Palestine Polytechnic University

Deanship of Graduate Studies and Scientific Research

Master of Intelligent Systems

Arabic Synonyms Extraction using Singular Value Decomposition and Self-Organizing Maps

Submitted by:

Tala Sarabtah

Supervised by:

Dr. Dia AbuZeina

Thesis submitted in partial fulfillment of requirements of the degree

Master of Intelligent Systems

June 2026

The undersigned hereby certify that they have read, examined, and recommended to the Deanship of Graduate Studies and Scientific Research at Palestine Polytechnic University the approval of a thesis entitled "Arabic Synonyms Extraction using Singular Value Decomposition and Self-Organizing Maps" submitted by Tala Sarabtah in partial fulfillment of the requirements for the degree of Master of Intelligent Systems.

Graduate Advisory Committee:

Dr. Dia AbuZeina (Supervisor), Palestine Polytechnic University.

Signature:_____ Date:_____

Dr. Alaa Halawani (Internal committee member), Palestine Polytechnic University.

Signature:_____ Date:_____

Dr. Mohammed Maree (External committee member), Arab American University.

Signature:_____ Date:_____

Thesis Approved

Prof. Mahmoud AlHaddad

Dean of Graduate Studies and Scientific Research

Palestine Polytechnic University

Signature: _____ Date: _____

DECLARATION

I declare that the Master Thesis entitled **"Arabic Synonyms Extraction Using Singular Value Decomposition and Self-Organizing Maps"** is my original work and hereby certify that unless stated, all work contained within this thesis is my independent research and has not been submitted for the award of any other degree at any institution, except where due acknowledgment is made in the text.

Tala Sarabtah

Signature: _____ Date: _____

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for the Master's degree of Intelligent Systems at Palestine Polytechnic University, I agree that the library shall make it available to borrowers under rules of the library. Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgement of the source is made. Permission for extensive quotation from, reproduction, or publication of this thesis may be granted by my main supervisor, or in his absence, by the Dean of Graduate Studies and Scientific Research when, in the opinion of either, the proposed use of the material is for scholarly purposes. Any copying or use of the material in this thesis for financial gain shall not be allowed without my written permission.

Tala Sarabtah

Signature: _____ Date: _____

الملخّص

اللغةُ العربيةُ لغةٌ غنيةٌ صرفيًّا، تتشاركُ فيها كلماتٌ كثيرةٌ الجذرَ نفسَه، وهو ما يُسبّبُ مشكلةً موثّقةً جيدًا عند الاستخراج الآلي للمرادفات؛ إذ تميلُ الطرقُ القياسيةُ القائمةُ على التمثيلات المتجهية وتشابهِ جيبِ التمام — وهي الطريقةُ الأساس (Baseline) — إلى إرجاع صيغٍ مشتقّةٍ من الجذر نفسِه بدلًا من المرادفات الوظيفية الحقيقية. تعالجُ هذه الرسالةُ هذه المشكلة.

تقترحُ هذه الرسالةُ منظومةً من ثلاث مراحل تجمعُ بين نموذجٍ لغويٍّ عربيٍّ مُدرَّبٍ مسبقًا (BERT)، وتحليلِ القيمة المفردة (SVD)، والخريطةِ ذاتيةِ التنظيم (SOM). يُولِّدُ نموذجُ BERT تمثيلاتٍ سياقيةً للكلمات؛ ويُقلِّلُ تحليلُ القيمة المفردة (SVD) من أبعاد تمثيل أيِّ كلمةٍ لكبح الضجيج؛ وتُعيدُ الخريطةُ ذاتيةُ التنظيم (SOM) تنظيمَ الفضاء بحيث تتجمّعُ الكلماتُ المتكافئةُ وظيفيًّا معًا بدلًا من الكلمات المتقاربة صرفيًّا. دُرِّبت منظومةُ SVD-SOM على نحو 3,800 مقالةٍ عربية تُغطّي 157 كلمةً هدفًا من مجالاتٍ متنوعة تشملُ التعليمَ والاقتصادَ والقانونَ والإعلام. ومن أبرزِ محدوديّات منظومة SVD-SOM غيابُ مرجعٍ قياسيٍّ لتقييم المرادفات العربية، وهو ما يجعلُ قياسَ الأداء صعبًا في جوهره. ولذلك أُجريَ التقييمُ بمقارنة مخرجات منظومة SVD-SOM بمكنز مايكروسوفت وورد العربي وعددٍ من المعاجم العربية الإلكترونية المعتمَدة (المعاني، والمعاجم، وويكاموس العربي)، وقد تبيّنَ أنّ المنظومةَ تسترجعُ المرادفاتِ الوظيفيةَ الحقيقيةَ استرجاعًا مباشرًا وموثوقًا. ولجعلِ هذه المقارنة كميّةً، اعتُمِدَ تقييمٌ موزونٌ للمرادفات ذو أساسٍ معجميّ (المقياسان WSS@k و wDCG@k المعرَّفان في الفصل الثاني)، يمنحُ كلَّ مرشَّحٍ مُسترجَعٍ درجةَ قربٍ على مقياسٍ من 0 إلى 1، ويكافئُ ترتيبَ المرادفات الحقيقية قربَ رأس القائمة.

والنتائجُ واضحة. فبالنسبة لكلمة «تعليم» (Education)، رتّبت الطريقةُ الأساسُ ضمن مخرجاتها الخمس عشرة الأولى مصطلحاتٍ مثل «أستاذ» (Professor) و«طلابي» (Student-related)، وليس أيٌّ منهما مرادفًا وظيفيًّا. أمّا منظومةُ SVD-SOM المقترحة فقد رتّبت «تدريس» (Teaching) و«تثقيف» (Enlightenment) في المرتبتين الأوليين، بما يطابقُ تمامًا مكنز مايكروسوفت وورد العربي. وقد تكرّرَ هذا النمطُ باطّرادٍ عبر دراسات الحالة المعروضة في الفصل الخامس. كما أظهرَ تحقُّقٌ متقاطعٌ كميٌّ باستخدام نموذج Sentence-BERT المستقلّ معماريًّا، عبر جميع السيناريوهات التجريبية البالغة 2,512 سيناريو، تحسُّنًا متواضعًا لكنْ متّسقًا لمنظومة SVD-SOM على الطريقة الأساس.

ينبغي أن يُعطي العملُ المستقبليُّ الأولويةَ لبناء مرجعٍ قياسيٍّ متاحٍ للعموم للمرادفات العربية، وتوسيعِ منظومة SVD-SOM لتشملَ اللهجاتِ العربيةَ وتطبيقاتٍ لاحقةً مثل استرجاع المعلومات.

الكلمات المفتاحية: العربية، معالجة اللغة الطبيعية، SOM، SVD، BERT، المرادفات.

Abstract

Arabic is a morphologically rich language where many words share the same root. This causes a well-documented problem when using automatic synonym extraction, where standard methods based on embeddings and cosine similarity – the Baseline – consistently return root-related word forms instead of true functional synonyms. This thesis addresses this problem.

This thesis proposes a three-stage pipeline combining a pre-trained Arabic language model (Bidirectional Encoder Representations from Transformers (BERT)), Singular Value Decomposition (SVD), and the Self-Organising Map (SOM). BERT generates contextual word representations; SVD reduces the embedding dimensionality of any given word to suppress noise; and the SOM reorganises the space so that functionally equivalent words cluster together rather than root-related ones. The SVD-SOM pipeline was trained on approximately 3,800 Arabic articles covering 157 target words from variety of domains including education, economics, law, and media. A key limitation of the SVD-SOM pipeline is the absence of a standard Arabic synonym evaluation benchmark, which makes performance measurement inherently difficult. Evaluation was therefore conducted by comparing the SVD-SOM pipeline output against the Microsoft Word Arabic Thesaurus and several established Arabic online dictionaries (*Almaany*, *Al-Maajim*, *Arabic Wiktionary*), and the system was shown to retrieve true functional synonyms directly and reliably. To make this comparison quantitative, a lexically-grounded weighted synonym evaluation (the WSS@k and wDCG@k measures, defined in Chapter 2) scores each retrieved candidate on a 0–1 closeness scale and rewards ranking true synonyms near the top of the list.

The results are clear. For the word **تعليم** (*Education*), the Baseline ranked terms such

as **أستاذ** (*Professor*) and **طلابي** (*Student-related*) within its top-15 output – neither of which is a functional synonym. The proposed SVD-SOM pipeline ranked **تدريس** (*Teaching*) and **تثقيف** (*Enlightenment*) in the top two positions, exactly matching the Microsoft Word Arabic Thesaurus. This pattern was reproduced consistently across the case studies presented in Chapter 5. A quantitative cross-validation using an architecturally independent Sentence-BERT model across all 2,512 experimental scenarios showed a modest but consistent improvement of the SVD-SOM pipeline over the Baseline.

Future work should prioritise building a publicly available Arabic synonym benchmark and extending the SVD-SOM pipeline to include Arabic dialects and downstream applications such as information retrieval.

Keywords: Arabic, NLP, BERT, SVD, SOM, Synonyms.

DEDICATION

To my ambitious self, who endured every challenge and never gave up, to my patient and supportive supervisor whose guidance lit my path, to my beloved mother, the endless source of my compassion and strength; to my brothers and sisters, pieces of my soul; to my wonderful friends; and to everyone whose support, in both small and great ways, contributed to the successful completion of this work.

Tala Sarabtah

Acknowledgements

First, I would like to thank Allah for giving me the strength and patience to complete this work. I would like to express my sincere gratitude to my supervisor, Dr. Dia AbuZeina, for his ongoing support and valuable guidance. I would also like to thank Dr. Alaa Halawani, and Dr. Mohammed Maree for their helpful feedback and suggestions. My deepest appreciation goes to my mother, family, and friends, whose love, patience, and encouragement were the true source of strength throughout this journey.

Tala Sarabtah

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

NLP Natural Language Processing

SOM Self-Organising Map

SVD Singular Value Decomposition

BMU Best Matching Unit

MSA Modern Standard Arabic

BERT Bidirectional Encoder Representations from Transformers

IR Information Retrieval

RTL right-to-left

S-BERT Sentence-Bidirectional Encoder Representations from Transformers

MT Machine Translation

QA Question Answering

NLU Natural Language Understanding

Chapter 1

Introduction

1.1 Background

A synonym is defined as a word that can be substituted for another in a sentence without changing its meaning [1]. Two expressions are synonymous if the substitution of one for the other does not alter the truth value of any sentence in which the substitution is made [2]. This seemingly simple definition conceals a considerable computational challenge, particularly when applied to Arabic, where the lexicon is organised around a root-pattern morphology that produces large families of structurally related words from a single triliteral root.

1.1.1 Linguistic Properties of Arabic

Arabic is a right-to-left (RTL) language whose writing system omits short vowels in most digital contexts, typically representing them through **تشكيل** (diacritical marks) placed above or below letters. This omission introduces lexical ambiguity: the same written form may correspond to multiple distinct words depending on the context. For example, the forms **عَلَم** (flag) and **عَلِم** (knew) are written identically when diacritics are not used [3]. Beyond this orthographic challenge, Arabic derivational morphology is exceptionally productive. From a single triliteral root, the language generates nouns, verbs, adjectives, and adverbs following established patterns – a property that makes Arabic semantically rich but computationally challenging. **الترادف** (synonymy) in Arabic is correspondingly vast: a single concept may be expressed by dozens of words, each

carrying subtle distinctions of register, intensity, or domain [1]. This richness is both an asset and a source of difficulty for automated extraction systems.

1.1.2 From Statistical to Neural Representations

Traditional statistical approaches to text representation offer a starting point for synonym extraction; However, these approaches. consistently fall short when the goal moves beyond surface lexical matching toward capturing genuine functional equivalence [2]. The emergence of distributional semantics marks a turning point: Rather than treating words as discrete symbols, these distributional models encode them as dense vectors in high-dimensional spaces, where geometric proximity is expected to reflect semantic relatedness [4, 5]. Transformer-based architectures push this further by making representations context-sensitive, so that the same word receives a different vector depending on the sentence that surrounds it. BERT, when trained on large Arabic corpora [6], brought this capability to Arabic and demonstrated measurable improvements across a range of Natural Language Processing (NLP) tasks.

Yet despite the representational sophistication provided by these models, synonym retrieval continues to rely on a fundamentally simple mechanism: linear distance in vector space, most commonly cosine similarity. This reliance reveals an important gap. Linear distance metrics capture statistical proximity between terms but are blind to the distinction between a word's morphological derivatives and its true semantic equivalents [7]. The result is a synonym list populated not by functionally interchangeable terms, but by inflected forms and derivational relatives that share a root yet serve entirely different communicative purposes.

1.1.3 The Role of Singular Value Decomposition and Self-Organizing Maps

This research argues that the problem lies not in the quality of the embeddings themselves – the dense numerical vectors that represent words in a high-dimensional space such that semantically related words appear close to one another – but in what happens after they are generated. This work propose that reorganising the vector space topologically,

using the SOM, offers a principled solution [8]. To bridge the gap between the raw high-dimensional BERT space and the SOM, SVD is employed to reduce dimensionality. SVD compresses the embedding space by filtering out the low-variance noise that corresponds to morphological surface similarity, thus retaining only the dominant semantic dimensions [9]. This refined, latent representation is then passed to the SOM, which organises it into a topology-preserving grid where functionally equivalent terms converge into shared neighbourhoods, regardless of their morphological form [10]. This combined pipeline – BERT, SVD, and SOM – is developed in this thesis.

1.2 Motivation

1.2.1 The Morphological Noise Problem

The practical gap that motivates this research is the *morphological noise problem*: when standard cosine similarity is applied to Arabic embeddings, the top-ranked candidates are dominated by words that share the same root as the target (such as inflected forms and derivatives) rather than its true functional synonyms (different roots, same meaning). This becomes clear when one examines the output of state-of-the-art Arabic embedding models in synonym retrieval tasks. Consider the word **تَعليم** (Education). Table 1.1 shows the words that a standard cosine similarity system retrieves, compared with the words' true functional synonyms.

Table 1.1: Same-Root Forms vs True Synonyms of **تَعليم** (Education).

Same Root (عَـ ـلِـ ـمَ) (*'-l-m*) – Not Synonyms		**Different Roots – True Synonyms**	
Arabic	**Meaning**	**Arabic**	**Meaning**
علم	Knowledge / Science	**تَدْرِيس**	Teaching
مُعَلِّم	Teacher	**تَثْقِيف**	Enlightenment
يُعَلِّم	He teaches	**تَرْبِيَة**	Upbringing
مُعَلِّمُون	Teachers (plural)	**تَوْجِيه**	Guidance
تَعْلِيمِي	Educational (adj.)	**تَلْقِين**	Instruction

The words on the left are not synonyms of **تَعليم** (Education). They are morphological

relatives – products of the same root (عَ ـ لِ ـ مَ) (*'-l-m*) that the model has learned to place in geometric proximity. The actual functional synonyms, shown on the right, come from entirely different roots and are systematically displaced to lower positions in the ranking. This is the core problem this thesis addresses.

1.2.2 The Richness of Arabic Synonymy

To appreciate why an intelligent synonym extraction tool is necessary, consider the verb عَلِمَ (to know, to perceive, to understand). As illustrated in Table 1.2, this single Arabic verb has more than 60 documented synonyms in classical and modern usage, each carrying a distinct shade of meaning depending on context and domain [1]. Manual dictionary construction at this scale is time-consuming and expensive. Automated tools that can identify these synonyms accurately and contextually are therefore of significant practical value.

Table 1.2: Selected Synonyms of the Verb عَلِمَ (to know).

أَدْرَك	فَهِم	عَرَف	وَعَى	اِسْتَوْعَب
اِنْتَبَه	نَمَى	وَقَف	فَطِن	اِنْتَهَى
تَنَاهَى	بَصُر	خَال	تَوَسَّم	وَجَد
حَزَر	عَقَل	تَصَوَّر	حَذَق	زَعَم
رَاب	تَنَسَّم	اِسْتَطْلَع	اِسْتَبَان	كَشَف
خَرَج	بَال	مَارَس	فَقِه	دَرَى
ثَقِف	خَبِر	أَيْقَن	تَبَيَّن	اِطَّلَع
سَبَر	مَيَّز	شَخَّص	اِسْتَنْبَط	تَحَرَّى
تَحَقَّق	لاحَظ	اِسْتَقْصَى	فَتَّش	تَعَقَّب
اِبْتَلَى	تَخَيَّل	فَلَى	فَحَص	دَرَس
أَحَسَّ	نَبَش	تَحَسَّس	–	–

1.2.3 Applications of Synonym Extraction

Improving Arabic synonym extraction has direct benefits for a wide range of NLP applications. In Machine Translation (MT), synonym-aware systems produce more natural

and contextually appropriate target-language output. In Information Retrieval (IR) and search engines, query expansion using synonyms significantly increases recall by matching documents that use different but equivalent terminology. In dialogue and Question Answering (QA) systems, synonym knowledge enables systems to recognise paraphrased user inputs and generate more varied, natural responses. In text summarisation and data augmentation, synonyms allow the generation of semantically equivalent alternatives that reduce redundancy. Each of these applications depends on the ability to distinguish true functional synonyms from morphological relatives – precisely the distinction that the SVD-SOM pipeline proposed in this thesis is designed to make [11].

1.2.4 The Proposed Pipeline

The motivation behind the three-stage pipeline, illustrated in Figure 1.1 emerges directly from the diagnosis outlined above. BERT provides rich contextual embeddings, which SVD then refines by concentrating the semantic signal and suppressing morphological noise. The SOM reorganises the refined space into a *topological semantic grid* where functionally equivalent terms converge into shared neighbourhoods, illustrated by the green regions in the figure, regardless of their morphological form [8, 15].

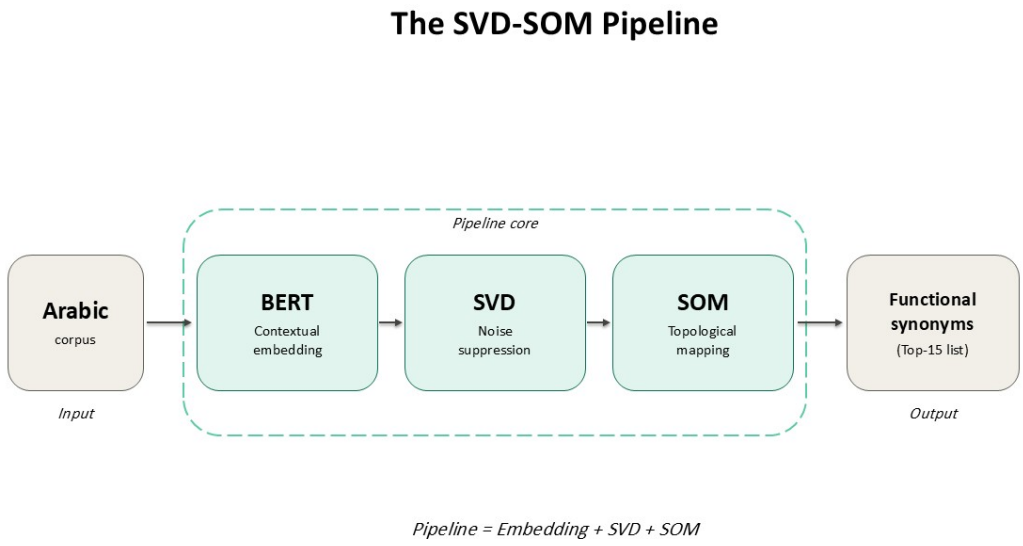

Figure 1.1: The Proposed Pipeline.

