אוניברסיטת חיפה
UNIVERSITY OF HAIFA
جامعة حيفا

Research Proposal

Combining Classical Rhetorical Theory with State-of-the-Art AI Techniques for Discovering Hidden Tashbeeh in Arabic Poetry as a Case Study

Faisal Omari – 325616894

Prof. Tsvi Kuflik, The Department of Information Systems

Prof. Ali Hussein, Department of Arabic Language and Literature

Department of Computer Science

May 2025

Formatted: Font: 14 pt, Complex Script Font: 14 pt

Formatted: Font: 14 pt, Complex Script Font: 14 pt

Commented [TK1]: Much more generic and ambitious, but still, correct…

Deleted: Using NLP and Deep Learning

Deleted: Full

Deleted: Assoc.

TABLE OF CONTENTS

Deleted: *5*
Deleted: 5
Deleted: 5
Deleted: 5
Deleted: 6
Deleted: 6
Deleted: 7
Deleted: 7
Deleted: *10*
Deleted: 10
Deleted: 10
Deleted: *12*
Deleted: 12
Deleted: 14
Deleted: *15*
Deleted: 15
Deleted: 15
Deleted: 16
Deleted: *18*
Deleted: *19*
Deleted: 19
Deleted: 19
Deleted: *21*

Discovering Hidden Tashbeeh in Arabic Poetry Using NLP and Deep Learning

1. Introduction

Arabic poetry is one of the rich literary traditions, deeply intertwined with the culture, language, and identity of the Arab people. Within this poetic heritage, *tashbīh* (تشبيه), or simile, stands out as a central rhetorical device, used to create vivid imagery, evoke emotions, and clarify abstract concepts. Traditionally, *tashbīh* is expressed through explicit linguistic markers such as “كأن” (as if), “مثل” (like), or “كما” (as). For instance, the poet may write: "كأنّ سيوفهم نجومٌ في الليلة الظلماء" "As if their swords were stars in the dark night."

In this line, the simile is clearly marked by the word “كأنّ” and draws a comparison between swords and stars, emphasizing their brilliance and sharpness.

However, beyond these overt constructions lies a subtler and more complex form of rhetorical expression— which is called *tashbīh balīgh* (implicit simile). These similes lack explicit comparison words yet they maintain the poetic function of comparison. For example, a poet may say:

"سيوفهم نجوم في الليل" "Their swords are stars in the night"

While the line does not include a formal simile particle, the imagery implicitly evokes the same comparison to stars. These hidden similes are more difficult to detect, both for readers and for computational methods, as they require semantic and contextual understanding rather than keyword spotting.

Previous work on Arabic rhetoric, including research conducted on the Arabic Poetry Corpus developed at the University of Haifa, has focused on detecting explicit tashbīh using rule-based approaches—typically via handcrafted if-else conditions that look for tashbīh markers in poems. This method, while useful, is limited to only surface-level structures and fails to capture deeper, implicit rhetorical meanings.

This thesis aims to take a step forward by addressing the detection of implicit *tashbīh* using artificial intelligence. We propose to build machine learning models—

Commented [קצ2]: Nice. Ali?

Deleted: est

Deleted: in the world

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Deleted: hidden tashbīh

Deleted: (التشبيه الخفي)

Deleted: analogy

Deleted: تلمع سيوفهم في الليل

Deleted: gleam in the night.

Deleted: word

Deleted: significant

Deleted: hidden

Formatted: Font: Italic, Complex Script Font: Italic

specifically natural language processing (NLP) models fine-tuned on large pre-trained Arabic transformers such as AraBERT—that are capable of learning what makes a poem "contain *tashbīh*," even when the *tashbīh* is not explicitly marked.

In the initial phase of this research, we have:

- Converted the legacy Java-based rule system to Python for scalability and integration with modern NLP pipelines;
- Used the existing rule-based output to generate a binary labeled dataset of poems with and without *tashbīh* word;
- Preprocessed the poems by removing diacritics and cleaning the text;
- Fine-tuned the AraBERT model on this dataset, achieving a 96% accuracy in identifying *tashbīh* poems.

This result demonstrates the feasibility of training models to identify the rhetorical role of *tashbīh* in classical Arabic poetry.

The next challenge, and the core focus of this thesis, is to move from detecting obvious *tashbīh* to identifying hidden *tashbīh*: poetic lines where the comparison is embedded within context and meaning rather than structure. Our long-term goal is to build a model capable of identifying rhetorical richness in Arabic poetry even when surface cues are missing. To this end, we plan to:

- Manually transform a sample of explicit *tashbīh* poems into hidden ones (29265 poems);
- Investigate what features distinguish hidden *tashbīh* from non-*tashbīh* lines;
- Train and evaluate models that can generalize to these deeper forms of rhetorical expression.

Ultimately, this research contributes to the field of digital Arabic literary analysis by combining classical rhetorical theory with state-of-the-art AI techniques,

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Deleted: tashbīh

Formatted: Font: Italic, Complex Script Font: Italic

Deleted: strong

Deleted: teaching

Deleted: "understand"

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Commented [עח5]: Do you want to mention the number of the explicit similes that you have?

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: 12 pt, Complex Script Font: 12 pt

opening new avenues for understanding and interpreting the depth of Arabic poetic language.

