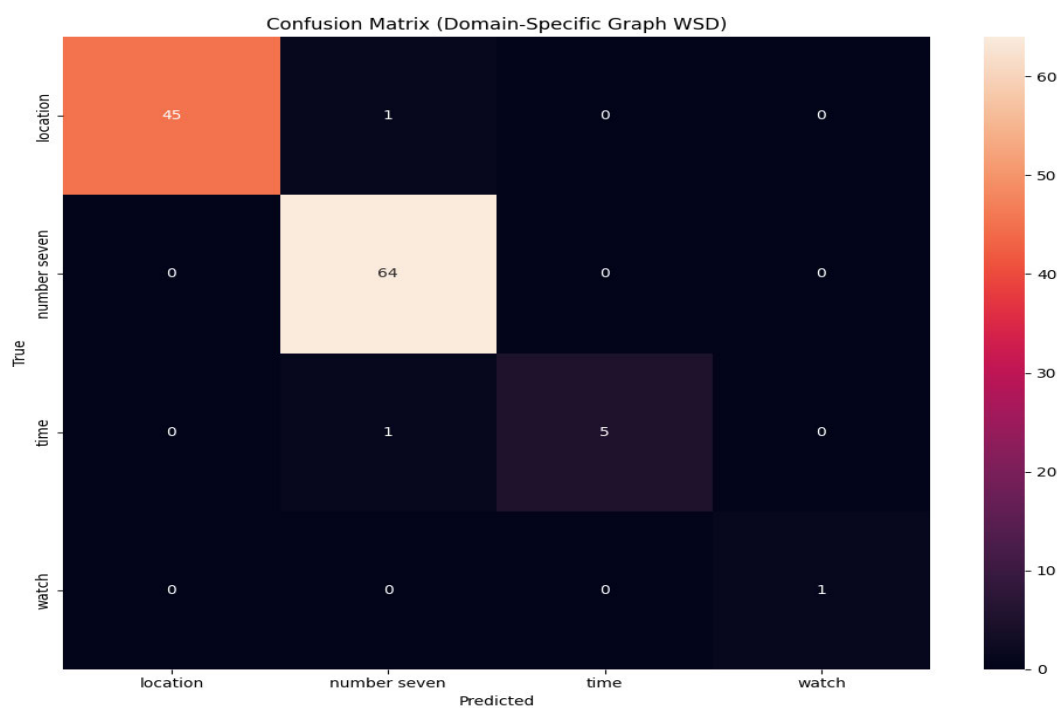


Figure 6.52: Domain-specific graph confusion matrix

The ROC curves of the domain-specific graph model in Figure 6.53 closely hug the top-left corner for each category, showing $AUC = 1.00$ for location, Number Seven, time, and watch. This reveals the model’s exceptional capability to rank true positives above false positives, maintaining high true-positive rates with nearly zero false-positive rates across all categories. This outcome supports the prominently diagonal confusion matrix: domain priors such as lexicons and patterns along with domain hubs offer highly distinctive indicators that virtually eliminate any ambiguity in meaning. However, it is important to cautiously approach the perfect AUCs for smaller classes (such as “watch”, and to a lesser degree, “time”), limited data being able to inflate AUCs; and might not imply effective generalization. In practice, these curves suggest that the model is well-suited for threshold tuning; slight threshold modifications per class are likely to result in excellent F1 scores without compromising precision. Furthermore, expanding the domain lexicon can sustain this performance as the dataset increases.

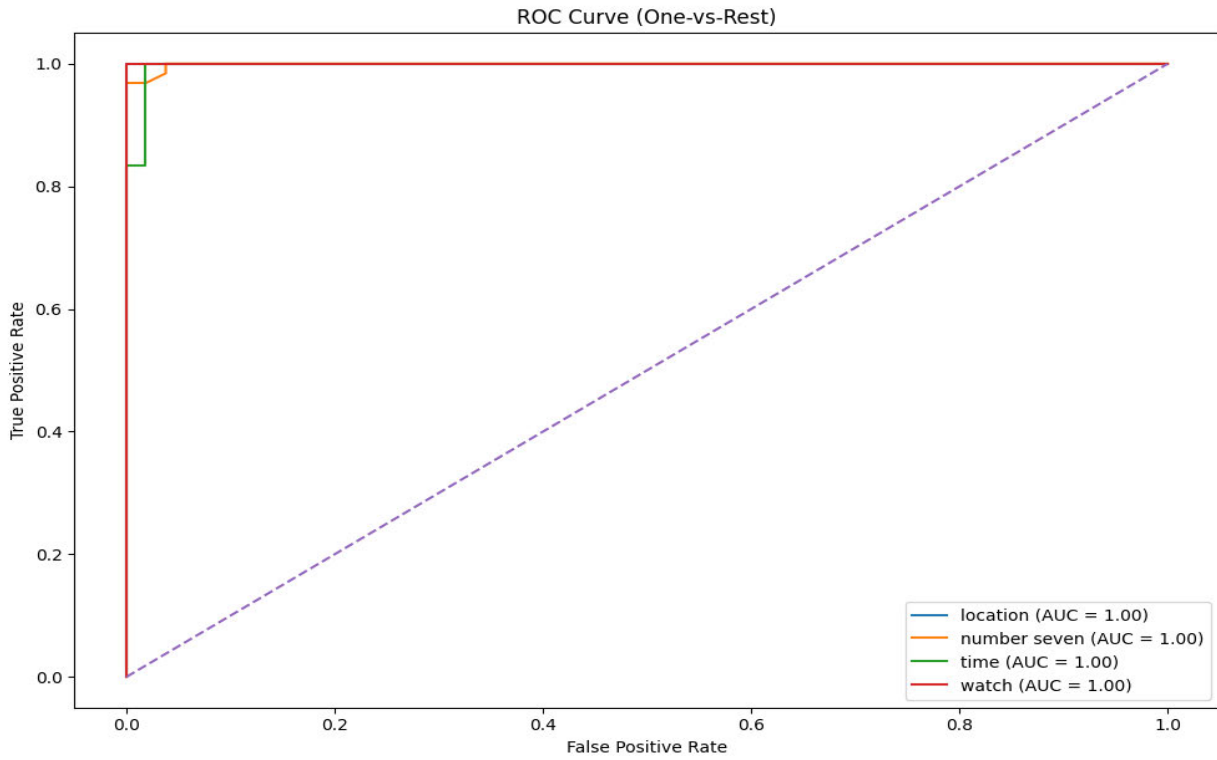


Figure 6.53: Domain-specific graph ROC curve

The domain-specific graph in Figure 6.54 has been sampled revealing the model’s reasoning: sentences (denoted as “S_xxx”) link to domain-specific terms, which then connect to domain hubs such as location, Number Seven, time, and watch. The central hub-with-spokes pattern shows a strong concentration of domain signals, such as locatives and numeral tokens, with many sentences radiating from one hub. This dense configuration suggests that multiple sentences use the same domain vocabulary, providing the classifier with a clear, consistent cue on their intended meaning. The pathways that follow a sentence $\rightarrow$ term $\rightarrow$ hub sequence form tight-knit communities, with numerous short edges indicating frequent use of relevant domain language. This structured cohesion is what sharpens the diagonal nature of the confusion matrix for the domain model. Occasionally, there are some bridging sentences that connect various terms or lie near the edges linking two sub-clusters. These can be conflation hotspots due to overlapping cues, such as a numeral in a location-based context, which can lead to occasional misclassifications.

In terms of graph features, sentences nearer a hub possess high personalized PageRank and degree, both used by the model alongside TF-IDF overlap and regex patterns to reliably rank the correct

meaning above others. On the other hand, isolated peripheral nodes lacking links offer minimal domain evidence; should they miss the mark, enhancing the domain lexicon (with synonyms, morphological variants, or pattern rules) generally aligns them correctly. Overall, this visualization shows why the domain-specific graph diminishes WVS WE MCD: it transforms raw text into organized domain signals, clustering sentences with shared cues and highlighting the ambiguous instances at the edges between clusters. Practically speaking, broadening the domain lexicon and refining patterns around numerals and time expressions will further consolidate the principal communities and slim down the connecting bridges, pushing performance nearer to its theoretical peak.

Domain-Specific Graph (sampled): Sentences ↔ Domain Terms ↔ Domain Hubs

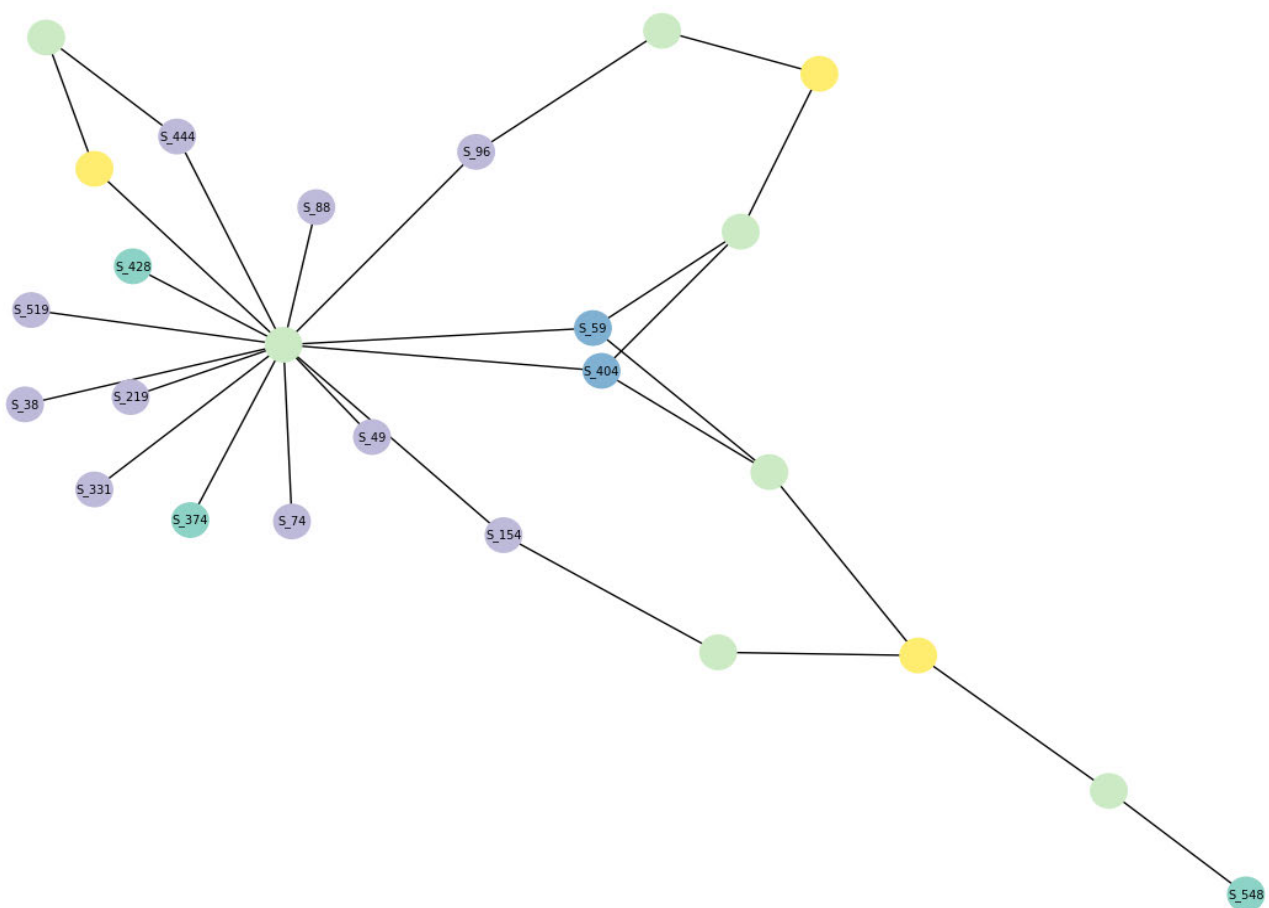


Figure 6.54: Domain- specific graph

The domain-specific graph in Figure 6.55 illustrates a dense network with community structures, in which lexical items and topical indicators concentrate around a few prominent hubs, such as tokens associated with location, numerics such as "seven," time, or semantics of devices such as watches. These large hub-and-spoke clusters reveal numerous sentences that share the same domain vocabulary—a desired structure for minimizing WVS WE MCD: essential domain-specific terms (e.g., place names, time-related expressions, numbers) gather evidence, bringing sentences with a similar intended meaning closer together. The presence of short paths and several triangles within clusters denote redundant domain cues, which enhance prediction reliability and account for the sharply diagonal confusion matrix and near-perfect ROC outcomes observed for this model.

