Automatic Classification of Figurative and Literal Usage in Hebrew Idioms

Igor Nazarenko (ID. 322029158), Yuval Amit (ID. 203839907)

Submitted as final project report for the NLP course, IDC, 2025

1 Introduction

Idiomatic expressions are prevalent in natural language and often convey meanings that diverge from their literal form. For NLP systems, distinguishing figurative from literal usage remains a persistent challenge, especially in morphologically rich, low-resource languages like Hebrew, impacting tasks such as sentiment analysis, machine translation, and dialogue understanding.

Hebrew presents a unique case: its idioms are highly context-dependent, yet existing annotated resources for figurative vs. literal distinctions are scarce. Despite its linguistic richness, this task has received limited attention in Hebrew NLP, making it a compelling testbed for evaluating both multilingual and Hebrew-specific models.

In this work, we investigate how well pre-trained Transformer models can distinguish between literal and figurative usages of Hebrew idioms. We construct a balanced dataset, compare transfer learning strategies (fine-tuning vs. frozen feature extraction), and evaluate multiple model types. Our analysis includes quantitative metrics and qualitative insights via confusion matrices, attention visualizations, and interpretability tools.

This study addresses two key questions:

1. To what extent can pre-trained Hebrew and multilingual Transformer models distinguish literal from figurative idiom usage?
2. How do different adaptation strategies influence model performance on this nuanced classification task?

2 Background and Related Work

Figurative language, including idioms and metaphors, presents a longstanding challenge in NLP due to its non-compositional meaning and reliance on contextual or cultural cues. Models trained on literal data often misinterpret figurative intent, degrading performance in tasks like translation, sentiment analysis, and dialogue understanding.

This difficulty stems from ambiguity, many idioms ("שבר שתיקה") have both literal and figurative interpretations and from sparsity, as such expressions are infrequent in most corpora, especially in morphologically rich and low-resource languages like Hebrew.

Early approaches detected idiomaticity using compositionality scores (1) and discourse cohesion features (2). More recent work has applied attention-based transformers for idiom classification (3), including models that incorporate discourse-aware structures (4). Yet, even advanced multilingual models often struggle in zero-shot figurative tasks (5).

In Hebrew, research remains sparse. Tennen et al. (6) showed that Hebrew-specific models like AlephBERT can classify metaphors when fine-tuned. However, no prior work has addressed literal vs. figurative idiom classification in Hebrew. Our work fills this gap by evaluating multilingual and Hebrew models using transfer learning and interpretability tools.

3 Dataset Construction

We curated a dataset of 41 Hebrew idiomatic expressions, each exhibiting both literal and figurative usage. Initial candidates were generated using ChatGPT and refined manually based on linguistic plausibility, dual-usage potential, and stylistic diversity. The final set spans varied grammatical forms, topics (emotion, action, conflict), and tones, providing broad coverage of modern Hebrew idiomaticity.

For each expression, we generated 15 literal and 15 figurative sentences using ChatGPT, then manually adapted them to sound fluent, contextually realistic, and stylistically varied across domains such as workplace, family, and daily life. This resulted in 1,230 sentences (615 literal, 615 figurative), all reviewed for clarity and annotated with a binary label (0 = literal, 1 = figurative). The dataset is perfectly balanced across classes and expressions.

Sentences were grouped by idiom and partitioned using stratified GroupShuffleSplit into training (70%), validation (15%), and test (15%) sets, ensuring no idiom leakage between splits. Labels were unified under a single column `labels`, and each split was saved to disk as a CSV.

The dataset includes idioms across a range of lengths (5–18 tokens; average 9.9 tokens per sentence). Table 3 presents one literal and one figurative example:

Literal	קפץ למים כדי להתרענן ביום חם.
Figurative	הוא קפץ למים ופתח עסק משלו ללא ניסיון קודם.

For preprocessing, we used HuggingFace's `datasets` and `transformers` libraries. Each split was loaded into a `DatasetDict` and tokenized using the pretrained tokenizer of the evaluated model (AlephBERTGimmel, AlephBERT, DictaBERT, mBERT, XLM-RoBERTa). Sentences were padded or truncated to 64 tokens, with outputs including `input_ids`, `attention_mask`, and `labels`, all formatted into PyTorch tensors. A `DataCollatorWithPadding` ensured efficient batching during training.