Furthermore, the methodological framework developed in this work–particularly the approach to identifying implicit similes–can potentially be adapted to other languages and literary traditions, laying the foundation for broader cross-linguistic applications of rhetorical detection in computational literature studies.

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Formatted: Normal, Indent: Before: 1.27 cm, First line: 0.63 cm, No bullets or numbering

Deleted: Ultimately, this research contributes to the field of digital Arabic literary analysis by combining classical rhetorical theory with state-of-the-art AI techniques, opening new avenues for understanding and interpreting the depth of Arabic poetic language.

Formatted: Font: 12 pt, Complex Script Font: 12 pt

2. Background and Related Work

Commented [צק7]: Ali, please check/comments on the language aspects

Arabic Language

Commented [TK8]: Faisal, in this chapter, every claim you make requires a reference to support it, I briefly commented where references are needed, as an example – add them everywhere

Arabic is a Semitic language characterized by rich morphology, complex syntactic structures, and deep-rooted rhetorical traditions [reference?]. It features a templatic root-and-pattern system, where most words are derived from triliteral roots, enabling high lexical productivity [reference?]. Arabic is also diglossic, with a significant difference between Modern Standard Arabic (MSA) and its spoken dialects, adding another layer of linguistic complexity[reference?].

Arabic Poetry

Arabic poetry, particularly pre-modern and classical forms, is a core cultural and literary heritage in the Arab world. Poems are rich in rhetorical devices, emotional depth, and symbolic expression. They are written in a highly stylized and metrically constrained form, which enhances their expressive power but also introduces challenges for computational analysis due to non-standard grammar and poetic licenses [reference?].

Deleted: (al-ʿarūḍ)

Natural language processing (NLP)

Commented [צק9]: A short paragraph about NLP in general – what it is, what it is used for etc (1/4-1/2 page)

Natural Language Processing (NLP) is a subfield of artificial intelligence (AI) that focuses on enabling computers to understand, interpret, generate, and manipulate human language. It lies at the intersection of linguistics and computer science and plays a critical role in applications such as machine translation, sentiment analysis, speech recognition, chatbots, question answering systems, and text classification [reference?]. NLP systems rely on a combination of rule-based algorithms, statistical methods, and machine learning techniques—particularly deep learning models—to process language at different levels, from word morphology and syntax to semantic understanding and discourse analysis [reference?]. With the advent of transformer-based models like BERT and GPT, NLP has achieved significant advances in performance, bringing machines closer to understanding

Formatted: Font: Not Bold, Complex Script Font: Not Bold

Formatted: Indent: First line: 1.27 cm

Commented [צק10]: Any relevant recent text book or review

Formatted: Font: Not Bold, Complex Script Font: Not Bold

Formatted: Font: Not Bold, Complex Script Font: Not Bold the nuances, ambiguity, and complexity of human communication. As such, NLP serves not only as a technical tool but also as a bridge between computational power and human expression [reference?].

Commented [צק11]: Search for a recent review or a recent text book as a reference

Formatted: Font: Not Bold, Complex Script Font: Not Bold

Formatted: Font: Not Bold, Complex Script Font: Not Bold

Arabic NLP – challenges and importance

Deleted: Why NLP in

Deleted: is Challenging

Natural Language Processing (NLP) in Arabic faces several unique challenges [reference?]:

- Morphological richness: A single Arabic word may encode extensive grammatical information (e.g., gender, number, tense), leading to data sparsity and ambiguity.
- Orthographic ambiguity: Arabic script lacks short vowels (diacritics) in most modern texts, causing lexical and syntactic ambiguities.
- Tokenization issues: Clitics (e.g., prepositions, conjunctions, articles) often attach to words, making segmentation non-trivial.
- Lack of high-quality labeled data: Compared to English, Arabic has fewer publicly available, large-scale annotated corpora for complex NLP tasks.

Deleted: Why Arabic NLP is Interesting and Important¶

Despite the challenges, Arabic NLP is a rich and underexplored field with high cultural, academic, and technological value. It offers opportunities for preserving literary heritage, enabling cross-linguistic computational research, and empowering Arabic-speaking communities with advanced AI tools [reference?]. Rhetorical structures like *tashbīh*, metaphor (including personification) are deeply embedded in Arabic expression, making their computational study both intellectually rewarding and practically useful for applications in education, search, and cultural preservation [reference?].

Formatted: Font: Italic, Complex Script Font: Italic

Deleted: , and

Tashbīh in Arabic Rhetoric

Formatted: Font: Italic, Complex Script Font: Italic

Arabic rhetoric (*balāgha*) treats *tashbīh* (simile) as a foundational stylistic device. It is used to compare two entities by highlighting a shared attribute. The components of a classical simile are [reference?]:

1) *mushabbah* (Tenor; i.e., the object being compared)

2) *mushabbah bihi* (Vehicle; i.e., the object to which it is compared)

3) *wajh al-shabah* (the point of similarity)

4) the simile particle (e.g., كأن، مثل، كما) [1].

• Explicit similes (*tashbīh muṣarraḥ mursal*) include the simile particle. When all four components of a simile—vehicle, tenor, particle, and point of similarity—are present, the simile is called *tashbīh tāmm* (complete simile). For example: 'Zayd is like a lion in bravery.' [1].

• Implicit simile (*tashbīh muʾakkad*) omits the simile particle and relies on context to imply similarity. For example, a poet might say: 'He is a lion in battle', implying courage without the word 'like' [2].