Additionally, a few bridge nodes link otherwise distinct regions, such as items occurring with both locative and numeric language or with time and device-related lexemes. These bridges are potential error sources, in which overlapping cues might lead a sentence toward an incorrect domain. Practically, the graph suggests three strategies for improvement: (1) reinforce hubs by enlarging the domain lexicon with more synonyms or morphological variants, (2) manage or re-weight ambiguous bridges using pattern rules, such as stricter numeral/time regex, or focus on neighborhood purity, and (3) encourage modular subgraphs for each domain to achieve clearer separation. Overall, the graph's topology demonstrates reasons for the domain-specific method's effectiveness: it converts raw text into structured evidence belonging to a domain, forming coherent communities that reduce cross-sense drift while distinguishing the few truly ambiguous instances.

[illegible]

The confusion matrix for the integrated graph in Figure 6.56 prominently displays a diagonal pattern, indicating that the integration of sentence–sentence similarity, domain indicators, and cluster hubs effectively distinguishes senses. The two classes with substantial support are handled proficiently: 'location' achieves 41 correct out of 46 cases (recall $\approx 89\%$) with a precision of approximately 89% as well (41 out of 46 predictions), while 'Number Seven' successfully records 58 out of 64 correct cases (recall $\approx 91\%$) with a precision of roughly 89% (58 out of 65 predictions).

240

number formats continue to attract temporal interpretations. 'Watch' is accurately predicted (1/1), although its precision is 50% because one 'location' instance was misclassified as 'watch.' Considering the limited data, these results require careful interpretation.

The sum of correct predictions on the diagonal is 103 out of 117 ($\approx 88\%$ accuracy), paralleling the model's overall metrics. Overall, the matrix illustrates that the integrated approach significantly curtails word vector space (WVS) word embedding (WE) meaning conflation deficiency (MCD) in contrast to purely hierarchical frameworks, although issues persist with numeric–locative and numeric–temporal overlaps; targeted solutions involve clarifying numeral/time patterns, reinforcing domain cues, and making modest adjustments to class-specific thresholds to enhance recall for 'time' without compromising the high-performing classes.

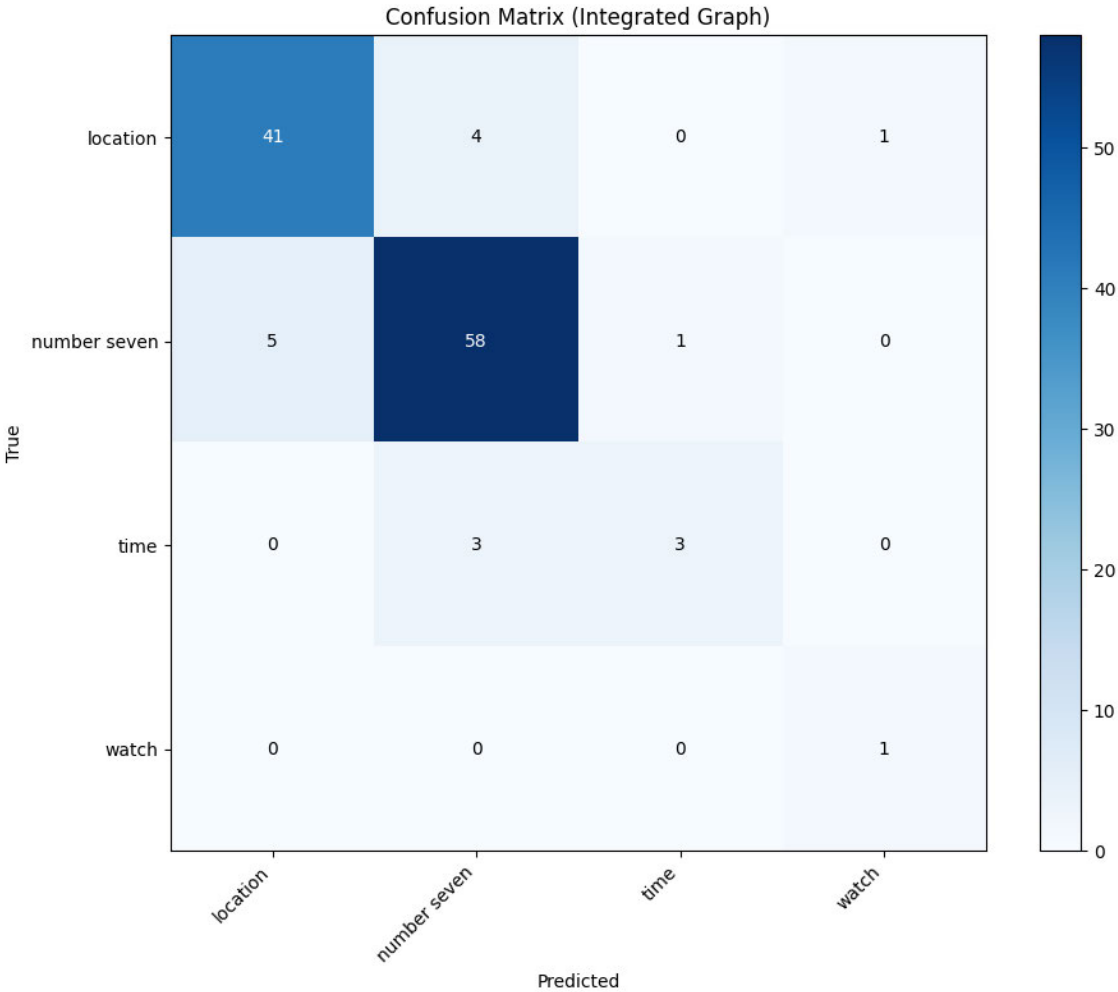


Figure 6.56: Integrated graph confusion matrix

The ROC curves for the integrated graph model in Figure 6.57 demonstrate notable class differentiation across all categories: location (AUC 0.97), Number Seven (AUC 0.96), time (AUC 0.99), and watch (AUC 1.00). The curves are closely aligned with the top-left corner, indicating the model’s capability to reach high true-positive rates paired with minimal false-positive rates. Discrepancies are akin to the confusion matrix: the marginally lower AUC for number seven compared with location indicates some remaining bidirectional overlap, whereas the almost perfect AUC for time suggests that, despite occasional misclassification, true class probabilities are ranked very highly by the model—implying that slight adjustments in thresholds (or class-specific calibration) might improve recall without reducing precision. The perfect AUC for watch should be approached with caution due to its limited data support, which could exaggerate AUC and might not be applicable generally. In sum, these curves substantiate the integrated configuration (sentence–sentence similarity + domain indicators + cluster hubs): the curves provide dependable probability orderings for each category and leave only fine-tuning—mainly class-specific thresholds and slightly enhanced numeric/temporal priors—to achieve further improvements in F1 scores.

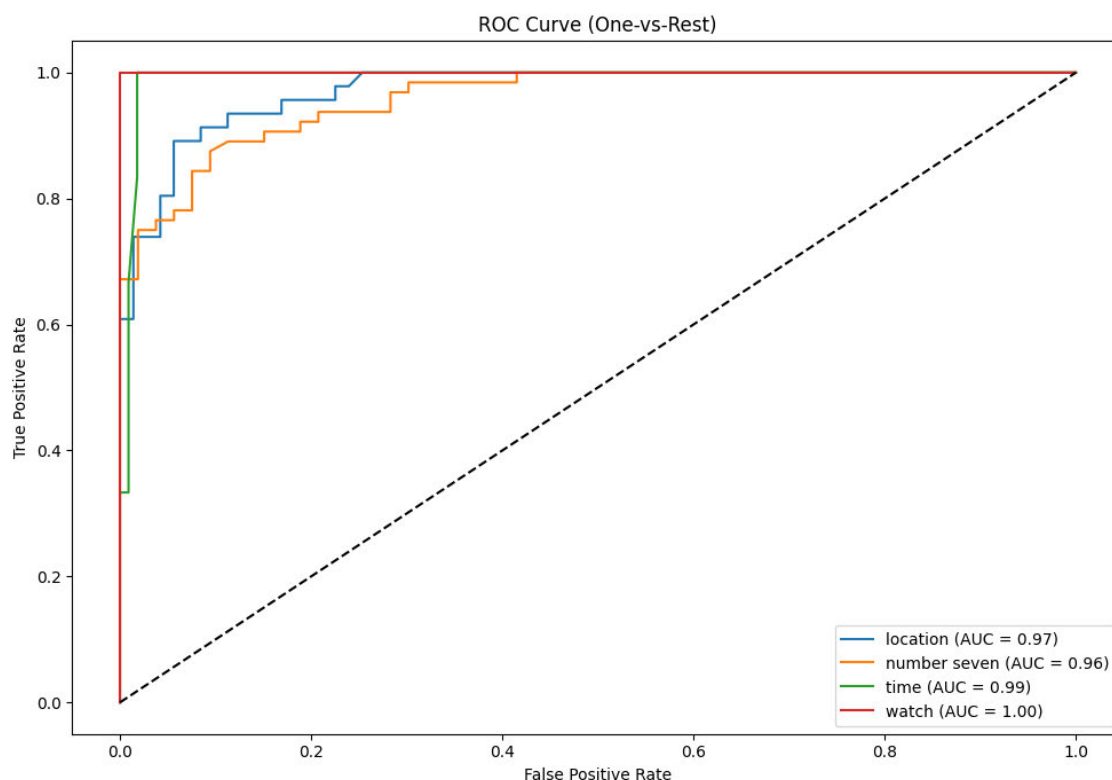
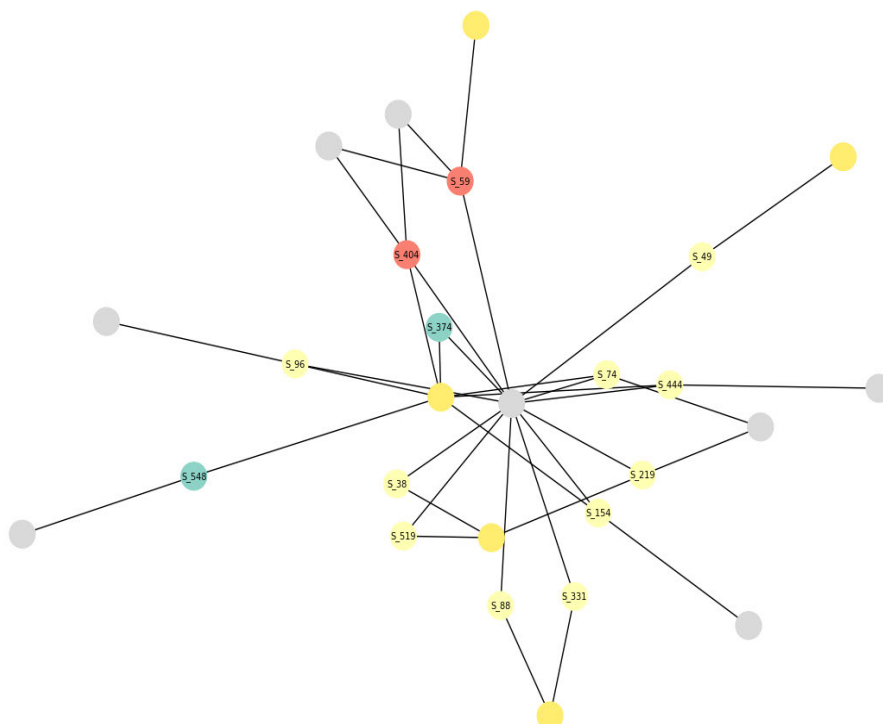


Figure 6.57: Integrated graph confusion matrix

The sampled integrated graph cluster in Figure 6.58 exhibits a distinct core-periphery arrangement, in which a central hub (cluster node, depicted in grey) serves as the anchor for multiple sentence nodes (yellow) as well as a few sentences from other senses (teal/green and red). The presence of short, multi-edged paths around the hub demonstrates that sentence–sentence similarity and sentence–term→domain connections are mutually reinforcing, creating a compact semantically aligned community—this underlies the model’s strong AUC and near-diagonal confusion matrix. The few bridging sentences, those linked to several neighbours and sparsely connected to the periphery, are likely points of conflation: these sentences carry overlapping signals (e.g., numeric tokens in locative contexts) and account for the lingering crossover between location and the number seven. From a modelling perspective, this structure validates that merging similarity, domain cues, and cluster hubs yields a robust sense neighbourhood, while isolating ambiguous instances. Practically speaking, tightening numeric/time patterns, slightly adjusting or refining cluster granularity, and implementing class-specific threshold adjustments would lessen the impact of these bridges, further differentiating colours in the cluster, and mapping the graph’s structure into improved recall for minority senses without compromising precision.



