Journal Title
XX(X):1–7
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

MEMpHIS: Towards a New Benchmark for Arabic Figurative Speech Classification

Zouheir Banou[1] Sanaa El Filali[1] El Habib Benlahmar[1] Fatima-Zahra Alaoui[1] and Laila Eljiani[1]

Abstract

The automatic detection and interpretation of figurative language remain critical challenges in Natural Language Processing (NLP), particularly for languages with limited annotated resources. In this work, we introduce 6- Figure, a novel dataset designed to facilitate the computational analysis of figurative speech in Arabic. Our dataset deals with figures of speech metaphors, idioms, similes, metonymy, hyperboles and euphemisms, annotated at sentence-level. The dataset is sourced from various previous research works, ensuring a good quality benchmark for Arabic language. We gather and translate benchmark figurative language datasets from English to Arabic, then evaluate their relevance by training and testing several NLP models on the resulting corpus. We provide baseline results using transformer-based models and recurrent models, highlighting key challenges and areas for future research. This dataset serves as a valuable resource for advancing figurative language understanding and improving NLP models in Arabic. The dataset and annotation guidelines will be publicly released to encourage further research.

Keywords

Figurative Speech, NLP, Dataset, Computational Linguistics, Annotation, Low-Resource Languages

Introduction

Figurative language (FL) is a fundamental aspect of human communication, allowing speakers to convey meaning beyond the literal interpretation of words or is when words, phrases, and expressions have meanings that are different from their literal ones [1]. Metaphors, idioms, similes, metonymy, hyperboles, euphemisms, and other expressions enhance language by highlighting abstract ideas, enhancing rhetorical depth, and improving communication [2].However, the complexity and contextual dependencies of figurative speech pose significant challenges for Natural Language Processing (NLP) systems, particularly in languages with limited computational resources [3]. While extensive research has been conducted on figurative language detection in English and other widely studied.

Existing research on Arabic figurative language is often hindered by the lack of well-annotated datasets, making it difficult to develop and evaluate robust computational models. While some studies have attempted to compile figurative language corpora, these resources are often limited in scope, annotation consistency, or accessibility. To bridge this gap, we introduce **MEMpHIS**, a novel dataset designed specifically for the computational analysis of figurative speech in Arabic. This dataset encompasses six major categories of figurative expressions—*metaphors, idioms, similes, metonymy, hyperboles,* and *euphemisms*—all annotated at the sentence level.

We construct the **MEMpHIS** dataset by compiling and translating existing benchmark datasets for figurative language detection from English to Arabic. We evaluate the resulting dataset using multiple NLP models to assess its quality and usability in Arabic figurative language processing. This enables scalability and ensures a consistent annotation framework. To further validate the dataset's quality, we provide baseline experimental results using both **transformer-based** and **recurrent models**, evaluating their performance on figurative speech classification.

The contributions of this work are as follows:

- We introduce **MEMpHIS** *, a high-quality Arabic dataset dedicated to figurative speech analysis, covering six key figurative categories.

We establish benchmark results using modern NLP models, including transformer-based architectures and recurrent neural networks (RNNs), highlighting key challenges in figurative speech classification.

We publicly release **MEMpHIS** and its annotation guidelines to facilitate future research in Arabic figurative language processing.

The rest of this paper is structured as follows: section reviews related work on figurative language detection and Arabic NLP resources. section presents the methodology, detailing dataset construction, atechniquesch- niques, and evaluation protocols. section provides experimental results and analysis, and section concludes the paper with a discussion of challenges and future directions.

[1] LIAS Laboratory - Faculty of Sciences Ben M'Sick, Hassan II University, Casablanca, Morocco

Corresponding author:
Zouheir Banou
Email: zouheir.banou4-etu@etu.univh2c.ma

*`https://github.com/ZOUHEIRBN/MEMPHIS-data`

Related Work

Research on figurative language processing in NLP has expanded across various domains, covering dataset construction, annotation techniques, and computational modeling. This section reviews key contributions, identifying challenges and positioning our work within the field.