This pipeline ensured clean, consistent, and model-agnostic preparation of the dataset for all downstream experiments.

4 Methodology: Models and Training

We framed the task as binary classification of figurative vs. literal idiom usage in Hebrew, using pretrained Transformer encoders extended with a classification head. Models were adapted via fine-tuning, frozen feature extraction, and zero-shot evaluation. We compared multilingual and Hebrew-specific models. This section details the models, training setup, evaluation procedures, and implementation.

4.1 Selected Models

To address idiomaticity prediction in Hebrew, we evaluated Transformer-based models differing in language coverage and specialization:

Model	**Motivation and Notes**
XLM-RoBERTa *Multilingual*	Trained on 100+ languages. Offers strong cross-lingual transfer and serves as a robust benchmark for multilingual performance.
mBERT *Multilingual*	Trained on Wikipedia across many languages. While generally weaker than XLM-RoBERTa, it is a widely used baseline for multilingual tasks.
AlephBERTGimmel *Hebrew-only*	Trained on a large, high-quality Hebrew corpus. Excels in capturing Hebrew morphology and syntax.
AlephBERT *Hebrew-only*	An earlier Hebrew model trained on a smaller corpus. Included to evaluate the impact of training scale and corpus quality.
DictaBERT *Hebrew (Legal Domain)*	Pretrained on legal and formal texts. Its exposure to complex sentence structures may help interpret figurative language.

This comparison helps assess whether wide language coverage or domain-specific pretraining leads to better idiomaticity classification in Hebrew.

4.2 Training Setup

All models were trained using the same pipeline, allowing direct comparison across variants. Training was performed on a single GPU, with hyperparameters optimized via the `Optuna` package. We searched over learning rate, batch size, warmup ratio, and weight decay. The best checkpoint per model was selected based on validation F1-score. All runs used `seed=42`, and AlephBERTGimmel was also assessed with 10-fold cross-validation (accuracy 0.918 ± 0.057).

The table below summarizes the final hyperparameters selected by Optuna for each model:

Model	Learning Rate	Batch Size	Warmup Steps (%)	Weight Decay	Label Smoothing	LR Scheduler
XLM-RoBERTa	$1e^{-5}$	8	5.15%	0.05	0.10	Cosine
mBERT	$5e^{-5}$	16	5.88%	0.05	0.10	Cosine
AlephBERTGimmel	$5e^{-5}$	8	22.49%	0.00	0.10	Linear
AlephBERT	$5e^{-5}$	8	28.10%	0.00	0.20	Cosine
DictaBERT	$1e^{-5}$	8	10.10%	0.00	0.00	Cosine

Table 1: Final hyperparameters selected by Optuna for each model (rounded to 2 decimal places).

Frozen and Zero-shot Variants: To evaluate transfer learning without full fine-tuning, we froze encoder weights and trained only the classification head on labeled data. For zero-shot evaluation, we used randomly initialized heads and passed test examples through the pre-trained models directly. Both approaches used the same pipeline and evaluation setup as in fine-tuning.

4.3 Evaluation Metrics

We evaluated performance using **accuracy**, **precision**, **recall**, and **F1-score**, along with confusion matrices (Figure 1) for per-class error inspection. We also report **ROC AUC** and **PR curves** to assess robustness under class imbalance.

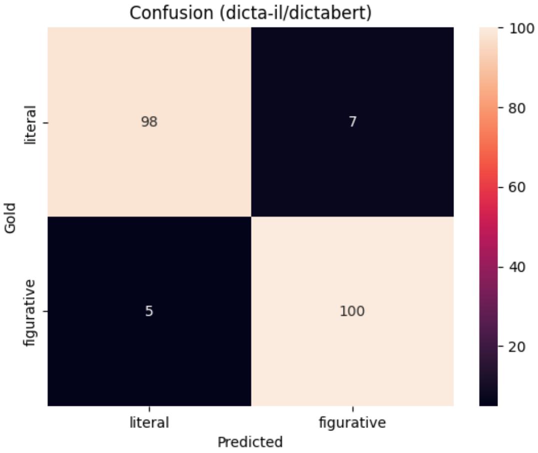


Figure 1: Example confusion matrix from DictaBERT model evaluation.