• Classical rhetoricians and lexicographers elaborated on various types of simile and their rhetorical effects [3][4].

Advances in NLP for Arabic

Recent Arabic NLP has benefited from large-scale pre-trained models such as AraBERT. These models can capture contextual nuances and semantic relationships, offering promise for detecting hidden similes [10].

In English NLP, similar challenges have been addressed through transformer-based models for metaphor and figurative language detection, such as DeepMet and MelBERT [ref?]. These models leverage contextual embeddings to capture nuanced semantic

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Deleted: *ī*

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Commented [עת12]: You previously used the term 'simile word.' Please be consistent in your terminology; I prefer the term 'simile particle'.

Deleted: comparison

Deleted: Explicit similes (*tashbīh muṣarraḥ*) state all four components clearly, e.g., 'Zayd is like a lion in bravery'

Deleted: s

Deleted: *ḍimnī*

Deleted: comparison

Deleted: y

Formatted: Font: Italic, Complex Script Font: Italic

Deleted: such as al-Jurjānī in 'Asrār al-Balāgha'

Deleted: such as al-Zubaidī in 'Tāj al-ʿArūs'

Deleted: Related Work¶
• The Arabic Rhetoric Identifier (REI), developed at the University of Haifa, is a rule-based system for detecting rhetorical devices in Arabic poetry, including *tashbīh*. It uses explicit pattern-matching rules combined with NLP tools [5]. ¶
• The system operates on a large Arabic Poetry Corpus containing over 26,000 poems spanning 10 centuries. Only a subset of this corpus has been manually annotated due to the complexity of rhetorical analysis [6]. ¶
The • Tools such as AlKhalil Morpho Sys 2 and MADAMIRA are integrated for morphological analysis and tokenization, which help detect comparative particles and relevant linguistic structures [7][8].¶
• Limitations: Rule rule-based approaches struggle with detecting implicit *tashbīh* due to their reliance on explicit markers. They require constant rule updates and are unable to generalize well to creative or indirect comparisons [5][9].¶

Deleted: • R

relationships that go beyond surface-level word usage. Their success in identifying implicit metaphors in English prose suggests that similar architectures are well-suited for handling the complexity and subtlety of Arabic poetic expressions [ref?]. This relevance stems from the shared need to interpret figurative meaning, handle syntactic variation, and model context-dependent rhetoric—core challenges in both languages [11][12].

Detailed Overview of Arabic NLP Tools for Morphological Analysis:

• AlKhalil Morpho Sys 2: This is a robust Arabic morphological analyzer developed to handle both vocalized and unvocalized text. It uses a large internal lexicon and a set of rules to provide information such as the root, pattern, POS tags, and affixes of Arabic words. Pros: High accuracy on Classical Arabic; extensive lexicon coverage. Cons: Performance may degrade with dialectal or modern informal texts [7].

• MADAMIRA: An evolution of MADA and AMIRA, MADAMIRA is a Java-based tool offering tokenization, POS tagging, lemmatization, and diacritization. It is optimized for Modern Standard Arabic and provides a user-friendly interface with both batch and web modes. Pros: Fast, comprehensive, and widely used in academic NLP projects. Cons: Less effective for historical or poetic texts where language is highly figurative or archaic [8].

• BAMA (Buckwalter Arabic Morphological Analyzer): This was one of the earliest tools for analyzing Arabic morphology. It provides possible morphological analyses for Arabic words based on a large dictionary of stems and affixes. Pros: Fundamental and influential in early Arabic NLP research. Cons: Generates many ambiguous outputs and lacks disambiguation capabilities; limited support for contextual inference [13].

• SAMA (Standard Arabic Morphological Analyzer): An improved version of BAMA developed by the Linguistic Data Consortium. It refines and expands the analysis with updated lexicons and additional grammatical coverage. Pros: Better performance than BAMA and integrated in many academic projects. Cons: Still suffers from ambiguity and requires external tools for disambiguation [14].

Commented [צק16]: why it is relevant?

Deleted: • In English NLP, similar challenges have been addressed with transformer-based models for metaphor detection, such as DeepMet and MelBERT. These methods inspire the application of similar architectures to Arabic poetic text

• Farasa: A fast and efficient Arabic segmentation and POS tagging tool designed especially for processing large-scale data quickly. Pros: Speed and scalability; good for real-time applications. Cons: Less accurate in deeper morphological analysis compared to tools like MADAMIRA or AlKhalil [15].

• AMIRA: A tool for Arabic tokenization and base phrase chunking, offering rapid POS tagging with reasonably high accuracy. Pros: Lightweight and good for simple NLP pipelines. Cons: Outperformed by MADAMIRA and lacks support for complex morphological tasks [16].

Related Work

The Arabic Rhetoric Identifier (REI), developed at the University of Haifa, is a rule-based system for detecting rhetorical devices in Arabic poetry, including *tashbīh*. It uses explicit pattern-matching rules combined with NLP tools [5]. The system operates on a large Arabic Poetry Corpus containing over 29,000 poems spanning 10 centuries. Only a subset of this corpus has been manually annotated due to the complexity of rhetorical analysis [6]. The rule-based approaches struggle with detecting implicit *tashbīh* due to their reliance on explicit markers. They require constant rule updates and are unable to generalize well to creative or indirect comparisons [5][9].

Commented [צק17]: You need to search google scholar for related studies, this is not enough.