The *Waterloo Rhetorical Figure Corpus* provides a structured XML-based annotation scheme for rhetorical figures, enabling computational analysis of figurative speech [1]. This dataset includes annotated instances of various rhetorical figures such as metaphor, antithesis, and mesodiplosis, offering a valuable resource for studying patterns in literary and argumentative texts. However, the dataset's reliance on manual annotation limits its scalability, making it challenging to extend to larger corpora automatically. The *Figurative Corpus for Arabic Language (FCAL)* introduces a manually annotated dataset dedicated to Arabic figurative expressions, particularly hyperboles and similes [2]. Designed primarily for sentiment analysis, this resource provides labeled data for computational models to differentiate between literal and exaggerated statements. However, the dataset's scale remains relatively small, restricting its applicability in large-scale deep learning applications. Additionally, no deep learning models were benchmarked on it, leaving its practical effectiveness for modern NLP architectures unexplored. The *Arabic Idiomatic and Literal Statements Dataset* provides a structured resource for studying idiomatic expressions in Arabic, distinguishing between their figurative and literal meanings [3]. This dataset was developed using a deep learning-based annotation approach, automating the classification of idioms within different contexts. While this method improves efficiency and consistency in annotation, the dataset's coverage of dialectal variations remains limited, affecting its generalizability across diverse Arabic-speaking regions. The *Arabic Figurative Sentiment Analysis (AFSA) Corpus* provides a sentiment-annotated dataset for Arabic, focusing on hyperboles, similes, and sarcasm [4]. The dataset was manually annotated and used to evaluate classical machine learning classifiers such as Naïve Bayes and Logistic Regression. While these models demonstrated reasonable performance, no deep learning models were tested, leaving unexplored the potential benefits of modern NLP architectures in figurative sentiment analysis. The *Chiasmus and Antimetabole Corpus* investigates structured rhetorical figures in German texts, focusing on chiasmus and antimetabole patterns [5]. The dataset was constructed using a combination of rule-based techniques and syntactic parsing to identify instances of these rhetorical devices. While the approach is effective in structured literary texts, it suffers from high false positive rates due to syntactic ambiguities, limiting its reliability for broader linguistic applications. The *Euphemism in the Quran: A Corpus-based Study* presents an Arabic dataset annotated for euphemisms, focusing on linguistic and cultural nuances within religious texts [6]. This resource provides valuable insights into how euphemisms are employed in sacred discourse, making it a significant contribution to Arabic figurative language studies. However, the dataset has not been evaluated using computational models, limiting its immediate applicability in NLP tasks requiring automated figurative language detection. The *FigMemes Dataset* introduces a multimodal corpus for analyzing figurative language in memes, integrating image-text pairs annotated for sarcasm, metaphor, and exaggeration [7]. This dataset broadens the scope of figurative language research beyond traditional text-based resources, allowing for the study of how visual and textual elements interact in conveying figurative meaning. However, the dataset presents challenges in aligning visual and textual cues, which complicates automatic detection and interpretation using NLP models. The *FLUTE Dataset* presents a corpus designed to provide textual explanations for figurative language, covering multiple categories such as sarcasm, idioms, metaphors, and similes [8]. By offering human-written explanations alongside figurative instances, the dataset enhances interpretability in NLP tasks, allowing models to learn both classification and reasoning aspects of figurative speech. However, despite its comprehensive coverage of figurative categories, the dataset lacks multilingual support, limiting its applicability to non-English contexts. The *Multilingual and Multicultural Figurative Language Dataset* introduces a cross-linguistic corpus covering various forms of figurative speech across multiple languages [9]. By enabling comparative analysis of figurative expressions in diverse cultural contexts, this dataset contributes to a broader understanding of language-specific figurative patterns. However, maintaining annotation consistency across different linguistic structures remains a key challenge, which may impact the dataset's effectiveness in cross-lingual NLP applications. The *ArSarcasm Dataset* introduces a publicly available Arabic sarcasm corpus, comprising 10,547 tweets, with 1,682 labeled as sarcastic and 8,865 as non-sarcastic [10]. The dataset was constructed by reannotating existing sentiment datasets, namely SemEval 2017 and ASTD, using Figure-Eight crowd-sourced workers for manual binary annotation (sarcastic vs. non-sarcastic). The *German Antithesis Corpus* explores the use of German Wiktionary antonym resources to construct an antithesis detection dataset [11]. The dataset integrates rule-based detection techniques using PoS tagging and antonym retrieval to identify rhetorical antithesis in German texts. While this approach provides structured annotations, it is constrained by the incompleteness of Wiktionary's antonym resources and the complexity of German morphology, which impacts detection accuracy. The *WIMCOR Dataset* is a corpus for metonymy detection, introducing sequence labeling as an alternative to traditional classification approaches [12]. The dataset focuses on metonymy resolution using Conditional Random Fields (CRF) and GloVe embeddings, compared against deep learning models. While effective in structured text, the dataset lacks large-scale manually annotated corpora and does not evaluate transformer-based architectures such as BERT, limiting its applicability to modern NLP tasks. [13] presents an approach to detecting chiasmus and antimetabole using a linguistically informed ranking model. The study introduces the Annotated Chiasmus Corpus, a manually annotated dataset that ranks chiasmus candidates based on syntactic and lexical features. While this method improves precision in detecting rhetorical figures, it remains time-consuming and requires linguistic expertise. Moreover, the absence of large-scale automated systems for chiasmus detection limits its applicability in broader NLP tasks.