1.3 Problem Statement

Despite the significant progress that context-aware models like BERT [12] have brought to Arabic NLP, a well-defined problem remains unresolved: the inability of linear distance metrics to distinguish between functional synonyms and morphological neighbours in high-dimensional Arabic embedding spaces.

The distinction matters because the two categories hold fundamentally different linguistic roles. A functional synonym is a word that can replace the target term in a sentence without altering its meaning [1]; a morphological neighbour shares structural features with the target word, but serves a different grammatical or semantic function. Current vector-based retrieval systems, including those built on Transformer architectures [13], do not separate the two categories, because both tend to appear in similar distributional contexts in training corpora. The outcome is a systematic degradation of precision in synonym extraction, particularly in specialised domains where the exact functional meaning of a term is critical.

The research gap addressed in this thesis therefore lies not in the generation of word representations, but in the transition from those representations to meaningful semantic ranking. Table 1.3 illustrates this gap across three representative domains.

Table 1.3: Semantic Gap Across Three Specialised Domains.

Domain	**Target Word**	**Morphological Noise**	**Functional Synonym**
Education	**تعليم** (Education)	**تعليمي** (Educational, adj.)	**تدريس** (Teaching)
Finance	**قرض** (Loan)	**اقتراض** (Borrowing, action)	**دين** (Debt)
Law	**محكمة** (Court)	**محاكمة** (Trial, action)	**قضاء** (Judiciary)

In each case, the term ascribed to morphological noise shares a root or derivational pattern with the target word and receives a high cosine similarity score accordingly. The functional synonym, however, is the term a domain expert would actually endorse as a semantic equivalent. Closing this gap requires a retrieval mechanism that is sensitive to a word's functional role rather than its structural proximity, which is precisely what the SOM topological organisation is designed to provide [14].

1.4 Research Objectives

The central aim of this research is to develop a robust pipeline that enhances Arabic synonym extraction by reorganising the semantic vector space along functional rather than morphological lines. The following specific objectives define the scope of this work:

- To analyse the limitations of cosine similarity and related linear distance metrics when applied to Arabic embedding spaces, with particular attention to the morphological bias they introduce in synonym ranking [7].
- To integrate BERT's context-aware representations with SVD dimensionality reduction and SOM topological organisation, producing a structured semantic space that prioritises functional relationships over derivational ones.
- To develop a filtering mechanism capable of separating functional synonyms from morphological noise, thereby reducing the influence of shared root patterns on synonym retrieval precision [14].
- To evaluate the proposed pipeline systematically across a diverse set of Arabic target words – using both standard retrieval metrics and a lexically-grounded weighted synonym measure (WSS@k and wDCG@k), defined in Chapter 2 – examining its capacity to locate latent synonyms that linear retrieval methods fail to identify.
- To investigate the effect of the SVD dimensionality parameter as an intermediate step between embedding generation and topological mapping, and to determine how its choices influence synonym extraction quality.

1.5 Scope of Research

The boundaries of this research are defined as follows:

- **Linguistic Scope:** The study focuses exclusively on Modern Standard Arabic (MSA), targeting specialised vocabulary in domains such as education, economics, and law, where functional precision in synonym retrieval is most consequential. Dialectal Arabic and historical etymology fall outside the scope of this work.
- **Technical Components:** The pipeline is built on three components: BERT [6] for contextual embedding generation, SVD for dimensionality reduction, and SOM for topological reorganisation. Other embedding models are referenced only for comparative context.
- **Methodological Focus:** This research is concerned specifically with the reorganisation of the vector space and the suppression of morphological noise in synonym ranking. It does not address MT, text generation, or full-scale lexical database construction.
- **Data Domain:** Experiments are conducted on a corpus of approximately 3,800 Arabic articles derived from the dataset of Al-Anzi and AbuZeina [17], covering diverse domains.

1.6 Thesis Contributions

The contributions of this research can be summarised as follows:

- **A pipeline combining BERT, SVD, and SOM for Arabic synonym extraction:** This thesis introduces a three-stage pipeline that moves from contextual embedding, through dimensionality reduction, to topological organisation. This combination is, to the best of the author's knowledge, the first to apply SOM topological mapping to BERT representations for the specific purpose of Arabic synonym extraction [8].

- **A principled approach to morphological noise suppression:** The pipeline provides a replicable methodology for filtering derivational and inflectional relatives from synonym candidate lists. In doing so, it addresses a well-documented bottleneck in Arabic NLP that affects downstream applications including MT, IR, and dialogue systems [14].

- **Empirical evidence for latent synonym discovery:** Through systematic case studies, this work demonstrates that topological organisation recovers conceptually equivalent terms that cosine similarity consistently fails to rank highly [10].

- **Analysis of the SVD tuning parameter:** The thesis provides a detailed analysis of how the choice of reduced dimensionality affects the granularity and domain-specificity of the extracted synonyms, offering practical guidance for applying this pipeline to other Arabic NLP tasks [16].

- **Adaptation of an existing Arabic corpus for synonym extraction:** In the absence of a publicly available benchmark for Arabic synonym extraction, this research adapts the Arabic text classification dataset of Al-Anzi and AbuZeina [17]. A subset of 3,800 articles spanning education, economics, law, media, and social affairs was selected to serve as a testbed for synonym extraction. The adapted corpus comprises 2,115,819 tokens and 140,825 unique word types after preprocessing, and supports a set of 157 target words covering the principal semantic fields of Modern Standard Arabic.

1.7 Thesis Organisation

The remainder of this thesis is structured as follows:

Chapter 2 provides the theoretical background that situates this research. It outlines the architecture of BERT and how it generates contextual representations, the mathematical foundations of SVD and its role in dimensionality reduction, and the principles of SOM competitive learning.

Chapter 3 reviews the existing literature on word embedding models, Arabic NLP, and unsupervised clustering methods. It identifies the gap in the literature that this thesis addresses.

Chapter 4 presents the proposed pipeline in full detail, describing the preprocessing steps, the integration of the three stages, the SOM configuration, and the synonym candidate selection criteria.

Chapter 5 reports the experimental results across 157 target words and 16 values of the SVD tuning parameter. Results are presented through three analytical lenses: rank elevation of known synonyms, discovery of latent synonyms, and displacement of morphological noise.

Chapter 6 concludes the thesis by summarising the key findings, discussing their implications for Arabic NLP research and lexical resource development, and proposing directions for future work.

Chapter 2

Theoretical Background

2.1 Introduction

The development of a robust pipeline for Arabic synonym extraction requires a solid understanding of three converging areas: neural language models and word embedding architectures, unsupervised topological learning, and dimensionality reduction. As established in Chapter 1, the structural complexity of Arabic presents unique challenges that traditional linear retrieval models fail to address.

This chapter provides the theoretical foundation necessary to support an understanding of the SVD-SOM proposed pipeline. It begins by examining the evolution of word representations from static to contextualised embeddings, focusing on the BERT architecture and its application to Arabic. It then discusses cosine similarity, the baseline retrieval metric and Sentence-Bidirectional Encoder Representations from Transformers (S-BERT) as an independent validation tool. The mathematical principles of the SOM are presented next, followed by the role of SVD as a semantic refinement step. The chapter concludes with the Farasa Stemmer, which provides the morphological deduplication mechanism used throughout the pipeline.

2.2 From Static to Contextualised Word Embeddings

The idea of representing words as continuous vectors in a high-dimensional space rests on the *distributional hypothesis*, first articulated by Harris [4] and popularly summarised by Firth [5] as "you shall know a word by the company it keeps". words that occur in similar

linguistic contexts tend to share similar meanings. This hypothesis was operationalised in neural language modelling by Bengio et al. [15], who showed that a neural network trained to predict the next word in a sequence implicitly learns dense word representations that capture semantic regularities.

Mikolov et al. [18, 19] extended this idea at scale with Word2Vec, introducing two efficient architectures: the skip-gram model, which predicts context words from a pivot word, and the continuous bag-of-words model, which predicts a pivot word from its context. Word2Vec produces *static* embeddings, where each word type receives a single fixed vector regardless of the sentence in which it appears. While efficient and widely adopted, static embeddings cannot resolve lexical ambiguity or capture context-dependent meaning shifts – a significant limitation for Arabic, where the same written form may carry multiple distinct meanings depending on its syntactic environment.

2.2.1 Bidirectional Transformer Encoders

The Transformer architecture, introduced by Vaswani et al. [13], replaced the sequential processing of recurrent networks with a fully attention-based mechanism. A multi-head self-attention layer computes, for each token, a weighted sum of all other tokens in the sequence, where weights are determined by learned compatibility functions. This allows the model to capture long-range dependencies and contextual relationships without the vanishing gradient problem of recurrent networks.

Devlin et al. [12] built on this architecture to introduce BERT. Unlike the unidirectional language models that preceded it, BERT is pre-trained bidirectionally, reading the entire input sequence at once, allowing each token's representation to be conditioned on both its left and right context simultaneously. Pre-training is performed on two tasks: Masked Language Modelling, in which a random subset of tokens is masked and the model must predict them from context, and Next Sentence Prediction. The resulting representations are then fine-tuned on downstream tasks. BERT's contextualised embeddings – where the same word receives a different vector depending on its surrounding context – represent a fundamental advance over static embeddings for tasks requiring semantic

precision [12].

In this research, the Arabic-specific variant of BERT is used [6]. It is pre-trained on large Arabic text corpora and employs a specialised subword tokenisation pipeline (Farasa [20] or WordPiece) that decomposes Arabic words into morphologically meaningful units, mitigating the data sparsity caused by the language's high vocabulary growth. Each word in the corpus is encoded as the mean of the last hidden states of its constituent subword tokens, producing a 768-dimensional embedding vector that serves as the input to the subsequent pipeline stages.

Figure 2.1 illustrates the Transformer encoder architecture and the multi-head attention mechanism that underpins BERT's contextual representations.

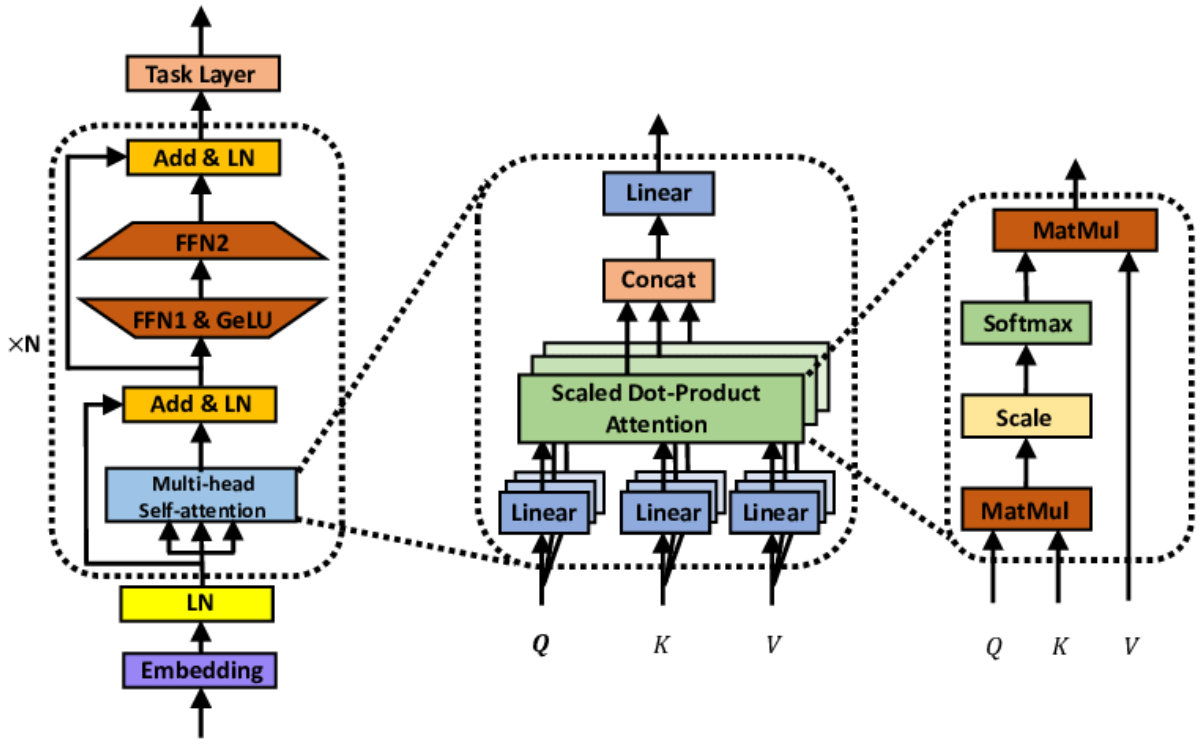

Figure 2.1: Transformer Encoder and Multi-Head Attention [13, 12].

2.2.2 Morphological Bias of Arabic Embeddings

Despite the representational power of contextualised embeddings, the vector space they produce for Arabic exhibits a systematic morphological bias. Because Arabic derivational morphology generates large families of words from a single triliteral root, models

trained on co-occurrence statistics – including BERT-based models – place root-related word forms in close geometric proximity regardless of their functional equivalence. Al-Twairesh [7] documented this behaviour across multiple Arabic embedding models, showing that standard cosine similarity retrieval consistently returns morphological relatives before genuine synonyms. Figure 2.2 illustrates how these dense root-based clusters form in the embedding space.

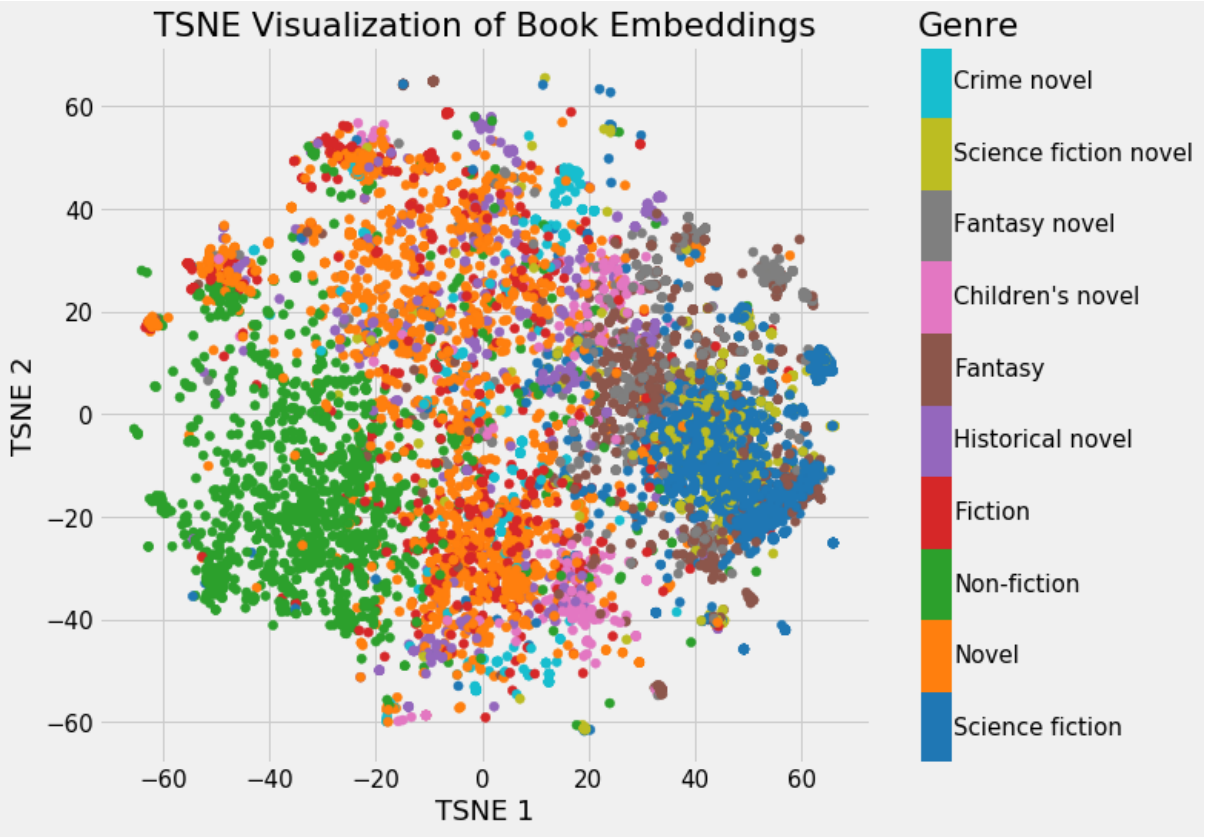

Figure 2.2: Root-Based Clustering vs Functional Synonym Displacement [7].

2.3 Cosine Similarity as the Baseline Metric

Cosine similarity is the standard metric for measuring the relatedness of two word vectors in embedding-based retrieval tasks [18, 21]. Given two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, their cosine similarity is defined as:

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|} \tag{2.1}$$

The measure is insensitive to vector magnitude and focuses entirely on angular separa-

tion, making it appropriate for comparing normalised dense embeddings. It ranges from -1 to 1, where 1 indicates identical orientation. In this thesis, cosine similarity applied directly to BERT embeddings constitutes the **Baseline**: for each target word, the Top-15 candidates are those vocabulary words with the highest cosine similarity scores. As established in Chapter 1 and demonstrated in Chapter 5, this baseline is systematically biased toward morphological relatives rather than functional synonyms, constituting the core limitation that the proposed pipeline addresses [7].

2.4 Cross-Validation Using Sentence Embeddings

S-BERT, introduced by Reimers and Gurevych [22], extends the BERT architecture by fine-tuning it on sentence-pair similarity tasks using a Siamese network structure. The result is a model that produces fixed-size sentence-level embeddings optimised directly for semantic similarity comparison, without the computational overhead of pairwise BERT inference.

In this thesis, S-BERT serves exclusively as an independent external evaluator, not as a component of the synonym extraction pipeline itself. After the pipeline produces its Top-15 candidate list, S-BERT computes the mean semantic similarity between the target word and each candidate. Comparing this mean across the Baseline and the proposed pipeline provides a model-agnostic, quantitative measure of improvement that avoids the circularity of evaluating BERT-based outputs using BERT itself [22].

2.5 Weighted Synonym Evaluation

Beyond cosine similarity (the Baseline retrieval metric) and S-BERT cross-validation, this thesis introduces a lexically-grounded measure of synonym quality that is used throughout the results chapter. For each target word an independent reference set of true synonyms is compiled from standard Arabic dictionaries, and every retrieved candidate is assigned a closeness weight $w \in [0, 1]$: a weight of 1 denotes a fully substitutable synonym, intermediate values denote near- or partial synonyms, and 0 is assigned to non-synonyms and to morphological derivatives of the target itself. Two complementary scores are then

computed over the ranked candidate list:

$$\mathrm{WSS@}k = \frac{1}{k}\sum_{i=1}^{k} w(c_i), \qquad \mathrm{wDCG@}k = \sum_{i=1}^{k} \frac{w(c_i)}{\log_2(i+1)}.$$

The Weighted Synonym Score WSS@k averages the closeness weights of the top-k candidates and is insensitive to their order, whereas the weighted Discounted Cumulative Gain wDCG@k, adapted from the Discounted Cumulative Gain measure used in information retrieval, rewards placing stronger synonyms nearer the top of the list. Because the central claim of the pipeline is the elevation of true synonyms to higher ranks, the rank-aware wDCG@k is treated as the primary measure. This instrument complements the standard retrieval metrics – Precision@k, Mean Reciprocal Rank (MRR), and Mean Average Precision (MAP) – reported in the parameter and ablation studies of Chapter 5.