Calibration and Brier Score: To assess probability calibration, we plotted reliability curves (Figure 2). A perfectly calibrated model follows the diagonal, where predicted probabilities match observed frequencies. The Brier score quantifies this calibration quality by computing the mean squared error between predicted probabilities and true outcomes (lower is better). Our results show the model is well-calibrated, with only minor deviations at low-confidence predictions.

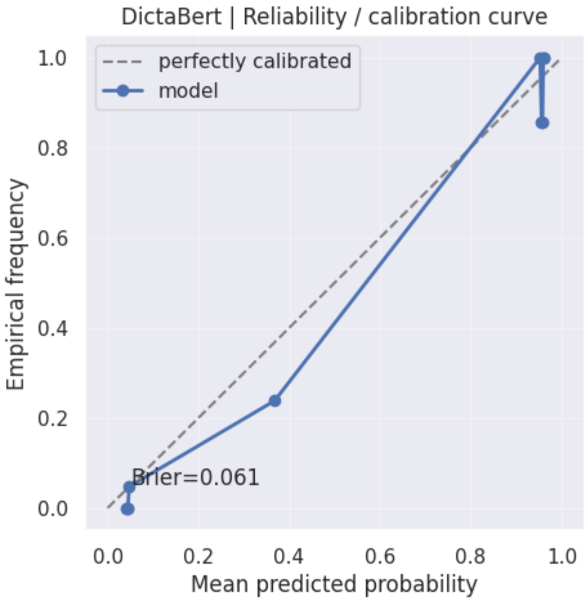


Figure 2: Calibration curve for the DictaBERT model.

4.4 Implementation

We used the Hugging Face `transformers` library and its `Trainer` API for fine-tuning and evaluation. Initial experiments were run locally on an M1 Mac and CPU laptop. For larger models, we transitioned to Google Colab's free GPU tier and later to a rented RTX 4080 GPU on VastAI, significantly accelerating training. All models used the same codebase and reproducible setup (`seed=42`).

5 Experiments, Results, and Analysis

This section presents the results of our experiments, including model comparison, error analysis, and embedding visualization. We evaluated multiple Transformer-based models on a held-out 15% test set and analyzed their strengths and limitations through both quantitative metrics and qualitative insights.

5.1 Overall Performance

Model	Class	Precision	Recall	F1-Score	Support
XLM-RoBERTa	Literal	0.89	0.95	0.92	105
	Figurative	0.95	0.89	0.92	105
	Total Accuracy		**0.92**		210
mBERT	Literal	0.94	0.87	0.90	105
	Figurative	0.88	0.94	0.91	105
	Total Accuracy		**0.90**		210
AlephBERTGimmel	Literal	0.97	0.93	0.95	105
	Figurative	0.94	0.97	0.95	105
	Total Accuracy		**0.95**		210
AlephBERT	Literal	0.97	0.91	0.94	105
	Figurative	0.92	0.97	0.94	105
	Total Accuracy		**0.94**		210
DictaBERT	Literal	0.95	0.93	0.94	105
	Figurative	0.93	0.95	0.94	105
	Total Accuracy		**0.94**		210

Table 2: Precision, Recall, F1-Score, and Total Accuracy results for all models on the idiomaticity prediction task.

AlephBERTGimmel achieved the best results across all metrics, with 0.918 (±0.057) accuracy in 10-fold cross-validation. A McNemar's test found no significant difference among Hebrew models, but a significant advantage over the weakest multilingual baseline mBERT ($p < 0.05$).

Fine Tuned vs Frozen Models: To assess the impact of fine-tuning, we compared results from fully trained models against their frozen counterparts, where only the classification head was trained while transformer weights remained unchanged. As shown in Table 3, performance decreased noticeably for XLM-RoBERTa and AlephBERTGimmel (accuracy drops of 7 to 10%), highlighting the importance of full model adaptation to the idiomaticity classification task. DictaBERT showed the smallest gap, retaining high accuracy (0.94) even when frozen, likely due to its strong pre-training on Hebrew legal text closely aligned with the dataset domain.