The *Rhetorical Figures in Argumentation Dataset* presents a collection of rhetorical figures from argumentation texts, primarily focused on detecting instances of antithesis, chiasmus, and parallelism [14]. This dataset provides a structured annotation scheme for rhetorical devices and enables computational models to analyze their usage in persuasive writing. However, the dataset is limited in size and focuses mainly on formal argumentative texts, which restricts its applicability to other discourse genres, such as social media and everyday conversations.

The *RetFig Ontology Corpus* introduces an ontological framework for rhetorical figures, incorporating a structured annotation scheme to capture instances of antithesis, chiasmus, parison, epanaphora, and epistrophe [15]. This dataset was constructed using parliamentary transcripts and political debates, with manual annotation guided by the RetFig ontology. The study focuses on knowledge-based reasoning and rule-based classification, highlighting challenges in modeling rhetorical figures computationally. However, the dataset remains limited in scope, excluding broader figurative language categories such as metaphor and irony.

On another axis, several computational models were applied to figurative speech detection tasks, [16] explores word-level metaphor detection using transformer-based embeddings, including BERT and XLNet. The study evaluates these models on the VU Amsterdam Metaphor Corpus (VUA) and the ETS Corpus of Non-Native Written English, demonstrating improvements in metaphor classification over traditional BiLSTM approaches. However, the dataset remains monolingual (English), limiting its applicability to multilingual NLP. Additionally, distinguishing between metaphorical and literal meanings remains a key challenge, particularly in cases of subtle metaphorical shifts.

Research conducted by [17] presents the *TEDB system*: a transformer-based approach for detecting figurative language in text, leveraging fine-tuned embeddings for improved contextual understanding. This system achieves strong performance on benchmark datasets, demonstrating the effectiveness of deep learning architectures in capturing figurative expressions. However, its application to low-resource languages remains unexplored, raising questions about its adaptability in diverse linguistic settings. The *Modelling Sarcasm in Twitter: A Novel Approach* introduces a computational model for sarcasm detection in tweets, leveraging linguistic features instead of simple keyword-based methods [18]. The study constructs the Twitter Sarcasm Corpus, manually annotated for sarcasm detection. The model utilizes a decision tree classifier and is compared against traditional machine learning methods such as Random Forest. While the feature-based approach improves sarcasm detection, the dataset is limited to English tweets, and the method does not explicitly distinguish sarcasm from irony, which remains a persistent challenge in figurative language processing.

The study in [19] investigates the use of antithesis and other contrastive relations in argumentation. The study introduces a taxonomy of argumentative uses of contrastive relations, providing a theoretical framework rather than a computational model. While the work enhances the understanding of how contrastive relations function in argumentation, it does not introduce a dataset or implement machine learning techniques, limiting its direct applicability to NLP tasks. Authors in [20] explore the role of rhetorical figures in argumentation, discussing the need for computationally annotated corpora of rhetorical devices such as symploce, epanaphora, epistrophe, chiasmus, and antithesis. The study suggests integrating computational methods to identify and analyze rhetorical figures in argumentative texts. However, it highlights the lack of large-scale annotated corpora for rhetorical figures, making it difficult to develop and evaluate automated rhetorical figure detection models.