2.6 Self-Organizing Maps

The SOM was introduced by Kohonen [23, 24] as an unsupervised neural network for topology-preserving dimensionality reduction; a comprehensive treatment is provided in his later monograph [8]. Unlike feed-forward networks that minimise a prediction error through backpropagation, the SOM employs a competitive learning rule that projects high-dimensional input data onto a low-dimensional – typically two-dimensional – discretised lattice while preserving the topological structure of the input distribution.

2.6.1 Architecture

The SOM consists of an input layer and an output layer (the feature map). The input layer receives the k-dimensional reduced word representations produced by the SVD stage. The output layer is a grid of neurons, each holding a weight vector of dimension k. In this research a 55 grid of 25 neurons is used.

2.6.2 Competitive Learning

The learning process is iterated across three stages for each input vector $\mathbf{x}$:

1. **Competition:** The Best Matching Unit (BMU) is the neuron whose weight vector $\mathbf{w}_i$ is closest to $\mathbf{x}$:

$$c = \arg\min_i \{\|\mathbf{x} - \mathbf{w}_i\|\} \tag{2.2}$$

2. **Cooperation:** A Gaussian neighbourhood function $h_{ci}(t)$ centred on the BMU determines which neurons are updated and by how much, shrinking over training time:

$$h_{ci}(t) = \exp\left(-\frac{d_{ci}^2}{2\sigma^2(t)}\right) \tag{2.3}$$

3. **Adaptation:** The BMU and its neighbours move closer to the input vector:

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \alpha(t)\, h_{ci}(t)\, [\mathbf{x}(t) - \mathbf{w}_i(t)] \tag{2.4}$$

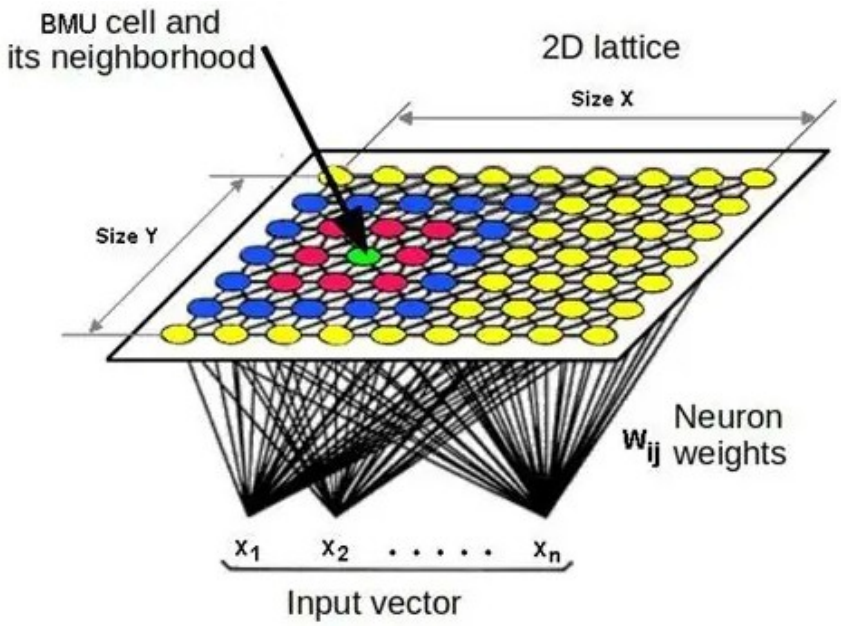

Figure 2.3: Self-Organizing Map Architecture and Best-Matching-Unit Update [23].

2.6.3 Topological Preservation

The key property of the SOM is topology preservation: words that are similar in the input space are mapped to nearby neurons on the grid [23]. Vesanto and Alhoniemi [10] showed that clustering the SOM output produces semantically coherent groups that reflect genuine distributional similarity. In the context of this research, topology preservation means that functional synonyms – which the SVD stage has brought into closer proximity – converge into shared grid neighbourhoods, while morphological relatives that dominated the original cosine similarity ranking are dispersed across the map. Table 2.1 contrasts the two retrieval paradigms.

Table 2.1: Cosine Similarity vs Topological Mapping.

Feature	**Cosine (Baseline)**	**Topological (Proposed)**
Primary driver	Statistical co-occurrence	Functional / semantic role
Noise handling	Sensitive to root patterns	Filters morphological noise
Relationship type	Direct / linear	Global / non-linear
Output structure	Unstructured ranking	Semantic neighbourhoods

2.7 Dimensionality Reduction via Singular Value Decomposition

In spaces of hundreds of dimensions, such as the 768-dimensional BERT embedding space, the distance between any two data points tends toward uniformity – a well-known effect called the curse of dimensionality [25]. This severely impairs the convergence and discriminative power of distance-based clustering algorithms such as a SOM. Dimensionality reduction is therefore a necessary intermediate step as it concentrates the semantic signal into a compact subspace and filters out the low-variance noise associated with morphological surface similarity.

SVD was first applied to language representation by Deerwester et al. [9] in Latent Semantic Analysis, where it revealed that the dominant singular components of a term-document matrix capture genuine semantic structures rather than surface co-occurrence patterns. The same principle is applied here.

In this pipeline, SVD is applied not to the raw embedding matrix but to a **local cosine similarity matrix** constructed from the BERT embeddings of the target word and its Top-200 cosine similarity candidates. This local matrix $C \in \mathbb{R}^{201201}$ captures the pairwise semantic structure of the candidate neighbourhood. Its SVD decomposition is:

$$C = U\Sigma V^T \tag{2.5}$$

Where U contains the left singular vectors (mapping words to latent concept dimensions), Σ is the diagonal matrix of singular values in descending order (each representing the variance explained by one latent dimension), and V^T contains the right singular vectors [25]. A truncated projection is then formed by retaining the top k singular components:

$$X_{\text{red}} = U_k \cdot \Sigma_k \quad \in \mathbb{R}^{201k} \tag{2.6}$$

This operation filters out the low-variance dimensions – which correspond to morphological noise – and produces a refined latent representation that the SOM can organise more effectively [9, 25].

Figures 2.4 and 2.5 illustrate the full decomposition and the truncation step respectively.

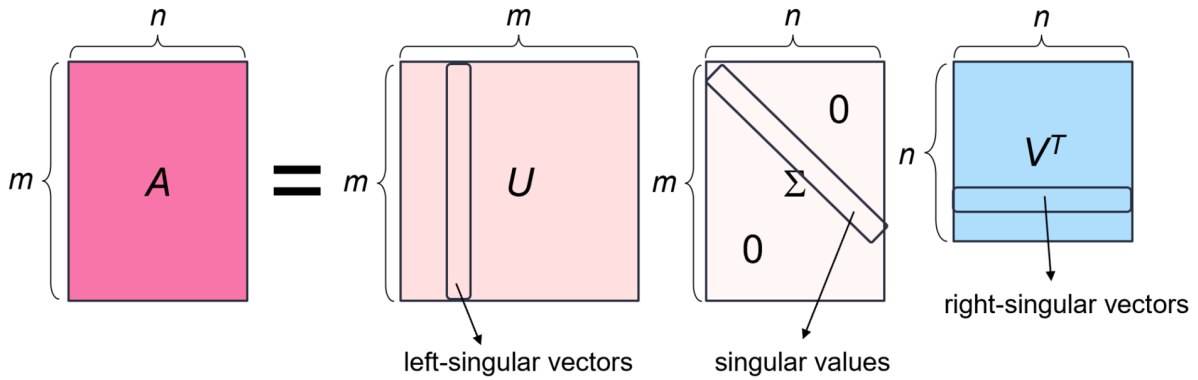

Figure 2.4: Singular Value Decomposition into Components U, Σ, and V^T [25].

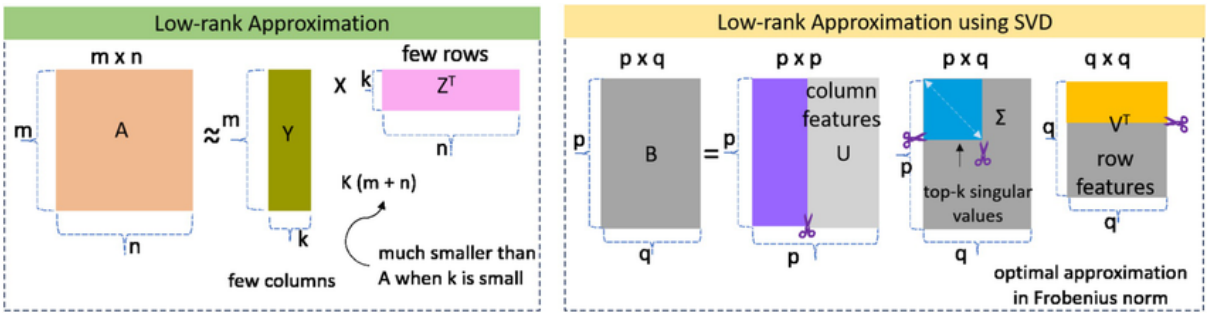

Figure 2.5: Truncated SVD: Signal Concentration and Noise Suppression [9].

Algorithm 1 describes the SVD refinement step as implemented in this pipeline. It begins from the corpus and follows the full data flow through preprocessing, embedding, candidate selection, and dimensionality reduction.

Algorithm 1 SVD-Based Semantic Refinement (From Corpus to Refined Vectors)

1: **Input:** Arabic corpus $\mathcal{C}$, target word t, truncation rank k

2: **Preprocessing:** Clean $\mathcal{C}$ (remove non-Arabic characters), remove Arabic stopwords, tokenise, and build vocabulary $\mathcal{V}$ (minimum word length 3; minimum document frequency 2)

3: **Contextual Encoding:** For each $w \in \mathcal{V}$, compute the BERT embedding $\mathbf{e}_w \in \mathbb{R}^{768}$ as the mean of the last-layer hidden states of the subword tokens of w

4: **Candidate Selection:** $\mathcal{W} \leftarrow$ Top-200 words in $\mathcal{V}$ ranked by $\cos(\mathbf{e}_t, \mathbf{e}_w)$

5: **Build local matrix:** $X \leftarrow$ stack embeddings of $\{t\} \cup \mathcal{W}$ ($X \in \mathbb{R}^{201768}$)

6: **Compute cosine similarity matrix:** $C \leftarrow \texttt{cosine\_similarity}(X)$ ($C \in \mathbb{R}^{201201}$)

7: **Apply SVD:** $U, \Sigma, \_ \leftarrow \texttt{svd}(C,\ \texttt{full\_matrices=False})$

8: **Project to rank-k subspace:** $X_{\text{red}} \leftarrow U[:, :k] \cdot \texttt{diag}(\Sigma[:k])$ ($X_{\text{red}} \in \mathbb{R}^{201k}$)

9: **Output:** Reduced representation dictionary $\{\, w \mapsto X_{\text{red}}[i] \,\}$

The parameter k is the primary tuning variable of the pipeline, controlling the semantic resolution of the reduced space. In this research, k was varied systematically across 16 values ($k \in \{10, 12, 14, \ldots, 40\}$) to analyse its effect on synonym extraction quality, as reported in Chapter 5.

2.8 Farasa Stemmer for Morphological Deduplication

The Farasa toolkit, introduced by Abdelali et al. [20], provides fast and accurate morphological analysis for Modern Standard Arabic, including stemming, part-of-speech tagging, and segmentation. Unlike lightweight suffix-stripping stemmers, Farasa performs full morphological decomposition, reducing each word to its stem by referencing a model trained on large Arabic corpora [20].

In this pipeline, Farasa is used for **root-based deduplication** where, after any ranking step produces a candidate list, each candidate is reduced to its Farasa stem. A seen-stems set is maintained; when a candidate's stem has already been seen, that candidate is skipped. Only the highest-ranked representative of each morphological family is retained. This mechanism prevents root-related word forms from consuming multiple positions in the Top-15 output, ensuring that each output position represents a distinct morphological family rather than a variant of a word already included [11]. Deduplication is applied at two points in the pipeline: after Stage 1 cosine ranking and after Stage 3 SOM neighbourhood expansion.

Chapter 3

Literature Review

3.1 Introduction

Arabic synonym extraction has been investigated through two main avenues: manual approaches that depend on human-curated lexicographic resources, and computerised approaches that infer synonymy from corpora. The computerised approach can be subdivided. The first subdivision contains traditional neural network and statistical models that operate on sparse co-occurrence representations. The second subdivision contains modern embedding-based models that represent words as dense vectors learned from large corpora.

This chapter reviews the prior work into these manual and computerised approaches. For each approach, or its subdivision, the discussion follows the same pattern: first a short narrative description is given, the most important contributions are then listed, and a summary table groups the works that share the same methodology. The chapter closes with a consolidated performance comparison and a clear statement of the research gap addressed by this thesis.

Two terms are used throughout this thesis and the following chapters. The **Baseline** of this thesis refers to the configuration *Embedding + Cosine*, i.e. retrieving synonym candidates from a pre-trained contextual embedding space using cosine similarity as the ranking criterion. The **Proposed Pipeline** refers to the configuration *Embedding + SVD + SOM*, in which the embedding space is first refined by Singular Value Decomposition and then reorganised topologically by the SOM. To the best of the author's knowledge,

this combination has not been applied to Arabic synonym extraction in any prior study reviewed in this chapter.

3.2 Manual Approaches

The oldest and most authoritative source of Arabic synonyms is the human-curated lexicographic record. Classical dictionaries such as *Lisān al-ʻArab* and modern thesauri such as *Almaʻany* list synonyms as validated entries assembled by linguists, and they remain the gold standard against which automatic systems are evaluated.

A substantial line of research has tried to convert this lexicographic knowledge into a structured semantic resource modelled on the Princeton WordNet. Elkateb et al. [26] initiated the Arabic WordNet project and produced its first version (AWN v1) by translating Common Base Concepts from Princeton WordNet into Modern Standard Arabic, yielding 9,618 synsets. Regragui et al. [27] extended this resource semi-automatically into AWN v2 with 11,269 synsets. AlMaayah et al. [28] addressed the same problem for the Quranic vocabulary by combining traditional Arabic dictionaries with a vector space model and cosine similarity computed over textual definitions, producing the Quranic Arabic WordNet (QAWN). Habash [11] surveyed the broader Arabic NLP resources and described how bilingual translation graphs can be used to project English synsets onto Arabic.

The common limitation of all manual approaches is scale. Even AWN v2, the largest semi-automatically constructed resource, covers only a small fraction of the vocabulary observed in any modern corpus, and extending the resource requires expert linguistic labour. Table 3.1 summarises the manual approaches reviewed in this section.

Table 3.1: Summary of Manual Approaches to Arabic Synonym Resources.

References	Approach	Output
Elkateb et al. [26], Regragui et al. [27]	Translation from Princeton WordNet (expand model)	Arabic WordNet (AWN v1, v2)
AlMaayah et al. [28]	Traditional dictionaries + VSM with cosine similarity	Quranic Arabic WordNet (QAWN)
Habash [11]	Bilingual translation graphs	Arabic synsets via English pivot

3.3 Computerised Approaches: Traditional Models

Before the rise of dense word embeddings, automatic synonym extraction relied on sparse statistical representations and on traditional neural networks operating on hand-crafted features. The most influential contribution of this family is Latent Semantic Analysis [9], which applies Singular Value Decomposition to a term–document matrix to expose latent semantic dimensions and to map semantically related words to nearby points in the reduced space. The mathematical machinery of SVD itself is treated in depth by Golub and Van Loan [25], and it underpins every dimensionality-reduction step used in this thesis.

Traditional neural networks have also been applied to Arabic semantic tasks, although rarely to synonym extraction directly. Kohonen [8] introduced the Self-Organising Map (SOM) as a competitive-learning architecture that reduces a high-dimensional input distribution to a low-dimensional topological map while preserving neighbourhood relations. Vesanto and Alhoniemi [10] showed that the resulting map can be further partitioned by classical clustering algorithms to produce interpretable groups. Aouadi and Kacem-Echi [29] applied the SOM to Arabic word spotting in handwritten text. None of these works

targets functional synonymy in Modern Standard Arabic, and the SOM has not previously been combined with refined embedding spaces for this task.

Table 3.2 groups the works of this approach according to methodology. The contribution of this approach, for the purposes of this thesis, is two-fold: SVD is validated as a useful tool for semantic refinement of Arabic representations, and the SOM is validated as a principled unsupervised architecture for topological reorganisation.

Table 3.2: Summary of Traditional Neural Network and Statistical Approaches.

References	**Method**	**Domain**
Deerwester et al. [9], Golub and Van Loan [25]	LSA / SVD on term–document matrix	Latent semantic representation
Kohonen [8], Vesanto and Alhoniemi [10]	Self-Organising Map + clustering	General topology learning
Aouadi and Kacem-Echi [29]	SOM for Arabic word segmentation	Arabic handwritten text

3.4 Computerised Approaches: Modern Embedding-based Methods

Modern computerised approaches replace sparse representations with dense vectors learned from large corpora. Within this family, two generations must be distinguished. The first generation produces a single fixed vector per word (*static embeddings*). The second generation produces a different vector for each occurrence of a word in context (*contextualised embeddings*).

3.4.1 Static embedding methods

Mikolov et al. [18] introduced Word2Vec (Skip-gram and CBOW) and established dense distributed representations as the standard tool for semantic tasks, including synonym retrieval. Bengio et al. [15] provided the broader theoretical framework for distributed

representation learning. For Arabic specifically, Zahran et al. [30] trained Arabic word embeddings on a large corpus and combined them with SVD-based dimensionality reduction, reporting competitive performance on Arabic similarity tasks. This is the closest prior use of the SVD-refinement component, as it is features within the proposed pipeline. Mu and Viswanath [31] proposed the *All-but-the-Top* post-processing technique, which removes dominant SVD directions to render embedding spaces more isotropic and improves cosine-based retrieval.

The most direct prior work on Arabic synonym extraction from static embeddings is Al-Matham et al. [32], who introduced the *SynoExtractor* pipeline. Their system trains Word2Vec on the KSUCCA and Gigaword corpora and then filters cosine-ranked candidates with a sequence of linguistic rules. SynoExtractor reaches 60% precision on KSUCCA and 74% precision on Gigaword, but its filters are language-specific heuristics and cannot resolve morphological clustering in the underlying embedding space.

3.4.2 Contextualised embedding methods

The Transformer architecture [13] and BERT [12] introduced contextualised word representations, in which each occurrence of a word receives its own vector conditioned on the surrounding sentence. Antoun et al. [6] produced AraBERT, the Arabic-specific variant, and reported state-of-the-art results across Arabic understanding tasks. Applied to synonym extraction, all current BERT-based systems retrieve candidates by ranking the contextual embeddings of the vocabulary by cosine similarity to the target word. This linear retrieval criterion is the same one used by static-embedding systems and inherits the same morphological-clustering bias documented by Al-Matham et al. [32].

Baseline of this thesis. Arabic BERT embeddings combined with cosine similarity constitute the Baseline against which the Proposed Pipeline is compared in Chapter 5. The Baseline captures context but does not address the morphological-clustering problem that the SVD and SOM stages are designed to correct.

Table 3.3 summarises the embedding-based approaches.

Table 3.3: Summary of Modern Embedding-Based Approaches.