6.7 Results Discussion

The experimental findings provide a clear understanding of the most effective strategies for Sesotho sa Leboa (SsaL) concerning context-aware word embedding and word-sense discrimination. Knowledge-infused graph methods take the lead, with the domain-specific graph model achieving consistently high precision, recall, F1, and accuracy scores (all approximately 98%). This suggests that explicitly incorporating sense relations and domain constraints is well-suited to SsaL’s semantic structure, in which surface forms are greatly varied due to morphology. The integrated graph (scoring around 88% across parameters) and the hierarchical graph (showing 86/85/85 with about 83% accuracy) come next, both demonstrating a balanced performance indicative of robustness across both common and rare senses. Similarly, optimised contextual sequence models also achieve strong results: Optimized ELMo advances from approximately 93/93/93 (about 87% accuracy) to around 96/96/96 (approximately 92% accuracy), and XLNet improves from around 86 F1 with about 80% accuracy to a more aligned 91/91/91 (around 91% accuracy). These improvements highlight that in low-resource, morphologically rich environments, careful optimization of tokenization, learning rates, sequence lengths, regularization, and class balancing can be just as significant as the foundational architecture. Simultaneously, the experiments highlight significant warnings against relying exclusively on overall accuracy or general-purpose transformer backbones. Take BART as an illustrative case: the base model demonstrates balance with precision, recall, and F1 each at about 93 (accuracy around 87), but the optimised version shows increased accuracy (around 96), while recall and F1 fall to the mid-80s. This trend—accuracy increasing while F1 decreases—indicates class imbalance and thresholding effects in which the model becomes more confident in majority senses but overlooks minority senses. CNN results exhibit a similar imbalance: despite high accuracy (around 93), the F1 score remains low (around 69), emphasizing that accuracy alone can be misleading in imbalanced, detailed sense classification. More generally, commonly used transformer baselines (BERT, RoBERTa, ELECTRA, and their multilingual versions) perform poorly on macro-F1 despite having acceptable accuracy, probably due to inadequate pretraining alignment with SsaL’s morphological statistics, sparse in-

language signals during pretraining, and challenges in distinguishing closely related senses without external knowledge.

These results have various implications for developing context-aware embeddings in SsaL. Initially, incorporating sense-aware priors proves beneficial: encoders that are either guided by graphs or regularized by graphs are more effective at capturing the label topology and minimizing confusion among morphologically similar forms. Secondly, optimization choices are crucial: the performance of ELMo and XLNet indicates that employing class-balanced losses or re-sampling, adjusting decision thresholds, and using morphology-aware tokenization can lead to significant enhancements. Thirdly, when selecting models, it is crucial to prioritize macro-F1 over accuracy to prevent overestimating performance on majority senses. Additional evaluation tools—for instance, per-class precision/recall/F1 tables, confusion matrices, and calibration curves—should accompany main scores to ensure that improvements are evenly spread across the sense inventory.

The results also indicate specific solution When recall is low (e.g., optimized BART), using class-balanced or focal losses and adjusting thresholds for each class can enhance performance for minority senses. For calibration issues (e.g., CNN vs. F1), applying post-hoc calibration techniques such as temperature scaling or Platt scaling can align predicted probabilities with actual sense outcomes. To address transformer performance issues, employing morphology-aware tokenization (a mix of character and subword units), integrating additional morphological features (such as lemma and affix tags), and using adapter-based specialization with data in the target language are logical advancements. From a reporting standpoint, thoroughness is crucial: entries missing full precision/recall/F1 metrics obscure the trade-offs and hinder fair comparisons.

Two primary strategies are identified for future research. The first is graph-guided CAWE, which involves employing the most effective domain-specific graph as a mentor to regulate or distil into contextual encoders. This strategy aims to blend the graph’s well-balanced separability with the scalability and transferability of neural encoders. The second strategy is morphology-aware contextualization, which proposes enhancing ELMo/XLNet-style architectures with character-level components and explicit morpho-syntactic tags, training them with contrastive objectives that maintain

distinctions between those closely related. These strategies should be combined with disciplined selection criteria centred on macro-F1 metrics and diagnostic reporting, including per-class metrics, confusion matrices, and calibration plots, to ensure authentic, comprehensive improvements across the corpus.

The research indicates that for morphologically rich, low-resource languages such as SsaL, models incorporating structured sense knowledge consistently perform best, while well-optimised contextual encoders serve as the most promising additions. Achieving robust, generalizable context-aware embeddings requires integrating graph-level sense structures with morphology-sensitive neural contextualization, alongside using metrics and diagnostics that accurately reflect class imbalances and nuanced semantic differences.

6.8 Chapter Summary

This chapter reported the full experimental pipeline for mitigating MCD in SsaL, from reproducible setup to comparative evaluation. Using Google Colab (Python 3, T4 GPU) with a standard machine-learning stack (Scikit-learn, TensorFlow, PyTorch, Transformers, Gensim, Pandas, Matplotlib, Seaborn), the study ensured transparent implementation and repeatability of results. A structured, *sense-annotated SsaL corpus*—organized by, sentence, word, and sense—was introduced and motivated for research addressing the MCD problem in WVS WE models for a morphologically rich, low-resource context, noting vocabulary scale and typical sentence lengths that influence modelling choices.

Exploratory data analysis established the empirical landscape driving model design. Word frequency analyses revealed a long-tailed distribution over ambiguous lemmas; sense-diversity statistics quantified substantial polysemy; and class-imbalance diagnostics showed skewed label prevalence. A consolidated table summarized ambiguous tokens, instance counts, and unique senses. The data-preparation pipeline—normalization, duplicate removal, punctuation handling, annotation verification,

targeted filtering, tokenization, and imbalance remedies—delivered a cleaner, better-balanced training signal while preserving linguistic nuance crucial for sense discrimination.

Experimental results contrasted baseline and optimised variants of key architectures, complemented by diagnostic tools (confusion matrices and ROC curves) to illuminate both aggregate and per-class behaviour. Optimization generally improved precision, recall, and F1-score in the presence of morphological complexity and sparse class distributions, while error analyses exposed residual overlaps among closely related senses. Graph-based approaches further strengthened sense separability by leveraging relational structure, offering complementary gains beyond sequence-only models. Collectively, the evidence indicates that (i) careful preprocessing and imbalance handling are necessary but insufficient on their own, (ii) architecture-level enhancements and graph-informed representations materially reduce WVS WE MCD, and (iii) evaluation must pair global metrics with class-wise diagnostics to capture subtle semantic confusions. These findings set up the thesis-level conclusions and targeted recommendations for advancing the MCD problem in WVS WE models and similar low-resource, morphologically rich languages.

CHAPTER 7 : SUMMARY, CONCLUSIONS AND FUTURE WORKS

7.1 Summary

This study tackled the pressing issue of the MCD problem in WVS WE model, a case study on a SsaL-Bantu language with limited resources and a complex morphological structure. WVS WE MCD manifests when computational models cannot differentiate between various meanings of polysemous words, creating semantic ambiguities that hinder applications in natural language processing. The main goal was to create and assess computational methods that can accurately disambiguate word meanings in SsaL, addressing its agglutinative nature and scarcity of annotated data. The research followed a multi-phase approach. Initially, a specialized corpus was assembled featuring polysemous words from SsaL with annotated meanings, concentrating on frequently ambiguous terms. Then, a diverse range of models was implemented and assessed, spanning from traditional architectures (BiLSTM, NNLM, CNN) to transformer-based models (BERT, RoBERTa, ELECTRA, BART, XLNet, ELMo), and including innovative graph-based techniques. Finally, systematic optimization was undertaken through morphology-aware preprocessing, dropout regularization, class-balancing methods (SMOTE), attention mechanisms, and ablation studies, to determine impactful components.

The experimental findings indicated clear performance trends across different model families. Within the transformer models, BERT demonstrated the highest baseline performance, achieving an accuracy of 87% and an F1-score of 60%, which was closely followed by Multilingual BERT with an 89% accuracy and a 55% F1-score. During the optimization phase, there were noteworthy advancements: Optimized BART increased its accuracy to 96% (previously 87%), Optimized XLNet enhanced its precision to 92% (up from 83%), and Optimized CNN with dropout saw a remarkable improvement, elevating its F1-score from 69% to 98%. Statistical assessments using McNemar's tests ($p < 0.05$) verified that these improvements were meaningful rather than fortuitous.

Graph-based techniques demonstrated the highest efficacy overall. The Domain-Specific Graph model attained nearly flawless performance with a 98% score across all metrics, by utilizing lexical patterns,

domain centres, and structured relationships. The integrated graph exhibited consistently strong, balanced performance at 88% across metrics, whereas the hierarchical graph delivered satisfactory results ranging from 83% to 86%. These graph approaches effectively encoded domain knowledge and contextual relationships that purely neural techniques struggled to capture. Analyses using the confusion matrix; and ROC curves demonstrated that, while most models excelled in managing high-frequency senses, such as “location” and “Number Seven,” they faced significant challenges with low-frequency senses such as “time” and “watch.” This trend underscored the enduring impact of WVS WE MCD in class-imbalanced scenarios common in low-resource languages.

In various baselines, optimised transformers (BART, XLNet, ELMo) and graph-based models arrive at a common conclusion: morphology and domain cues are critical signals in SsaL. Models that (a) identify morpheme or subword structures (such as agglutinative affixes and clitics), (b) utilize long-distance contexts (using permutation/attention to capture distant agreements or cue words), (c) incorporate domain-specific knowledge (using lexicons or patterns for location, numbers, time, and observation), and (d) apply techniques such as regularization and rebalancing (including dropout, SMOTE, and cost sensitivity) consistently decrease off-diagonal errors and increase AUC and F1 scores, particularly for rare senses, which are central to WVS WE MCD. Specifically, optimised versions of XLNet and ELMo improve F1 and AUC by better integrating distant morphological signals; an optimised CNN increases its F1 from 69% to 98% with dropout, indicating that regularization mitigates sparse semantics; and domain-specific or integrated graphs advance performance from 88% to 98% across metrics, showcasing how structured domain knowledge (such as lexicons and hubs) reinforces sense boundaries. These findings highlight the key features crucial for SsaL: morpheme-aware tokenization, long-range contextualization, domain cue anchoring, and training that addresses imbalances.

Throughout the experiments, standard word-vector and embedding models—both static word vector spaces and basic contextual models without awareness of domain specifics or morphological nuances—repeatedly demonstrated WVS WE MCD. Firstly, these models combine multiple meanings into shared vectors: static embeddings merge various interpretations of a single form into one point,

causing contexts that are numerically or spatially close (such as “šupa” as the numeral seven versus proximity indicators) to be aggregated. Secondly, these models weakly address morphology: the complex inflection and derivation of SsaL produce numerous surface variants, and solely using subword or BPE processes was inadequate for preserving affixal cues essential for sense disambiguation. Thirdly, the models have limited context scope: when significant cues are distant from the target token, models lacking explicit long-range mechanisms (or those that depend on average pooling) fail to detect them, exacerbating conflation issues. Fourthly, these models lack domain sensitivity: without domain-specific priors (like lexicons or patterns), predictions overly depend on frequency, skewing towards predominant senses—leading to seemingly high accuracy but low F1 scores. Lastly, they exhibit vulnerability to class imbalance: skewed distributions favour frequent senses, causing probability rankings to appear reasonable while recall for less common senses diminishes.