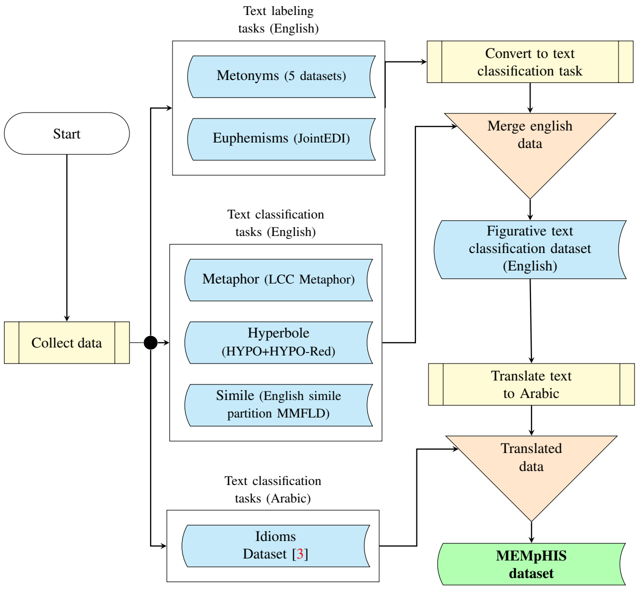
Text labeling tasks (English)
Metonyms (5 datasets)
Euphemisms (JointEDI)
Start
Convert to text classification task
Merge english data
Text classification tasks (English)
Metaphor (LCC Metaphor)
Collect data
Hyperbole (HYPO+HYPO-Red)
Simile (English simile partition MMFLD)
Figurative text classification dataset (English)
Translate text to Arabic
Translated data
Text classification tasks (Arabic)
Idioms Dataset [3]
MEMpHIS dataset

Figure 1. Dataset construction flowchart

Methodology

This section details the construction, annotation, and evaluation of the **MEMpHIS** dataset for figurative speech detection in Arabic. Our methodology consists of the following key steps: dataset collection and adaptation, trans- lation, and model benchmarking. The full process is described by Figure 1and model benchmarking. The full process is described by Figure 1

Dataset Collection

The process starts by collecting textual data from a range of sources. This stage is key to ensuring a variety of content and representative representation of the figures of speech under study. The texts collected may come from existing corpora, linguistic databases or other relevant textual sources. To build a high-quality Arabic dataset for figurative speech detection, we gathered benchmark datasets originally designed for English, which we sub- sequently translated to Arabic using Deepl API[†].

[†] https://www.deepl.com/

Text labeling tasks – English Datasets used for metonyms and euphemisms were originally designed for text sequence labeling rather than full text classification, which aim to train models to identify where exactly the respective figures of speech were employed. To standardize our tasks, we have reduced these datasets into text classification tasks, by splitting larger texts into smaller sentences, then labeling sentences containing at least one figurative span as positive samples, while the rest are negative samples. This step is crucial to making the data homogeneous and ready for integration into subsequent steps

Text classification tasks – English Following that initial transformation, the previous datasets are then combined with 3 other datasets originally designed for classification tasks, which aim to categorize the expressions into figurative and non-figurative. This category of datasets contains three distinct style figures. The LCC Metaphor dataset is used to classify the metaphors, whereas the HYPO-I and HYPO-Red databases are used to classify the hyperboles. Finally, the comparisons (similes) are extracted from the MMFLD dataset partition. This categorization makes it possible to better organize the data according to several linguistic categories.

Text classification tasks – Arabic In parallel with English figurative text classification datasets, similar works have been carried out for Arabic. Unlike metaphors, hyperboles, and similes classified in English texts, the focus here is on idioms, which are fixed expressions whose meaning cannot be directly inferred from the individual words they contain. These idioms are extracted from various textual sources and compiled into a specific dataset published by [3]. Identifying and classifying idioms requires a specialized approach, as these expressions lose their idiomatic connotation when translated to a different language, thus, we chose to include an originally Arabic dataset for this figure, to keep the original idiomatic meaning of sentences.