Commented [צק18]: should be after the "advances in arabic NLP" and extended with real related work - search google scholar for automatic identification of rhetoric elements in Aarabic - see https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=automatic+identification+of+rhetoric+elements+in+Aarabic&btnG=

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Formatted: Normal, No bullets or numbering

Belinkov and Glass [17] analyze the application of deep learning techniques to Arabic natural language processing, focusing on the challenges introduced by Arabic's morphological complexity and orthographic variability. Their study explores how segmentation strategies—such as tokenization, affix separation, and normalization—affect model performance across various NLP tasks. They highlight the impact of diglossia and the difficulty in capturing semantic context when dealing with inflected forms, which often results in sparse data representations. The authors stresses the importance of careful pre-processing and morphological analysis to improve deep learning model outcomes when working with Arabic texts [17].

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Deleted: paper

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Abdul-Raof Hussein [18] provides a comprehensive rhetorical and pragmatic analysis of *tashbīh* in Arabic, distinguishing it from other figurative devices through a structured framework. His work outlines the four components of *tashbīh* —the likened-to (*al-mushabbah*), the likened (*al-mushabbah bihi*), the *tashbīh* feature, and the *tashbīh* element—and classifies different types such as implicit, reverse, compound, and imaginary forms. He emphasizes the rhetorical power of succinct and stylistically elevated expressions, especially those that omit explicit comparison markers. This typology not only deepens the understanding of *tashbīh* in classical Arabic discourse but also serves as a valuable reference for formulating linguistic rules to support automated *tashbīh* detection in poetry [18].

Ibrahim and Elghazaly [19] proposed a hybrid Arabic text summarization model that combines Rhetorical Structure Theory (RST) [reference?] with the Vector Space Model (VSM) [reference?] to enhance semantic extraction and paragraph ranking. The first component of their model applies RST to segment text and identify rhetorical relationships (e.g., justification, contrast), classifying text units as nucleus or satellite. The second component uses cosine similarity in a vector space representation to rank and refine the most relevant segments. Their evaluation of Arabic news articles showed a precision improvement from 56.3% using RST alone to 71.6% using the hybrid method. This work demonstrates the value of rhetorical analysis and cue-based segmentation in Arabic NLP tasks, which can be informative for constructing linguistically aware models in poetic analysis [19].

Khan and Rahman [20] conducted a large-scale comparative study between Quranic verses and pre-Islamic Arabic poetry using modern NLP methods. They applied CAMeLBERT-based embeddings and clustering algorithms to analyze semantic and stylistic patterns across over 10,000 Quranic sentences and 38,000 poetic verses. Their results show that even the most poetic Quranic verses (Makki) form distinct clusters from classical Arabic poetry, supporting the claim of the Quran's rhetorical uniqueness. This

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Deleted: (2013)

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Deleted: on

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Commented [צק19]: או שתציין שנה בכל מקום או שלא,
תהיה עקבי

Deleted: (2025)

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Formatted: Indent: First line: 0 cm

work provides computational evidence for a long-standing debate and highlights the power of embedding-based analysis in modeling Classical Arabic texts. The clustering-based methodology may also inspire future tasks in Arabic literary analysis, such as distinguishing rhetorical devices like tashbīh, and could inform the design of machine learning pipelines for identifying implicit structures in poetry [20].

3. Research Goals and Questions

The central goal of this research is to develop, demonstrate, and evaluate automatic methods for identifying *implicit tashbīh* in classical Arabic poetry. This task lies at the intersection of classical Arabic rhetoric and modern natural language processing (NLP), aiming to bridge centuries-old poetic tradition with contemporary computational techniques.

Research Questions

The proposed research seeks to answer the following key questions:

- RQ1: How can machine learning models accurately identify *tashbīh* in Arabic poetry beyond rule-based lexical pattern matching?

To address the above research question, we will address the following questions

- RQ1.1: What are the linguistic and semantic characteristics that distinguish implicit *tashbīh* from other poetic statements such as metaphors?
- RQ1.2: How effective is a transformation-based method in creating high-quality examples of implicit *tashbīh* for training data augmentation?
- RQ1,3: To what extent can contextual language models like AraBERT generalize to detect implicit rhetorical devices in highly stylized, classical Arabic text?
- RQ1.4: What are the limitations and challenges in modeling deep figurative meaning computationally in the context of Arabic rhetoric?

Deleted: hidden

Deleted: (implicit similes)

Deleted: , where the comparison is not signaled by explicit linguistic markers

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Commented [צק21]: You defined one abstract research questions - these are sub questions that addressing them will provide an answer to the research question. in the research plan - make sure to plan tasks that provide answers to each specific research question

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Formatted: Normal, Indent: Before: 1.27 cm, No bullets or numbering

Commented [צק22]: I did not see a tsak that answers this question

Deleted: hidden

Deleted: or

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Commented [צק23]: The tasks in the plan answer these questions

Moved down [1]: Research Goals¶
To achieve this overarching aim, the research will pursue the following specific objectives: ¶
Construct a labeled dataset of Arabic poems containing both explicit and hidden tashbīh, based on an initial rule-based classifier and manual verification. ¶
Fine-tune transformer-based Arabic language models (e.g., AraBERT) to distinguish poems with tashbīh from those without, using both explicit and implicit instances. ¶
Design a transformation framework to manually or semi-automatically convert explicit tashbīh examples into implicit ones by removing or hiding the comparison marker while preserving semantic meaning and poetic structure. ¶
Develop a classification model capable of identifying tashbīh ḍimnī based on semantic features, syntactic context, and poetic discourse patterns. ¶
Evaluate the models' performance using standard metrics (accuracy, precision, recall, F1-score), alongside qualitative analysis of correctly and incorrectly classified examples.¶

