Idiom Detection in Turkish and Italian

Muhammet Öztürk
AI and Data Engineering Department
Istanbul Technical University
Istanbul, Turkey
ozturkmu20@itu.edu.tr

Ulaş Polat
AI and Data Engineering Department
Istanbul Technical University
Istanbul, Turkey
polatul20@itu.edu.tr

1 Introduction

Understanding idiomatic expressions remains a significant challenge in natural language processing (NLP). Idioms are phrases whose meanings cannot be inferred from the literal interpretation of their individual words. For example, the Italian expression *"aprire gli occhi"* (literally "to open one's eyes") means "to become aware," while the Turkish idiom *"gözden düşmek"* (literally "to fall from the eye") means "to lose favor." These expressions are deeply tied to culture and context, making them difficult to detect and interpret accurately using standard NLP methods.

The inability to recognize idiomatic expressions can lead to errors in a wide range of applications, including machine translation, sentiment analysis, information extraction, and dialogue systems. Despite their importance, idioms remain underrepresented in most multilingual NLP pipelines, particularly for morphologically rich languages like Turkish and Italian. Recent work has shown that idiom detection is a challenging task across multiple languages, especially due to cultural and syntactic variability, and current models often struggle to generalize beyond memorization of specific idioms (Giulianelli et al., 2023)(1).

To address this problem, we developed a multilingual idiom detection system that automatically identifies idiomatic expressions in Turkish and Italian texts. Our approach is based on **XLM-RoBERTa**, a transformer-based language model well-suited for cross-lingual understanding. The system detects whether an input sentence contains an idiomatic expression, precisely identifies the span of the idiom within the sentence, and handles structural and contextual variation across languages.

To further improve model performance and training generalization, we extended our dataset by leveraging **GPT-4.1-mini**. This model was used to construct a vocabulary of idioms in both Turkish and Italian, distinct from those present in the training and evaluation sets. Based on this vocabulary, we generated approximately 6,000 Turkish sentences containing idiomatic expressions to enrich the dataset and enhance model robustness. Additionally, GPT-4.1-mini was used to tokenize and reformat idiomatic expressions in alignment with the structure of the (2) dataset, ensuring consistent and scalable data augmentation.

Our work makes several key contributions to the field of multilingual idiom processing. First, we develop a robust idiom detection pipeline tailored for two under-resourced languages, Turkish and Italian. By fine-tuning XLM-RoBERTa, we are able to accurately detect idioms and precisely localize their spans within text. To further enhance our training data, we augment existing resources by processing additional idiomatic expressions using GPT-4.1-mini. Our approach achieved F1 scores of 0.91 for Italian and 0.92 for Turkish, which underscores the effectiveness of our methods. Collectively, these contributions advance the state of multilingual idiom processing and support improved cross-cultural language understanding within natural language processing.

2 Dataset

We use a multilingual dataset for idiom detection, focusing on Turkish and Italian. The dataset combines manually annotated data and GPT-generated idiomatic expressions to support both classification and span prediction tasks.

2.1 Source and Format

The base dataset is derived from the DoDiOM project, which is part of an academic study on gamified crowdsourcing for idiom corpus construction(3). It includes both Turkish and Italian idiomatic and literal sentences, annotated with idiom spans and usage categories. The training

set (`train.csv`) contains 5,483 Turkish sentences and 6,029 Italian sentences. The evaluation set (`eval.csv`) includes 686 Turkish and 751 Italian sentences. Each entry consists of the following fields:

- **id**: Unique identifier.
- **language**: Language code (`tr` or `it`).
- **sentence**: Raw sentence text.
- **tokenized_sentence**: Tokenized version of the sentence.
- **expression**: The idiomatic or literal expression.
- **category**: Usage type—`idiomatic` or `literal`.
- **indices**: Token span indicating the location of the expression (e.g., `[3, 4]`). A span of `[-1]` indicates no idiomatic expression.

2.2 Dataset Augmentation

To expand the training data and improve generalization, we generated additional examples using **GPT-4.1-mini**. Specifically, we:

- Constructed a new vocabulary of idioms for both Turkish and Italian, distinct from those in the training and evaluation sets,
- Generated approximately 6,000 Turkish sentences containing these idioms using a custom generation pipeline available at `idiom-generation-with-vocab`,
- Tokenized and annotated these sentences with idiom spans and usage categories, maintaining the same structure as the original dataset.