The English data collected in the earlier steps are translated into Arabic. This translation makes it possible to expand the dataset's scope and increase the potential for cross-linguistic analysis. The translated data are then grouped together and then merged with the idiom dataset, forming the final version of MEMpHIS.

This data's application for natural language processing (NLP) tasks is made easier by its consolidation. By offering a more uniform and representative corpus of many types of English figurative language, a well-structured dataset improves the performance of machine learning models. Additionally, this fusion ensures higher consistency within the dataset by streamlining data pretreatment processes like cleaning, normalization, and annotation. Furthermore, by exposing NLP models to a large variety of figurative language samples, creating a combined dataset enhances their capacity for generalization. This is especially crucial for applications like automated figure-of-speech recognition in literary or journalistic texts, semantic analysis, and enhancing machine translation and text production systems. The foundation for further study and more efficient advancements in the area of figurative language processing in English is laid by organizing and consolidating all of this data into a single, useable corpus.

MEMpHISdataset Finally, the English and Arabic data from the various stages are combined to form the final dataset, called the 'MEMpHISdataset'. This dataset includes figurative expressions in English and Arabic, facilitating the study and analysis of figures of speech in a multilingual context. This stage marks the completion of the dataset construction process, which can now be used for classification, linguistic recognition, and advanced semantic analysis tasks.

Baseline Models

To benchmark the newly created Arabic dataset, we evaluate the performance of hybrid models that combine transformer-based embeddings with recurrent neural networks (RNNs). The architecture follows a two-step process, the first of which is Feature Extraction followed by Sequence Classification :

Feature Extraction: Pre-trained transformer models leverage deep neural networks to generate contextualized embeddings, capturing semantic and syntactic nuances of each sentence. Unlike traditional word embeddings, these models consider the surrounding words to create dynamic representations that adapt to different contexts. This enables a more accurate understanding of figurative language, idiomatic expressions, and polysemous words. These embeddings serve as input for downstream NLP tasks such as classification, translation, and sentiment analysis. For feature extraction, we used the XLM-RoBERTa *(FacebookAI/xlm-roberta-base)* which is a pre-trained model based on the Transformer architecture, a multilingual version of RoBERTa by pre-training on 2.5 TB of common crawled data containing 100 languages. It aims to improve comprehension and expressiveness in several languages [4], the mT5-Small and mT5-Base are transformer-based multilingual variant of "Text-to-Text Transfer Transformer" (T5) [21] that was pre-trained on a new Common Crawl-based dataset covering 101 languages. As for BERT-based models, they are language representation models which rely on an encoder-based architecture with attention mechanisms. For the sake of this study, we have used 2 Arabic variants of BERT-models: AraBERTv2 [22], and Qarib-BERT [23].

Sequence Classification: The extracted embeddings are passed to recurrent neural network architectures for classification. The tested models include:

- The Bidirectional Gated Recurrent Unit (BiGRU) is made up of both forward-propagating and backward-propagating GRU units, forming a bidirectional neural network [5].
- Bidirectional Long Short-Term Memory (BiLSTM) that is consists of a forward LSTM and a backward LSTM in which data can be processed in both forward and backward directions. Reverse processing allows the capture of features and hidden patterns in the data that are usually ignored by LSTMs [6]
- Convolutional BiLSTM(CNN-BiLSTM).

By combining **contextual embeddings** from transformers with **sequential learning** via RNNs, we aim to capture both deep semantic meaning and long-range dependencies in figurative expressions.

Evaluation Metrics

To evaluate the model's performance, we present a detailed analysis of its accuracy (Acc), Precision (P), Recall (R), Macro F1-score and Area Under ROC (AUC) for class balance analysis. Values for True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are used to calculate them.

- **Accuracy:**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:**

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:**

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:**

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Area Under the ROC Curve (AUC):** AUC is computed as the area under the Receiver Operating Characteristic curve, which plots the True Positive Rate (Recall) against the False Positive Rate:

$$\text{FPR} = \frac{FP}{FP + TN}$$

AUC does not have a closed-form equation like the others but is typically computed using numerical integration.

Results are compared to determine the effectiveness of different architectures.

Results and Discussion

This section presents the quantitative results obtained from the translation quality evaluation and model benchmarking on the Arabic figurative speech dataset.