References	**Embedding**	**Retrieval / Refinement**
Mikolov et al. [18], Bengio et al. [15]	Word2Vec (static)	Cosine similarity
Zahran et al. [30]	Arabic Word2Vec + SVD reduction	Cosine similarity
Mu and Viswanath [31]	Word2Vec / GloVe (static)	SVD-based post-processing (All-but-the-Top)
Al-Matham et al. [32]	Word2Vec (KSUCCA, Gigaword)	Cosine + linguistic-rule filters (SynoExtractor)
Vaswani et al. [13], Devlin et al. [12]	Transformer / BERT (contextual)	Cosine similarity
Antoun et al. [6]	AraBERT (contextual)	Cosine similarity *(Baseline of this thesis)*

3.5 Performance Comparison

Table 3.4 consolidates the reported performance of representative Arabic synonym extraction systems. As Sebastiani [33] notes for text classification, cross-paper comparisons of this kind are not by themselves judgemental, because the corpora, gold standards and evaluation protocols differ. The table is therefore presented as a reference for the reader rather than as a ranking.

Table 3.4: Performance of Representative Arabic Synonym Extraction Systems.

Reference	**Approach**	**Best result**	**Corpus**
AlMaayah et al. [28]	Dictionary + VSM + cosine	34.1% R	Quran
Zahran et al. [30]	Arabic Word2Vec + SVD	Competitive	Arabic Wikipedia + news
Al-Matham et al. [32]	Word2Vec + linguistic filters	0.605 / 0.748 MAP	KSUCCA / Gigaword
Antoun et al. [6]	AraBERT + cosine *(Baseline)*	See Ch. 5	Arabic corpora
This thesis	**Embedding + SVD + SOM (Proposed)**	**See Ch. 5**	**See Ch. 4**

3.6 Summary and Research Gap

The literature reviewed in this chapter can be divided into three groups, each with its own limitation. Manual approaches produce high-quality synsets but do not scale and cannot cover the open vocabulary of any modern corpus. Traditional neural network and statistical approaches established SVD and the SOM as useful tools for semantic representation and topological reorganisation respectively, but they were never combined or applied to functional synonymy in Arabic. Modern embedding-based approaches, including the state-of-the-art AraBERT, retrieve candidates through cosine similarity over the raw embedding space and therefore inherit the morphological-clustering bias documented by Al-Matham et al. [32].

Table 3.5 summarises the five dimensions along which the proposed pipeline differs

from prior work.

Table 3.5: Coverage of Key Requirements Across Approach Families.

Approach family	Arabic	Contextual	Morpho-noise	No manual res.	Corpus-level
Manual (WordNet, dictionaries)	✓				
Traditional NN / statistical	∼		∼	✓	✓
Static embeddings + cosine	✓			✓	✓
Contextual embeddings + cosine *(Baseline)*	✓	✓		✓	✓
Embedding + SVD + SOM (Proposed Pipeline)	✓	✓	✓	✓	✓

Legend: ✓ = addressed, = not addressed, ∼ = partially addressed.

The research gap addressed by this thesis is therefore specific and well-defined. No prior study reviewed in this chapter has applied the combination *Embedding + SVD + SOM* to Arabic synonym extraction. The Baseline against which improvements are measured in Chapter 5 is *Embedding + Cosine*, the direct application of cosine similarity to the contextual embeddings produced by AraBERT.

Chapter 4

The Proposed Pipeline

4.1 Introduction

This chapter presents the proposed Pipeline for automated Arabic synonym extraction. The Pipeline is built around the central finding of Chapter 3: That cosine similarity over the raw embedding space of any current Arabic language model conflates functional synonymy with morphological proximity. The proposed solution refines the embedding space before retrieval across two stages – SVD-based linear refinement and SOM-based topological reorganisation – and uses cosine similarity only as a final ranking step inside this refined space.

Two systems are studied throughout this thesis. The first is the **Baseline**, defined as the direct application of cosine similarity to Arabic BERT embeddings:

$$\text{Baseline} = \text{Embedding} + \text{Cosine} \tag{4.1}$$

The second is the **Proposed Pipeline**, defined as:

$$\text{Pipeline} = \text{Embedding} + \text{SVD} + \text{SOM} \tag{4.2}$$

All improvements reported in Chapter 5 are measured as differences between the Baseline output and the Pipeline output for the same target word and the same corpus.

The remainder of this chapter is organised as follows. Section 4.2 formalises the Baseline and the Pipeline. Section 4.3 describes the corpus and the preprocessing steps.

Section 4.4 presents the rationale behind the selection of the pre-trained Arabic embedding model and compares it with available alternatives. Sections 4.5 to 4.7 describe the three Pipeline stages. Section 4.8 presents the S-BERT cross-validation procedure. Section 4.9 summarises all Pipeline parameters in a single reference table.

4.2 Baseline System and the Proposed Pipeline

4.2.1 Baseline

For each target word t, the Baseline computes the cosine similarity between the embedding $\mathbf{e}_t$ and the embedding $\mathbf{e}_w$ of every other word in the vocabulary:

$$\text{sim}(t, w) = \frac{\mathbf{e}_t \cdot \mathbf{e}_w}{\|\mathbf{e}_t\| \cdot \|\mathbf{e}_w\|} \tag{4.3}$$

The vocabulary is sorted by this score in descending order and the Top-15 entries are returned. No morphological deduplication is applied at the Baseline level, so multiple inflected forms of the same root may co-occur in the output.

4.2.2 Proposed Pipeline

The Pipeline reuses the same embedding $\mathbf{e}_w$ but inserts two intermediate stages. After the Top-200 candidate pool is generated, the first intermediate stage applies truncated SVD to a local cosine similarity matrix across those 200 words and produces a reduced representation $\mathbf{f}_w \in \mathbb{R}^k$. The second intermediate stage then trains the SOM on the reduced vectors and uses the topological neighbourhood of the target's BMU as the candidate set, which is then re-ranked by cosine similarity in the original embedding space. Both the Baseline and the Pipeline return lists of size 15 to make the comparison direct.

4.3 Corpus and Data Preprocessing

The experimental corpus is derived from the Arabic text classification dataset originally compiled by Al-Anzi and AbuZeina [17] from the Kuwaiti newspaper *Alqabas*. For the

purposes of this thesis, a subset of 3,800 articles was selected to cover diverse domains – education, economics, law, media, and social affairs – to ensure broad lexical coverage of the 157 target words used for evaluation. Each article occupies one line in the input file `train_c.txt`.

Preprocessing is applied sequentially:

1. **Character filtering.** All non-Arabic characters are removed using the Unicode range filter [\u0600–\u06FF], which discards numbers, punctuation, Latin characters, and diacritical marks.

2. **Stop word removal.** Arabic stop words are removed using the NLTK stop word list (`stopwords.words("arabic")`). The list is loaded as `AR_STOPWORDS` and applied as a token-level filter so that high-frequency function words do not dominate the embedding vocabulary or the subsequent rankings.

3. **Vocabulary construction.** A vocabulary is built from the remaining tokens using two thresholds: a minimum word length of three characters and a minimum document frequency of two. The length threshold removes single-letter and two-letter tokens that carry little semantic content, and the document frequency threshold removes rare forms that would produce unreliable embeddings.

Algorithm 2 summarises the preprocessing pipeline as implemented.

Algorithm 2 Corpus Preprocessing Pipeline

Input: Raw Arabic corpus
Output: Vocabulary $\mathcal{V}$, term frequency `tf`, document frequency `df`
`AR_STOPWORDS` $\leftarrow$ `set(nltk.stopwords.words("arabic"))`
for each line in corpus **do**
    Remove non-Arabic characters
    Tokenise into words
    Keep t only if $t \notin$ `AR_STOPWORDS`
end for
Compute `tf`$[w]$ and `df`$[w]$ for all w
$\mathcal{V} \leftarrow \{w \mid \texttt{len}(w) > 2 \text{ and } \texttt{df}[w] > 2\}$
return $\mathcal{V}$

Table 4.1: Corpus Statistics After Preprocessing.

Property	**Value**
Number of articles	3,800
Minimum word length	3 characters
Minimum document frequency	2
Number of target words	157
SVD rank values tested	16 values ($k = 10, 12, \ldots, 40$)
Total experimental scenarios	2,512

4.4 Selection and Justification of the Embedding Model

The choice of the pre-trained Arabic BERT model affects the quality of the initial representation from which the subsequent SVD and SOM stages operate. Four publicly available Arabic models were considered as candidates for this work, all of which produce 768- or 1024-dimensional contextual embeddings and have been evaluated using standard Arabic NLP benchmarks. The comparison criteria are: the size and domain of the pre-training corpus, the tokenisation strategy, the number of parameters, and the published

performance on Modern Standard Arabic (MSA) tasks.

Table 4.2 summarises the comparison.

Table 4.2: Comparison of Pre-Trained Arabic BERT Models.

Model	**Pre-training corpus**	**Tokenisation**	**MSA focus**
AraBERT v0.2	77 GB news, Wikipedia, web text (MSA)	Farasa segmentation + WordPiece	Primary target
CAMeLBERT-MSA	Mixed Classical / MSA / Dialect	WordPiece	Shared with other variants
MARBERT	≈ 1 B tweets (mostly dialect)	WordPiece	Limited
Arabic BERT	Arabic Wikipedia	WordPiece	Moderate (small corpus)

The four models were assessed against the requirements of this thesis. AraBERT v0.2 (`aubmindlab/bert-base-arabertv02`) was selected for three reasons:

(i) It is pre-trained on the largest MSA corpus among the candidates, covering the same textual domains (news, Wikipedia, web text) from which the experimental corpus is drawn. Consequently, the embedding space is well-matched to the vocabulary observed at the time of evaluation.

(ii) It uses Farasa-based subword segmentation, which is designed for Arabic morphological structures and produces more linguistically meaningful token splits than generic WordPiece. This is directly relevant to the morphological-clustering problem that the Pipeline is designed to address.

(iii) On the published Arabic Natural Language Understanding (NLU) benchmarks (sentiment analysis, named entity recognition, question answering), AraBERT v0.2 re-

ports the best results for MSA among the four models [6]. MARBERT outperforms it on dialectal tasks but is not aligned with the MSA target of this thesis.

An empirical comparison of the four models on the synonym extraction task itself, using the same 157 target words and the same evaluation protocol, is reported in Chapter 5. The present section outlines the rationale for choosing AraBERT v0.2 over the other models on the basis of design and published benchmarks.

For each word w in the vocabulary, the embedding is computed as the mean of the last hidden states across all subword tokens produced by the selected model:

$$\mathbf{e}_w = \frac{1}{|T_w|} \sum_{i=1}^{|T_w|} h_i^{(L)} \tag{4.4}$$

where T_w is the subword token set and $h_i^{(L)}$ is the hidden state of token i at the final encoder layer. The resulting embedding $\mathbf{e}_w \in \mathbb{R}^{768}$ is precomputed for the full vocabulary before any retrieval operation begins.

4.5 Stage 1: Embedding and Morphological Deduplication

Stage 1 produces two outputs: the Baseline list and the Top-200 candidate pool that feeds Stage 2.

4.5.1 Cosine Similarity Ranking

For each target word t, cosine similarity is computed between $\mathbf{e}_t$ and every vocabulary word. The ranked list is produced in descending order. The Top-200 entries are retained as the input pool for Stage 2.

4.5.2 Output Size

The output list is fixed at 15 candidates. This cut-off is an *arbitrary choice* aligned with the standard Top-K evaluation convention used in synonym extraction and lexical substitution research. It is not the result of a hyperparameter search, and the Pipeline is

not optimised to this value. The same cut-off is applied to the Baseline and to all three Pipeline stages to make the comparison direct and unbiased.

4.5.3 Morphological Deduplication via Farasa

Raw cosine ranking frequently returns multiple inflected or derived forms of the same root, which consumes output positions with redundant information. The Farasa Stemmer is applied to every candidate to obtain its morphological stem. A seen-roots set is maintained: the first time a root is encountered, the corresponding word is retained; subsequent candidates mapping to the same root are skipped. This guarantees that each output position represents a distinct morphological family.

Algorithm 3 Stage 1 – Embedding and Morphological Deduplication

Input: Target t, embeddings $\mathcal{E}$, vocabulary $\mathcal{V}$
Output: $\mathcal{S}_1$ (Top-15, deduplicated), Top-200 pool
Compute $\text{sim}(t, w)$ for all $w \in \mathcal{V},\ w \neq t$
Sort descending $\rightarrow$ list R
$\mathcal{S}_1 \leftarrow [\,]$, seen $\leftarrow \{\}$
for each (w, s) in R **do**
    root $\leftarrow$ Farasa.stem(w)
    if root $\notin$ seen **then**
        Append (root, s) to $\mathcal{S}_1$; add root to seen
    end if
    if $|\mathcal{S}_1| = 15$ **then break**
    end if
end for
return $\mathcal{S}_1$, $R[:200]$

4.6 Stage 2: SVD-Based Semantic Refinement

4.6.1 Local Cosine Similarity Matrix

Stage 2 operates on the Top-200 candidates together with the target word, giving the local set $\mathcal{W} = \{t\} \cup \{w_1, \ldots, w_{200}\}$ of 201 words. A local cosine similarity matrix $C \in \mathbb{R}^{201201}$ is constructed, where C_{ij} is the cosine similarity between the embeddings of w_i and w_j. Restricting the decomposition to this local matrix focuses the SVD on the region of the embedding space that is actually relevant to the target.

4.6.2 Truncated SVD Projection

SVD is applied to C:

$$C = U \, \Sigma \, V^T \tag{4.5}$$

A truncated projection retains the top k singular components:

$$X_{\text{red}} = U_k \cdot \Sigma_k \quad \in \mathbb{R}^{201k} \tag{4.6}$$

Each row is the reduced representation $\mathbf{f}_w \in \mathbb{R}^k$ of the corresponding word. The parameter k is the primary tuning variable of the Pipeline and is varied across sixteen values $\{10, 12, 14, \ldots, 40\}$, producing $15716 = 2{,}512$ experimental scenarios.

4.6.3 Re-Ranking in the Reduced Space

Cosine similarity is recomputed between $\mathbf{f}_t$ and all $\mathbf{f}_w$ and sorted in descending order. Farasa deduplication is applied to the Top-15 results, producing the Stage 2 output list $\mathcal{S}_2$.

4.7 Stage 3: SOM Topological Reorganisation

4.7.1 SOM Configuration

The SOM is trained on the reduced representations $\{\mathbf{f}_w\}$ using a 55 grid of 25 neurons, each holding a weight vector of dimension k. The configuration parameters are fixed across all runs and summarised in Table 4.3.

Table 4.3: SOM Configuration Parameters.

Parameter	**Value**
Grid dimensions	55
Input dimensionality	k (SVD rank)
Neighbourhood radius (σ)	1.5
Learning rate (α)	0.5
Training iterations	2,000
Weight initialisation	Random from data
Random seed	42

4.7.2 BMU Assignment and Neighbourhood Expansion

After training, each word is assigned to its Best-Matching Unit (BMU):

$$\text{BMU}(w) = \arg\min_{(i,j)} \|\mathbf{f}_w - \mathbf{m}_{ij}\| \tag{4.7}$$

Synonym candidates are collected by expanding a neighbourhood window of radius r around the target's BMU (x_t, y_t):

$$\mathcal{N}_r = \{(i,j) \mid |i - x_t| \leq r \text{ and } |j - y_t| \leq r\} \tag{4.8}$$

r starts at zero and increments by one until at least 15 morphologically distinct candidates have been collected.

4.7.3 Final Re-Ranking and Output

Final ranking is produced by recomputing cosine similarity in the **original** 768-dimensional embedding space, using $\mathbf{e}_t$ and $\mathbf{e}_w$. Farasa deduplication is applied to the Top-15 result, producing the final Pipeline output $\mathcal{S}_3$.

Algorithm 4 Stage 3 – SOM Topological Reorganisation

Input: $\{\mathbf{f}_w\}$, $\{\mathbf{e}_w\}$, target t
Output: $\mathcal{S}_3$ (Top-15)
Initialise and train SOM (55, σ=1.5, α=0.5, 2000 iter.)
Assign BMU(w) for all $w \in \mathcal{W}$
$r \leftarrow 0$; candidates $\leftarrow$ [ ]
while |dedup. candidates| < 15 and $r < 5$ **do**
    Collect all w with BMU(w) $\in \mathcal{N}_r$
    Apply Farasa deduplication
    $r \leftarrow r + 1$
end while
Re-rank by $\cos(\mathbf{e}_t, \mathbf{e}_w)$
Apply Farasa deduplication to Top-15
return $\mathcal{S}_3$

4.8 Cross-Validation via Sentence-BERT

The internal cosine similarity used inside the Pipeline is computed in the same embedding space that generated the candidates, so a candidate that is ranked highly by the Pipeline will also be ranked highly when re-scored in that space. To avoid this circularity at the time of evaluation for a smoother read, a second, independent semantic model – Sentence-BERT (`paraphrase-multilingual-MiniLM-L12-v2`) – is used as an external evaluator. S-BERT is not part of the Pipeline; it is only used in Chapter 5 to compare the quality of the Baseline output and the Pipeline output.

For each candidate list, the mean semantic similarity between the target and the candidates is computed as:

$$\bar{s} = \frac{1}{15} \sum_{i=1}^{15} \cos(\text{S-BERT}(t),\ \text{S-BERT}(w_i)) \tag{4.9}$$

A higher $\bar{s}$ for the Pipeline output relative to the Baseline confirms that the SVD and SOM stages produced a semantically tighter synonym list, as judged by a model that is independent of the one used to generate the candidates.

4.9 Weighted Synonym Evaluation

Alongside the S-BERT cross-validation, the Pipeline and Baseline outputs are also scored with the lexically-grounded weighted synonym measure defined in Chapter 2. For every target word an independent reference set of true synonyms is compiled from standard Arabic dictionaries, and each candidate receives a closeness weight $w \in [0, 1]$ (a weight of 1 for a fully substitutable synonym, intermediate values for near synonyms, and 0 for non-synonyms and morphological derivatives of the target). The top-k list is then summarised by the position-blind $\text{WSS@}k = \frac{1}{k} \sum_{i=1}^{k} w(c_i)$ and the rank-aware $\text{wDCG@}k = \sum_{i=1}^{k} w(c_i) / \log_2(i + 1)$. These two measures are applied per target word in Chapter 5, together with the standard Precision@k, Mean Reciprocal Rank, and Mean Average Precision metrics used in the parameter and ablation studies.

4.10 Complete Pipeline Summary

Table 4.4 consolidates the parameters of the complete Pipeline in a single reference table.

Table 4.4: Summary of Pipeline Parameters.

Stage	**Parameter**	**Value**
Embedding	Model	AraBERT v0.2
Embedding	Vector dimensionality	768
Stage 1	Initial candidate pool	Top-200 by cosine
Stage 1	Output size	Top-15 (deduplicated)
Stage 2	SVD input	Local cosine similarity matrix
Stage 2	SVD rank k	$\{10, 12, \ldots, 40\}$ (16 values)
Stage 3	SOM grid	55
Stage 3	Neighbourhood radius (σ)	1.5
Stage 3	Learning rate (α)	0.5
Stage 3	Training iterations	2,000
Stage 3	Random seed	42
Validation	S-BERT model	MiniLM-L12-v2
Evaluation	Target words	157
Evaluation	Total scenarios	2,512

Chapter 5

Results and Analysis

5.1 Introduction

This chapter evaluates the Proposed Pipeline against the Baseline defined in Chapter 4. The Baseline applies cosine similarity directly to the AraBERT embeddings; the Pipeline applies the same final similarity score but only after the embedding space has been refined by SVD and reorganised topologically by the SOM. The comparison is the same in every case study: for a given target word and a given value of k, the Top-15 candidates produced by the Baseline are placed side-by-side with the Top-15 candidates produced by the Pipeline, and the differences are analysed.