In addition, we incorporated data from the idiom-corpus-llm GitHub repository, which contains GPT-generated idiomatic content. We used GPT-4.1-mini to re-tokenize and index this dataset to ensure compatibility and consistency with our existing data.

2.3 Train Set Structure

The final `train.csv` file comprises **18,912 labeled examples**, formatted consistently across all sources. For Turkish, the training data consists solely of 5,483 manually annotated sentences provided in the original dataset. For Italian, the training set includes 6,029 human-annotated examples and an additional 7,400 sentences sourced from the idiom-corpus-llm repository. These additional Italian sentences were processed using **GPT-4.1-mini** to ensure consistent tokenization and idiom span indexing in line with the existing format.

2.4 Evaluation and Test Sets

Evaluation set shares the same structure as the training set and is used for validation. It contains 686 Turkish sentences and 751 Italian sentences. A version excluding the `expression` and `category` fields is also used for blind evaluation during development.

Test set for Codabench Evaluation contains only `id`, `language`, `sentence`, and `tokenized_sentence`, with the true labels withheld to support final blind testing.

2.5 Preprocessing Pipeline

Before training, all dataset entries undergo a standardized preprocessing pipeline to ensure compatibility with transformer-based models and accurate span labeling. Each sentence is tokenized at the word level, and then further split into subword tokens using the XLM-RoBERTa tokenizer. During this process, word-level idiom spans are mapped to the corresponding subword indices, and BIO (Begin, Inside, Outside) labels are assigned to each subword: the first subword of an idiom span receives a `B` tag, subsequent subwords within the same idiom receive an `I` tag, and all other tokens are labeled as `O`. Special tokens (such as `[CLS]` and `[SEP]`) and padding are assigned a label of `-100` to be ignored during loss computation. The preprocessing also includes dynamic padding and truncation to a fixed maximum sequence length, ensuring efficient batching. For training, class weights are computed based on label frequencies to address class imbalance. This preprocessing pipeline guarantees that both original and augmented data are consistently formatted for model input and evaluation.

3 Related Work

Detecting idiomatic expressions has long been a challenging subtask in Natural Language Processing (NLP) due to their non-compositional semantics and strong dependence on cultural and contextual cues. With the emergence of large-scale pre-trained language models, idiom processing has

attracted renewed interest, especially in multilingual and low-resource settings.

Most prior research on idiom detection has approached the task at the **sentence level**, where the model predicts whether an input sentence contains an idiomatic expression or not. This binary classification setup, while useful for downstream tasks such as sentiment analysis or machine translation, does not address the more granular problem of **idiom span identification**, which is critical for tasks like semantic parsing, explainability, and text simplification.

Briskilal and Subalalitha (2022) explored this sentence-level classification task by proposing an ensemble model that combines predictions from BERT and RoBERTa (4). Their approach demonstrates that ensemble methods leveraging multiple transformer-based models can boost idiom classification performance. However, the model is limited to sentence-level outputs and does not provide token-level span annotations, which restricts its applicability for fine-grained language understanding.

Building on the idea of leveraging diverse transformer representations, **Abarna et al. (2022)** introduced a stacking ensemble that incorporates K-BERT—a knowledge-enhanced version of BERT capable of integrating external semantic graphs—alongside BERT and RoBERTa (5). These base learners are combined using a logistic regression meta-classifier. By incorporating structured knowledge, their system improves idiom detection accuracy, particularly in distinguishing between literal and figurative uses of ambiguous phrases. Still, their work does not attempt to extract the exact location of idioms in context.

Despite these advances, existing systems are largely limited to English and rarely evaluate performance in morphologically rich or low-resource languages, where idioms often exhibit greater syntactic variation and cultural dependency. Furthermore, the lack of token-level predictions limits the interpretability and applicability of these models for more nuanced language tasks.