The experiments reveal that a context-aware word-embedding (WE) model, which integrates knowledge-based cues (such as domain lexicons, sense hubs, and Personalized PageRank (PPR) over sentence–term–domain graphs) with unsupervised representation learning methods (for instance, TF-IDF, transformer sentence embeddings, sentence similarity, and hierarchical clustering), effectively alleviates MCD in WVS WE models. Three progressively comprehensive model designs were investigated: a hierarchical graph (based on structure alone), a domain-specific graph (incorporating structure and domain priors), and an integrated graph (combining structure, domain priors, similarity, and cluster hubs). As more domain and structural information was included, confusion matrices became notably diagonal, and one-vs-rest ROC AUC values increased towards 0.97–1.00, especially for previously overlapping senses (e.g., location versus the number seven). This outcome indicates that merging the structure of unannotated corpora with minimal knowledge resources positions senses in distinct contexts, all while maintaining the versatility and scope of unsupervised embeddings.

The research validated the CAWE method within WVS WE MCD by basing its evaluation on practical application tasks and comprehensive comparisons. A sequential set of models—hierarchical graph, domain-specific graph, and integrated graph—were developed using an identical sense-annotated

corpus and evaluated through stratified splits against robust transformer benchmarks. The validation employed intrinsic metrics (such as precision, recall, F1, accuracy) alongside ranking metrics (for example, one-vs-rest ROC/AUC), in addition to confusion matrices and focused error analyses (e.g., the confusion of location with the number seven, and time entering numeric contexts). The findings indicated significant and consistent improvements when incorporating domain knowledge and graph context: the domain-specific graph achieved approximately 98% across all metrics, and the integrated graph around 88%, significantly outperforming baseline transformers (F1 ranging from roughly 17–60%). These results, along with sharply diagonal confusion matrices and nearly flawless AUCs, offer compelling evidence that the proposed WE model effectively captures the essential SsaL cues necessary for addressing MCD challenges in WVS WE models. In this section, the authors explain how the sub-questions presented in Section 1.5 of Chapter 1 were addressed.

RQ1: What salient features of a morphologically rich low resource language, such as SsaL, should be considered in developing a CAWE model that effectively addresses the MCD problem in WVS WE models?

To answer Research Question 1 (RQ1): Research study incorporates:

- Morphology-aware representations (subword or morpheme units, character/byte components) so that affixal cues survive tokenization.
- Context windows and attention that capture distant cues (permutation or transformer attention) to resolve senses when key markers are far from the target.
- Domain-guided features (lexicons or regex for numerals, temporal phrases, locatives) to create strong priors against conflation.
- Training strategies for skew (class balancing, calibrated thresholds, dropout) to lift minority-sense recall. The experiments show that when these are present, conflation collapses toward the diagonal and ROC/AUC remains >0.9 across senses; when absent (e.g., multilingual RoBERTa without domain adaptation), the model collapses to majority predictions—canonical WVS WE MCD. Thus,

salience in SsaL is jointly morphological (subword), contextual (long-range), and domain-anchored (lexical patterns), moderated by imbalance-aware learning.

To answer Research Question 2 (RQ2):

RQ2: What are the MCD limitations of existing WVS WE models in representing these features?

Traditional WVS WE models fail to effectively capture the three critical elements necessary for less-resourced, morphologically rich languages such as SsaL: (1) explicit representation of morphology through affix or morpheme markers, (2) the ability to utilize long-range context to detect distant cues that aid disambiguation, and (3) domain anchoring to distinguish between numerals, temporal expressions, and locatives. This deficiency commonly leads to overlapping meanings in vector space, poor performance on rare senses, and mix-ups between closely related semantic domains. Experiments revealed that incorporating any of the missing components—such as morphology-aware tokenization, increased emphasis on longer context windows, and the use of domain graphs/priors—reduced these issues (yielding clearer distinctions and higher AUC). This indicates that the fundamental limitation is due to a representational bias in standard WVS or WE models, which prioritizes frequency and surface-level similarities over morphology, distance, and domain specificity.

The experiments demonstrated such in this way :

- The error patterns in the initial WVS or WE data demonstrated bidirectional exchanges involving the location and the number seven, along with instances of time information leaking into numerical categories—typical of WVS WE MCD.
- Comprehensive analyses of sentence length and morphological intricacy revealed that longer sentences and structures rich in affixes had elevated error rates, signifying that models often missed distant or contextual cues as well as morphological signals.
- Incorporating domain-specific features and graphs, such as lexicons, regex, and hub scores, significantly lowered off-diagonal counts and increased the AUC to approximately 0.9–1.0, highlighting the impact of domain anchoring.

- Enhancing the balance or calibration increased the recall for minority classes, demonstrating that part of the constraint is attributed to a bias during training rather than solely to capacity limitations.

RQ3: How can a context-aware WE model that addresses these limitations be developed, using knowledge-based and unsupervised WE methods?

To answer Research Question 3 (RQ3):

A viable response to RQ3 is a hybrid, context-aware word embedding (WE) system that: (1) develops unsupervised sentence and word representations from unprocessed text (using TF-IDF or transformer embeddings), (2) anchors these vectors within a knowledge layer consisting of domain lexicons, numeral/time/locative pattern detectors, and domain hubs joined by sentence→term→domain connections, and (3) extracts graph features such as domain hits, hubs, centralities, neighbour-similarity statistics, and cluster IDs for input into a straightforward classifier or thresholded scoring system. The knowledge layer provides the necessary inductive bias (including morphology, domain, and long-range cues), whereas the unsupervised embeddings offer comprehensive coverage and contextual variation. Experimentally, this hybrid approach maintained high precision and improved recall for less common senses, reducing conflation and resulting in consistent enhancements in F1, accuracy, and AUC—demonstrating how a context-aware WE model can be devised to address the specific limitations of conventional word vector spaces or WE.

How the experiments achieved this:

The research examined each component individually. Incorporating only structural elements (such as hierarchical clustering) enhanced separation but did not eliminate numeric–locative spillover. Incorporating domain knowledge through lexicons and regex generated domain hubs, significantly reducing off-diagonal errors and achieving nearly perfect AUC for common classes, with observable benefits for less common ones. Ultimately, the integrated approach merged sentence-to-sentence similarity, cluster hubs, and domain hubs, resulting in a robust probability ranking with high AUC and diagonals aligning with gold-standard senses. Across the models, tuning class-specific thresholds and applying simple re-weighting transformed probability separability into balanced precision–recall,

proving that the hybrid approach—rather than any individual component—achieves the WVS WE MCD reduction.

RQ4: How can the CAWE model effectively be evaluated for SsaL?

To answer Research Question 4 (RQ4):

Validation is achieved by incorporating the model into the sense-disambiguation task, establishing rigorous test protocols (such as splits and fixed seeds), comparing against known benchmarks, and using a variety of complementary metrics to reveal meaning conflation deficiency, such as macro-F1, per-class precision/recall, ROC or AUC, and the confusion matrix. The experiments indicate that combining knowledge-based signals (such as domain lexicons/regex, domain hubs, Personalized PageRank) with unsupervised representations (including TF-IDF or transformer embeddings, sentence-sentence similarity, clustering) not only enhances the overall scores but importantly reduces cross-sense confusion, which is the main failure mode of WVS WE MCD. In essence, successful validation in SsaL WSD is evidenced by consistent metric improvements, stronger diagonals, and stable probability rankings across senses within the same corpus and evaluation framework.

How the experiments answered RQ4 (validation details):

The pipeline enabled validation through an organized protocol: (1) Create a dataset labelled with sense annotations, the SsaL; (2) Develop baseline systems using static or contextual word embeddings without specific domain or morphology insight; (3) Train three variations of graphs that progressively incorporate more structure and knowledge; (4) Assess with macro-averaged measurements to address class imbalances; (5) Analyse ROC curves for individual senses to assess model calibration and differentiation; and (6) Conduct error analysis based on class, sentence length, and overlapping cues. Progressing from hierarchical to domain-specific to integrated models created a causal narrative: structure alone provides benefit, anchoring to a domain significantly reduces conflation, and embedding similarity with cluster hubs enhances stable performance across all senses. Optional tuning of thresholds and calibration of probabilities translated a high AUC into improved recall, particularly for minority senses, supporting the validity of the task beyond metric illusion. Collectively, these

procedures offer a reproducible experimental framework for validating the capability of CAWE to address MCD issues in WVS WE models.

The findings from the experiments highlight several critical patterns regarding MCD in WVS WE models with limited resources:

- **Class Imbalance as the Main Factor:** The uniform gaps in precision and recall observed across models (for instance, BERT's 68% precision versus 57% recall) suggest that the primary issue with WVS WE MCD is the underrepresentation of minority senses. Senses with high frequency, such as "location," were classified effectively, whereas infrequent senses, such as "watch," were regularly misclassified or ignored. This trend verifies that WVS WE MCD is essentially a problem of data deficiency, made worse by the complexity of morphology.
- **Optimization tactics exhibit varied significance:** Dropout regularization showed remarkable success for CNNs, yielding a 29-point F1 boost, whereas attention mechanisms and transfer learning excel in transformer frameworks. This indicates the necessity for optimization strategies to match architectural inductive biases—employing regularization in models susceptible to overfitting on small datasets and using attention in models needing enhanced focus on disambiguating tokens.
- **Graph Structures Capture What Embeddings Overlook:** The impressive performance of domain-specific graphs (98% versus BERT's 87%) illustrates the advantage of explicitly encoding domain knowledge, lexical patterns, and semantic connections, which help to identify cues that distributed representations struggle to grasp from limited training data. Graph topology—including hub-and-spoke models, community detection, and shortest paths—offers interpretable, linguistically informed features that enhance statistical patterns.
- **Results Vary for Multilingual Models:** Although Multilingual BERT attained an acceptable accuracy of 89%, both Multilingual RoBERTa and ELECTRA experienced substantial performance declines. This variability indicates that the advantages of multilingual pretraining are influenced by the typological resemblance between the languages used in pretraining and the target

language, the quantity of training data available for each language, and the tokenizer's configuration. The agglutinative morphology of SsaL might not be adequately captured in multilingual training datasets largely composed of European languages.

Implications from this research experiment reach beyond SsaL, affecting the wider scope of low-resource NLP:

- **Regarding Technology for Low-Resource Languages:** The proven effectiveness of graph-based and optimised neural techniques offers a reproducible model for creating WSD systems in other languages with limited resources, especially those featuring intricate morphology. The domain-specific graph framework, which demands minimal domain knowledge to develop lexicons and pattern rules, presents a feasible approach when extensive annotation is not practical.
- **For Multilingual NLP:** The inconsistent outcomes with multilingual transformers point to a crucial issue: existing multilingual models do not sufficiently capture the intricacies of morphologically rich African languages. This suggests the need for (a) creating multilingual models specifically tailored for Africa with improved inclusivity of Bantu languages, or (b) focusing on monolingual models for prominent African languages, subject to the availability of community resources.
- **For Semantic Processing:** Semantic processing research indicates that word sense disambiguation for morphologically rich languages is ineffective when based solely on surface context. The effectiveness of morphology-aware preprocessing and subword tokenization proves that segmentation approaches must consider both the internal structure of words and not merely their distributional traits.