Deleted: Research Questions¶
The proposed research seeks to answer the following key questions: ¶
RQ1: Can machine learning models accurately ident ... [1]

4. Tools and Methods

This section outlines the methodological approach and computational tools used to achieve the research goals of identifying implicet *tashbīh* in Arabic poetry. The methodology integrates rule-based extraction, data preprocessing, model fine-tuning, and evaluative analysis, all grounded in the principles of Arabic rhetoric and modern NLP.

Deleted: hidden

Deleted: (similes)

Formatted: Font: Italic, Complex Script Font: Italic

Overview of the Research Pipeline

Commented [צק25]: How do you answer RQrq1.1?

Commented [fo26R25]: See stage 5

The research pipeline consists of the following stages:

1) Initial Rule-Based Detection of Tashbīh:
   - A rule-based system (converted from prior Java code to Python) was employed to detect explicit *tashbīh* in classical Arabic poems.
   - These rules searched for explicit simile particles (e.g., كأن، مثل، كما، كـ) and comparison-related verbs (e.g., شبه، ظن، خال) to identify overt similes.

Formatted: Font: Italic, Complex Script Font: Italic

2) Dataset Construction and Labeling:
   - Based on the output of the rule-based detection, a binary-labeled dataset was built: poems with *tashbīh* vs. poems without *tashbīh*.
   - Manual verification was conducted to ensure quality and correctness, especially for edge cases.

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

3) Text Cleaning and Preprocessing:
   - Diacritics (*harakāt*) was removed for consistency.
   - Poems was normalized (e.g., converting "ى" to "ي", removing punctuation).
   - Tokenization was applied using tools such as Farasa and MADAMIRA for further segmentation and linguistic analysis.

Deleted: were

Formatted: Font: Italic, Complex Script Font: Italic

Deleted: were

Deleted: was

4) Model Training Using AraBERT:
   - The cleaned and labeled dataset was used to fine-tune AraBERT, a transformer-based model pre-trained on large Arabic corpora.

Deleted: was

- A subset of poems was used for training, while the rest (over 28,000 poems) served as a test set.
- The fine-tuned classifier achieved around 96% accuracy, suggesting strong discriminative power in identifying tashbīh-laden verses.

5) Linguistic Guidance on Implicit Similes:
- In collaboration with experts from the Arabic Language Department at the University of Haifa, a focused analysis was conducted to identify the key linguistic and semantic features that distinguish implicit tashbīh from metaphors and literal statements.
- The results of this collaboration helped define annotation criteria and guided the construction of a labeled dataset for further modeling.

Formatted: Font: 12 pt, Complex Script Font: 12 pt

6) Implicit Tashbīh Generation and Classification:
- An automatical process will be used to convert explicit *tashbīh* into implicit forms, removing the simile particle while preserving meaning and rhetorical impact.
- These transformed examples were added to the dataset to train a second model focused on detecting implicit *tashbīh*.
- Feature extraction involved semantic embeddings, syntactic patterns, and context windows.