In contrast to these approaches, our work tackles the idiom detection problem from a **token-level perspective**, enabling the model to precisely locate idiomatic expressions within a sentence. We build on **XLM-RoBERTa**, a multilingual transformer model, to address the idiom span detection task in both **Turkish and Italian**, two typologically and morphologically rich languages. Unlike prior methods, we also integrate synthetic data generated by **GPT-4.1-mini** to improve generalization and robustness, especially in low-frequency idiom cases. Our architecture—comprising Weighted Loss, MLP classifier, and CRF decoding—provides structured span predictions aligned with the BIO tagging scheme, thereby offering a more detailed and language-agnostic approach to idiom identification.

4 Methodology

We developed two specialized models for idiom detection, one for Italian and one for Turkish, based on the XLM-RoBERTa architecture. Both models are framed as token-level classification systems, where each token is labeled as idiomatic (1) or non-idiomatic (0). In this section, we provide technical details for each.

4.1 Italian Idiom Detection Model

4.1.1 Model Architecture

Our Italian idiom detection system (see appendix figure-1) builds upon **XLM-RoBERTa Large**, a multilingual transformer with approximately 550 million parameters, pre-trained on 2.5TB of filtered CommonCrawl data across 100 languages. We augment this with a lightweight classification head—a single linear layer followed by a sigmoid activation—that outputs token-level probabilities indicating idiomatic usage.

4.1.2 Input Processing

Input sentences are tokenized using XLM-RoBERTa's subword tokenizer. Word-level idiomatic annotations are aligned to subword tokens. Sequences are padded or truncated to a fixed length of 128 tokens.

4.1.3 Training Setup

The model is trained using Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss) and optimized with AdamW. Key hyperparameters include a learning rate of 2e-5 with linear warmup over 500 steps, batch size of 16, dropout rate of 0.2, gradient clipping with max norm 1.0, and 16 epochs. We set a fixed random seed to ensure reproducibility.

4.1.4 Output and Post-processing

Token-level predictions are aggregated from subwords back to word-level spans, producing final idiomatic expression boundaries as token indices.

4.1.5 Dataset and Augmentation

Training data consists of idiomatic and literal Italian sentences from the idiom-corpus-llm repository. To augment this, we utilized **GPT-4.1-mini** to tokenize and annotate additional idiomatic examples in the same structured format. Each sample includes the original and tokenized sentences, idiomatic expression labels, and token-level span indices.

4.1.6 Performance

Our model achieves a token-level F1 score of **0.92** on the evaluation set and **0.91** on the test set, considering a prediction correct only when the full idiomatic span matches exactly.

4.2 Turkish Idiom Detection Model

4.2.1 Model Architecture

Our Turkish idiom detection system(see appendix figure-2) employs a more sophisticated architecture compared to the Italian model, utilizing **XLM-RoBERTa Large** as the backbone with additional sequence modeling components. The model incorporates a BiLSTM layer for enhanced sequence understanding, followed by a Multi-Layer Perceptron (MLP) classifier and a Conditional Random Field (CRF) layer for structured prediction. This architecture is specifically designed to handle the complex morphological structure of Turkish language and the BIO (Beginning, Inside, Outside) tagging scheme for idiom detection.

4.2.2 Input Processing

Similar to the Italian model, input sentences are tokenized using XLM-RoBERTa's subword tokenizer. However, the Turkish model processes these tokens through a BiLSTM layer (hidden dimension of 768 per direction) to capture sequential dependencies before classification. The MLP classifier, with a hidden dimension of 512, processes these enriched representations to produce token-level predictions.

4.2.3 Training Setup

The model is trained using a weighted loss function to handle class imbalance, with the following key hyperparameters:

- Learning rate: 2e-5 with linear warmup
- Batch size: 16
- Dropout rate: 0.1
- Number of epochs: 15

The training process incorporates several advanced techniques:

- CRF layer for structured sequence prediction
- Weighted loss function to address class imbalance
- MLP classifier for enhanced feature representation
- Optional BiLSTM layer for capturing sequential dependencies

4.2.4 Output and Post-processing

The model outputs BIO (Beginning, Inside, Outside) tags for each token, which are then processed through the CRF layer to ensure valid tag sequences. The final output is converted to token indices representing the boundaries of idiomatic expressions.