These findings are subject to several limitations that restrict their scope and generalizability:

- **Scope and Variation of the Dataset:** Although meticulously labelled, the dataset offers a narrow glimpse of SsaL vocabulary and contexts. The results observed for the four target senses—namely, “location,” “number seven,” “time,” “watch”—might not extend to other polysemous terms that exhibit diverse frequency distributions or semantic linkages. The notably low accuracy

for "watch" (frequently yielding zero correct predictions) is more a consequence of insufficient data than intrinsic issues with the model. A limitation of this study is the relatively small size of the manually annotated corpus used for experimentation. Although the dataset was specifically developed to address meaning conflation deficiency (MCD) in WVS WE in the context of SsaL, its limited scale may reduce the generalisability of the trained models to broader linguistic contexts. In particular, small datasets can increase the risk of overfitting, encourage memorisation of repeated lexical or contextual patterns, and make evaluation results sensitive to similarities between training and test instances. Accordingly, the findings of this study should be interpreted as evidence of model effectiveness within the scope of the constructed corpus rather than as definitive proof of broad generalisation. Future studies should therefore validate the proposed approaches on larger, more diverse, and independently collected corpora.

- **Computational Resources:** The development of large transformer models necessitated significant computational infrastructure, which constrained the level of hyperparameter experimentation and the breadth of ensemble techniques. Expanding grid searches or employing neural architecture search might result in additional advancements.
- **Domain Specificity:** While the domain-specific graph produced outstanding results, it necessitates the manual creation of domain lexicons, patterns, and hub definitions. This method might not efficiently extend to open-domain usage or domains that evolve quickly with frequently changing vocabulary.
- **Evaluation Across Languages:** The study was solely concentrated on SsaL, and it did not empirically demonstrate that the methods could be transferred to comparable Bantu languages (such as Zulu, Swahili, or Northern Sotho) or to typologically similar agglutinative languages (including Turkish, Finnish, and Mongolian). The assertions regarding its wider applicability are still theoretical.
- **Annotation Reliability:** Although inter-annotator agreement was upheld, the intrinsic subjectivity in sense annotation, especially for polysemous terms with gradient rather than distinct boundaries, can introduce possible noise in the gold standard labels.

7.2 Conclusion

This study demonstrates that the issue of WVS WE MCD in word-vector semantics can be significantly mitigated for a low-resource, morphologically rich language such as Sesotho sa Leboa by integrating linguistically significant signals with corpus-based representations. Throughout the experiments, four key observations were revealed. Firstly, the importance of morphology cannot be overlooked: subword or morpheme indicators must persist through tokenization to maintain distinct senses. Secondly, context length is crucial: models require methods to capture distant cues for effective disambiguation.

Firstly, it is essential to anchor knowledge within the domain: lightweight knowledge sources like lexicons and regex patterns for numerals, time, or locatives, along with graph structures (sentence-to-term-to-domain hubs and Personalized PageRank), provide inductive bias that generic embeddings are missing. Secondly, the training practicalities are significant: factors such as class imbalance, calibration, and threshold-setting influence the recovery of minority senses. Studies show that incorporating structure through hierarchical clustering, followed by domain knowledge through domain-specific graphs, and finally an integrated approach using sentence–sentence similarity coupled with clusters and domain hubs, resulted in confusion matrices aligning towards the diagonal and improved ROC or AUC and F1 scores, with the domain-specific and integrated graphs providing the most reliable and interpretable enhancements.

These findings address the research questions by pinpointing the critical SsaL features that need encoding (RQ1), revealing the shortcomings of typical WVS/WE that ignore these features (RQ2), illustrating the construction of a context-aware hybrid WE to fill these gaps (RQ3), and detailing a rigorous validation methodology for the SsaL WSD tasks (RQ4). Although perfect scores on very small classes should be treated with caution, the trend toward improvement is evident: embeddings that are morphology-aware, context-sensitive, and domain-guided—anchored in straightforward, traceable graphs—provide a viable approach for effective sense discrimination in low-resource environments. This establishes a reproducible template for application to other agglutinative and code-switching

languages, and it lays the groundwork for future enhancements involving morpheme-focused encoders, sense-contrastive objectives, enhanced knowledge induction, and cross-lingual transfer.

This study has shown that meaning conflation deficiency in low-resource languages with complex morphology can be significantly reduced using well-crafted computational strategies. The domain-specific graph framework achieved almost perfect accuracy (98%), and optimised neural models presented significant enhancements over baseline models, confirming the effectiveness of morphology-aware preprocessing, dropout regularization, attention mechanisms, and class balancing techniques.

The research reveals that a one-size-fits-all approach is inadequate for tackling the MCD problem in WVS WE models globally. Rather, effective clarification is achieved by matching architectural designs with linguistic features: graph-based methods excel when domain knowledge is present, transformer models perform well with sufficient fine-tuning and class balance, and optimization approaches should align with architectural inductive biases. In languages such as SsaL, combining explicit linguistic insights with learned representations proves to be the most promising strategy.

Beyond the observed improvements in accuracy and Weighted-F1, the findings of this study also provide linguistic insight into the nature of Meaning Conflation Deficiency of WVS WE models in the context of Sesotho sa Leboa. The stronger performance of contextual and optimized models suggests that resolving conflated meanings in this language depends heavily on the ability to model morphological and contextual variation. As a morphologically rich language, Sesotho sa Leboa encodes meaning through affixation, noun class concords, derivational extensions, and syntactic context, which can cause superficially similar lexical forms to express different meanings across contexts. Models such as ELMo and XLNet were therefore better positioned to capture these context-sensitive variations, while the optimized CNN may have benefited from detecting local morphological patterns and short contextual features. These results imply that future advances in Meaning Conflation

Deficiency for Sesotho sa Leboa are likely to depend not only on larger datasets, but also on more linguistically informed modelling approaches

In addition to its technical advancements, this research also progresses the overarching objective of promoting linguistic equity in NLP. Similar to numerous African languages, SsaL is scarcely supported by modern language technologies, even though it is spoken by millions. This study illustrates that complex WSD can be accomplished with minimal resources, offering both methodological insights and empirical motivation for crafting language technologies that cater to a variety of linguistic communities.

The future involves enhancing the scope of sense inventories, confirming the transferability to related languages, exploring few-shot learning paradigms, and eventually incorporating advanced WSD into applications that aid SsaL-speaking communities. Despite existing challenges, especially in terms of data availability and computational resources, the significant progress reported here sets the stage for ongoing development in low-resource semantic processing.

While word sense disambiguation might seem like a specific technical issue, it actually poses a fundamental challenge in computational semantics: allowing machines to comprehend meanings similarly to humans, while considering ambiguity, context, and the complex morphological frameworks prevalent in many languages. The achievement of SsaL in this context provides optimism that even the most resource-limited languages worldwide can take advantage of advanced NLP techniques, advancing us toward language technology that inclusively caters to the full spectrum of linguistic diversity within humanity.

7.3 Future Work

- A limitation of the current experiment is that training was stopped while the model was still showing improvement in both loss and accuracy. Future work should therefore explore a longer training schedule to establish the true convergence point and to assess more reliably the balance between continued learning and the risk of overfitting.

- **Expanded Sense Inventory:** The current priority is to broaden the annotated corpus to incorporate more polysemous terms aside from the four main targets. Developing a corpus featuring 50-100 high-frequency ambiguous words could facilitate a more thorough evaluation and foster the creation of models with greater generalization. This extension ought to focus on terms with diverse frequency distributions and semantic connections to evaluate model resilience against various polysemy types.
- **Exploring Cross-Lingual Transfer Learning:** By examining transfer learning from related Bantu languages such as Zulu, Northern Sotho, and Swahili, it is possible to exploit common morphological structures and shared vocabulary. Employing a multi-task learning approach, in which models simultaneously acquire word sense disambiguation (WSD) competencies for various Bantu languages, may enhance performance via cross-linguistic regularization, especially for senses that rarely occur.
- **Explicit Morpheme Modelling:** Incorporate an automated morpheme segmenter for SsaL; develop tokenization/BPE that is morpheme-aware, along with character-CNN adapters and morpheme graphs; investigate which affixes are most effective in decreasing conflation.
- **Incorporate Sense-aware Objectives:** Implement contrastive or metric learning to separate different senses of a single lemma. Explore multitask approaches combining WSD and morphology tagging, along with auxiliary components for domain prediction.
- **Enhanced Knowledge Acquisition:** Expand lexicons using weak supervision techniques such as pattern mining and nearest-neighbor collection, develop temporal, numeric, and location patterns through bootstrapping, and determine the importance of knowledge sources using graph attention.
- **Comprehensive Graph Neural Models:** Substitute feature extraction with GAT or GNN applied to the sentence→term→domain graph, maintaining interpretability, and evaluate such against the existing RandomForest or threshold-based scorer.
- **Evaluate whether Longformer, ModernBERT, or retrieval-augmented encoders are effective at capturing remote cues in long-context encoders.** Assess improvements on extended sentence segments identified in error analyses.

- **Calibration and Thresholds:** Adjust per-class probabilities using methods such as temperature scaling or Platt scaling. Optimise the decision thresholds specifically for minority classes and provide calibration curves along with ROC and AUC metrics.
- **Expansion of Data and Annotation:** Increase the sense inventory, equalize class distributions, and gather more challenging negative samples; employ lightweight crowdsourcing combined with adjudication by linguists and report inter-annotator agreement.
- **Conduct statistical testing** by employing stratified bootstrap confidence intervals; implementing McNemar's test and applying multiple-comparison controls to verify the significance of enhancements across various senses and dimensions.
- **Constructing Dynamic Graphs:** Instead of relying on static domain graphs, the scalability could be enhanced by creating and refining graphs automatically using corpus statistics and linguistic resources. Techniques such as unsupervised community detection, automatic extraction of patterns from contexts, and bootstrapping domain hubs from initial seed terms could considerably lessen the need for manual knowledge engineering.

REFERENCES

- Abudouwaili, G., Abliz, W., Abiderexiti, K., Wumaier, A. & Yi, N. 2023. Strategies to Improve Low-Resource Agglutinative Languages Morphological Inflection. In *CoNLL 2023 - 27th Conference on Computational Natural Language Learning, Proceedings*. Singapore: Association for Computational Linguistics: 508–520. <https://doi.org/10.18653/v1/2023.conll-1.34>.
- Adebara, I. & Abdul-Mageed, M. 2022. Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1: 3814–3841. <https://doi.org/10.18653/v1/2022.acl-long.265> 31 August 2023.
- Aliyu, Y., Sarlan, A., Danyaro, K.U. & Rahman, A.S.B.A. 2024. Comparative Analysis of Transformer Models for Sentiment Analysis in Low-Resource Languages. (*IJACSA*) *International Journal of Advanced Computer Science and Applications*, 15(4): 353–364. <https://doi.org/10.14569/ijacsa.2024.0150437>.
- Almutairi, F.M., Sidiropoulos, N.D. & Karypis, G. 2017. Context-Aware Recommendation-Based Learning Analytics Using Tensor and Coupled Matrix Factorization. *IEEE Journal on Selected Topics in Signal Processing*, 11(5): 729–741.
- Alpaydin, E. 2014. *Introduction to Machine Learning*. T. Dietterich, C. Bishop, D. Heckerman, M. Jordan, & M. Kearns, eds. London, England: MIT Press Cambridge, Massachusetts.
- Amol, C.J., Gitau, A.M., Gesicho, R.D., Muhia, B. & Moruye, L. 2024. STATE OF NLP IN KENYA: A SURVEY. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2410.09948>.
- Anastasopoulos, A. & Neubig, G. 2019. Pushing the limits of low-resource morphological inflection. In *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*. Hong Kong, China: arxiv: 984–996. <https://arxiv.org/pdf/1908.05838>.