Model	Class	Precision	Recall	F1-Score	Support
XLM-RoBERTa (Frozen)	Literal	0.90	0.78	0.84	105
	Figurative	0.81	0.91	0.86	105
	Total Accuracy		**0.85**		210
AlephBERTGimmel (Frozen)	Literal	0.79	0.89	0.84	105
	Figurative	0.87	0.77	0.82	105
	Total Accuracy		**0.83**		210
DictaBERT (Frozen)	Literal	0.96	0.92	0.94	105
	Figurative	0.93	0.96	0.94	105
	Total Accuracy		**0.94**		210

Table 3: Precision, Recall, F1-Score, and Total Accuracy results for frozen models.

Zero Shot Performance: Zero shot evaluation yielded near-random accuracy (50%) for both models. XLM-RoBERTa defaulted to predicting only the literal class, while AlephBERT produced balanced but low accuracy predictions. These results highlight the necessity of fine tuning for effective idiomaticity classification.

Model	Class	Precision	Recall	F1-Score	Support
XLM-RoBERTa (Zero-Shot)	Literal	0.50	1.00	0.67	105
	Figurative	0.00	0.00	0.00	105
	Total Accuracy		**0.50**		210
AlephBERT (Zero-Shot)	Literal	0.49	0.46	0.48	105
	Figurative	0.50	0.53	0.51	105
	Total Accuracy		**0.50**		210

Table 4: Precision, Recall, F1-Score, and Total Accuracy results for zero-shot models.

5.2 Error Analysis

To better understand model limitations, we examined misclassified examples from the test set using sequence classification explainer and attention heatmaps. Most errors involved confusing figurative expressions with literal ones, particularly when contextual cues were weak or ambiguous.

Error distribution: For most models, false negatives (*figurative* → *literal*) were slightly more frequent than false positives. This suggests that the models tended to be conservative, requiring stronger contextual signals to predict a figurative meaning.

Error Type	Count	Percentage
Literal → Figurative	7	3.3%
Figurative → Literal	5	2.4%

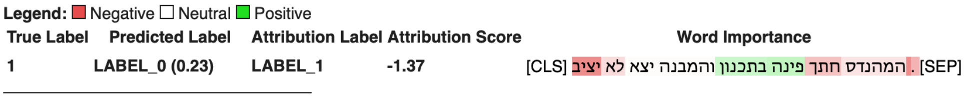
Table 5: Types of misclassifications for DictaBERT (out of 210 samples).

Example misclassifications:

- **Gold: Figurative** – "המהנדס חתך פינה בתכנון והמבנה יצא לא יציב"
  Predicted: Literal – The idiom "חתך פינה" (cut corners) was taken literally.

  Explanation: This sentence is tricky even for Hebrew speakers, since the idiom "cut corners" could easily be interpreted as a physical action without broader context. The model focused on literal cues like "המהנדס" and "יציב", and gave weak or mixed importance to the actual idiom, which led to misclassification.

  Attention weights — false negative: ignored idiom, focused on literal words

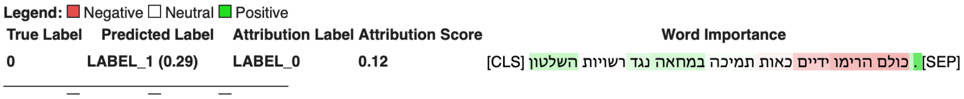
  Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	LABEL_0 (0.23)	LABEL_1	-1.37	[CLS] המהנדס חתך פינה בתכנון והמבנה יצא לא יציב . [SEP]

- **Gold: Literal** – "כולם הרימו ידיים כאות תמיכה במחאה נגד רשויות השלטון"
  Predicted: Figurative – The idiom "הרימו ידיים" (raised hands) was taken figuratively.

  Explanation: The idiom "הרימו ידיים" usually signals surrender, but here it describes a literal act of protest. The model strongly focused on the idiom itself and assigned it high negative attribution, which misled the classification toward a figurative interpretation. In contrast, it gave relatively weaker positive weight to disambiguating context words like "מחאה" and "תמיכה", resulting in misclassification despite the overall literal meaning.