4.2.5 Performance

The Turkish model achieves an impressive token-level F1 score of **0.93** on the evaluation set and F1 score of **0.93** on the test set, with the best performance observed at epoch 12 out of 16 epoch. This performance demonstrates the effectiveness of the enhanced architecture in handling Turkish idiom detection, particularly considering the language's complex morphological structure.

5 Experiments & Results

5.1 Experimental Setup

To evaluate our idiom detection system, we conducted extensive experiments on both Turkish and Italian datasets. We used a sequence labeling approach with BIO tagging, leveraging the multilingual capabilities of **XLM-RoBERTa Large** as our base encoder. Training was performed on an NVIDIA A100-SXM4-40GB GPU using PyTorch and HuggingFace Transformers. The final model architecture includes optional **BiLSTM**, **MLP**, and **CRF** layers on top of the transformer embeddings, depending on the experimental configuration.

5.2 Training Configuration

For Italian, the best-performing model was trained for 16 epochs with a learning rate of $2e^{-5}$, dropout rate of 0.2, and batch size of 16. The model was evaluated after every epoch on a validation set of 751 instances, and the checkpoint with the highest

F1 score was saved. The best evaluation F1 score for Italian was **0.9221**, achieved at epoch 16.

For Turkish, we used 15 epochs of training with a batch size of 16 and dropout rate of 0.1. Unlike the Italian model, we excluded the BiLSTM layer in our final Turkish model for efficiency and generalization purposes. The best F1 score for Turkish was **0.93**, obtained at epoch 12.

5.3 Ablation Study

We conducted a series of ablation experiments to investigate the impact of model architecture and data augmentation on performance. Table 1 summarizes our results across different model configurations and training regimes.(see appendix table for results)

The transition from binary to token-level BIO classification substantially improved model performance in Turkish language. Adding the WeightedLoss function and MLP+CRF layers further increased accuracy, likely due to better sequence modeling and structured prediction. Augmenting the dataset with LLM-generated idiomatic expressions led to gains, especially in Italian, where idiom variety and data sparsity were more severe. The best final configuration used only human-annotated data for Turkish, and a mix of human and synthetic data for Italian, resulting in our highest overall performance.

5.4 Error Analysis

CRF decoding significantly reduced inconsistent BIO tagging (e.g., "I" tags without preceding "B"), enhancing the reliability of span predictions. The weighted loss approach also contributed to better handling of class imbalance between idiomatic and non-idiomatic tokens.

6 Discussion

Our idiom detection models demonstrated strong performance for both Turkish and Italian, with F1 scores exceeding 0.90 in span-level evaluation. This confirms the effectiveness of XLM-RoBERTa in capturing idiomatic expressions, even in underresourced languages.

Interestingly, increasing model complexity—such as adding a BiLSTM layer on top of XLM-RoBERTa did not consistently improve performance. In fact, the inclusion of a BiLSTM led to a noticeable drop in F1 scores. We hypothesize that this is due to the already rich contextualized representations produced by XLM-RoBERTa; additional sequential modeling may introduce redundancy or overfitting, particularly given the relatively small size of the dataset. These findings suggest that simpler classification heads are often sufficient when leveraging strong pretrained multilingual encoders for idiom detection.

Data augmentation using GPT-4.1-mini yielded mixed results. While the Italian model slightly benefited from the addition of LLM-generated examples, the Turkish model's performance marginally decreased. This suggests that simply expanding the dataset with synthetically generated sentences containing different idioms does not always lead to gains, and can even introduce distributional noise that affects generalization—especially in smaller training regimes.

Despite these limitations, even a basic binary classification variant of our model—without span prediction—achieved 0.89 accuracy, highlighting the robustness of XLM-RoBERTa for idiom classification tasks.

Challenges remain in handling rare idioms and cases with ambiguous semantics, and subword-level alignment continues to introduce complexity during both training and evaluation. Future work could explore the integration of external idiom lexicons, dynamic curriculum learning strategies for rare patterns, and extending the model to additional typologically diverse languages to further assess generalization.

7 Conclusion

In this work, we presented a multilingual idiom detection framework that operates at the token level, with a focus on two under-resourced languages: Turkish and Italian. Unlike prior approaches that typically treat idiom detection as a sentence-level classification task, our model precisely localizes idiomatic expressions within text, enabling more fine-grained linguistic analysis.