Dataset Structure

The final dataset consists of **56,531** annotated sentences, evenly balanced between positive and negative samples for each figure, distributed as shown in Table 1

For each figure of speech, a sentence is labeled as either `0` or `1`, forming a balanced dataset for classification.

Model Performance on Translated Dataset

To evaluate model performance on the Arabic dataset, we fine-tune various NLP models, categorized as Hybrid Models which includes Transformer embeddings (XLM-RoBERTa, mT5, AraBERT) combined with BiGRU, BiLSTM, or CNN-BiLSTM classifiers, and End-to-End Transformer Models with directly fine-tuning models such as AraBERT on the translated dataset. Table 2 presents the evaluation results of benchmark experiments that were conducted on the translated dataset.

Table 2 displays the effectiveness of various models applied to a translated dataset, assessed across distinct figurative styles: Metonym, Euphemism, Metaphor, Hyperbole, Idiom, and Simile. Examination of the findings reveals that the **Simile** and **Idiom** styles are much more successfully identified by the models, with notably high scores across all evaluated metrics. **Simile** achieves precision, recall, F1-score, and accuracy of 94.20%, accompanied by very low variability and an impressive AUC of 97.84%. **Idiom** closely trails, also obtaining strong results, just slightly lower than Simile, particularly with an F1-score of 93.38% and an AUC of 97.40%.

In comparison, the **Euphemism** and **Metaphor** styles present more significant difficulties. Their accuracy rates are around 66.5%, with similar F1-scores and AUCs falling below 75% for **Euphemism** and about 70% for **Metaphor**. Additionally, these styles show relatively high standard deviations, indicating inconsistent results across different samples. Hyperbole, on the other hand, performs moderately well, achieving a solid accuracy of 74.33% and demonstrating improved consistency, whereas **Metonym** remains closer to the lower performance tiers observed for **Metaphor** and **Euphemism**.

From Table 2, we can see a significant performance gap between idioms and similes, with scores ranging between 90% and 99%. While other figurative styles have seen their scores range between 55% and 75%. The high performance recorded on **similes** likely stems from their simple patterns, namely using specific comparison particles such as "كـ" and "مثل". **Idioms** are always used within different contexts as such, unlike other figures of speech, which can overlap with literal meaning, thus, they are easier for the model to identify.

Contrastively, performance scores for the remaining tasks vary between 65% and 70%, due to the nature of each figure of speech. **Euphemisms** and **Metonyms** for example, require additional knowledge on specific entities within the sentence, whether being attenuative expressions or associative relationships.

Metaphors and **Hyperboles** consist more of words being placed out of context, such as associating space with poverty in "**اليوم رأيت ما يحدث عندما تقتحم الفرصة المساحة التي يحتلها الفقر**", thus, using pre-trained language models which were initially trained for *Next-Word Prediction* is not likely to return high performance. This opens up research opportunities for exploring novel techniques for Arabic figurative language classification.

Conclusion

This paper presents the *MEMpHIS* dataset, a novel resource for figurative speech detection in Arabic, constructed by compiling and translating benchmark English datasets. The dataset encompasses six key figurative speech categories: metaphors, idioms, similes, metonymy, hyperboles, and euphemisms, offering a comprehensive benchmark for evaluating Arabic figurative language processing. Using hybrid architectures that combine transformer-based embeddings and RNN classifiers, as well as fine-tuned transformers, we conducted extensive benchmarking to evaluate model performance.

Table 1. Dataset sample counts

Figurative Style	Test		Train		Valid	
	Fig	Non-Fig	Fig	Non-Fig	Fig	Non-Fig
Metonym	606	1020	2874	4247	606	1014
Euphemism	253	235	906	788	255	241
Metaphor	1997	1990	6053	6035	1998	1991
Hyperbole	149	150	1650	1619	50	49
Idiom	436	438	1520	1531	219	219
Simile	150	150	6465	6232	2492	2457

Table 2. Performance of models on the translated dataset.