  Attention weights — false positive: focused on idiom, missed context

  Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	LABEL_1 (0.29)	LABEL_0	0.12	[CLS] כולם הרימו ידיים כאות תמיכה במחאה נגד רשויות השלטון . [SEP]

Takeaways: These examples illustrate recurring failure modes:

- **Ambiguous contexts** where literal and figurative interpretations are both plausible.
- **Token-level overfitting**, models focus on the idiom token itself without adequate integration of surrounding context.
- **Shallow pragmatic reasoning**, which limits the model's ability to disambiguate subtle figurative uses.

These patterns align with the broader limitations discussed in Section 6.

Opportunities for Improvement: Addressing these issues may involve:

- Training with extended or multi-sentence context.
- Incorporating external lexical resources (idiom dictionaries, usage examples).
- Refining training objectives to penalize high-confidence misclassifications more explicitly.

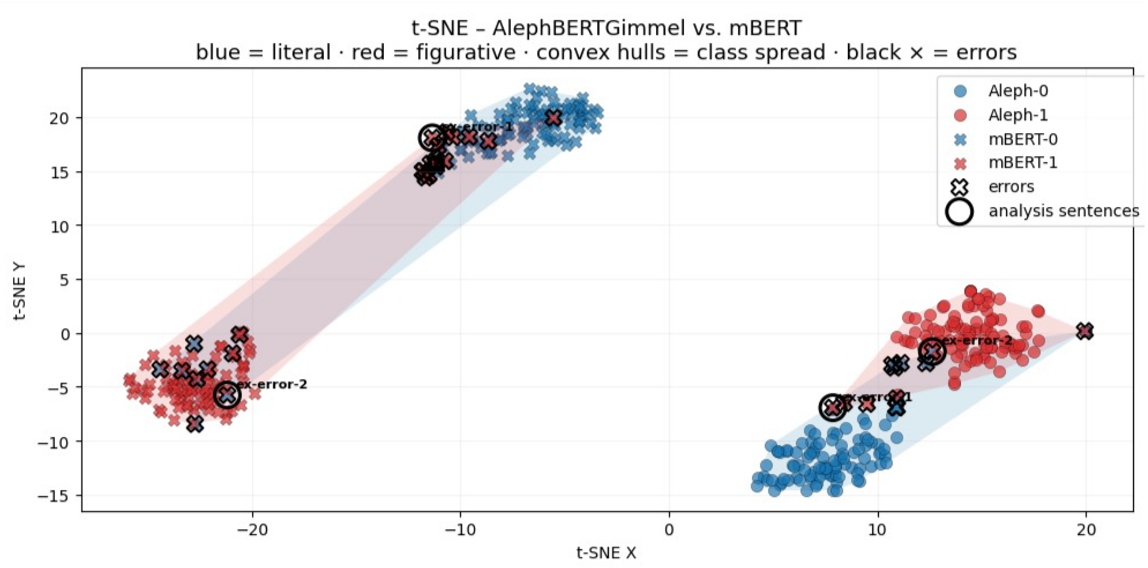
5.3 t-SNE Embeddings and Model Behavior

Interpretation: To analyze how models distinguish figurative from literal meaning, we visualized [CLS] sentence embeddings using t-SNE for AlephBERTGimmel and mBERT.

Both models produced distinct class clusters, with AlephBERTGimmel forming more compact and well-separated regions indicating stronger idiomatic encoding. Misclassifications (black Xs) typically appear near decision boundaries, reflecting genuine ambiguity rather than random noise.

The two circled examples in the figure below correspond to the misclassified sentences discussed in the error analysis. These highlight how contextually ambiguous idioms remain difficult, even for fine-tuned models.

Overall, the visualization supports that AlephBERTGimmel encodes figurative/literal distinctions more effectively than mBERT, though edge cases still pose challenges.



6 Discussion and Conclusion

Key Insights: This work tackles the underexplored challenge of figurative language understanding in Hebrew, a low-resource and morphologically complex language. Our findings demonstrate that:

- **Hebrew-specific models consistently outperform multilingual ones.**
- **AlephBERTGimmel** achieved the highest scores across all metrics, with compact, separable [CLS] embeddings between literal and figurative classes, though differences from other Hebrew models were not significant.
- **Attention visualizations and error patterns** confirmed its superior contextual sensitivity and focus on idiomatic cues.