We fine-tuned **XLM-RoBERTa Large** for span-level idiom identification, leveraging its strong multilingual capabilities. To enhance the training data, we employed **GPT-4.1-mini** to generate additional annotated examples aligned with the structure of existing idiom corpora, thereby increasing idiomatic coverage and diversity.

Empirical results demonstrate the effectiveness of our approach, with token-level F1 scores of **0.92** for Turkish and **0.91** for Italian. These findings

underscore the potential of pretrained multilingual models for idiom detection and suggest promising applications in downstream tasks such as machine translation, language education, and cultural language analysis.

Overall, this study highlights the importance of token-level modeling for figurative language and sets the stage for further research on idiom understanding across diverse languages and linguistic typologies.

8 Acknowledgments

In the preparation of this report, We used generative AI tools only to assist with rephrasing some sentences and improving clarity in a few sections. However, all core ideas, analysis, project design, interpretation of results, and critical discussions were developed independently by us. The AI-generated content was carefully reviewed and edited to ensure it reflects our own understanding and work.

References

[1] Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. *A Hard Nut to Crack: Idiom Detection with Conversational Large Language Models*. In *Proceedings of the 4th Workshop on Figurative Language Processing (FLP)*, pages 35–44, June 2024. https://aclanthology.org/2024.flp-1.4

[2] Doğukan Arslan, Hüseyin Anıl Çakmak, Gülşen Eryiğit, and Joakim Nivre. *Using LLMs to Advance Idiom Corpus Construction*. In Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025), pages 21–31, Albuquerque, New Mexico, U.S.A., 2025. Association for Computational Linguistics.

[3] Gülşen Eryiğit, Ali Şentaş, and Johanna Monti. *Gamified crowdsourcing for idiom corpora construction*. Natural Language Engineering, 29(4):909–941, 2023. https://doi.org/10.1017/S1351324921000401

[4] J. Briskilal and C. N. Subalalitha. *An ensemble model for classifying idioms and literal texts using BERT and RoBERTa*. Information Processing & Management, 59(6):103041, 2022. Elsevier. https://doi.org/10.1016/j.ipm.2022.103041

[5] S. Abarna, J. I. Sheeba, and S. P. Devaneyan. *An ensemble model for idioms and literal text classification using knowledge-enabled BERT in deep learning*. Measurement: Sensors, 24:100434, 2022. Elsevier. https://doi.org/10.1016/j.measen.2022.100434