Figurative Style	**Accuracy**	**Precision**	**Recall**	**F1**	**AUC**
Metonym	67.21 ± 4.77	67.17 ± 4.91	67.0 ± 4.63	66.47 ± 6.41	72.71 ± 5.63
Euphemism	66.51 ± 7.44	69.09 ± 7.31	67.13 ± 7.4	65.61 ± 8.01	74.28 ± 8.23
Metaphor	66.66 ± 7.0	66.76 ± 6.89	66.66 ± 7.01	66.42 ± 7.51	69.93 ± 7.97
Hyperbole	74.33 ± 5.27	74.88 ± 5.26	74.33 ± 5.26	74.17 ± 5.37	81.57 ± 6.27
Idiom	**93.39 ± 3.39**	93.45 ± 3.34	93.39 ± 3.39	**93.38 ± 3.40**	97.40 ± 2.96
Simile	**94.20** ± **1.66**	94.29 ± 1.67	94.20 ± 1.66	**94.20** ± **1.66**	97.84 ± 1.15

The experimental results reveal significant performance variation across figurative styles. Idioms and similes exhibit the highest performance scores, likely due to their distinct lexical patterns and less ambiguous usage. In contrast, categories like euphemisms and hyperboles show relatively lower performance, highlighting the challenges of capturing subtle and context-dependent figurative meanings. Furthermore, the translation process introduces semantic shifts and cultural mismatches that likely contributed to model misclassifications, particularly for more complex figurative forms.

While the *MEMpHIS* dataset provides a valuable resource for Arabic NLP, several limitations remain. First, the dataset relies entirely on machine translation for most categories, with no manual post-editing or verification. This reliance can lead to inconsistencies in the translated texts, particularly for nuanced figurative expressions. Second, idioms, which were directly collected in Arabic, exhibit higher classification performance, raising questions about whether other categories would benefit from native Arabic data collection rather than translation. Finally, the lack of dialectal variations in the dataset restricts its applicability across diverse Arabic-speaking populations.

Given these findings, we emphasize the need for incorporating manual annotation processes to improve dataset quality and enhance model performance. While automatic translation is efficient, manual annotation is essential for:

- Verifying the preservation of figurative meaning during translation.
- Correcting semantic inaccuracies introduced by machine translation tools.
- Creating a more balanced and representative dataset for all figurative styles, including underperforming categories like metaphors and euphemisms.

Future work will focus on addressing these limitations by expanding the dataset with manually verified and annotated examples, particularly for the most challenging figurative categories. Additionally, we plan to incorporate additional techniques to fully annotate all the texts for every figurative task. Finally, further research should explore advanced models that integrate cultural and contextual understanding to better handle the complexities of Arabic figurative speech.

By releasing the *MEMpHIS* dataset and our evaluation pipeline, we aim to foster further research in Arabic NLP, enabling the community to tackle the challenges of figurative language detection and enhance computational understanding of one of the world's most linguistically rich languages.

References

[1] Randy Allen Harris et al. "An Annotation Scheme for Rhetorical Figures". In: *Argument & Computation* (2018). DOI: 10.3233/aac-180037.

[2] Nouh Sabri Elmitwally and Saad Alanazi. "Arabic Corpus for Figurative Sentiment Analysis". In: *International Journal of Advanced Science and Technology* 29.3 (2020).

[3] Hanen Himdi. "Arabic Idioms Detection by Utilizing Deep Learning and Transformer-based Models". In: *Procedia Computer Science*. 6th International Conference on AI in Computational Linguistics 244 (Jan. 2024), pp. 37–48. ISSN: 1877-0509. DOI: 10.1016/j.procs.2024.10.176. (Visited on 01/26/2025).

[4] Nouh Sabri Elmitwally and Ahmed Alsayat. "CLASSIFICATION AND CONSTRUCTION OF ARABIC CORPUS: FIGURATIVE AND LITERAL". In: . *Vol.* 19 (2005).

[5] Felix Schneider et al. "Data-Driven Detection of General Chiasmi Using Lexical and Semantic Features". In: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Ed. by Stefania Degaetano-Ortlieb et al. Punta Cana, Dominican Republic (online): Association for Computational Linguistics, Nov. 2021, pp. 96–100. DOI:

10.18653/v1/2021.latechclfl-1.11. (Visited on 01/14/2025).

[6] Sameer Naser Olimat. “Euphemism in the Qur’an: A Corpus-based Linguistic Approach”. In: *International Journal of Computational Linguistics (IJCL)* 10.2 (2019), pp. 16–32.