Transfer Learning:

A central goal of our study was to assess whether pretrained language models could generalize to idiomatic Hebrew. Our analysis revealed the following:

- In **zero-shot settings**, all models performed poorly. Hebrew-specific models produced near-random predictions, while multilingual models collapsed into predicting a single class. This suggests idiomatic understanding is absent in general-purpose representations without task-specific adaptation.
- Comparing **fine-tuning vs. frozen feature extraction**:
  - **Fine-tuned models consistently outperformed frozen ones**, both in classification metrics and embedding space (Figure 5.3), confirming the value of task-specific adaptation.
  - Interestingly, **frozen DictaBERT matched some fine-tuned Hebrew models**, likely due to its strong Hebrew-specific pretraining corpus. This raises questions about architecture robustness versus fine-tuning effectiveness in low-data regimes.
- These results highlight the **limitations of frozen transfer** under domain and language mismatch, and validate the use of full fine-tuning even with modest Hebrew datasets.

Limitations: While overall performance was strong, several challenges remain:

- Misclassifications frequently occurred in **ambiguous contexts**, especially where idioms could plausibly be interpreted literally.
- We observed three recurring model weaknesses:
  1. **Shallow contextual reasoning**

   2. **Token-level overfitting**, as seen in attention heatmaps
   3. **Bias toward the more frequent figurative sense**

- **Limited computational resources** constrained:
  - Batch sizes and training duration
  - Hyperparameter search
  - Use of ensemble methods or multi-fold validation (only AlephBERTGimmel was cross-validated)

Future Work: Several promising directions arise from our findings:

- **Expand the dataset** with diverse idioms from natural contexts (news, dialogue, literature).
- **Apply k-fold fine-tuning** to better evaluate generalization and reduce overfitting in low-resource settings.
- **Investigate frozen DictaBERT's strong performance** on larger datasets to disentangle pre-training effects from fine-tuning gains.
- **Incorporate external lexical resources**, such as idiom dictionaries or Wiktionary, to aid disambiguation.
- **Adopt span or token-level supervision** for more granular idiom understanding.
- **Explore prompt-based or instruction-tuned models** to improve zero-shot generalization.
- **Evaluate on downstream tasks** (sentiment analysis, QA, translation) to assess real-world utility.
- **Extend to joint idiom detection and classification** in free-form text without pre-marked spans.

7 Code

The full implementation, including the training notebook and the idiomaticity dataset, is available in the following GitHub repository.

References

[1] A. Fazly, P. Cook, and S. Stevenson, "A lexical and syntactic approach to identifying idiomatic expressions in context," in *Proceedings of the 47th Annual Meeting of the ACL*, 2009, pp. 48–56. [Online]. Available: https://aclanthology.org/J09-1005.pdf

[2] C. Sporleder and L. Li, "Unsupervised recognition of literal and non-literal use of idiomatic expressions," in *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 2009, pp. 754–762. [Online]. Available: https://aclanthology.org/E09-1086.pdf

[3] H. Zeng, H. Ruan, F. Liang, J. Li, and M. Sun, "Are transformer-based models smart enough to detect and understand idioms?" in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2998–3013. [Online]. Available: https://aclanthology.org/2021.findings-acl.264.pdf

[4] R. Yayavaram, G. Krishna, and S. Chakrabarti, "Cohesion-based idiom classification with transformer representations," *arXiv preprint arXiv:2402.00752*, 2024. [Online]. Available: https://aclanthology.org/2024.mwe-1.26.pdf

[5] V. Shwartz and I. Dagan, "Still a pain in the neck: Evaluating text representations on lexical composition," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 149–161. [Online]. Available: https://aclanthology.org/Q19-1027.pdf

[6] T. Tennen, N. Slonim, and Y. Goldberg, "Few-shot hebrew metaphor classification using language model probing," in *Proceedings of the 2024 Annual Conference of the Association for Computational Linguistics (ACL)*, 2024, to appear. [Online]. Available: https://aclanthology.org/2024.eacl-short.39.pdf#:~:text=4.3%20Transformer,hy%02perparameters%20can%20in%20found%20in