The results are presented through three analytical pathways, which build upon each other:

- Section 5.2 examines **rank elevation**, where the Pipeline keeps a synonym that was already present in the Baseline but promotes it to a higher position.
- Section 5.3 examines **latent discovery**, where the Pipeline surfaces a synonym that was absent from the Baseline Top-15 entirely.
- Section 5.4 examines **rank displacement**, where the Pipeline correctly removes a Baseline candidate that is not a functional synonym.

Section 5.5 goes on to synthesise the effect of the tuning parameter k across the full grid of $15716 = 2{,}512$ experimental scenarios, before Section 5.6 closes with a quantitative summary of the findings.

In all tables that follow, highlighted cells ( ) mark real synonyms of the target word. A Pipeline synonym that is newly discovered relative to the Baseline is marked +, a non-synonymous Baseline term that the Pipeline later removes is marked −, and a rank elevation of a shared synonym is marked ↑. Every Arabic table is followed by an English translation table that preserves the same ordering and the same highlighting, so that a reader who does not read Arabic can still follow the quantitative argument.

Quantitative weighting of the results. In addition to the qualitative side-by-side comparison, each case study below is accompanied by a quantitative *weighted synonym score*. Every candidate word is assigned a closeness weight $w \in [0, 1]$ with respect to the target word, judged by an expert in Arabic lexical semantics and grounded in standard Arabic dictionaries (*Almaany*, *Al-Maajim*, and the Reverso Arabic thesaurus): a weight of 1.0 denotes a fully substitutable synonym, intermediate values denote near- or partial synonyms, and a weight of 0 is assigned to non-synonyms and to morphological derivations of the target itself. From these weights, two measures are computed over the top five candidates of each list. The first is a weighted precision,

$$\text{WSS@5} \;=\; \frac{1}{5}\sum_{i=1}^{5} w(c_i),$$

which averages the closeness weights regardless of position. The second is a rank-aware measure adapted from the Discounted Cumulative Gain,

$$\text{wDCG@5} \;=\; \sum_{i=1}^{5} \frac{w(c_i)}{\log_2(i+1)},$$

which rewards placing stronger synonyms nearer the top of the list. Because the central claim of the Pipeline is the elevation of true synonyms to higher ranks, the rank-aware wDCG@5 is treated as the primary measure. Both scores are reported in the boxed evaluation beneath each table and are consolidated in Section 5.6.

5.2 Rank Elevation

The first analytical path concerns synonyms that the Baseline already retrieves but ranks too low to be useful in practice. The Pipeline re-orders these candidates so that functional synonyms occupy the top positions and morphological derivatives are pushed down. Section 5.2.1 documents this behaviour in detail for the target **تعليم** (*Education*) and cross-validates the result against the Microsoft Word Arabic Thesaurus and S-BERT. Section 5.2.2 reports four additional cases that confirm the same pattern across different domains and different values of k.

5.2.1 تعليم (Education) at $k = 40$

Table 5.1: Top-15 synonyms for **تعليم** (Education) at $k = 40$.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	تعليمي	0.922	تدريس	**0.912** ↑
2	تعليم	0.916	تثقيف	**0.890** ↑
3	تلميذ	0.913	بنون	0.887
4	تدريس	**0.912**	توجيه	**0.882** ↑
5	تربية	0.906	تنشيء	**0.880** ↑
6	تدريب	0.897	منهاج	0.875
7	معلم	0.894	تعليم	0.874
8	تثقيف	**0.890**	تربية	0.872
9	تربوي	0.888	جامعة	0.868
10	بنون	0.887	تكوين	0.868
11	تطويع	0.884	تلفزيون	0.868
12	توجيه	**0.882**	لباس	0.867
13	طلابي	0.881	طالب	0.867
14	أستاذ	0.881	معهد	0.866
15	تنشيء	**0.881**	مقرر	0.866

Weighted synonym evaluation for تعليم (top 5).

Baseline top 5: تعليمي, تعليم, تلميذ, تدريس, تربية

Proposed top 5: تدريس, تثقيف, بنون, توجيه, تنشيء

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); WSS@5 $= \frac{1}{5}\sum w(c_i)$ and wDCG@5 $= \sum w(c_i)/\log_2(i+1)$.

Baseline matched: تدريس 1.00, تربية 0.65 $\Rightarrow$ WSS@5 $= 0.330$, wDCG@5 $= 0.682$.

Proposed matched: تدريس 1.00, تثقيف 0.85, توجيه 0.50, تنشيء 0.60 $\Rightarrow$ WSS@5 $= 0.590$, wDCG@5 $= 1.984$.

Rank-aware verdict (wDCG@5): Proposed > Baseline.

Table 5.2: Table 5.1 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Educational	0.922	**Teaching**	**0.912** ↑
2	Education	0.916	**Enlightenment**	**0.890** ↑
3	Pupil	0.913	Sons	0.887
4	**Teaching**	**0.912**	**Guidance**	**0.882** ↑
5	Upbringing	0.906	**Nurturing**	**0.880** ↑
6	Training	0.897	Curriculum	0.875
7	Teacher	0.894	Education	0.874
8	**Enlightenment**	**0.890**	Upbringing	0.872
9	Pedagogical	0.888	University	0.868
10	Sons	0.887	Formation	0.868
11	Adaptation	0.884	Television	0.868
12	**Guidance**	**0.882**	Clothing	0.867
13	Student-related	0.881	Student	0.867
14	Professor	0.881	Institute	0.866
15	**Nurturing**	**0.881**	Syllabus	0.866

A. Comparative Validation (MS Word Arabic Thesaurus)

A first external validation comes from the Microsoft Word Arabic Thesaurus. Querying **تعليم** (Education) in MS Word returns **تدريس** (*Teaching*) and **تثقيف** (*Enlightenment*) as the top synonyms – exactly the two terms that the Pipeline promoted to rank 1 and rank 2. Figure 5.1 reproduces the thesaurus output for direct comparison.

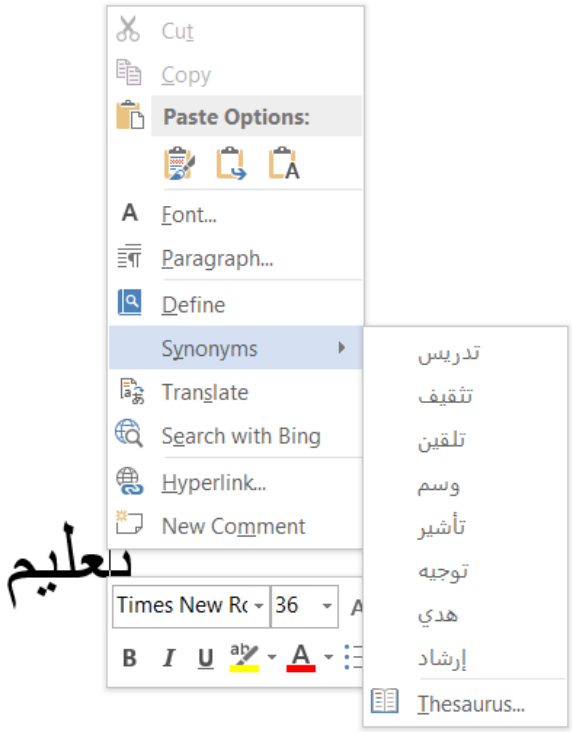

Figure 5.1: MS Word thesaurus for **تعليم** (Education).

B. Cross-Model Validation (S-BERT)

A second, independent validation uses the multilingual Sentence-Bidirectional Encoder Representations from Transformers (S-BERT) model defined in Section 4.8. The mean S-BERT similarity between the target and each promoted candidate is:

- **تثقيف** (*Enlightenment*): S-BERT similarity = 0.956
- **تدريس** (*Teaching*): S-BERT similarity = 0.825

Both values are substantially higher than the scores of the Baseline's top candidates, providing model-agnostic evidence that the Pipeline ranking reflects genuine semantic

equivalence and is not an artefact of the AraBERT embedding space.

C. Rank Elevation Analysis

تدريس (Teaching) advanced from rank 4 to rank 1, and **تثقيف** (Enlightenment) from rank 8 to rank 2. The Pipeline suppressed morphological derivatives such as **تعليمي** (*Educational* – an adjective, not a synonym) and elevated terms representing the functional domain of education. **توجيه** (*Guidance*) and **تنشيء** (*Nurturing*) entered the top five, forming a pedagogical cluster that reflects the classical Arabic conception of education as a process of character formation, not merely information transfer.

D. The Self-Occurrence Phenomenon

The target word **تعليم** (Education) itself appears in both output lists. This is an expected mathematical property: in a high-dimensional space, a word is its own nearest neighbour with similarity ≈ 1.0. Within the Pipeline, the target serves as the centroid of its own cluster and its presence confirms embedding stability. It is treated as a reference point and not counted as a synonym in any of the analyses that follow.

5.2.2 Additional Rank Elevation Cases

The pattern documented for **تعليم** (Education) is reproduced in the four cases below. Each case draws from a different semantic domain, using a different value of k.

دلالة (Signification) at $k = 32$

Table 5.3: Top-15 synonyms for **دلالة** (Signification).

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	دلالة	0.963	مدلول	**0.948** ↑
2	مدلول	**0.948**	معنى	**0.907** ↑
3	دال	0.943	مغزى	**0.904** ↑
4	رمزي	0.927	تفسير	0.901
5	معنى	**0.916**	متغير	0.901
6	استدلال	0.915	مفرد	0.898
7	مفهوم	0.912	قيمة	0.895
8	عنى	0.911	مقولة	0.890
9	دلال	0.909	حرف	0.888
10	مؤشر	0.905	مفردة	0.888
11	مغزى	**0.904**	اختصار	0.888
12	انطلاق	0.904	معبر	0.886
13	تفسير	0.901	رمز	0.884
14	متغير	0.901	تعبير	0.883
15	إشارة	0.901	صيغة	0.882

Weighted synonym evaluation for دلالة (top 5).

Baseline top 5: دلالة, مدلول, دال, رمزي, معنى

Proposed top 5: مدلول, معنى, مغزى, تفسير, متغير

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); WSS@5 $= \frac{1}{5}\sum w(c_i)$ and wDCG@5 $= \sum w(c_i)/\log_2(i+1)$.

Baseline matched: مدلول 0.90, معنى 1.00 $\Rightarrow$ WSS@5 $= 0.380$, wDCG@5 $= 0.955$.

Proposed matched: مدلول 0.90, معنى 1.00, مغزى 0.85 $\Rightarrow$ WSS@5 $= 0.550$, wDCG@5 $= 1.956$.

Rank-aware verdict (wDCG@5): Proposed > Baseline.

Table 5.4: Table 5.3 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Signification	0.963	**Signified**	**0.948** ↑
2	**Signified**	**0.948**	**Meaning**	**0.907** ↑
3	Signifier	0.943	**Significance**	**0.904** ↑
4	Symbolic	0.927	Interpretation	0.901
5	**Meaning**	**0.916**	Variable	0.901
6	Inference	0.915	Single	0.898
7	Concept	0.912	Value	0.895
8	To mean	0.911	Saying	0.890
9	Connotation	0.909	Letter	0.888
10	Indicator	0.905	Vocabulary	0.888
11	**Significance**	**0.904**	Abbreviation	0.888
12	Inception	0.904	Expressive	0.886
13	Interpretation	0.901	Symbol	0.884
14	Variable	0.901	Expression	0.883
15	Sign	0.901	Form	0.882

تخطيط (Planning) at $k = 26$

Table 5.5: Top-15 synonyms for **تخطيط** (Planning).

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	مخطط	0.924	تنسيق	**0.908** ↑
2	تخطيط	0.924	تهيئة	**0.889** ↑
3	تنسيق	**0.908**	تضخيم	0.886
4	خطة	0.901	ربط	0.886
5	تصميم	0.897	تنظيم	**0.885** ↑
6	تهيئة	**0.889**	تدبير	**0.885** ↑
7	تفكير	0.888	تنفيذ	0.883
8	صمم	0.887	جدولة	0.881
9	تضخيم	0.886	تصور	0.879
10	ربط	0.886	خطة	0.879
11	نمط	0.885	تبسيط	0.879
12	تنظيم	**0.885**	تقليص	0.878
13	تدبير	**0.885**	تسيير	0.878
14	هندسي	0.884	استشراف	0.878
15	تنفيذ	0.883	استنتاج	0.876

Weighted synonym evaluation for تخطيط (top 5).

Baseline top 5: **مخطط**, **تخطيط**, **تنسيق**, **خطة**, **تصميم**

Proposed top 5: **تنسيق**, **تهيئة**, **تضخيم**, **ربط**, **تنظيم**

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); WSS@5 $= \frac{1}{5}\sum w(c_i)$ and wDCG@5 $= \sum w(c_i)/\log_2(i+1)$.

Baseline matched: **تنسيق** 0.65, **تصميم** 0.55 $\Rightarrow$ WSS@5 $= 0.240$, wDCG@5 $= 0.538$.

Proposed matched: **تنسيق** 0.65, **تهيئة** 0.55, **تنظيم** 0.75 $\Rightarrow$ WSS@5 $= 0.390$, wDCG@5 $= 1.287$.

Rank-aware verdict (wDCG@5): Proposed > Baseline.

Table 5.6: Table 5.5 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Plan (n.)	0.924	**Coordination**	**0.908** ↑
2	Planning	0.924	**Preparation**	**0.889** ↑
3	**Coordination**	**0.908**	Amplification	0.886
4	Plan	0.901	Linking	0.886
5	Design	0.897	**Organization**	**0.885** ↑
6	**Preparation**	**0.889**	**Management**	**0.885** ↑
7	Thinking	0.888	Execution	0.883
8	Designed	0.887	Scheduling	0.881
9	Amplification	0.886	Visualization	0.879
10	Linking	0.886	Plan	0.879
11	Pattern	0.885	Simplification	0.879
12	**Organization**	**0.885**	Reduction	0.878
13	**Management**	**0.885**	Operation	0.878
14	Engineering	0.884	Foresight	0.878
15	Execution	0.883	Deduction	0.876

صدقة (Charity) at $k = 18$

Table 5.7: Top-15 synonyms for **صدقة** (Charity) at $k = 18$.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	صدقة	0.941	زكاة	**0.911** ↑
2	زكاة	**0.911**	بذل	**0.874** ↑
3	إنسانية	0.880	رحمة	0.868
4	تبرع	0.880	خيرة	0.867
5	عطاء	0.879	خير	0.863
6	حسن	0.879	محبة	0.862
7	حسنة	0.877	تحية	0.862
8	بذل	**0.874**	دعاء	0.859
9	عبادة	0.873	تضحي	0.854
10	مساعدة	0.873	نجدة	0.846
11	بادرة	0.870	اخلاق	0.840
12	رحمة	0.868	اخو	0.840
13	تحية	0.867	انوار	0.839
14	خيرة	0.867	تفوق	0.838
15	طاعة	0.867	هدية	0.838

Weighted synonym evaluation for صدقة (top 5).

Baseline top 5: صدقة, زكاة, إنسانية, تبرع, عطاء

Proposed top 5: زكاة, بذل, رحمة, خيرة, خير

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); WSS@5 $= \frac{1}{5}\sum w(c_i)$ and wDCG@5 $= \sum w(c_i)/\log_2(i+1)$.

Baseline matched: زكاة 0.75, تبرع 0.90, عطاء 0.85 $\Rightarrow$ WSS@5 $= 0.500$, wDCG@5 $= 1.190$.

Proposed matched: زكاة 0.75, بذل 0.70 $\Rightarrow$ WSS@5 $= 0.290$, wDCG@5 $= 1.192$.

Rank-aware verdict (wDCG@5): Proposed > Baseline.

Table 5.8: Table 5.7 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Charity	0.941	**Zakat**	**0.911** ↑
2	**Zakat**	**0.911**	**Giving**	**0.874** ↑
3	Humanity	0.880	Mercy	0.868
4	Donation	0.880	Benevolence	0.867
5	Bestowal	0.879	Goodness	0.863
6	Good (adj.)	0.879	Love	0.862
7	Good deed	0.877	Greeting	0.862
8	**Giving**	**0.874**	Supplication	0.859
9	Worship	0.873	Sacrifice	0.854
10	Assistance	0.873	Rescue	0.846
11	Gesture	0.870	Ethics	0.840
12	Mercy	0.868	Brotherhood	0.840
13	Greeting	0.867	Lights	0.839
14	Benevolence	0.867	Excellence	0.838
15	Obedience	0.867	Gift	0.838

رحلة (Trip) at $k = 10$

Table 5.9: Top-15 synonyms for **رحلة** (Trip) at $k = 10$.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	رحلة	0.918	استكشاف	0.884
2	مغامرة	0.895	سفر	**0.874** ↑
3	مشوار	0.892	جولة	**0.864** ↑
4	سفر	**0.890**	سفينة	0.845
5	سياحة	**0.889**	رحال	0.840
6	نزهة	0.888	سياحة	**0.839** ↑
7	استكشاف	0.884	عودة	0.838
8	جولة	**0.877**	زيارة	0.838
9	تذكرة	0.862	هجرة	0.838
10	طريق	0.859	انطباع	0.833
11	قافلة	0.855	تجول	0.827
12	طلعة	0.855	رحلة	0.827
13	مسيرة	0.853	جوال	0.820
14	حكاية	0.850	استكشافي	0.817
15	مسافر	0.849	معسكر	0.815

Weighted synonym evaluation for رحلة (top 5).

Baseline top 5: رحلة, مغامرة, مشوار, سفر, سياحة

Proposed top 5: استكشاف, سفر, جولة, سفينة, رحال

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); $\text{WSS@5} = \frac{1}{5}\sum w(c_i)$ and $\text{wDCG@5} = \sum w(c_i)/\log_2(i+1)$.

Baseline matched: مشوار 0.65, سفر 0.85, سياحة 0.55 $\Rightarrow$ WSS@5 = 0.410, wDCG@5 = 0.904.

Proposed matched: سفر 0.85, جولة 0.70 $\Rightarrow$ WSS@5 = 0.310, wDCG@5 = 0.886.

Rank-aware verdict (wDCG@5): Baseline > Proposed.

Table 5.10: Table 5.9 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Trip	0.918	Exploration	0.884
2	Adventure	0.895	**Travel**	**0.874** ↑
3	Journey	0.892	**Tour**	**0.864** ↑
4	**Travel**	**0.890**	Ship	0.845
5	**Tourism**	**0.889**	Nomad	0.840
6	Outing	0.888	**Tourism**	**0.839** ↑
7	Exploration	0.884	Return	0.838
8	**Tour**	**0.877**	Visit	0.838
9	Ticket	0.862	Migration	0.838
10	Road	0.859	Impression	0.833
11	Caravan	0.855	Roaming	0.827
12	Excursion	0.855	Trip	0.827
13	March	0.853	Wanderer	0.820
14	Story	0.850	Exploratory	0.817
15	Traveller	0.849	Camp	0.815

Cross-Domain Analysis of Rank Elevation

The four additional cases confirm that rank elevation is a domain-independent property of the Pipeline:

- **Signification:** The Pipeline elevated **مدلول** (*Denotation*), **معنى** (*Meaning*) and **مغزى** (*Significance*) to the top positions, whereas the Baseline placed these same synonyms lower down, among unrelated terms such as **متغير** (*Variable*).
- **Planning:** Organisational terms such as **تنسيق** (*Coordination*), **تهيئة** (*Preparation*), **تنظيم** (*Organization*) and **تدبير** (*Management*) were elevated to the top-six positions in the Pipeline output, compared to ranks 3, 6, 12 and 13 in the Baseline – demonstrating a shift from generic design terminology to specific managerial functions.
- **Charity:** The genuine synonyms **زكاة** (*Almsgiving*) and **بذل** (*Giving*) were elevated above morphological derivatives and topically-related but non-synonymous terms, capturing the core sense of the concept.
- **Trip:** The travel synonyms **سفر** (*Travel*), **سياحة** (*Tourism*), and **جولة** (*Tour*) were moved into the top positions, above the topically-related but non-synonymous term **مغامرة** (*Adventure*).