Deleted: manual

Deleted: was

Formatted: Font: Italic, Complex Script Font: Italic

Deleted: comparison

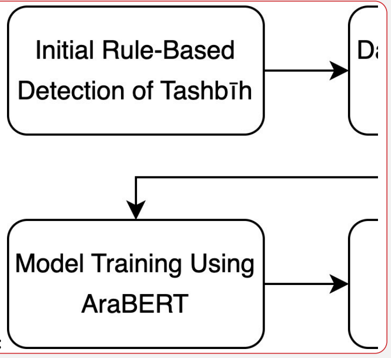
Deleted: word

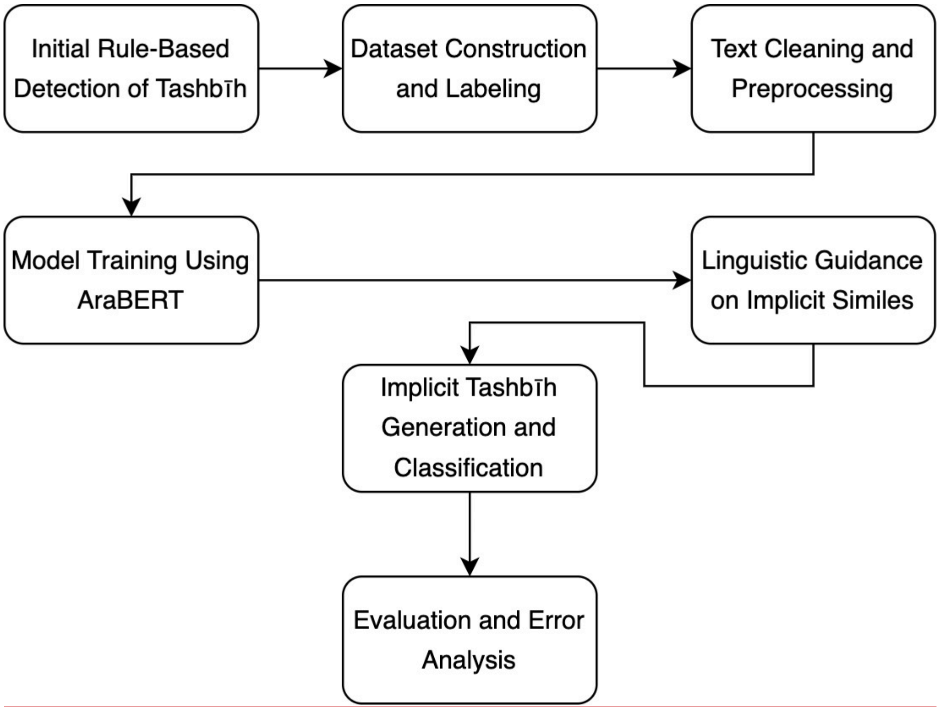
Deleted: hidden

Formatted: Font: Italic, Complex Script Font: Italic

7) Evaluation and Error Analysis:
- The model was evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix.
- Special attention was given to false positives and false negatives to analyze whether the model captured the poetic simile beyond surface-level cues.

Deleted:


Initial Rule-Based Detection of Tashbīh
Da
Model Training Using AraBERT


Initial Rule-Based Detection of Tashbīh
Dataset Construction and Labeling
Text Cleaning and Preprocessing
Model Training Using AraBERT
Linguistic Guidance on Implicit Similes
Implicit Tashbīh Generation and Classification
Evaluation and Error Analysis

Tools and Frameworks

Several key tools and libraries are employed in this research:

- AlKhalil Morpho Sys 2: For root extraction, morphological analysis, and POS tagging on classical Arabic poetry.
  Strength: Excellent performance on vocalized/unvocalized classical texts.
  Weakness: Limited support for dialectal or informal Arabic.
- MADAMIRA: A comprehensive tool for morphological disambiguation and tokenization of Modern Standard Arabic.
  Strength: Integrates multiple preprocessing steps into a unified tool.
  Weakness: Less robust on non-standard poetic constructions.

- Farasa: Lightweight Arabic segmenter and POS tagger, useful for fast preprocessing.
  Strength: High speed, scalable.
  Weakness: Lower precision in classical or rhetorical contexts.
- BAMA & SAMA: Classical tools from the Linguistic Data Consortium for Arabic morphological analysis.
  Strength: Foundational systems with wide lexical coverage.
  Weakness: Lack of disambiguation and contextual awareness.
- AraBERT: A deep transformer-based model pre-trained on Arabic corpora.
  Strength: Strong contextual understanding, suitable for nuanced NLP tasks.
  Weakness: Requires careful fine-tuning and large GPU resources.
- scikit-learn & PyTorch: Used for training, evaluating, and visualizing model outputs and confusion matrices.

5. Initial Experiments and Results

The initial experiments focused on evaluating the performance of a fine-tuned AraBERT model in classifying Arabic poems into two categories: poems that contain *tashbīh* and those that do not.

Deleted: (similes)

Formatted: Font: Italic, Complex Script Font: Italic

Dataset and Setup

- The dataset was constructed by applying a rule-based detection system to label Arabic poems as containing *tashbīh* or not.
- From the full dataset, only 5% of the data was used for training, while the remaining 95% was split for validation and testing. Specifically:

  ```
  train_df, temp_df = train_test_split(df, test_size=0.95, stratify=df['label'], random_state=42)
  val_df, test_df = train_test_split(temp_df, test_size=0.98, stratify=temp_df['label'], random_state=42)
  ```
- This small training set was chosen due to the large size of the dataset and the high computational cost of training transformer-based models like AraBERT.
- Despite using only a small fraction of the data for training, the model achieved high performance, indicating the strength of contextual embeddings in Arabic NLP.

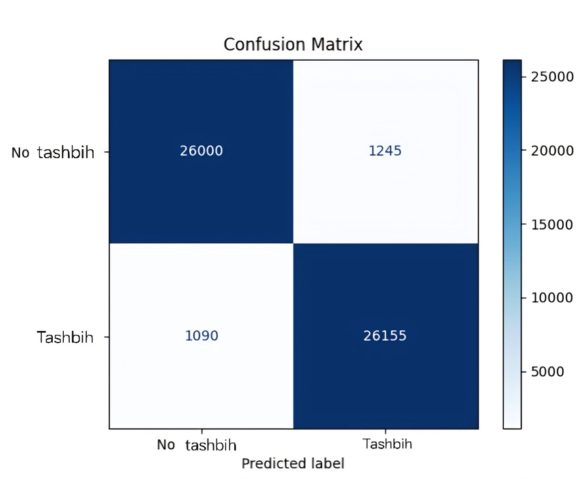
Formatted: Font: Italic, Complex Script Font: Italic

Results

The model's performance on the test set was evaluated using standard classification metrics:

- Accuracy: 95.71%
- Precision: 95.46%
- Recall: 95.99%

These scores show that the model is highly reliable in distinguishing between simile and non-simile poetry lines, even with limited training data.



Commented [עח27]: Correct the word tashbeeh into tashbih

The confusion matrix confirms the model's balanced ability to capture both classes with relatively low error rates.