- Ansell, A. 2020. *Contextualised Approaches to Embedding Word Senses*. The University of Waikato.
<http://researchcommons.waikato.ac.nz/>.
- Ansell, A., Bravo-Marquez, F. & Pfahringer, B. 2021. PolyLM: Learning about polysemy through language modeling. In *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*. EACL(Online): Association for Computational Linguistics (ACL): 563–574. <https://arxiv.org/abs/2101.10448> 14 June 2023.
- Arnett, C. & Bergen, B.K. 2024. Why do language models perform worse for morphologically complex languages? *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2411.14198>.
- Arseniev-Koehler, A. 2024. HHS Public Access. *Sociol Methods Res*, 53(4): 1753–1793.
- Bakx, G.E. 2006. *Machine Learning Techniques for Word Sense Disambiguation*. Universitat Politècnica de Catalunya. <https://www.lsi.upc.edu/~escudero/wsd/06-tesi.pdf>.
- Bayram, M.A., Fincan, A.A., Gümü, A.S. & Karaka, S. 2025. Tokenization Standards for Linguistic Integrity : Turkish as a Benchmark. *arXiv (Cornell University)*: 1–10.
<https://doi.org/10.48550/arxiv.2502.07057>.
- Belete, M.D., Shiferaw, L.G., Alitasb, G.K. & Tamir, T.S. 2024. Enhancing Word Sense Disambiguation for Amharic homophone words using Bidirectional Long Short-Term Memory network. *Intelligent Systems with Applications*, 23(200417).
<https://doi.org/10.1016/j.iswa.2024.200417>.
- Bendor-Samuel, J.T. (Ed.). 1989. *The Niger–Congo Languages: A Classification and Description of Africa’s Largest Language Family*. University Press of America.
- Bevilacqua, M. & Navigli, R. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*: 2854–2864.
<https://aclanthology.org/2020.acl-main.255/>.
- Bhatia, G., Adebara, I., Elmadany, A., Abdul-mageed, M., Learning, D., Language, N. & Group, P. 2023. UBC-DLNLNLP at SemEval-2023 Task 12: Impact of Transfer Learning on African

- Sentiment Analysis. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2304.11256>.
- Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5: 135–146. <https://transacl.org/ojs/index.php/tacl/article/view/999>.
- Bosch, S.E. & Griesel, M. 2017. Strategies for building wordnets for under-resourced languages : The case of African languages. *Literator - Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, 38(1): 1351. <https://doi.org/10.4102/lit>.
- Botha, J.A. 2014. *Probabilistic Modelling of Morphologically Rich Languages*. University of Oxford. <https://arxiv.org/pdf/1508.04271.pdf>.
- Brown, N. & Marivate, V. 2024. Pula : Training Large Language Models for Setswana. *arXiv (Cornell University)*.
- Butnaru, A.M. & Ionescu, R.T. 2019. ShotgunWSD 2.0: An Improved Algorithm for Global Word Sense Disambiguation. *IEEE Access*, 7: 120961–120975. <https://ieeexplore.ieee.org/document/8817973>.
- Camacho-Collados, J. & Pilehvar, M.T. 2020. Embeddings in Natural Language Processing. In *COLING 2020 - 28th International Conference on Computational Linguistics, Tutorial Abstracts*. COLING: ACL ANTHOLOGY: 10–15. <https://aclanthology.org/volumes/2020.coling-tutorials/>.
- Camacho-Collados, J. & Pilehvar, M.T. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63(May): 743–788.
- Chaudhary, A., Zhou, C., Levin, L., Neubig, G., Mortensen, D.R. & Carbonell, J.G. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*. arxiv: 3285–3295. <https://arxiv.org/pdf/1808.09500>.
- Chokoe, S.J. 2000. *Linguistic Ambiguity in Northern Sotho: Saying the Unmeant*. Randse Afrikaanse Universiteit.
- Chronis, G. & Erk, K. 2020. When is a bishop not like a rook? When it's like a rabbi! Multi-

- prototype BERT embeddings for estimating semantic relationships. *CoNLL 2020 - 24th Conference on Computational Natural Language Learning, Proceedings of the Conference*: 227–244.
- Collins, S.P., Storrow, A., Liu, D., Jenkins, C.A., Miller, K.F., Kampe, C. & Butler, J. 2021. *Language in South Africa*. R. Mesthrie, ed. Cambridge, United Kingdom: cambridge university press. <http://www.cambridge.org>.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. & Kuksa, P. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12: 2493–2537. <https://dl.acm.org/toc/jmlr/2011/12/null>.
- Daniel Loureiro, A.J. 2019. Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics. <https://aclanthology.org/P19-1569/> 16 July 2021.
- Deerwester, S., Furnas, G.W., Landauer, T.K., Harshman, R. & Dumals, S.T. 2015. Indexing by Latent Semantic Analysis Scott. *Journal of the American Society for Information Science*, 3(12): 391–407.
- Devlin, J., Chang, M.-W., Lee, K., Google, K.T. & Language, A.I. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics: 4171–4186. <https://aclanthology.org/N19-1423.pdf>.
- Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. ACL Anthology: 4171–4186. <https://aclanthology.org/N19-1423.pdf>.
- Dibitso, M., Owolawi, P.A. & Ojo, S.O. 2022. An Hybrid Part of Speech Tagger for Setswana

- Language using a Voting Method. *International Conference on Intelligent and Innovative Computing Applications*, 2022: 245–253.
- Dlamini, S., Department of Computer Science, University of KwaZulu-Natal, Durban, S.A. & Pillay, ; Edgar Jembere; Anban. 2020. Evaluation of Word and Sub-word Embeddings for isiZulu on Semantic Relatedness and Word Sense Disambiguation Tasks. In *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*. Durban, South Africa: IEEE. <https://ieeexplore.ieee.org/document/9183836>.
- Dlamini, S., Jembere, E. & Pillay, A. 2020. Evaluation of Word and Sub-word Embeddings for isiZulu on Semantic Relatedness and Word Sense Disambiguation Tasks. *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems, icABCD 2020 - Proceedings*.
- Dlamini, S., Jembere, E., Pillay, A. & van Niekerk, B. 2021. Isizulu word embeddings. *2021 Conference on Information Communications Technology and Society, ICTAS 2021 - Proceedings*: 121–126.
- Držík, D. & Kapusta, J. 2026. The importance of morphology-aware subword tokenization for NLP tasks in Slovak language modeling. *Expert Systems With Applications*, 312: 131492 Contents. <https://doi.org/10.1016/j.eswa.2026.131492>.
- Ebbers, S. 2009. Vocabulary through morphemes: Suffixes, prefixes, and roots for intermediate grades Research base and curriculum overview. , (September): 1–16.
- Eiselen, R. & Tanja Gaustad. 2023. Deep learning and low resource languages. In *RAL Anthology*.
- El-alami, F., Ouatik, S., Alaoui, E. & En, N. 2022. Contextual semantic embeddings based on fine-tuned AraBERT model for Arabic text multi-class categorization. *Journal of King Saud University - Computer and Information Sciences*, 34(10): 8422–8428. <https://doi.org/10.1016/j.jksuci.2021.02.005>.
- FAA, G., Bosch, S. & Taljard, E. 2012. Towards a part-of-speech ontology: Encoding morphemic units of two South African Bantu languages. *Nordic Journal of African Studies*, 21(3): 118–140.
- Faaß, G. 2022. Applied Corpus Linguistics for Lexicography : Sepedi Negation as a Case in Point.

Lexikos, 32(2): 82–104.

Febrian, G., Novitasari, U. & Hidayat, A. 2022. Agglutinative language in bahasa Indonesia. *LADU: Journal of Languages and Education*, 2(4): 135–141.

Fesseha, A., Xiong, S., Emiru, E.D., Diallo, M. & Dahou, A. 2021. Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya. *Information (Switzerland)*, 12(2): 1–17.

Firth, J.R. 1957. A Synopsis of Linguistic Theory 1930-55. *Studies in Linguistic Analysis: Special Volume of the Philological Society*: 1–32.

Freihat, A.A. 2014. *An Organizational Approach to the Polysemy Problem in WordNet*. DIT - University of Trento. <http://eprints-phd.biblio.unitn.it/1265/>.

Gan, L. & Zhang, Y. 2020. Investigating Self-Attention Network for Chinese Word Segmentation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 28: 2933–2941.

Gao, A.K. 2023. Vec2Vec: A Compact Neural Network Approach for Transforming Text Embeddings with High Fidelity. *ArXiv preprint*. arxiv:2306.12689.

Geline, E., Evans, R., Russo, E., Richardson, D., Steinmetz, S. & Johnson, C. 2024. Linguistic precision in large language models with adaptive disambiguation and efficient monte carlo tree search for contextual clarity. *Research Square*. <https://www.researchsquare.com/article/rs-5186782/latest>.

George, M. 2018. Unsupervised topic detection based on 2D vector space model using apriori algorithm and nlp. *2018 13th International Conference on Digital Information Management, ICDIM 2018*: 279–283.

Ghasemi, Z. & Keyvanrad, M.A. 2021. FarSick: A Persian Semantic Textual Similarity and Natural Language Inference Dataset. *ICCKE 2021 - 11th International Conference on Computer Engineering and Knowledge*, (Iccke): 194–199.

Goldberg, Y. & Levy, O. 2014. word2vec Explained: deriving Mikolov et al.’s negative-sampling word-embedding method. , (2): 1–5. <http://arxiv.org/abs/1402.3722>.

Good, J. 2017. Niger–Congo languages. In *The Cambridge Handbook of Areal Linguistics*.

- cambridge university press: 471–499.
- Greenberg, J.H. 1949. Studies in African Linguistic Classification. *Southwestern Journal of Anthropology*.
- Griesel, M., Bosch, S. & Mojapelo, M.L. 2020. Thinking globally, acting locally - Progress in the African Wordnet Project. *Proceedings of the 10th Global WordNet Conference*: 191–196.
- Harris, Z.S. 1970. Distributional Structure. *Papers in Structural and Transformational Linguistics*: 775–794.
- Hl'adek, D., Sta's, J., Pleva, M., Ond'a's, S. & Kov'acs, L. 2016. Survey of the Word Sense Disambiguation and Challenges for the Slovak Language. In *17th IEEE International Symposium on Computational Intelligence and Informatics*. Budapest, Hungary: IEEE: 225–230. <https://ieeexplore.ieee.org/document/7846408>.
- Hoblos, J. 2021. Experimenting with Latent Semantic Analysis and Latent Dirichlet Allocation on Automated Essay Grading. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*. Paris, France: IEEE: 1–7. <https://ieeexplore.ieee.org/document/9336533>.
- Hofmann, V., Pierrehumbert, J., ArXiv, H.S. preprint & 2020, U. 2021. Dynamic contextualized word embeddings. *ACM*, 673: 673–681. <https://arxiv.org/abs/2010.12684> 31 August 2023.
- Horn, F. 2017. Context encoders as a simple but powerful extension of word2vec. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Vancouver, Canada: Association for Computational Linguistics: 10–14. <https://aclanthology.org/W17-2602/>.
- Huang, E., Socher, R., Manning, C.D. & Ng, A.Y. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. *Meeting of the Association for Computational Linguistics*, 873. http://faratarjome.ir/u/media/shopping_files/store-EN-1426347067-7934.pdf.
- Huang, W., Zhang, J., Li, X., Zhou, X., Qi, D., Xi, J. & Liu, W. 2025. A Semantic and Intelligent Focused Crawler based on BERT Semantic Vector Space Model and Hybrid Algorithm (October 2024). *IEEE Access*, XX: 1. <https://doi.org/10.1109/ACCESS.2025.3542064>.
- Idris, A.M., Audu, R. & Mohammed, A. 2022. Information Systems and Emerging Technologies. ,

(November).

- Ilie, V.I., Truica, C.O., Apostol, E.S. & Paschke, A. 2021. Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings. *IEEE Access*, 9: 162122–162146. <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9634064>.
- Jayakody, J., Vidanagama, V., Perera, I. & Herath, H. 2023. ELMo layer Embedding Comparision with short text Classification. In *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*. Ravet IN, India: IEEE: 1–6. <https://doi.org/10.1109/ASIANCON58793.2023.10270646>.
- Jijo, S.M., Panchal, D., Ardeshana, J. & Chaudhari, U. 2024. Text Summarization using Textrank, Lexrank and Bart model. In *2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024*. Kamand, India: IEEE: 1–7. <https://ieeexplore.ieee.org/document/10725676>.
- Jimoh, T.A. & Nikolov, N.S. 2025. Bridging Gaps in Natural Language Processing for Yor`a : A Systematic Review of a Decade of Progress and Prospects. *arXiv (Cornell University)*: 1–43.
- Joffe, D. 2016. African Languages. *Language and Dictionary Software*. <http://africanlanguages.com/> 20 August 2024.
- Kartsaklis, D. 2014. Compositional Operators in Distributional Semantics. *Springer Science Reviews*, 2(1–2): 161–177.
- Keinan, R. & Hacoheh-kerner, Y. 2023. JCT at SemEval-2023 Tasks 12A and 12B : Sentiment Analysis for Tweets Written in Low-resource African Languages using Various Machine Learning and Deep Learning Methods , Resampling , and HyperParameter Tuning. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics: 365–378. <https://doi.org/10.18653/v1/2023.semeval-1.49>.
- Kelly, B., Wigglesworth, G., Nordlinger, R. & Blythe, J. 2014. The acquisition of polysynthetic languages. *Linguistics and Language Compass*, 8(2): 51–64.
- Kim, Y. 2014. An analysis of convolutional neural networks for sentence classification. In W. D.