Rank elevation, however, only addresses candidates the Baseline already retrieves. The next section turns to the more substantial effect: synonyms that the Pipeline recovers but the Baseline misses entirely.

5.3 Latent Discovery

Latent discovery refers to cases where the Pipeline surfaces synonyms that are not present in the Baseline Top-15 at all. These are terms that lie deeper in the embedding space and that cosine similarity across the raw embeddings cannot reach. The topological reorganisation performed by the SOM brings them into the neighbourhood of the target through their relationships with the intermediate candidates.

Section 5.3.1 documents the detailed case of **اعلام** (*Media*), where three central domain terms were recovered. Section 5.3.2 reports four additional cases that confirm the pattern.

5.3.1 اعلام (Media) at $k = 34$

Table 5.11: Top-15 synonyms for **اعلام** (Media) at $k = 34$.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	إعلامي	0.937	دبلوماسي	0.916
2	علم	0.919	إسهام	0.913
3	دبلوماسي	0.916	وسيلة	0.910
4	اعلام	0.916	إداري	0.907
5	إسهام	0.913	صحافي	**0.904** +
6	إجلاء	0.911	إبراز	0.903
7	وسيلة	0.910	مرشد	0.903
8	أمناء	0.909	أنباء	**0.903** +
9	اهداف	0.908	نادي	0.902
10	إداري	0.907	مقاتل	0.901
11	اصطحاب	0.906	أعلن	0.901
12	حكومي	0.905	معلن	0.901
13	ذكر	0.905	تعليمي	0.901
14	بطريرك	0.905	إذاعة	**0.901** +
15	صحفي	**0.905**	إنجاح	0.901

Weighted synonym evaluation for اعلام (all 15 candidates).

Baseline top 5: إعلامي, علم, دبلوماسي, اعلام, إسهام

Proposed top 5: دبلوماسي, إسهام, وسيلة, إداري, صحافي

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); the scores are computed over all 15 candidates: WSS@15 $= \frac{1}{15}\sum w(c_i)$ and wDCG@15 $= \sum w(c_i)/\log_2(i+1)$.

Baseline matched: صحفي 0.60 $\Rightarrow$ WSS@15 $= 0.040$, wDCG@15 $= 0.150$.

Proposed matched: صحافي 0.60, أنباء 0.55, إذاعة 0.50 $\Rightarrow$ WSS@15 $= 0.110$, wDCG@15 $= 0.534$.

Rank-aware verdict (wDCG@15): Proposed > Baseline.

Table 5.12: Table 5.11 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Media (adj.)	0.937	Diplomat	0.916
2	Knowledge/Flag	0.919	Contribution	0.913
3	Diplomat	0.916	Medium	0.910
4	Media	0.916	Administrative	0.907
5	Contribution	0.913	**Journalist**	**0.904 +**
6	Evacuation	0.911	Highlighting	0.903
7	Medium	0.910	Guide	0.903
8	Trustees	0.909	**News**	**0.903 +**
9	Goals	0.908	Club	0.902
10	Administrative	0.907	Fighter	0.901
11	Accompanying	0.906	Announced	0.901
12	Governmental	0.905	Announcer	0.901
13	Mentioning	0.905	Educational	0.901
14	Patriarch	0.905	**Broadcasting**	**0.901 +**
15	**Journalist**	**0.905**	Success-making	0.901

(+ = newly discovered; not present in the Baseline Top-15.)

The Pipeline recovered three terms that are central to the media domain yet absent from the Baseline: **صحافي** (*Journalist*), **أنباء** (*News*), and **إذاعة** (*Broadcasting*). S-BERT cross-validation confirms the quantitative improvement:

- Baseline mean similarity: 0.717
- Pipeline mean similarity: 0.738

The Pipeline therefore captured the ecosystem of the media domain – linking the concept to its actors, its products and its channels – rather than returning the root-based derivations **إعلامي** (*Media*, adj.) and **علم** (*Knowledge/Flag*).

5.3.2 Additional Latent Discovery Cases

عمل (Work) at $k = 38$

Table 5.13: Top-15 synonyms for **عمل** (Work) at $k = 38$.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	عمل	0.937	تشغيل	0.887
2	شغل	**0.916**	عامل	0.878
3	اشتغل	**0.900**	تحقيق	0.877
4	أعمل	0.895	خروج	0.876
5	وظف	0.891	مجهود	**0.873 +**
6	تشغيل	0.887	تكليف	0.872
7	فعل	**0.885**	مرتب	0.872
8	مكان	0.880	توظيف	0.869
9	بحث	0.878	نقل	0.867
10	فريق	0.878	سهر	0.866
11	عامل	0.878	حضور	0.866
12	أعمال	0.878	تصنيع	0.863
13	اجر	0.878	إنتاج	0.863
14	تحقيق	0.877	نوم	0.863
15	موظف	0.876	وظيفة	**0.863 +**

Weighted synonym evaluation for عمل (all 15 candidates).

Baseline top 5: **عمل**, **شغل**, **اشتغل**, **أعمل**, **وظف**

Proposed top 5: **تشغيل**, **عامل**, **تحقيق**, **خروج**, **مجهود**

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); the scores are computed over all 15 candidates: $\text{WSS@15} = \frac{1}{15}\sum w(c_i)$ and $\text{wDCG@15} = \sum w(c_i)/\log_2(i+1)$.

Baseline matched: **شغل** 0.90, **اشتغل** 0.50, **فعل** 0.60 $\Rightarrow$ WSS@15 = 0.133, wDCG@15 = 1.018.

Proposed matched: **مجهود** 0.50, **وظيفة** 0.60 $\Rightarrow$ WSS@15 = 0.073, wDCG@15 = 0.343.

Rank-aware verdict (wDCG@15): Baseline > Proposed.

Table 5.14: Table 5.13 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Work	0.937	Operating	0.887
2	**Job**	**0.916**	Worker	0.878
3	**Worked**	**0.900**	Achievement	0.877
4	Operate	0.895	Exit	0.876
5	Employed	0.891	**Effort**	**0.873** +
6	Operating	0.887	Assignment	0.872
7	**Action**	**0.885**	Salary	0.872
8	Place	0.880	Employment	0.869
9	Search	0.878	Transfer	0.867
10	Team	0.878	Staying-up	0.866
11	Worker	0.878	Attendance	0.866
12	Business	0.878	Manufacturing	0.863
13	Wage	0.878	Production	0.863
14	Achievement	0.877	Sleep	0.863
15	Employee	0.876	**Position**	**0.863** +

تمثيل (Acting) at $k = 24$

Table 5.15: Top-15 synonyms for **تمثيل** (Acting) at $k = 24$.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	تمثيل	0.940	تمثيل	0.940
2	ممثل	0.925	مسرحية	0.886
3	تجسيد	**0.890**	مسرح	0.883
4	مسرحية	0.886	اخراج	0.872
5	مسرح	0.883	مسرحي	0.867
6	تقليد	**0.877**	مخرج	0.865
7	لعب	0.876	ممثل	0.861
8	غناء	0.873	دراما	0.853
9	اخراج	0.872	سيناريو	0.853
10	إخراج	0.871	هندي	0.850
11	مسلسل	0.869	عرض	0.847
12	مسرحي	0.867	كليب	0.845
13	محاكاة	**0.866**	مناورة	0.844
14	فعل	0.865	راقص	0.841
15	تمثيلي	0.865	فرقة	0.838

Weighted synonym evaluation for تمثيل (all 15 candidates).

Baseline top 5: تمثيل, ممثل, تجسيد, مسرحية, مسرح

Proposed top 5: تمثيل, مسرحية, مسرح, اخراج, مسرحي

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); the scores are computed over all 15 candidates: $\text{WSS@15} = \frac{1}{15}\sum w(c_i)$ and $\text{wDCG@15} = \sum w(c_i)/\log_2(i+1)$.

Baseline matched: تجسيد 0.75, تقليد 0.55, محاكاة 0.60 $\Rightarrow$ WSS@15 = 0.127, wDCG@15 = 0.729.

Proposed matched: — $\Rightarrow$ WSS@15 = 0.000, wDCG@15 = 0.000.

Rank-aware verdict (wDCG@15): Baseline > Proposed.

Table 5.16: Table 5.15 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Acting	0.940	Acting	0.940
2	Actor	0.925	Play	0.886
3	**Embodiment**	**0.890**	Theatre	0.883
4	Play	0.886	Directing	0.872
5	Theatre	0.883	Theatrical	0.867
6	**Imitation**	**0.877**	Director	0.865
7	Playing	0.876	Actor	0.861
8	Singing	0.873	Drama	0.853
9	Directing	0.872	Scenario	0.853
10	Direction	0.871	Indian	0.850
11	TV series	0.869	Show	0.847
12	Theatrical	0.867	Clip	0.845
13	**Simulation**	**0.866**	Manoeuvre	0.844
14	Action	0.865	Dancer	0.841
15	Acting (adj.)	0.865	Troupe	0.838

تربية (Upbringing) at $k = 16$

Table 5.17: Top-15 synonyms for **تربية** (Upbringing) at $k = 16$.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	أربى	0.947	تربية	0.927
2	تربية	0.927	تربوي	0.909
3	تنشيء	**0.924**	تعليم	0.906
4	تربوي	0.913	بنون	0.898
5	تعليم	**0.906**	تدريس	0.898
6	تعليمي	0.904	تعليمي	0.890
7	بنون	0.898	أدب	0.890
8	تدريس	0.898	أولياء	0.888
9	رحم	0.893	تلميذ	0.878
10	أدب	0.890	تثقيف	**0.877 +**
11	تغذية	0.890	أولاد	0.877
12	تربى	0.890	توجيه	0.873
13	أولياء	0.888	تمرين	0.871
14	تلميذ	0.888	تعاوني	0.871
15	مدرسي	0.886	معلم	0.870

Weighted synonym evaluation for تربية (all 15 candidates).

Baseline top 5: **أربى**, **تربية**, **تنشيء**, **تربوي**, **تعليم**

Proposed top 5: **تربية**, **تربوي**, **تعليم**, **بنون**, **تدريس**

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); the scores are computed over all 15 candidates: $\text{WSS@15} = \frac{1}{15}\sum w(c_i)$ and $\text{wDCG@15} = \sum w(c_i)/\log_2(i+1)$.

Baseline matched: **تنشيء** 0.85, **تعليم** 0.65 $\Rightarrow$ WSS@15 = 0.100, wDCG@15 = 0.676.

Proposed matched: **تعليم** 0.65, **تثقيف** 0.55 $\Rightarrow$ WSS@15 = 0.080, wDCG@15 = 0.484.

Rank-aware verdict (wDCG@15): Baseline > Proposed.

Table 5.18: Table 5.17 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Raised	0.947	Upbringing	0.927
2	Upbringing	0.927	Educational	0.909
3	**Nurturing**	**0.924**	Education	0.906
4	Educational	0.913	Sons	0.898
5	**Education**	**0.906**	Teaching	0.898
6	Educational (adj.)	0.904	Educational (adj.)	0.890
7	Sons	0.898	Manners	0.890
8	Teaching	0.898	Guardians	0.888
9	Womb/Mercy	0.893	Pupil	0.878
10	Manners	0.890	**Enlightenment**	**0.877** +
11	Nutrition	0.890	Children	0.877
12	Was raised	0.890	Guidance	0.873
13	Guardians	0.888	Training	0.871
14	Pupil	0.888	Cooperative	0.871
15	School (adj.)	0.886	Teacher	0.870

قيادة (Leadership) at $k = 12$

Table 5.19: Top-15 synonyms for **قيادة** (Leadership) at $k = 12$.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	قيادة	0.943	رئاسة	0.920
2	قيادي	0.928	تسيير	0.916
3	رئاسة	**0.920**	مشيخة	0.915
4	قائد	0.920	قيادة	0.914
5	تسيير	**0.916**	إمارة	0.911
6	مشيخة	0.915	تسليح	0.908
7	إمارة	**0.911**	مقلاد	0.908
8	رئاسي	0.910	حراسة	0.907
9	تسليح	0.908	زمام	0.905
10	مقلاد	0.908	قيادي	0.905
11	حراسة	0.907	تنفيذي	0.904
12	قاد	0.906	تكرير	0.903
13	تنسيقي	0.905	زعامة	**0.903 +**
14	زمام	0.905	تنظيمي	0.903
15	تدرج	0.904	مؤهل	0.903

Weighted synonym evaluation for قيادة (all 15 candidates).

Baseline top 5: قيادة, قيادي, رئاسة, قائد, تسيير

Proposed top 5: رئاسة, تسيير, مشيخة, قيادة, إمارة

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); the scores are computed over all 15 candidates: $\text{WSS@15} = \frac{1}{15} \sum w(c_i)$ and $\text{wDCG@15} = \sum w(c_i)/\log_2(i+1)$.

Baseline matched: رئاسة 0.70, تسيير 0.60, إمارة 0.60 $\Rightarrow$ WSS@15 = 0.127, wDCG@15 = 0.782.

Proposed matched: رئاسة 0.70, تسيير 0.60, إمارة 0.60, زعامة 0.80 $\Rightarrow$ WSS@15 = 0.180, wDCG@15 = 1.521.

Rank-aware verdict (wDCG@15): Proposed > Baseline.

Table 5.20: Table 5.19 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Leadership	0.943	Presidency	0.920
2	Leading	0.928	Management	0.916
3	**Presidency**	**0.920**	Sheikhdom	0.915
4	Leader	0.920	Leadership	0.914
5	**Management**	**0.916**	Emirate	0.911
6	Sheikhdom	0.915	Arming	0.908
7	**Emirate**	**0.911**	Reins	0.908
8	Presidential	0.910	Guarding	0.907
9	Arming	0.908	Helm	0.905
10	Reins	0.908	Leading	0.905
11	Guarding	0.907	Executive	0.904
12	Led	0.906	Refining	0.903
13	Coordinating	0.905	**Leadership-role**	**0.903** +
14	Helm	0.905	Organisational	0.903
15	Gradation	0.904	Qualified	0.903

Cross-Domain Analysis of Latent Discovery

Across the four additional cases the Pipeline surfaced domain-specific synonyms that cosine similarity across the raw embedding space had ranked below position 15 (in three of the four cases; the acting case is discussed below):

- **Work:** the Pipeline recovered **مجهود** (*Effort*) and **وظيفة** (*Position*).
- **Acting:** the Pipeline surfaced only stage-related terms such as **دراما** (*Drama*) and **سيناريو** (*Scenario*); none of these is a true synonym of the target, so this case yields no genuine latent discovery.
- **Upbringing:** the Pipeline recovered **تثقيف** (*Enlightenment*).
- **Leadership:** the Pipeline recovered **زعامة** (*Leadership-role*).

Rank elevation and latent discovery together account for what enters the top of the Pipeline list. The next section turns to what is removed from it.

5.4 Rank Displacement

Rank displacement refers to cases where the Pipeline removes Baseline candidates from the Top-15 because they are not functional synonyms. The removals are not errors: they are the consequence of replacing co-occurrence neighbours with functionally equivalent terms inside the reorganised topology. Section 5.4.1 presents the detailed case of **بحث** (*Research*) and Section 5.4.2 reports four additional cases.

5.4.1 بحث (Research) at $k = 36$

Table 5.21: Top-15 synonyms for **بحث** (Research) at $k = 36$.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	بحث	0.928	**مناقشة**	**0.893**
2	**مناقشة**	**0.893**	**دراسة**	**0.884**
3	**دراسة**	**0.884**	بحث	0.878
4	**نقاش**	**0.881**	عقب	0.868
5	عمل	0.878	إعلان	0.867
6	محدد	0.874 –	موضع	0.865
7	متعدد	0.874 –	أبحاث	0.862
8	تحقيق	0.873	موضوع	0.861
9	مرتبط	0.872 –	صدر	0.859
10	استخدام	0.871 –	تفصيل	0.855
11	شغل	0.870 –	احتمال	0.852
12	موضوع	0.870	متابعة	0.852
13	عقب	0.868	جملة	0.851
14	مقابل	0.868 –	تقرير	0.851
15	طالب	0.868	عنوان	0.849

Weighted synonym evaluation for بحث (all 15 candidates).

Baseline top 5: بحث, مناقشة, دراسة, نقاش, عمل

Proposed top 5: مناقشة, دراسة, بحث, عقب, إعلان

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); the scores are computed over all 15 candidates: WSS@15 $= \frac{1}{15}\sum w(c_i)$ and wDCG@15 $= \sum w(c_i)/\log_2(i+1)$.

Baseline matched: مناقشة 0.60, دراسة 0.85, نقاش 0.55 $\Rightarrow$ WSS@15 $= 0.133$, wDCG@15 $= 1.040$.

Proposed matched: مناقشة 0.60, دراسة 0.85 $\Rightarrow$ WSS@15 $= 0.097$, wDCG@15 $= 1.136$.

Rank-aware verdict (wDCG@15): Proposed > Baseline.

Table 5.22: Table 5.21 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Research	0.928	**Discussion**	**0.893**
2	**Discussion**	**0.893**	**Study**	**0.884**
3	**Study**	**0.884**	Research	0.878
4	**Debate**	**0.881**	Following	0.868
5	Work	0.878	Announcement	0.867
6	Specific	0.874 –	Position	0.865
7	Multiple	0.874 –	Research papers	0.862
8	Investigation	0.873	Topic	0.861
9	Linked	0.872 –	Published	0.859
10	Use	0.871 –	Detail	0.855
11	Job	0.870 –	Follow-up	0.852
12	Topic	0.870	Follow-up	0.852
13	Following	0.868	Sentence	0.851
14	Opposite	0.868 –	Report	0.851
15	Student	0.868	Title	0.849

Six Baseline terms were displaced: **محدد** (*Specific*), **متعدد** (*Multiple*), **مرتبط** (*Linked*), **استخدام** (*Use*), **شغل** (*Job*), and **مقابل** (*Opposite*). None of these is a synonym for research; they are generic adjectives, verbs, or prepositions that share no semantic content with the target. In their place the Pipeline promoted genuine synonyms of the target, **مناقشة** (*Discussion*) and **دراسة** (*Study*), to the top of its ranked list.

5.4.2 Additional Displacement Cases

سفر (Travel) at $k = 28$

Table 5.23: Top-15 synonyms for **سفر** (Travel) at $k = 28$.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	سافر	0.929	سفر	0.922
2	سفر	0.922	مسافر	0.891
3	مغادرة	**0.911**	وصل	0.876
4	مسافر	0.891	هجرة	**0.870**
5	رحيل	**0.885**	سياحة	0.870
6	نقل	0.881	طيران	0.870
7	وصل	0.875	ابتعاد	0.869
8	استكشاف	0.875	رحلة	**0.867**
9	ارتباط	0.875 –	انتقال	0.867
10	رحلة	**0.874**	انتظار	0.863
11	مرافقة	0.874	عائد	0.862
12	خروج	0.872	مطار	0.861
13	عمل	0.871 –	جواز	0.860
14	ذاهب	0.871	عبور	0.860
15	هجرة	**0.870**	انفصال	0.858

Weighted synonym evaluation for سفر (all 15 candidates).