Interpretation and Potential

The results confirm that AraBERT and similar transformer models are capable of learning complex semantic patterns in Arabic poetry. Even though the current dataset consists mostly of explicit *tashbīh* (with visible markers such as كأن, مثل), the model's success in capturing these patterns provides strong evidence that it understands rhetorical context rather than relying solely on keyword spotting.

Formatted: Font: Italic, Complex Script Font: Italic

This outcome suggests that such models can potentially be extended and fine-tuned further to identify implicit *tashbīh* — where no clear simile particle exists — by learning more abstract patterns of simile. These results form a solid baseline and motivate

Deleted: (tashbīh ḍimnī)

Deleted: comparative

Deleted: word

Formatted: Font: Italic, Complex Script Font: Italic

Deleted: analogy

the next stage of the thesis: focusing on generating, labeling, and detecting implicit similes.

Deleted: hidden

6. Planned Research and Future Work

While the initial results demonstrate strong performance in detecting explicit *tashbīh*, the ultimate objective of this research is to extend this capability to implicit *tashbīh*—those similes that lack overt comparative markers and rely on deeper semantic or contextual inference.

To achieve the overarching aim and to answer the specific research questions, the research will include the following tasks:

- Construct a labeled dataset of Arabic poems containing both explicit and implicit *tashbīh*s, based on an initial rule-based classifier and manual verification.
  - Manual Transformation:

    A curated set of poems with explicit *tashbīh* will be manually rewritten to remove the simile marker (e.g., "مثل", "كأن") while preserving the core meaning and rhetorical structure. This will simulate how a poet might express a simile implicitly.
  - Rule-Based and Semi-Automatic Transformation:

    Exploratory methods will be developed to automatically remove simile markers from known *tashbīh* lines, creating a larger set of implicit-style examples for training.
  - Building a New Dataset:

    The transformed implicit *tashbīh* lines will be labeled and verified by Arabic language experts. A balanced dataset will be constructed containing explicit *tashbīh*, implicit *tashbīh*, and non-*tashbīh* lines.

Commented [צק28]:

Commented [צק29]: check the tasks and provide a note if a task was already (partially) done (as described in the initial results)

Moved (insertion) [2]

Formatted: Font: Italic, Complex Script Font: Italic

Deleted: or hidden

Deleted: (تشبيه ضمني)

Formatted: Font: Italic, Complex Script Font: Italic

Deleted: Research Goals¶

Deleted: is

Deleted: ,

Formatted: Level 2, Indent: First line: 0 cm

Deleted: pursue the following specific objectives

Deleted: hidden

Formatted: Font: Italic, Complex Script Font: Italic

Moved up [2]: <#>While the initial results performance in detecting explicit tashbih, the ultimate objective of this research is to extend this capability to implicit or hidden tashbīh (تشبيه ضمني)—those similes that lack overt comparative markers and rely on deeper semantic or contextual inference.¶

Deleted: <#>Fine-tune transformer-based Arabic language models (e.g., AraBERT) to distinguish poems with tashbīh from those without, using both explicit and implicit instances. ¶
Design a transformation framework to manually or semi-automatically convert explicit tashbīh examples into implicit ones by removing or hiding the comparison marker while preserving semantic meaning and poetic structure. ¶
Develop a classification model capable of identifying tashbīh ḍimnī based on semantic features, syntactic context, and poetic discourse patterns. ¶
Evaluate the models' performance using standard metrics (accuracy, precision, recall, F1-score), alongside qualitative analysis of correctly and incorrectly classified examples.¶

... [2]

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

- Fine-tune transformer-based Arabic language models (e.g., AraBERT) to distinguish poems with *tashbīh* from those without, using both explicit and implicit instances.

Formatted: Font: Italic, Complex Script Font: Italic

- Design a transformation framework to manually or semi-automatically convert explicit *tashbīh* examples into implicit ones by removing or hiding the comparison marker while preserving semantic meaning and poetic structure.

Formatted: Font: Italic, Complex Script Font: Italic

- Develop a classification model capable of identifying implicit *tashbīh* based on semantic features, syntactic context, and poetic discourse patterns.

Deleted: ḍimnī

Formatted: Font: Italic, Complex Script Font: Italic

- Developing Enhanced Detection Models:

  AraBERT or similar language models will be fine-tuned on the newly constructed dataset. Feature engineering will also be applied to enhance detection, such as sentence structure, syntactic trees, or rhetorical embeddings.

- Evaluate the models' performance using standard metrics (accuracy, precision, recall, F1-score), alongside qualitative analysis of correctly and incorrectly classified examples.

- Evaluation and Benchmarking:

  The model will be evaluated using accuracy, precision, recall, and confusion matrices. Qualitative analysis will also be conducted to interpret model decisions, particularly in ambiguous or misclassified cases

7. Contribution

This research aims to make significant contributions to the intersection of Arabic linguistics, digital humanities, and natural language processing. By focusing on the detection of both explicit and implicit tashbīhs in classical Arabic poetry, this work bridges centuries-old rhetorical theory with modern AI tools. Moreover, the methodologies and

Deleted: This

Formatted: Indent: First line: 0.63 cm models developed in this study hold the potential to be adapted for analyzing rhetorical patterns in other literary traditions, such as Hebrew and English, thereby opening the door to cross-linguistic and comparative literary research using NLP.