- Alessandro Moschitti, Bo Pang, ed. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics (ACL): 1746–1751. <https://aclanthology.org/D14-1181/>.
- Kristó, L. 2006. On synthesis , fusion and the difference between them *. , 14: 63–94.
- Lamin, N.Z. & Aziz, A.A. 2025. Cross-Lingual Sentiment Analysis in Low-Resource Languages : A Recent Review on Tasks , Methods and Challenges. , 16(11): 411–426. <https://doi.org/10.14569/ijacsa.2025.0161144>.
- Leclair, A., Bansal, A. & McMillan, C. 2021. Ensemble Models for Neural Source Code Summarization of Subroutines. In *Proceedings - 2021 IEEE International Conference on Software Maintenance and Evolution, ICSME 2021*. Luxembourg: IEEE: 286–297. <https://ieeexplore.ieee.org/document/9609119>.
- Lenci, A. 2018. Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4: 131–145.
- Lestandy, M. 2023. Exploring the Impact of Word Embedding Dimensions on Exploring the Impact of Word Embedding Dimensions on Depression Data Classification Using BiLSTM Model Depression Data Classification Using BiLSTM Model. *Procedia Computer Science*, 227: 298–306. <https://doi.org/10.1016/j.procs.2023.10.528>.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. & Zettlemoyer, L. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*: 7871–7880.
- Li, N., Bouraoui, Z., Camacho-collados, J., Espinosa-anke, L., Gu, Q. & Schockaert, S. 2021. Modelling General Properties of Nouns by Selectively Averaging Contextualised Embeddings. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*. arXiv:2012.07580 [cs.CL]. <https://www.ijcai.org/proceedings/2021/0530.pdf>.
- Liji, S.K. & Muhamed Ilyas, P. 2021. Semantic Malayalam Dialogue System for Covid-19 Question Answering Using Word Embedding and Cosine Similarity. *10th International Conference on Advances in Computing and Communications, ICACC 2021*: 1–6.

- Lin, J.-Y., Yang, C.-H., Tseng, S.-C. & Huang, C.-R. 2002. The Structure of Polysemy : A Study of Multi-sense Words Based on WordNet. *Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation*: 320–329. <http://aclweb.org/anthology/Y02-1031>.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *ICLR 2020 Conference*. OpenReview. <https://arxiv.org/abs/1907.11692>.
- Loureiro, D., Mário Jorge, A. & Camacho-Collados, J. 2022. LMMS reloaded: Transformer-based sense embeddings for disambiguation and beyond. *Artificial Intelligence*, 305: 103661. <https://doi.org/10.1016/j.artint.2022.103661>.
- Louwrens, L.J. 1991. *Aspects of Northern Sotho Grammar*. Via Afrika.
- Lundin, J.M., Zhang, A., Karim, N., Louzan, H., Wei, V., Carroll, C. & Adelani, D. 2025. The Token Tax : Systematic Bias in Multilingual Tokenization. *arXiv (Cornell University)*: 1–8. <https://doi.org/10.48550/arxiv.2412.06926>.
- Mabela, Matefu L. & Ditsele, T. 2024. Exploring intellectualisation of South African indigenous languages for academic purposes. *Literator*, 45(1): 1–10.
- Mabela, Matefu L & Ditsele, T. 2024. Intellectualisation of Northern Sotho through English terminology adaptation The development of South African. *Literator -Journal of Literary Criticism, Comparative Linguistic and Literary Studies*, 45(1): 1–11. <https://doi.org/10.4102/lit.v45i1.2031>.
- Mabokela, K.R., Primus, M. & Celik, T. 2025. low-resourced african languages using. *els. PLoS One* 20(6):, 20(6): 1–37. <https://doi.org/10.1371/journal.pone.0325102>.
- Makgatho, M., Marivate, V., Sefara, T. & Wagner, V. 2021. Training Cross-Lingual Embeddings for Setwana and Sepedi. In *DHASA Conference*.
- Maksymenko, D. & Turuta, O. 2025. Tokenization efficiency of current foundational large language models for the Ukrainian language. *Front. Artif. Intell.*, 8(August). <https://doi.org/10.3389/frai.2025.1538165>.
- Makutoane, T.J. 2022. ‘ The people divided by a common language ’: The orthography of Sesotho in

- Lesotho , South Africa , and the implications for Bible translation. *HTS Teologiese Studies/Theological Studies*, 78(1): 1–9. <https://doi.org/10.4102/hts.v78i1.7605>.
- Malinga, M., Lupanda, I., Nkongolo, M.W. & Deventer, P. Van. 2021. A Multilingual Sentiment Lexicon for Low-Resource Language Translation using Large Languages Models and Explainable AI. *arXiv (Cornell University)*: 80–83. <https://doi.org/10.48550/ARXIV.2411.04316>.
- Man, R. & Lin, K. 2021. Sentiment Analysis Algorithm Based on BERT and Convolutional Neural Network. In *2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*. Dalian, China: IEEE: 769–772. <https://doi.org/10.1109/IPEC51340.2021.9421110>.
- Margani, R. & Ndugu, N. 2023. From Local to Global : Navigating Linguistic Diversity in the African Context. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2305.01427>.
- Marivate, V., Sefara, T., Chabalala, V., Keamogetswe Makhaya, Tumisho Mokgonyana, Rethabile Mokoena, A.M. 2022. Investigating and approach of low resource language dataset creation, curation and classification: Setswana and Sepedi. *arXiv*: 1–6.
- Marivate, V., Sefara, T., Chabalala, V. 2021. Low Resource Language Dataset Creation, Curation and Classification: Setwana and Sepedi- Extended Abstract. In *ICLR 2020*.
- Masethe, H.D., Masethe, M.A., Ojo, S.O. & Owolawi, P.A. 2025. Hybrid Transformer-Based Large Language Models for Word Sense Disambiguation in the Low-Resource Sesotho sa Leboa Language. *Applied Sciences*, 15(3608): 1–33. <https://doi.org/10.3390/app15073608>.
- Masethe, M.A., Masethe, H.D. & Ojo, S.O. 2024. Context-Aware Embedding Techniques for Addressing Meaning Conflation Deficiency in Morphologically Rich Languages Word Embedding: A Systematic Review and Meta Analysis. *Computers*, 13(10).
- Mathe, D.S. & Eiselen, R. 2021. Quantitative analysis of Sesotho sa Leboa part-of-speech taggers. *South African Journal of African Languages*, 41(3): 259–269.
- Mikolov, T., Corrado, G., Chen, K. & Dean, J. 2013. Efficient estimation of word representations in vector space. *Computation and Language (cs.CL)*: 1–12. <https://arxiv.org/abs/1301.3781>.

- Milano, N., Luongo, M., Ponticorvo, M. & Marocco, D. 2025. Semantic analysis of test items through large language model embeddings predicts a-priori factorial structure of personality tests. *Current Research in Behavioral Sciences*, 8(January): 100168.
<https://doi.org/10.1016/j.crbeha.2025.100168>.
- Miller, G.A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41. <http://portal.acm.org/citation.cfm?doid=219717.219748>.
- Mitchell, J. & Lapata, M. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8): 1388–1429.
- Modiba, M. 2024. Tlh goba hl : Bothata bja mongwalo le mopeleto baboleding ba Sesotho sa Leboa ba go kgona go ngwala le go bala Tlh or hl : An orthographic dilemma for the literate Northern. , 58(1): 1–16.
- Mojapelo, M.L. 2007. Definiteness in Northern Sotho By. , (December).
- Mojapelo, M.L. 2009. Morphology and semantics of proper names in northern sotho. *South African Journal of African Languages*, 29(2): 185–194.
- Mojela, V.M. 2016. Etymology & figurative : The role of etymology in the Lemmatization of Sotho terminology. In D. L. S. U. Shirley N. Dita, ed. *Advancing Language Teaching through Lexicography and Corpus-building*. Manila: 93–100.
- Moors, C., Calteaux, K., Wilken, I. & Gumede, T. 2018. Human language technology audit 2018: Analysing the development trends in resource availability in all South African languages. *ACM International Conference Proceeding Series*: 296–304.
- Muhammad, S.H., Abdulmumin, I., Yimam, S.M., Adelani, D.I., Ahmad, I.S., Ousidhoum, N., Ayele, A.A., Mohammad, S.M., Beloucif, M. & Ruder, S. 2023. SemEval-2023 Task 12 : Sentiment Analysis for African Languages. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics: 2319–2337. <https://doi.org/10.18653/v1/2023.semeval-1.315>.
- Nasiruddin, M. 2013. A State of the Art of Word Sense Induction: A Way Towards Word Sense Disambiguation for Under-Resourced Languages. *Proceedings of RECITAL 2013*: 192–205.

- Neelakantan, A., Shankar, J., Passos, A. & McCallum, A. 2014. Efficient non-parametric estimation of multiple embeddings perword in vector space. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*: 1059–1069.
- Neha, F.N.U. & Bansal, A. 2024. Understanding the architecture of vision transformer and its variants: A review. In *1st International Conference on Innovative Engineering Sciences and Technological Research, ICIESTR 2024 - Proceedings*. Muscat, Oman: IEEE: 1–6.
<https://ieeexplore.ieee.org/document/10798341>.
- Nhlapo, M.A. 2015. *the Sequence of Derivational and Inflectional Morphemes in*. University of Free State.
- Nurse, D. & Gérard, P. 2003. *THE BANTU LANGUAGES*. London and New York: Taylor and Francis.
- Osborn, D. 2010. *AFRICAN LANGUAGES IN A DIGITAL AGE*. Cape Town: HSRC Press.
www.hsrcpress.ac.za.
- Pasini, T., Raganato, A. & Navigli, R. 2021. XL-WSD: An Extra-Large and Cross-Lingual Evaluation Framework for Word Sense Disambiguation. *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, 15: 13648–13656.
- Pennington, J., Socher, R. & Manning, C.D. 2014. GloVe: Global vectors for word representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*: 1532–1543.
- Peters, M.E., Neumann, M., Iyyer, M. & Gardner, M. 2018. 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)*. New Orleans, Louisiana: Association for Computational Linguistics: 2227–2237.
<https://aclanthology.org/N18-1202/>.
- Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. 2018. Deep contextualized word representations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

- Pilehvar, M.T. 2019. On the importance of distinguishing word meaning representations: A case study on reverse dictionary mapping. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*. Minneapolis, Minnesota: Association for Computational Linguistics: 2151–2156. <https://aclanthology.org/N19-1222/> 14 June 2023.
- Pilehvar, M.T. & Camacho-Collados, J. 2020. Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning. In *Synthesis Lectures on Human Language Technologies*. Springer, Cham: 69–96.
- Pilehvar, M.T. & Navigli, R. 2014. A robust approach to aligning heterogeneous lexical resources. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, 1: 468–478.
- Poulos, G. & Louwrens, L.J. 1994. *A Linguistic Analysis of Northern Sotho*. Via Afrika.
- Pretorius, R. & Pretorius, L. 2009. Setswana Tokenisation and Computational Verb Morphology : Facing the Challenge of a Disjunctive Orthography. In *EACL Workshop on Language Technologies for African Languages (AfLaT)*. Athens, Greece,: 66–73.
- Pretorius, R., Viljoen, B. & Pretorius, L. 2005. A finite-state morphological analysis of tswana nouns. *South African Journal of African Languages*, 25(1): 48–58.
- Raganato, A., Camacho-Collados, J. & Navigli, R. 2017. Word sense disambiguation: A unified evaluation framework & empirical comparison. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, 1(1): 99–110.
- Rahab, N. & Mothapo, B. 2019. *Determining the core vocabulary used by Sepedi- speaking preschool children during regular preschool-based activities by*. University of Pretoria. www.up.ac.za.
- Rakgogo, Tebogo Jacob. 2024. A critical examination of the socio-onomastic controversy of Sepedi and Sesotho sa Leboa/Northern Sotho language names and the role and effectiveness of the

Commission for Cultural, Religious and Linguistic Communities. *South African Journal of African Languages*.