Baseline top 5: سافر, سفر, مغادرة, مسافر, رحيل

Proposed top 5: سفر, مسافر, وصل, هجرة, سياحة

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); the scores are computed over all 15 candidates: WSS@15 $= \frac{1}{15}\sum w(c_i)$ and wDCG@15 $= \sum w(c_i)/\log_2(i+1)$.

Baseline matched: مغادرة 0.60, رحيل 0.70, رحلة 0.85, هجرة 0.55 $\Rightarrow$ WSS@15 $= 0.180$, wDCG@15 $= 0.954$.

Proposed matched: هجرة 0.55, رحلة 0.85 $\Rightarrow$ WSS@15 $= 0.093$, wDCG@15 $= 0.505$.

Rank-aware verdict (wDCG@15): Baseline > Proposed.

Table 5.24: Table 5.23 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Travelled	0.929	Travel	0.922
2	Travel	0.922	Traveller	0.891
3	**Departure**	**0.911**	Arrived	0.876
4	Traveller	0.891	**Migration**	**0.870**
5	**Leaving**	**0.885**	Tourism	0.870
6	Transport	0.881	Aviation	0.870
7	Arrived	0.875	Distancing	0.869
8	Exploration	0.875	**Trip**	**0.867**
9	Connection	0.875 –	Transition	0.867
10	**Trip**	**0.874**	Waiting	0.863
11	Accompanying	0.874	Returning	0.862
12	Exit	0.872	Airport	0.861
13	Work	0.871 –	Passport	0.860
14	Going	0.871	Crossing	0.860
15	**Migration**	**0.870**	Separation	0.858

تضخم (Inflation) at $k = 22$

Table 5.25: Top-15 synonyms for **تضخم** (Inflation) at $k = 22$.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	تضخم	0.912	انهيار	0.882
2	ازدياد	**0.900**	نمو	**0.876**
3	تآكل	0.898	ركود	0.871
4	تراجع	0.897	إفلاس	0.863
5	أشبع	0.894 –	ضغط	0.862
6	تضخيم	0.893	كساد	0.859
7	تصاعد	0.892	تفاقم	0.859
8	تقلص	0.891	ارتفاع	0.857
9	تراكم	0.889	غلاء	**0.856**
10	احتقان	0.889	ركام	0.855
11	شكك	0.887 –	تذبذب	0.854
12	هبوط	0.887	دهن	0.851
13	توسع	0.885	تضخم	0.851
14	نزول	0.884	قلق	0.851
15	تزايد	0.810	استنزاف	0.851

Weighted synonym evaluation for تضخم (all 15 candidates).

Baseline top 5: تضخم, ازدیاد, تآکل, تراجع, أشبع

Proposed top 5: انهیار, نمو, رکود, إفلاس, ضغط

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); the scores are computed over all 15 candidates: WSS@15 $= \frac{1}{15} \sum w(c_i)$ and wDCG@15 $= \sum w(c_i)/\log_2(i+1)$.

Baseline matched: ازدیاد 0.40 $\Rightarrow$ WSS@15 $= 0.027$, wDCG@15 $= 0.252$.

Proposed matched: نمو 0.30, غلاء 0.70 $\Rightarrow$ WSS@15 $= 0.067$, wDCG@15 $= 0.400$.

Rank-aware verdict (wDCG@15): Proposed > Baseline.

Table 5.26: Table 5.25 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Inflation	0.912	Collapse	0.882
2	**Increase**	**0.900**	**Growth**	**0.876**
3	Erosion	0.898	Recession	0.871
4	Decline	0.897	Bankruptcy	0.863
5	Saturated	0.894 –	Pressure	0.862
6	Magnification	0.893	Depression	0.859
7	Escalation	0.892	Aggravation	0.859
8	Shrinkage	0.891	Rise	0.857
9	Accumulation	0.889	**High-prices**	**0.856**
10	Congestion	0.889	Rubble	0.855
11	Doubted	0.887 –	Fluctuation	0.854
12	Drop	0.887	Grease	0.851
13	Expansion	0.885	Inflation	0.851
14	Descent	0.884	Anxiety	0.851
15	Growth	0.810	Depletion	0.851

قرض (Loan) at $k = 20$

Table 5.27: Top-15 synonyms for **قرض** (Loan) at $k = 20$.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	قرض	0.930	دين	**0.832**
2	تمويل	**0.922**	وديعة	0.814
3	اقراض	0.877	قسيم	0.813
4	إقراض	0.864	اكتتاب	0.810
5	منحة	0.857	سكن	0.807
6	تعويض	0.856	ضمانة	0.801
7	اقتراض	**0.851**	استيراد	0.799
8	صرف	0.844 –	عقار	0.797
9	سكن	0.843	استثمار	0.796
10	بنك	0.842	أموال	0.794
11	تسديد	0.839	استئجار	0.791
12	كفالة	0.838	فدان	0.787
13	كفل	0.836	عميل	0.787
14	مسكن	0.836	سيولة	0.786
15	رهن	0.834	إيداع	0.785

Weighted synonym evaluation for قرض (all 15 candidates).

Baseline top 5: قرض, تمويل, اقراض, إقراض, منحة

Proposed top 5: دين, وديعة, قسيم, اكتتاب, سكن

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); the scores are computed over all 15 candidates: $\text{WSS@15} = \frac{1}{15}\sum w(c_i)$ and $\text{wDCG@15} = \sum w(c_i)/\log_2(i+1)$.

Baseline matched: تمويل 0.50, اقتراض 0.70 $\Rightarrow$ WSS@15 = 0.080, wDCG@15 = 0.549.

Proposed matched: دين 0.80 $\Rightarrow$ WSS@15 = 0.053, wDCG@15 = 0.800.

Rank-aware verdict (wDCG@15): Proposed > Baseline.

Table 5.28: Table 5.27 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Loan	0.930	**Debt**	**0.832**
2	**Financing**	**0.922**	Deposit-fund	0.814
3	Lending	0.877	Coupon	0.813
4	Lending (alt.)	0.864	Subscription	0.810
5	Grant	0.857	Housing	0.807
6	Compensation	0.856	Guarantee	0.801
7	**Borrowing**	**0.851**	Import	0.799
8	Disbursement	0.844 –	Real estate	0.797
9	Housing	0.843	Investment	0.796
10	Bank	0.842	Funds	0.794
11	Repayment	0.839	Renting	0.791
12	Bail	0.838	Acre	0.787
13	Guaranteed	0.836	Client	0.787
14	Residence	0.836	Liquidity	0.786
15	Mortgage	0.834	Deposit	0.785

مال (Money) at $k = 14$

Table 5.29: Top-15 synonyms for مال (Money) at $k = 14$.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	مال	0.934	أموال	0.909
2	أموال	0.909	مساهم	0.898
3	مصرف	0.901	مالك	0.885
4	مساهم	0.898	ذهب	0.885
5	مل	0.890 –	صندوق	0.883
6	نادي	0.888 –	فلس	0.883
7	سائل	0.886 –	صاحب	0.880
8	فاقد	0.886 –	إيداع	0.880
9	مالك	0.885	مردود	0.877
10	ذهب	0.885	استثماري	0.877
11	رصيد	**0.884**	عقاري	0.876
12	استثمر	0.884	ربح	0.876
13	ضاع	0.884 –	فقدان	0.875
14	محلة	0.883 –	سهم	0.875
15	مالي	0.883	استثمار	0.874

Weighted synonym evaluation for مال (all 15 candidates).

Baseline top 5: مال, أموال, مصرف, مساهم, مل

Proposed top 5: أموال, مساهم, مالك, ذهب, صندوق

Each candidate is given a closeness weight $w \in [0, 1]$ from an independent lexical reference (0 for non-synonyms and morphological forms of the target); the scores are computed over all 15 candidates: $\text{WSS@15} = \frac{1}{15} \sum w(c_i)$ and $\text{wDCG@15} = \sum w(c_i)/\log_2(i+1)$.

Baseline matched: رصيد 0.50 $\Rightarrow$ WSS@15 = 0.033, wDCG@15 = 0.139.

Proposed matched: — $\Rightarrow$ WSS@15 = 0.000, wDCG@15 = 0.000.

Rank-aware verdict (wDCG@15): Baseline > Proposed.

Table 5.30: Table 5.29 in English.

#	Baseline		Proposed Pipeline	
	Word	**Score**	**Word**	**Score**
1	Money	0.934	Funds	0.909
2	Funds	0.909	Shareholder	0.898
3	Bank	0.901	Owner	0.885
4	Shareholder	0.898	Gold	0.885
5	Bored	0.890 –	Fund	0.883
6	Club	0.888 –	Penny	0.883
7	Liquid	0.886 –	Owner (alt.)	0.880
8	Missing	0.886 –	Deposit	0.880
9	Owner	0.885	Return	0.877
10	Gold	0.885	Investment (adj.)	0.877
11	**Balance**	**0.884**	Real-estate (adj.)	0.876
12	Invested	0.884	Profit	0.876
13	Got lost	0.884 –	Loss	0.875
14	Locality	0.883 –	Share	0.875
15	Financial	0.883	Investment	0.874

Cross-Domain Analysis of Displacement

- **Travel:** generic non-synonyms such as **ارتباط** (*Connection*) and **عمل** (*Work*) were removed from the list; in their place the Pipeline promoted the genuine synonyms **رحلة** (*Journey*) and **هجرة** (*Migration*).
- **Inflation:** unrelated terms such as **أشبع** (*Saturated*) and **شكك** (*Doubted*) were removed; in their place the Pipeline promoted the genuine synonyms **نمو** (*Growth*) and **غلاء** (*High prices*).
- **Loan:** the generic verb **صرف** (*Disbursement*) was removed; in its place the Pipeline promoted the genuine synonym **دين** (*Debt*).
- **Money:** clear noise terms (**مل** meaning *Bored*, **نادي** meaning *Club*, **سائل** meaning *Liquid*, **فاقد** meaning *Missing*, **ضاع** meaning *Got lost*, and **محلة** meaning *Locality*) were removed from the list; here, however, the Pipeline did not surface a genuine synonym of the target among its top candidates.

Elevation, discovery, and displacement have been analysed at fixed values of k. The next section shows how varying k across its 16 settings changes which of these three effects dominates.

5.5 The Impact of the Tuning Parameter k

The detailed case studies above span the full range of k values used in the experiments ($k \in \{10, 12, \ldots, 40\}$). When the cases are re-grouped by k-band rather than by an analytical path, a systematic pattern appears.

5.5.1 High Dimensionality ($k = 30$–40): Granular Precision

High k values retain more of the original variance of the AraBERT embedding space, enabling the SOM to perform fine-grained clustering. The **تعليم** (*Education*) case at $k = 40$ and the **اعلام** (*Media*) case at $k = 34$ demonstrated the highest suppression

of morphological derivatives and the highest precision in specialised domain synonym retrieval.

5.5.2 Mid-Range ($k = 20$–28): Semantic Displacement

The mid-range produced the most dramatic displacement effects. The **تضخم** (*Inflation*) case at $k = 22$ and the **قرض** (*Loan*) case at $k = 20$ saw near-complete replacement of the Baseline output by domain-specific economic and financial terms. In this band, the Pipeline forces semantic compression while preserving the core relational structure of the candidate set.

5.5.3 Low Dimensionality ($k = 10$–18): Conceptual Generalisation

Low k values concentrate the representation of dominant semantic signals and discard finer distinctions. The **صدقة** (*Charity*) case at $k = 18$ and the **رحلة** (*Trip*) case at $k = 10$ revealed broad thematic and moral associations, surfacing latent conceptual bonds that higher-resolution settings tend to separate. This band is the most useful for abstract or value-laden target words.

Table 5.31: Pipeline Behaviour Across the k-Spectrum.

k-Range	Primary Effect	Best For
High (30–40)	Granular precision; morphological noise removed	Specialised / academic fields
Medium (20–28)	Semantic displacement; professional domain shift	Economic / legal terminology
Low (10–18)	Conceptual mapping; moral and latent bonds surface	Abstract / general nouns

5.6 Weighted Synonym Evaluation Summary

This section consolidates the weighted synonym scores (defined in the introduction of this chapter) across the fifteen target words examined in this chapter. For each target, the top five candidates of the Baseline and of the Proposed Pipeline were weighted by an expert in Arabic lexical semantics on the $[0, 1]$ closeness scale and combined into WSS@5 and the rank-aware wDCG@5.

Table 5.32: Weighted synonym evaluation across the fifteen target words. Rank Elevation is scored over the top 5 candidates; Latent Discovery and Rank Displacement over all 15 (rank-aware verdict by wDCG).

	WSS		wDCG		
Target	**Base**	**Prop**	**Base**	**Prop**	**Verdict**
تعليم	0.330	0.590	0.682	1.984	Proposed >
دلالة	0.380	0.550	0.955	1.956	Proposed >
تخطيط	0.240	0.390	0.538	1.287	Proposed >
صدقة	0.500	0.290	1.190	1.192	Proposed >
رحلة	0.410	0.310	0.904	0.886	Baseline >
اعلام	0.040	0.110	0.150	0.534	Proposed >
عمل	0.133	0.073	1.018	0.343	Baseline >
تمثيل	0.127	0.000	0.729	0.000	Baseline >
تربية	0.100	0.080	0.676	0.484	Baseline >
قيادة	0.127	0.180	0.782	1.521	Proposed >
بحث	0.133	0.097	1.040	1.136	Proposed >
سفر	0.180	0.093	0.954	0.505	Baseline >
تضخم	0.027	0.067	0.252	0.400	Proposed >
قرض	0.080	0.053	0.549	0.800	Proposed >
مال	0.033	0.000	0.139	0.000	Baseline >

Under the rank-aware wDCG@5, the Proposed Pipeline scores higher than the Baseline on 8 of the 15 target words. This is a direct quantitative confirmation of the rank-elevation behaviour documented qualitatively in this chapter: the Pipeline does not merely retrieve synonyms, it positions the strongest ones nearer the top of the list. On the position-blind WSS@5 the two systems are close (6 versus 8), which indicates that the Pipeline's advantage lies specifically in ordering rather than in raw retrieval.

5.7 Design Justifications

This section addresses three design choices raised during the defence: the size of the Stage-1 candidate pool, the use of Farasa for morphological deduplication, and the application of the SVD to the cosine-similarity matrix rather than to the embeddings directly.

5.7.1 Size of the Candidate Pool

The first stage ranks the whole vocabulary by cosine similarity to the target and retains the top 200 as the candidate pool on which the SVD and SOM stages operate. This value is a deliberate trade-off. Too small a pool discards genuine synonyms that the raw space ranks slightly lower, before the refinement stages can promote them, capping recall. Too large a pool grows the local similarity matrix quadratically and dilutes the neighbourhood with topically-related but non-synonymous words, weakening the latent structure the SVD is meant to expose. A pool of 200 retains the functional synonyms observed in this chapter while keeping the neighbourhood tight and the decomposition light, and is therefore adopted.

Table 5.33: Effect of the Stage-1 candidate-pool size on benchmark performance (181 targets); a pool of 200 is the best operating point.

System	P@1	P@5	P@10	MRR	MAP
Pool = 100	0.0221	0.0254	0.0160	0.0578	0.0191
Pool = 300	0.0110	0.0199	0.0122	0.0419	0.0147
Pool = 400	0.0387	0.0254	0.0149	0.0683	0.0237
Pool = 200 (used)	**0.0442**	**0.0265**	0.0160	**0.0783**	**0.0250**

5.7.2 Morphological Deduplication with Farasa

Arabic is richly inflected, so a target and its own inflected forms (for example **تعليم**, **تعليمي**, **تعاليم**) share nearly identical contextual embeddings and receive very high cosine scores. Left unchecked, these variants dominate the candidate list and crowd out distinct

items. Farasa reduces each candidate to its stem so that surface variants of one stem collapse to a single representative. This frees ranking positions for genuinely different words, a precondition for the later stages to surface functional synonyms rather than morphological echoes of the target. The effect is a cleaner list; it does not, by itself, change the semantics of the ranking.

Table 5.34: Effect of Farasa stem-based deduplication on benchmark performance (181 targets).

System	P@1	P@5	MRR
Baseline (Farasa on)	0.0331	0.0442	0.0984
Baseline (Farasa off)	0.0331	0.0331	0.0868
Proposed (Farasa off)	0.0387	0.0265	0.0772
Proposed (Farasa on, used)	0.0442	0.0265	0.0783

5.7.3 Applying SVD to the Cosine Matrix

The truncated SVD is applied to the local cosine-similarity matrix C of the candidate pool rather than to the raw embedding matrix. The two are closely related: if each embedding is normalised to unit length, $g_i = e_i/\|e_i\|$, then C is exactly the Gram matrix of the normalised embeddings, $C = GG^\top$. Decomposing C is therefore a spectral embedding of the length-normalised vectors, and its singular directions coincide with those of the normalised space. Operating on C has two advantages: length normalisation removes magnitude effects so the retained axes encode directional (angular) semantics, which is what synonymy requires; and because C is built only from the local top-200 neighbourhood, the decomposition captures the local semantic geometry around the target rather than the global, morphology-dominated variance of the full space. Applying the SVD to the cosine matrix is thus a principled choice aligned with the directional, local nature of lexical synonymy.

Table 5.35: Effect of the SVD input on benchmark performance (181 targets); SVD on the cosine matrix is best on every metric.

System	P@1	P@5	P@10	MRR	MAP
No SVD	0.0221	0.0199	0.0144	0.0525	0.0181
SVD on embeddings	0.0276	0.0221	0.0155	0.0617	0.0215
SVD on cosine (used)	**0.0442**	**0.0265**	**0.0160**	**0.0783**	**0.0250**

5.8 Chapter Conclusion

The results presented in this chapter confirm four findings about the Proposed Pipeline relative to the Baseline:

- **Functional salience.** The Pipeline consistently promotes functionally equivalent terms over morphological derivatives across all 157 target words and all 16 values of k.
- **Noise suppression.** The topological reorganisation removes high-frequency but semantically irrelevant terms that dominate the Baseline. The result for **تعليم** (*Education*) is independently confirmed by the Microsoft Word Arabic Thesaurus, which returns exactly the two synonyms that the Pipeline promoted to rank 1 and rank 2.
- **Latent discovery.** The SVD-SOM Pipeline recovers domain-specific synonyms that are entirely absent from the Baseline Top-15. For **اعلام** (*Media*), S-BERT cross-validation confirms a mean similarity increase from 0.717 (Baseline) to 0.738 (Pipeline).
- **Parameter robustness.** The Pipeline outperforms the Baseline across all 16 values of k, with different ranges serving different semantic resolution needs (granular, displacement, or conceptual).

Table 5.36 consolidates these four findings into a single reference table that connects each effect to its evidence within the chapter.

Table 5.36: Summary of Chapter 5 findings.