9 Appendix

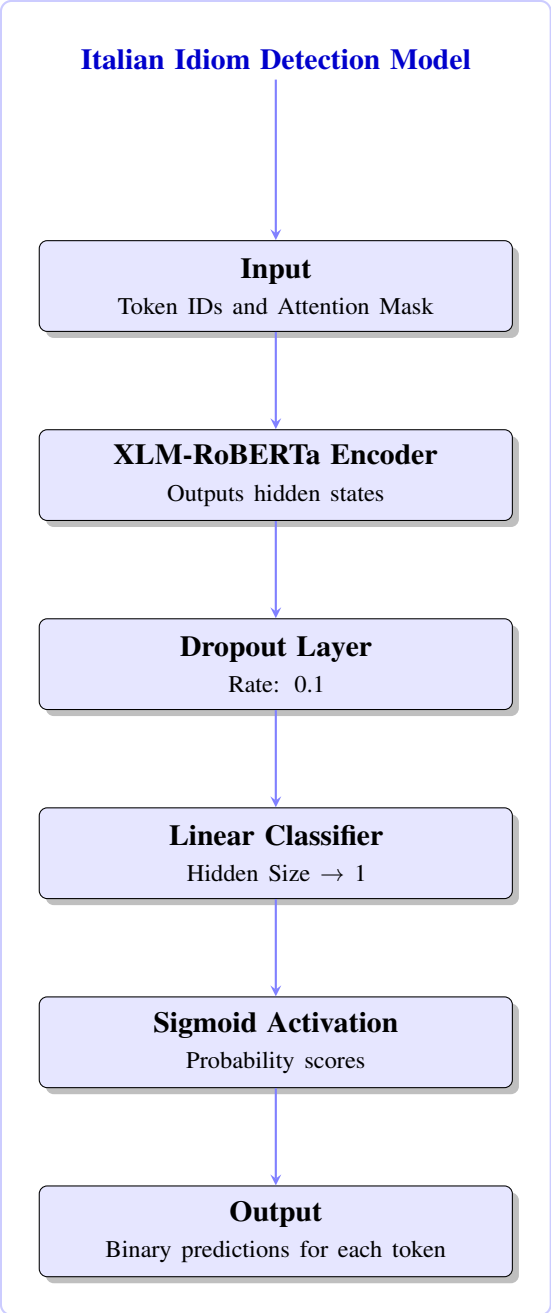

Figure 1: Architecture of the Idiom Detection Italian Model. The model processes input tokens through XLM-RoBERTa encoder, followed by dropout regularization and a linear classifier to predict binary labels for each token. The sigmoid activation converts logits to probabilities, which are then thresholded to make final predictions.

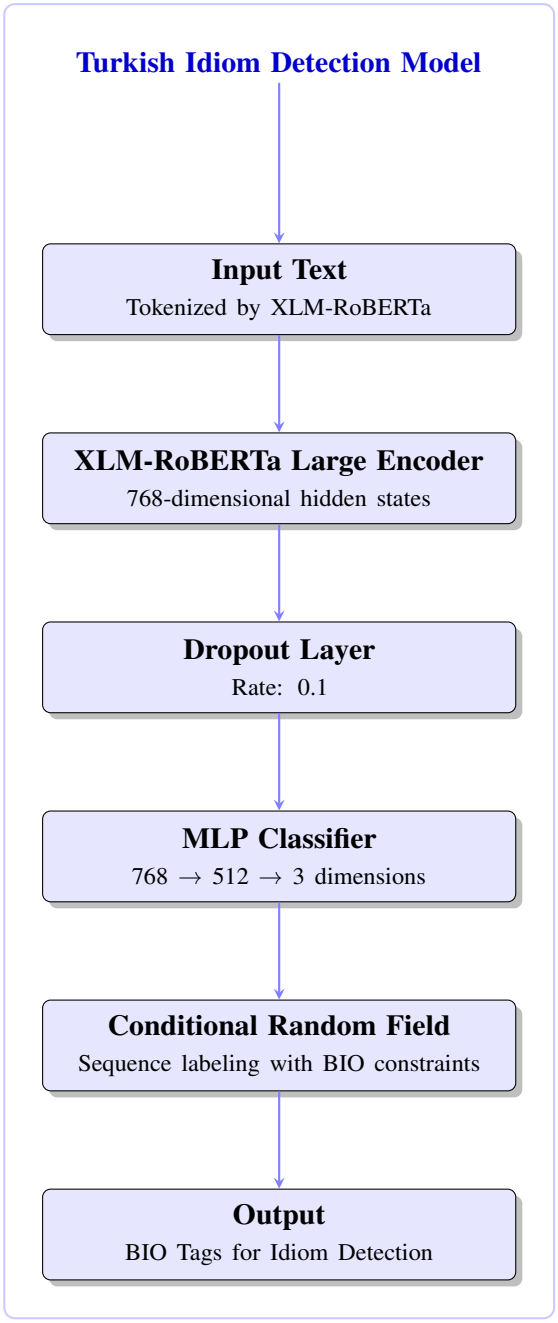

Figure 2: Architecture of the Turkish Idiom Detection Model. The model processes input text through XLM-RoBERTa Large encoder, followed by dropout regularization, MLP classification, and CRF-based sequence labeling to predict BIO tags for idiom detection.

Configuration	F1-TR	F1-IT	Avg. F1
Binary Classification (BCE loss)	0.8915	0.8697	0.8806
BIO Classification + BiLSTM + MLP + CRF	0.9110	0.9071	0.9091
+ LLM-Generated Data (Augmentation)	0.9156	0.9073	0.9115
Final Setup (Human-only TR, Hybrid IT)	**0.9224**	**0.9104**	**0.9164**

Table 1: F1 scores across different model and data configurations.