Linguistic and Theoretical Contributions

- Modeling implicit Rhetoric: The thesis explores and formalizes the concept of *tashbīh mu'akkad* (implicit simile) in computational terms—an area that has received very limited attention in Arabic NLP or literary analysis.
- Dataset Curation: It introduces a novel dataset of Arabic poetry annotated for both explicit and implicit similes, offering a resource that can support future research in figurative language processing for Arabic.
- Augmentation of Classical Theory: Through the transformation of explicit into implicit *tashbīh*, this work contributes to the study of Arabic *balāgha* by demonstrating new rhetorical equivalencies grounded in semantic preservation.

Technical and Computational Contributions

- Pipeline for *tashbīh* Detection:

  The research presents a hybrid methodology that combines rule-based techniques with deep learning models (e.g., AraBERT) to achieve high-accuracy classification of poetic verses.
- Benchmarking NLP Models on Arabic Poetry:

  It evaluates the performance of transformer-based models on a uniquely complex domain—classical Arabic poetry—which is morphologically rich and syntactically free-form.
- Demonstrating Semantic Learning:

  By achieving high accuracy even on a small training subset, the model shows that transformer models can meaningfully capture rhetorical intent in Arabic, setting the foundation for detecting deeper forms of metaphor and simile.

Commented [עח31]: Mention that the results may be developed to match other literatures in other languages; such as Hebrew and English.

Deleted: research aims to make significant contributions to the intersection of Arabic linguistics, digital humanities, and natural language processing. By focusing on the detection of both explicit and implicit *tashbīh*s (similes) in classical Arabic poetry, this work bridges centuries-old rhetorical theory with modern AI tools.

Deleted: Hidden

Deleted: ḍimnī

Formatted: Font: Italic, Complex Script Font: Italic

Deleted: *h*

Formatted: Font: Italic, Complex Script Font: Italic

Formatted: Font: Italic, Complex Script Font: Italic

Deleted: *T*

- Open-Source Codebase and Tools:

  The work includes clean, reusable Python code (available on GitHub) for processing, training, and evaluating Arabic poetry classification tasks—serving the community and future researchers.

8. References

1. al-Hāshimī, A. (1998). Jawāhir al-Balāghah. Dar al-Kutub al-ʿIlmiyya.
2. al-Jurjānī, ʿA. (1954). Asrār al-Balāgha. Istanbul.
3. al-Zubaidī, M. M. (1975). Tāj al-ʿArūs. Kuwait.
4. Ali, A. & Ahmad, H. (2015). Classical Arabic rhetoric. Journal of Arabic Literature.
5. Abd Alhadi, H., Hussein, A. A., & Kuflik, T. (2023). Automatic identification of rhetorical elements. Digital Humanities Quarterly.
6. Hussein, A. A. (2015). The Rhetorical Fabric of the Traditional Arabic Qaṣīda. Wiesbaden: Harrassowitz Verlag.
7. Boudchiche, M. et al. (2017). AlKhalil Morpho Sys 2. Journal of King Saud University.
8. Pasha, A. et al. (2014). MADAMIRA. Proceedings of LREC.
9. Hussein et al. (2023). Limitations of REI. Digital Humanities Quarterly.
10. Antoun, W. et al. (2020). AraBERT. Workshop on Open-Source Arabic Tools.
11. [Su, Y. et al. (2020). DeepMet. ACL.
12. Leong, C. et al. (2020). MelBERT. EMNLP.
13. Buckwalter, T. (2002). Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania.
14. Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2010). Standard Arabic Morphological Analyzer (SAMA) Version 3.1 (LDC2010L01). Linguistic Data Consortium.
15. 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 (pp. 11‑16). Association for Computational Linguistics.
16. Diab, M. (2009). Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

Commented [צק32]: Please order the references alphabetically

Deleted: [1]

Formatted: Font: 12 pt, Complex Script Font: 12 pt, Italian

Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.63 cm + Indent at: 1.27 cm

Deleted: [2]

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Deleted: [3]

Formatted: Font: 12 pt, Complex Script Font: 12 pt, Italian

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Deleted: [4]

Deleted: [5]

Deleted: [6]

Deleted: [7]

Deleted: [8]

Formatted: Font: 12 pt, Complex Script Font: 12 pt, Italian

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Deleted: [9]

Deleted: [10]

Deleted: 11]

Formatted: Font: 12 pt, Complex Script Font: 12 pt, Italian

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Deleted: [12]

Deleted: [13]

Deleted: [14]

Deleted: [15]

Formatted: Font: 12 pt, Complex Script Font: 12 pt, Italian

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Deleted: [16]

17. Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics, 7, 49-72.
18. Abdul-Raof, H. (2006). Arabic rhetoric: A pragmatic analysis. Routledge.
19. Ibrahim, A., & Elghazaly, T. (2013, November). Improve the automatic summarization of Arabic text depending on Rhetorical Structure Theory. In 2013 12th Mexican International Conference on Artificial Intelligence (pp. 223-227). IEEE.
20. Khan, R. S., & Rahman, A. (2025, April). Computationally Distinguishing Quran and Pre-Islamic Arabic Poetry. In 2025 Eighth International Women in Data Science Conference at Prince Sultan University (WiDS PSU) (pp. 1-6). IEEE.

Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.63 cm + Indent at: 1.27 cm

Formatted: Font: 12 pt, Complex Script Font: 12 pt

Page 13: [1] Deleted צבי קופליק **6/1/25 9:25:00 AM**

Page 21: [2] Deleted צבי קופליק **6/1/25 9:29:00 AM**