[7] Chen Liu et al. “FigMemes: A Dataset for Figurative Language Identification in Politically-Opinionated Memes”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Dec. 2022, pp. 7069–7086. DOI: 10.18653/v1/2022.emnlp-main.476. (Visited on 02/01/2025).

[8] Tuhin Chakrabarty et al. “FLUTE: Figurative Language Understanding through Textual Explanations”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Dec. 2022, pp. 7139–7159. DOI: 10.18653/v1/2022.emnlp-main.481. (Visited on 02/01/2025).

[9] Anubha Kabra et al. *Multi-Lingual and Multi-cultural Figurative Language Understanding*. https://arxiv.org/abs/2305.16171v1. May 2023. (Visited on 02/01/2025).

[10] Ibrahim Abu Farha et al. “From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset”. In: *OSACT* (2020). DOI: null.

[11] Ramona Kuehn, Jelena Mitrović, and Michael Granitzer. “Hidden in Plain Sight: Can German Wiktionary and Wordnets Facilitate the Detection of Antithesis?” In: *Proceedings of the 12th Global Wordnet Conference*. Ed. by German Rigau, Francis Bond, and Alexandre Rademaker. University of the Basque Country, Donostia - San Sebastian, Basque Country: Global Wordnet Association, Jan. 2023, pp. 106–116. (Visited on 01/09/2025).

[12] Kevin Alex Mathews and M. Strube. “Impact of Target Word and Context on End-to-End Metonymy Detection”. In: *ArXiv* (Dec. 2021). (Visited on 01/25/2025).

[13] Marie Dubremetz et al. “Rhetorical Figure Detection: The Case of Chiasmus”. In: *null* (2015). DOI: 10.3115/v1/w15-0703.

[14] John Lawrence, Jacky Visser, and Chris Reed. “Harnessing Rhetorical Figures for Argument Mining: A Pilot Study in Relating Figures of Speech to Argument Structure”. In: *Argument & Computation* 8.3 (Jan. 2017). Ed. by Randy Allen Harris and Chrysanne Di Marco, pp. 289–310. ISSN: 1946-2166, 1946-2174. DOI: 10.3233/aac-170026. (Visited on 01/14/2025).

[15] Jelena Mitrović et al. “Ontological Representations of Rhetorical Figures for Argument Mining”. In: *Argument & Computation* (2017). DOI: 10.3233/aac-170027.

[16] Jerry Liu et al. “Metaphor Detection Using Contextual Word Embeddings From Transformers”. In: *Proceedings of the Second Workshop on Figurative Language Processing*. Ed. by Beata Beigman Klebanov et al. Online: Association for Computational Linguistics, July 2020, pp. 250–255. DOI: 10.18653/v1/2020.figlang-1.34. (Visited on 01/09/2025).

[17] Peratham Wiriyathammabhum. “TEDB System Description to a Shared Task on Euphemism Detection 2022”. In: *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*. Ed. by Debanjan Ghosh et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 1–7. DOI: 10.18653/v1/2022.flp-1.1. (Visited on 01/25/2025).

[18] Francesco Barbieri et al. “Modelling Sarcasm in Twitter, a Novel Approach”. In: *null* (2014). DOI: 10.3115/v1/w14-2609.

[19] Nancy L. Green and Nancy L. Green. “The Use of Antithesis and Other Contrastive Relations in Argumentation”. In: *Argument & Computation* (2022). DOI: 10.3233/aac-210025.

[20] Randy Allen Harris et al. “Rhetorical Figures, Arguments, Computation”. In: *Argument & Computation* (2017). DOI: 10.3233/aac-170030.

[21] Linting Xue et al. “mT5: A massively multilingual pre-trained text-to-text transformer”. In: *CoRR* abs/2010.11934 (2020). arXiv: 2010.11934. URL: https://arxiv.org/abs/2010.11934.

[22] Wissam Antoun, Fady Baly, and Hazem Hajj. “AraBERT: Transformer-based Model for Arabic Language Understanding”. In: *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, p. 9.

[23] Ahmed Abdelali et al. “Pre-Training BERT on Arabic Tweets: Practical Considerations”. In: (2021). arXiv: 2102.10684 [cs.CL].