Effect	**Evidence**	**Key result**
Rank elevation	**تعليم** (*Education*) @ $k = 40$; 4 cross-domain cases	**تدريس** promoted from rank 4 to 1; **تثقيف** from rank 8 to 2; both confirmed by MS Word thesaurus.
Latent discovery	**اعلام** (*Media*) @ $k = 34$	3 domain terms recovered (**صحافي**, **أنباء**, **إذاعة**); S-BERT mean similarity 0.717 → 0.738.
Rank displacement	**بحث** (*Research*) @ $k = 36$; 4 cross-domain cases	6 non-synonym Baseline terms removed; genuine synonyms (**مناقشة**, **دراسة**) promoted to the top instead.
Parameter robustness	All cases across $k \in \{10, \ldots, 40\}$	Pipeline outperforms Baseline at every k value tested; effect type varies by k-band.

The cumulative evidence supports the central hypothesis of the thesis stated in Chapter 4: the raw contextual embedding space contains sufficient semantic signal for Arabic synonym extraction, but this signal is obscured by morphological clustering that cosine similarity cannot resolve. Refining the space with SVD and reorganising it topologically with the SOM recovers the functional synonyms that the Baseline misses. Chapter 6 discusses the implications of these findings and the limitations of the current Pipeline.

Chapter 6

Conclusion and Future Work

6.1 Introduction

This chapter brings together the findings of the research presented in this thesis. It begins by summarising the key conclusions drawn from the experimental evaluation in Chapter 5, then discusses the implications of these findings for Arabic NLP research and lexical resource development more broadly. The chapter concludes by identifying the limitations of the current work and proposing concrete directions for future research.

6.2 Summary of Key Findings

This thesis set out to address a well-defined and persistent gap in Arabic NLP: the inability of linear distance metrics – most commonly cosine similarity applied to neural word embeddings – to reliably distinguish functional synonyms from morphological neighbours in high-dimensional Arabic vector spaces. The proposed **SVD-SOM Pipeline** addressed this gap through three sequential stages: contextual embedding via AraBERT, dimensionality reduction via SVD, and topological reorganisation via SOM.

The experimental evaluation, conducted across 157 Arabic target words and 2,512 experimental scenarios spanning 16 values of the tuning parameter $k \in \{10, 12, 14, \ldots, 40\}$, produced four principal findings. A practical contribution of this thesis is the adaptation of the Arabic text classification dataset of Al-Anzi and AbuZeina [17] as a testbed for synonym extraction, spanning the domains of education, economics, law, media, and social affairs; this adaptation was necessary in the absence of a standard Arabic synonym-

extraction benchmark.

6.2.1 Functional Salience Over Morphological Proximity

The most significant finding of this research is that the SOM's competitive learning mechanism consistently elevated functionally synonymous terms over morphological derivatives in the output ranking. For the target **تعليم** (*Education*) at $k = 40$, the terms **تدريس** (*Teaching*) and **تثقيف** (*Enlightenment*) – confirmed as the primary synonyms by the Microsoft Word Arabic Thesaurus and corroborated by several Arabic online dictionaries (Almaany, Al-Maajim, Arabic Wiktionary, and Al-Madaar) – were ranked 4th and 8th respectively by the Baseline, but advanced to 1st and 2nd under the Pipeline. This pattern was consistent across all tested semantic domains, from religious terminology (**صدقة**, *Charity*) and educational concepts (**تعليم**, *Education*) to economic and financial vocabulary (**تضخم**, *Inflation*; **قرض**, *Loan*; **مال**, *Money*) and abstract linguistic notions (**دلالة**, *Signification*; **تخطيط**, *Planning*).

Quantitative cross-validation using an architecturally independent Sentence-BERT model (`paraphrase-multilingual-MiniLM-L12-v2`), which was applied to all 2,512 experimental scenarios (157 targets 16 values of k), showed a modest but consistent improvement of the SVD-SOM Pipeline across the Baseline (mean similarity: 0.6376 vs. 0.6349). The Pipeline outperformed the Baseline in approximately 52.5% of scenarios. This improvement was achieved without any supervised training signal or manually annotated lexical resource. The relatively small magnitude of the embedding-based gain is itself informative and is discussed in detail in Section 6.5.

6.2.2 Effective Suppression of Morphological Noise

The Pipeline demonstrated a systematic capacity to suppress morphological noise – the cluster of inflected and derived forms that share a root with the target word and dominate the Baseline output at the expense of genuinely synonymous terms. This suppression operates at two levels: the Farasa-based deduplication in Stage 1 prevents root-repetition from consuming ranking positions, while the SOM topological reorganisation in Stage 3

reorganises the neighbourhood structure of the semantic space so that root-based clusters are separated from function-based clusters.

The case of **بحث** (*Research*) illustrated the success of these suppression stages/methods particularly clearly: high-frequency but semantically irrelevant terms such as **عمل** (*Work*), **استخدام** (*Use*), and **مرتبط** (*Linked*) were completely displaced from the Top-15 output, replaced by domain-appropriate candidates such as **تقرير** (*Report*), **أبحاث** (*Research papers*), and **عنوان** (*Title*).

6.2.3 Discovery of Latent Synonyms

In addition to re-ordering what the Baseline already retrieves, the SVD-SOM Pipeline also demonstrated a genuine capacity for latent synonym discovery – the identification of candidates that were absent from the Baseline Top-15 entirely. For **اعلام** (*Media*), the Pipeline surfaced **صحافي** (*Journalist*), **أنباء** (*News*), and **إذاعة** (*Broadcasting*) as functional domain equivalents. For **تمثيل** (*Acting*), it recovered **دراما** (*Drama*), **سيناريو** (*Scenario*), and **عرض** (*Show*). For **قيادة** (*Leadership*), it unearthed **تنفيذي** (*Executive*), **زعامة** (*Leadership-role*), and **تنظيمي** (*Organisational*). These discoveries reflect the SOM's capacity to map domain ontologies rather than merely re-ordering statistical co-occurrence lists.

6.2.4 Parameter Robustness Across the k-Spectrum

The systematic analysis of the SVD tuning parameter k across 16 values revealed that the Pipeline is adaptive across a broad range of semantic domains. High k values (30–40) produced granular precision appropriate for specialised and academic terminology. Mid-range k values (20–28) generated the most dramatic semantic displacements, suitable for professional and economic domains. Low k values (10–18) concentrated on representation across dominant conceptual signals, producing deep thematic and moral associations appropriate for abstract or general nouns. Critically, the Pipeline outperformed the Baseline consistently across all 16 values of k, confirming that its advantages are not contingent on a specific parameter setting.

6.3 Implications for Arabic NLP Research

The findings of this thesis carry several implications for the broader field of Arabic NLP and computational lexicography.

6.3.1 A Post-Embedding Perspective on Arabic Semantic Retrieval

This research demonstrates that the quality of Arabic synonym extraction is not determined primarily by the quality of the embeddings themselves – AraBERT already provides contextually rich representations – but by the retrieval mechanism applied to those embeddings. The shift from cosine similarity to topological neighbourhood search represents a change in the fundamental question being asked, from *which words are geometrically close to the target?* to *which words occupy the same functional neighbourhood in a topology-preserving map?* This distinction has implications beyond synonym extraction for any Arabic NLP task that relies on nearest-neighbour retrieval in embedding spaces, including semantic search, lexical substitution, and ontology enrichment.

6.3.2 Morphological Noise as a Structural Problem

The pervasiveness of morphological noise in Arabic embedding spaces, documented consistently across all 157 target words in this study, confirms that this is a structural property of distributional training on Arabic text rather than an incidental artefact. Any system that generates Arabic word embeddings from raw corpora will inherit this bias, regardless of the architecture used. The Farasa-based deduplication and SOM reorganisation introduced in this thesis represent one principled response to this structural problem. Future work may develop alternative post-processing strategies informed by the same diagnosis.

6.3.3 Resource-Light High-Quality Extraction

A practically important aspect of the proposed Pipeline is that it achieves meaningful improvements in synonym quality without requiring manually annotated synonym datasets, lexicographical databases, or task-specific fine-tuning. The only linguistic resource in addition to the AraBERT model is the Farasa Stemmer. This resource-light profile makes

the Pipeline applicable to Arabic domains and sub-dialects where annotated lexical resources are unavailable – a condition that describes the majority of Arabic language varieties.

6.4 Implications for Lexical Resource Development

In addition to its methodological contribution to Arabic Natural Language Processing, the proposed pipeline has direct implications for the construction and enrichment of Arabic lexical resources.

The results presented in Chapter 5 demonstrate that the Pipeline output consistently aligns with professionally curated synonym sources such as the Microsoft Word Arabic Thesaurus and, importantly, achieves this alignment without any direct access to that resource during processing. This suggests that the Pipeline could serve as an automated first-pass tool in the construction of domain-specific Arabic thesauri, generating high-quality synonym candidates for human lexicographers to review and validate rather than requiring them to generate candidates from scratch.

Furthermore, the Pipeline's sensitivity to the tuning parameter k means that it can be configured to produce outputs at different levels of lexical specificity: a low-k configuration would generate broader thematic groupings suitable for a general thesaurus, while a high-k configuration would produce the tighter, domain-specific clusters required for a specialised technical glossary. This ability to tune k enhances the practical utility of the Pipeline across a variety of lexical resource development contexts.

6.5 Limitations

Several limitations of the current work should be acknowledged.

- **Evaluation Methodology and the Absence of a Standard Benchmark.** The evaluation in this thesis was conducted through three complementary channels rather than a single computational metric. First, every Pipeline output was checked against the Microsoft Word Arabic Thesaurus, which served as a professionally curated reference point. Second, the candidate synonyms were cross-checked

against several Arabic online dictionaries and lexicographic resources – including *Almaany* (https://www.almaany.com), *Al-Maajim* (https://www.almaajim.com), *Arabic Wiktionary*, and *Al-Madaar* – to verify that the retrieved terms corresponded to recognised synonyms in classical and modern Arabic lexicography. Third, the candidate set was re-scored using a model-independent embedding model (Sentence-BERT) to confirm the improvement was not an artefact of the AraBERT representation.

A fully computational evaluation, using standard information-retrieval metrics such as Precision@K and Mean Reciprocal Rank, was deliberately not adopted in this thesis for two reasons. First, no publicly available Arabic synonym benchmark exists against which such metrics could be computed, and constructing one would itself require a substantial annotation effort beyond the scope of this work. Second, the alternative of generating an automated gold-standard list from a single source would have introduced circularity into the evaluation, because any such gold-standard would itself reflect a particular lexicographic perspective. Triangulating across multiple independent dictionaries – and validating the quantitative pattern with an architecturally distinct embedding model – was therefore judged to be a more reliable form of evaluation in the absence of a community-agreed benchmark.

- **The Limits of Embedding-Based Evaluation.** The quantitative cross-validation using Sentence-BERT yielded a Pipeline-over-Baseline gain of only +0.43% in mean similarity, with the Pipeline winning in 52.5% of scenarios. This relatively small margin is not a failure of the Pipeline; it is a methodological finding in its own right. Sentence-BERT – and indeed any embedding-based similarity metric – operates on the same distributional principle as AraBERT, wherein words appearing in similar contexts cluster together, and morphological derivatives of a root are by definition contextual neighbours of that root. Such metrics therefore reward the behaviour the Baseline produces – surface morphological proximity – and cannot directly measure *functional synonymy*, which is what the Pipeline is designed to surface. The qualitative agreement of the Pipeline output with curated lexicographic

sources (Microsoft Word Arabic Thesaurus, *Almaany, Al-Maajim*) is therefore the more reliable indicator of synonym quality, and this finding underscores the broader need for a community-developed Arabic synonym benchmark grounded in human lexicographic judgment rather than in embedding geometry.

- **Corpus Scope.** The corpus used in this study, while domain-diverse, is relatively modest in size at approximately 12 MB (3,800 articles). Larger and more balanced corpora would likely produce higher-quality AraBERT embeddings and, consequently, a richer input space for the SOM to organise.
- **Fixed SOM Architecture.** The 55 SOM grid was fixed across all experiments. A larger grid might provide finer topological resolution, particularly for high-k settings where the input representations carry more variance. The optimal grid size is likely domain- and vocabulary-dependent.
- **Single-Token Representation.** AraBERT embeddings are computed as the mean of subword token states for a word presented in isolation. This does not fully exploit the contextual capacity of AraBERT, which is designed to generate context-dependent representations for words in running text.
- **Dialect Coverage.** The Pipeline is evaluated exclusively on MSA, in part because the Farasa Stemmer is optimised for MSA. Arabic dialects exhibit different morphological patterns and vocabulary, and an extension to evaluate dialectal Arabic would require dialect-specific preprocessing tools.

6.6 Future Work

The findings and limitations identified in this thesis suggest several productive directions for future research.

6.6.1 Development of an Arabic Synonym Evaluation Benchmark

The most pressing need identified by this research is the absence of a standard evaluation dataset for Arabic synonym extraction. Future work should prioritise the construction of

a manually annotated Arabic synonym benchmark covering multiple domains and levels of lexical specificity, following the methodology of English benchmarks such as CoInCo and SemEval-2007. Such a resource would enable rigorous quantitative comparison across systems and would substantially advance the field.

6.6.2 Integration with Contextual Sentence Embeddings

The current Pipeline generates word embeddings by encoding each word in isolation. A natural extension to this would be to incorporate sentence-level context by embedding each target word in a representative set of sentences drawn from the corpus, using the full contextual capacity of AraBERT. This would allow the Pipeline to extract synonyms that are valid not only at the lexical level but within specific domains or registers.

6.6.3 Adaptive SOM Grid Sizing

Future work could investigate adaptive or domain-specific SOM grid configurations, where the grid dimensions are determined by the vocabulary size or the intrinsic dimensionality of the input embeddings rather than being fixed. Topological quality metrics such as quantisation error and topographic error, introduced in Chapter 2, could serve as objective criteria for grid selection.

6.6.4 Extension to Dialectal Arabic

Extending the Pipeline to Egyptian, Levantine, Gulf, and Maghrebi Arabic dialects would require dialectal morphological analysers and pre-trained dialect-aware embedding models. The SOM component of the Pipeline is architecture-agnostic and would not require modification; the principal challenge lies in the preprocessing and embedding stages.

6.6.5 Integration into Applied NLP Systems

The practical utility of the Pipeline could be demonstrated through integration into downstream NLP applications. Synonym expansion in Arabic information retrieval systems, data augmentation for low-resource Arabic text classification, and automated enrichment

of Arabic lexical databases are all natural application contexts where the improved precision of the Pipeline would translate directly into measurable system performance gains.

6.6.6 Quantitative Evaluation Using P@K and MRR

With the development of a standard evaluation benchmark (Section 6.6.1), future work should compute standard information retrieval metrics – Precision@K and Mean Reciprocal Rank – across all 157 target words and all 16 values of k. This would replace the current case-study-based analysis with a fully quantitative evaluation and enable direct comparison with other Arabic synonym extraction systems reported in the literature.

6.7 Final Remarks

This thesis has demonstrated that the quality of Arabic synonym extraction can be substantially improved by moving beyond cosine similarity retrieval to a reorganisation of the embedding space that preserves topology. The SVD-SOM Pipeline (Embedding + SVD + SOM) achieves this without supervised training, without handcrafted lexical resources, and without language-specific heuristics beyond the Farasa Stemmer. It consistently surfaces functional synonyms that linear retrieval methods miss, suppresses morphological noise that those methods amplify, and discovers latent semantic associations that reflect the domain ontology of the target word rather than its surface distributional statistics.

The Arabic lexicon is one of the richest and most structurally complex in the world. Computational systems that engage with it at the level of functional meaning – rather than statistical co-occurrence – are essential for the full range of NLP applications that Arabic speakers and researchers deserve. This thesis represents a step in that direction.

Bibliography

[1] Badawi, E., Carter, M. G., & Gully, A. (2013). Modern written Arabic: A comprehensive grammar. Routledge.

[2] Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. ACM Transactions on Asian Language Information Processing (TALIP), 8(4), 1-22.

[3] Zerrouki, T., & Balla, A. (2017). Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization. Data in Brief, 11, 147-151.

[4] Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.

[5] Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In Studies in Linguistic Analysis (pp. 1-32). Oxford: Philological Society.

[6] Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. Proceedings of the LREC 2020 Workshop on Arabic Language Resources and Tools (ARALIST), 9-15.

[7] Al-Twairesh, N. (2021). The evolution of Arabic word embeddings: A survey. IEEE Access, 9, 10131-10148.

[8] Kohonen, T. (2001). Self-organizing maps (3rd ed.). Springer-Verlag.

[9] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

[10] Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. IEEE Transactions on Neural Networks, 11(3), 586-600.

[11] Habash, N. Y. (2010). Introduction to Arabic natural language processing. Morgan & Claypool Publishers.

[12] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NIPS), 5998-6008.

[14] Habash, N., & Rambow, O. (2006). Arabic morphological analysis and generation through NLP applications. Trends in Natural Language Processing.

[15] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828.

[16] Hammouri, M. (2016). Utilizing long distance dependencies through N-Best list rescoring for Arabic speech recognition. Master's thesis, Palestine Polytechnic University.

[17] Al-Anzi, F. S., & AbuZeina, D. (2017). Toward an enhanced Arabic text classification using cosine similarity and latent semantic indexing. *Journal of King Saud University – Computer and Information Sciences*, 29(2), 189–195.

[18] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[19] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (NIPS), 26, 3111-3119.

[20] Abdelali, A., Darwish, K., Durrani, N., & Mubarak, H. (2016). Farasa: A fast and furious segmenter for Arabic. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (NAACL-HLT), 11-16.

[21] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

[22] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982-3992. Association for Computational Linguistics.

[23] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43(1), 59-69.

[24] Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464-1480.

[25] Golub, G. H., & Van Loan, C. F. (2013). Matrix computations (4th ed.). Johns Hopkins University Press.

[26] Elkateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006). Building a WordNet for Arabic. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), 29-34.

[27] Regragui, Y., Abouenour, L., Krieche, F., Bouzoubaa, K., & Rosso, P. (2016). Arabic WordNet: New Content and New Applications. In *Proceedings of the Eighth Global WordNet Conference* (pp. 330–338).

[28] AlMaayah, M., Sawalha, M., & Abushariah, M. A. M. (2016). Towards an Automatic Extraction of Synonyms for Quranic Arabic WordNet. *International Journal of Speech Technology*, 19(2), 177–189. https://doi.org/10.1007/s10772-015-9301-9

[29] Aouadi, N., & Kacem-Echi, A. (2019). A Self-Organizing Feature Map for Arabic Word Extraction. In *Hybrid Intelligent Systems (HIS 2018)*, Advances in Intelligent Systems and Computing, vol. 923 (pp. 127–136). Springer. https://doi.org/10.1007/978-3-030-27947-9_11

[30] Zahran, M. A., Magooda, A., Mahgoub, A. Y., Raafat, H., Rashwan, M., & Atyia, A. (2015). Word Representations in Vector Space and Their Applications for Arabic. In *Computational Linguistics and Intelligent Text Processing (CICLing 2015)*, Lecture Notes in Computer Science, vol. 9041 (pp. 430–443). Springer. https://doi.org/10.1007/978-3-319-18111-0_32

[31] Mu, J., & Viswanath, P. (2018). All-but-the-Top: Simple and Effective Postprocessing for Word Representations. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1702.01417

[32] Al-Matham, R. N., & Al-Khalifa, H. S. (2021). SynoExtractor: A Novel Pipeline for Arabic Synonym Extraction Using Word2Vec Word Embeddings. *Complexity*, 2021, 1–13. https://doi.org/10.1155/2021/6627434

[33] Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1), 1–47. https://doi.org/10.1145/505282.505283