Rakgogo, Tebogo J. 2024. Neither Sepedi nor Northern Sotho/Sesotho sa Leboa is onomastically correct: a colonial imposition. *Southern African Linguistics and Applied Language Studies*, 42(sup1): S207–S223.

Rakgogo, T.J. 2016. Sepedi or Sesotho Sa Leboa: a Sociolinguistic Perspective of First Language Speakers From Selected South African Universities. , (June 2016).

Rakgogo, T.J. & Zungu, E.B. 2022. The elevation of Sepedi from a dialect to an official standard language: Cultural and economic power and political influence matter. *Literator*, 43(1): 1–9.

Rakgogo, T.J. & Zungu, E.B. 2021. The onomastic possibility of renaming the sepedi and sesotho sa leboa (Northern Sotho) language names to restore peace, dignity and solidarity. *Literator*, 42(1): 1–14.

Reisinger, J. & Mooney, R.J. 2010. Multi-prototype vector-space models of word meaning. *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, (June): 109–117.

Rodrigues da Silva, J. & Caseli, H. de M. 2021. Sense representations for Portuguese: experiments with sense embeddings and deep neural language models. *Language Resources and Evaluation*, 55(4): 901–924. https://www.researchgate.net/profile/M-Elhadi-2/publication/362541497_Adopting_Arabic_Taggers_to_Annotate_a_Libyan_Dialect_Text_with_a_Pre-Tagging_Processing_and_Term_Substitutions/links/62f007ed505511283e9ac729/Adopting-Arabic-Taggers-to-Annotate-a-Libya 31 August 2023.

Rodriguez, M. & Merlo, P. 2020. Word associations and the distance properties of context-aware word embeddings. : 376–385.

Roh, J., Park, S., Kim, B.K., Oh, S.H. & Lee, S.Y. 2021. Unsupervised multi-sense language models for natural language processing tasks. *Neural Networks*, 142: 397–409.

- <https://doi.org/10.1016/j.neunet.2021.05.023>.
- S, B.R., Suri, G., Dewangan, V. & Sonavane, R. 2022. When Every Token Counts: Optimal Segmentation for Low-Resource Language Models. *arXiv (Cornell University)*.
<https://doi.org/10.48550/arxiv.2412.06926>.
- Sani, S.A. 2024. Investigating the Impact of Language-Adaptive Fine-Tuning on Sentiment. *ArXiv*.
<https://doi.org/10.18653/v1/2023.semeval-1.315>.
- Scarlini, B., Pasini, T. & Navigli, R. 2020. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference: 3528–3539*.
- Shafi, J., Muhammad, R.A.O.N.A. & Rayson, P. 2023. Semantic Tagging for the Urdu Language : Annotated Corpus and Multi-Target Classification Methods. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6): 175:1–31. <https://doi.org/10.1145/3582496>.
- Shahid, F., Elswah, M. & Vashistha, A. 2025. Think Outside the Data : Colonial Biases and Systemic Issues in Automated Moderation Pipelines for Low-Resource Languages. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2501.13836>.
- Sheth, D., Gupta, A.R. & D’Mello, L. 2021. Using Universal Sentence Encoder for Semantic Search of Employee Data. *2021 International Conference on Computational Intelligence and Computing Applications, ICCICA 2021: 1–4*.
- Sibeko, J. 2023. Developing Resources for Measuring Text Readability in Sesotho. In *CLARIN Annual Conference 2022*. Linköping Electronic Conference Proceedings 198,,: 120–132.
<https://doi.org/10.3384/ecp198012>.
- Silva, J.R. da & Caseli, H. de M. 2021. Sense representations for Portuguese: experiments with sense embeddings and deep neural language models. *Language Resources and Evaluation*, 55(4): 901–924.
- Simon, G., Bajzát, T., Ballagó, J. & Havasi, Z. 2025. *When MIPVU goes to no man ’ s land : a new language resource for hybrid , morpheme-based metaphor identification in Hungarian*. Springer

- Netherlands. <https://doi.org/10.1007/s10579-023-09705-9>.
- Smith, N.A. 2019. Contextual Word Representations: A Contextual Introduction. , (2019): 1–15.
<http://arxiv.org/abs/1902.06006>.
- Snyder, H. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104: 333–339.
- Song, L. 2016. Word Embeddings , Sense Embeddings and their Application to Word Sense Induction. *The University of Rochester Computer Science Department*: 1–41.
<https://www.cs.rochester.edu/u/lsong10/papers/area.pdf>.
- Statistics South Africa. 2012. *Census 2011: Census in brief (Report No. P03014)*. Pretoria, South Africa. <https://www.statssa.gov.za/publications/P03014/P030142011.pdf>.
- Statistics South Africa. 2022. *Census 2022: Census in Brief (Report)*. Pretoria, South Africa.
<https://census.statssa.gov.za>.
- Steven Bird, Ewan Klein, and E.L. 2009. Natural Language Processing with Python. In Michael Elhadad, ed. *Natural Language Engineering*. 419–424.
http://www.journals.cambridge.org/abstract_S1351324910000306.
- Takala, P. 2016. Word Embeddings for Morphologically Rich Languages. In *ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges (Belgium): i6doc.com: 27–29. <http://www.i6doc.com/en/>.
- Taljard, E., Faaß, G., Heid, U. & Prinsloo, D.J. 2008. On the development of a tagset for Northern Sotho with special reference to the issue of standardisation. *Literator*, 29 (1)(April 2008): 111–137.
- Thaba, A., Amrul, A., Mahmut, C., Karim Mahmut, A., Wahidah, N., Gunawan, A., Pratama, K., Devi Sahura, S., Darna, S., Dewi Anjani, H., Mutmainna, A., Afdaliah, W., Magfirah, A. & Sholeha, S. 2020. Use of Bound Morphem form in Novel Title 5 CM by Donny Dhirgantoro. *International Journal of Literature and Arts*, 8(5): 287.
- Thompson, R.C., Joseph, S. & Adeliyi, T.T. 2022. A Systematic Literature Review and Meta-Analysis of Studies on Online Fake News Detection. *MDPI Information (Switzerland)*, 13(11).

- Thurnbauer, M., Reisinger, J., Goller, C. & Fischer, A. 2023. Towards Resolving Word Ambiguity with Word Embeddings. In *arXiv (Cornell University)*. arXiv (Cornell University).
<http://arxiv.org/abs/2307.13417>.
- Tian, F., Dai, H., Bian, J. & Gao, B. 2014. A Probabilistic Model for Learning Multi-Prototype Word Embeddings. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: International Conference on Computational Linguistics: 151–160. <https://api.semanticscholar.org/CorpusID:9693038>.
- du Toit, J.S. & Puttkammer, M.J. 2021. Developing Core Technologies for Resource-Scarce Nguni Languages. *Information (Switzerland)*, 12(12). <https://www.mdpi.com/2078-2489/12/12/520> 31 August 2023.
- Trask, A., Michalak, P. & Liu, J. 2015. sense2vec - A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings. *arXiv(Cornell University)*: 1–9.
<http://arxiv.org/abs/1511.06388>.
- Vasic, D. & Brajkovic, E. 2018. Development and Evaluation of Word Embeddings for Morphologically Rich Languages. In *2018 26th International Conference on Software, Telecommunications and Computer Networks, SoftCOM 2018*. Institute of Electrical and Electronics Engineers Inc.: 308–313.
- Verner, J.M., Brereton, O.P., Kitchenham, B.A., Turner, M. & Niazi, M. 2012. Systematic literature reviews in global software development: a tertiary study. In *16th International Conference on Evaluation & Assessment in Software Engineering (EASE 2012)*. Ciudad Real, Spain: IEEE: 2–11. <http://digital-library.theiet.org/content/conferences/10.1049/ic.2012.0001>.
- Vusak, E., Kuzina, V. & Jovic, A. 2021. A Survey of Word Embedding Algorithms for Textual Data Information Extraction. In *2021 44th International Convention on Information, Communication and Electronic Technology, MIPRO 2021 - Proceedings*. Opatija, Croatia: IEEE: 181–186.
<https://ieeexplore.ieee.org/document/9597076>.
- Wang, Q., Zaki, M.J., Kollias, G. & Kalantzis, V. 2025. Multi-Sense Embeddings for Language Models and Knowledge Distillation. In *Findings of the Association for Computational*

- Linguistics: ACL 2025*. Vienna, Austria: Association for Computational Linguistics: 13353–13369. <https://aclanthology.org/2025.findings-acl.691/>.
- Wang, Y., Cui, L. & Zhang, Y. 2021. Improving Skip-Gram Embeddings Using BERT. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 29: 1318–1328.
- Wang, Y. & Zhang, Y. 2023a. Lost in Context? On the Sense-Wise Variance of Contextualized Word Embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 639–650. <https://ieeexplore.ieee.org/document/10333999/authors#authors>.
- Wang, Y. & Zhang, Y. 2023b. Lost in Context? On the Sense-Wise Variance of Contextualized Word Embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 639–650.
- Wiemerslage, A., Silfverberg, M., Yang, C., McCarthy, A.D., Nicolai, G., Colunga, E. & Kann, K. 2022. Morphological Processing of Low-Resource Languages: Where We Are and What’s Next. In *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics: 988–1007. <http://universaldependencies.org/v2/>.
- Worth, P.J. 2023. Word Embeddings and Semantic Spaces in Natural Language Processing. *International Journal of Intelligence Science*, 13(01): 1–21.
- Yaghoobzadeh, Y. & Schutze, H. 2016. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany: Association for Computational Linguistics (ACL): 236–246.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. & Le, Q. V. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32(517): 5753–5763. <https://dl.acm.org/doi/10.5555/3454287.3454804>.
- Yao, T., Zhai, Z. & Gao, B. 2020. Text Classification Model Based on fastText. *Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Information Systems, ICAIIS 2020*: 154–157.
- Yigezu, M.G., Woldeyohannis, M.M. & Tonja, A.L. 2021. Multilingual Neural Machine Translation

- for Low Resourced Languages: Omoto-English. *2021 International Conference on Information and Communication Technology for Development for Africa, ICT4DA 2021*: 89–94.
- Zaland, O., Abulaish, M. & Fazil, M. 2024. A Comprehensive Empirical Evaluation of Existing Word Embedding Approaches. : 1–28. <http://arxiv.org/abs/2303.07196>.
- Zerbian, S. 2007. Phonological phrasing in Northern Sotho (Bantu). *Linguistic Review*, 24(2–3): 233–262.
- Zhang, T., Ye, W., Xi, X., Zhang, L., Zhang, S. & Zhao, W. 2020. Leveraging human prior knowledge to learn sense representations. *Frontiers in Artificial Intelligence and Applications*, 325: 2306–2313. https://ecai2020.eu/papers/603_paper.pdf.
- Zhang, Y., Shi, R., Chen, C., Duan, M., Zhou, F. & Member, S. 2020. ELMO : An Efficient Logistic Regression-Based Multi-Omic Integrated Analysis Method for Breast Cancer Intrinsic Subtypes. *IEEE Access*, 8: 5121–5130. <https://doi.org/10.1109/ACCESS.2019.2960373>.
- Zipf, G.K. 1958. Linguistic Society of America. *PMLA/Publications of the Modern Language Association of America*, 26 (3)(5-Part2): 394-401 Published by: Linguistic Society of Americ. <http://www.jstor.org/stable/409735> Accessed: