Metaphor in Context: Transformer-Based Token Classification with the VU Amsterdam Metaphor Corpus

Aidana Akkaziyeva

Natural Language Processing.
Data Science for Economics, University of Milan.

aidana.akkaziyeva@studenti.unimi.it

Abstract

This project explores the automatic detection of metaphorical language in text using advanced natural language processing techniques. Leveraging a token classification approach with a RoBERTa-based transformer model, we address the challenge of class imbalance through class weighting and empirical optimization. The dataset is preprocessed to align word-level metaphor annotations with tokenized inputs, and part-of-speech (POS) information is incorporated to analyze model performance across linguistic categories. The impact of class balancing and POS on metaphor detection was evaluated, reporting detailed metrics for nouns, verbs, adjectives, and other categories. The results demonstrate that class-balanced training improves metaphor identification, particularly for verbs and less frequent POS types, while highlighting ongoing challenges with adjective metaphors.

Keywords: metaphor detection, non-literal language, VU Amsterdam Metaphor Corpus, token classification, transformers, DistilRoBERTa, class imbalance, natural language processing

1 Introduction

Metaphor plays a central role in human language, shaping the way abstract concepts are understood through more concrete domains. Automatic metaphor detection has therefore become an important area of research, with applications in figurative language understanding, sentiment analysis, and computational linguistics.

Early approaches relied on feature-engineered classical machine learning models, but recent advances have shifted toward deep learning and transformer-based architectures. Models like BERT and RoBERTa have demonstrated competitive performance in token-level sequence labeling, making them strong candidates for metaphor detection tasks [1] [2].

This project investigates metaphor detection using the VU Amsterdam Metaphor Corpus (VUAMC) [3]. The corpus provides metaphor annotations at the word level, aligned with part-of-speech (POS) tags. The study focuses on transformer-based models (DistilRoBERTa) with a special attention to analyzing performance across different POS categories (nouns, verbs, adjectives, and others).

2 Research question and methodology

2.1 The central research question is:

How effectively can a transformer-based model (DistilRoBERTa) detect metaphorical language in the VU Amsterdam Metaphor Corpus, and how does performance vary across parts of speech?

2.2 Corpus Characteristics and Data Preparation

For this study, we use the **VU Amsterdam Metaphor Corpus (VUAMC)**, which is the largest available corpus hand-annotated for all metaphorical language use, regardless of lexical field or source domain [3]. The main features of the VUAMC are as follows:

- **Annotation Protocol**: The corpus was annotated according to the **MIPVU** (Metaphor Identification Procedure Vrije Universiteit), a systematic and explicit method for identifying metaphorical language.
- **Size**: The dataset contains approximately **190,000 lexical units**, providing extensive coverage of metaphorical usage in English.
- **Registers**: The data is drawn from four broad registers of the BNC-Baby, including academic texts, conversation, fiction and news texts.
- **Annotation Categories**: Each lexical unit is annotated for its relation to metaphor: literal or metaphor-related word.

2.3 Preprocessing and Simplification

For the purposes of this project, the labeling scheme was simplified to focus exclusively on **metaphor-related words (mrw)**. All metaphor categories were collapsed into a single **binary label**:

- 0 = literal
- 1 = metaphorical

Furthermore, only sentences containing at least one metaphor annotation are retained. In the project's scope, keeping only metaphor-containing sentences was found a reasonable design choice:

- It reduced dataset size (important for efficiency).

- It guaranteed the model saw enough positive examples of metaphors, which are relatively rare.

Sentences shorter than three words were removed, as these typically lacked sufficient context for meaningful analysis. The dataset is transformed from document-level to sentence-level, with each row representing a sentence and its word-level annotations. POS tags are simplified into categories: 'verb', 'noun', 'adj', 'other', or 'na' (for non-metaphorical words). After preprocessing, the final dataset consisted of **8,193 sentences**. These sentences form the basis for subsequent training, validation, and testing in our experiments.

2.4 Tokenization and Dataset Construction

A custom MetaphorSentenceDataset class is implemented to:
- Align word-level labels and POS tags to subword tokens.
- Assign label -100 to subwords (except the first subword of each word) and special tokens, so they are ignored during loss calculation and evaluation.
- Store tokenized inputs, attention masks, aligned labels, and POS tags for each sentence.

2.5 Data Splitting

To prevent data leakage, the split is performed at the document level.

- 70% training (5725 sentences);
- 15% validation (1416);
- 15% test (1052).

For computational efficiency, for the baseline model with no optimized weights, a sampled subset was used (2000, 500, 500 corresponding to training, validation and test sets).

2.6 Model Architecture

The model is based on RobertaForTokenClassification with two output classes (literal, metaphor). The classification head is fine-tuned on the metaphor detection task. Class weights are computed based on the distribution of literal and metaphor tokens in the training set. A custom WeightedTrainer subclass of HuggingFace's Trainer is used to apply these weights in the loss function.

2.7 Training Procedure

Training Arguments:
- Training uses mixed precision (fp16), and saves the best model based on F1 score.
- The batch size (64), learning rate (2e-5), and number of epochs (3-4) are set for efficient training.
Loss Function:
- Weighted cross-entropy loss is used, with class weights to address class imbalance.
- Only tokens with labels 0 or 1 are included in the loss; tokens with label -100 are ignored.

2.8 Evaluation

Evaluation and Metrics: The model predicts on the test set, outputting logits for each token. Only the first subword of each word is evaluated. Metrics include accuracy, precision, recall, and F1 score, both overall and per class.
POS-Conditioned Evaluation: Performance is further analyzed by grouping tokens according to their simplified POS tags (e.g., noun vs. literal, verb vs. literal). Classification reports are generated for each POS group.
Error Analysis: Token-level predictions are aligned back to original words for interpretability. Example outputs are displayed for qualitative analysis.

3 Experimental results and Discussion

3.1 Baseline Model (no class balancing)

The project began with a relatively straightforward RoBERTa baseline: training without any special handling of class imbalance. Unsurprisingly, this model performed well on the majority class (literal tokens), reaching 94% F1 for literals. However, it struggled with metaphors, achieving 63% F1, with a noticeable gap between precision (0.75) and recall (0.54). In practical terms, the baseline model adopted a conservative strategy of labeling most tokens as literal, which boosted its accuracy on the majority class but led to many missed metaphors.

3.2 Class-Balanced Model (initial weights)

To address the issue above, class weights (Literal=0.59, Metaphor=3.34) to counteract imbalance were aplied. The new model traded some precision for a substantial boost in recall: metaphor recall jumped from 0.54 $\rightarrow$ 0.88, while precision fell from 0.75 $\rightarrow$ 0.53. The metaphor F1 improved to 0.66, and overall macro averages also increased. This shift confirmed that the model was previously too conservative; weighting encouraged it to "take more risks" in identifying metaphorical tokens.

3.3 Optimized Class Weights (multiplier=0.4)

For the final model, several weight multipliers for the metaphor class were tested. The multiplier that yielded the highest F1 score for metaphors on the validation set was selected for use in the final model. This approach allowed us to empirically balance the trade-off between precision and recall for the minority class without incurring high computational costs. On the full test set, the model achieved 0.74 F1 for metaphors, which is a notable improvement compared to the earlier subsampled experiments. At the same time, Literal tokens remain highly precise and stable (0.96 precision, 0.94 recall), while metaphors lean toward higher recall (0.79) than precision (0.69). In other words, the model tends to over-predict metaphors, capturing many true cases but introducing slightly more false positives. This is a deliberate trade-off: without balancing, the model was overly conservative and missed many metaphors; with balancing, it is more tuned to identify non-literal language.

Table 1: Overall performance comparison

Model	Class	Precision	Recall	F1-score	Support
Baseline	Literal	0.93	0.96	0.94	7571
	Metaphor	0.73	0.57	0.64	1352
	Accuracy			0.90	8923
Class-Balanced	Literal	0.97	0.86	0.91	7571
	Metaphor	0.53	0.88	0.66	1352
	Accuracy			0.86	8923
Optimized Model + All Data	Literal	0.96	0.94	0.95	15567
	Metaphor	0.69	0.79	0.74	2763
	Accuracy			0.92	18330

3.4 POS-Specific Results

3.4.1 Nouns:

The F1-score for noun metaphors increased from 0.26 to 0.38 compared to the baseline model with less data and with no optimized weights. To be exact, precision improved from 0.16 to 0.29, while recall decreased from 0.74 to 0.54. This suggests the final model is more conservative in predicting noun metaphors, reducing false positives at the expense of some true positives.

3.4.2 Adjectives:

Similar to nouns, the model now makes fewer incorrect metaphor predictions for adjectives, though it misses more true cases. For adjectives, the F1-score rose from 0.13 to 0.22, with precision doubling from 0.07 to 0.14 and recall dropping from 0.77 to 0.49. A better balance between precision and recall is obtained in the last model.

3.4.3 Verbs:

The verb metaphor F1-score improved substantially from 0.34 to 0.54. Precision nearly doubled (from 0.21 to 0.41), without recall going too low (from 0.94 to 0.77). This shows the model is now much better at correctly identifying verb metaphors without over-predicting them.

3.4.4 Other POS:

The "other" category saw the largest gain, with F1 rising from 0.46 to 0.70 and precision from 0.30 to 0.58. Recall remained high, indicating robust performance for less common POS types.

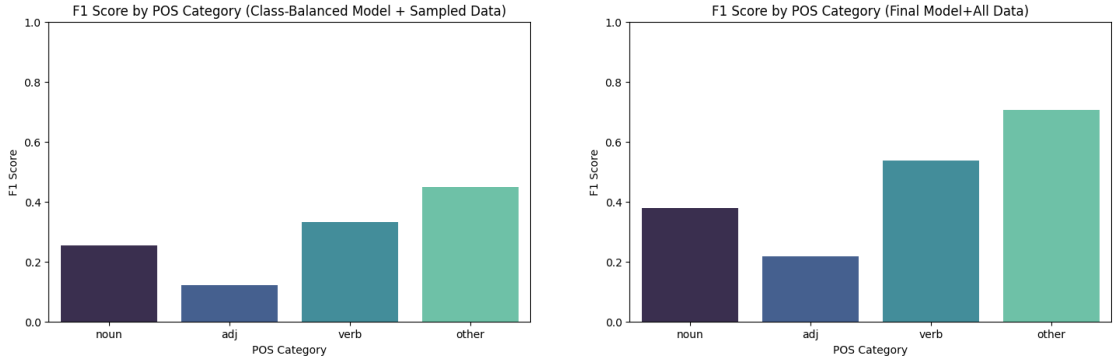

Fig. 1: F1 Score by POS Category

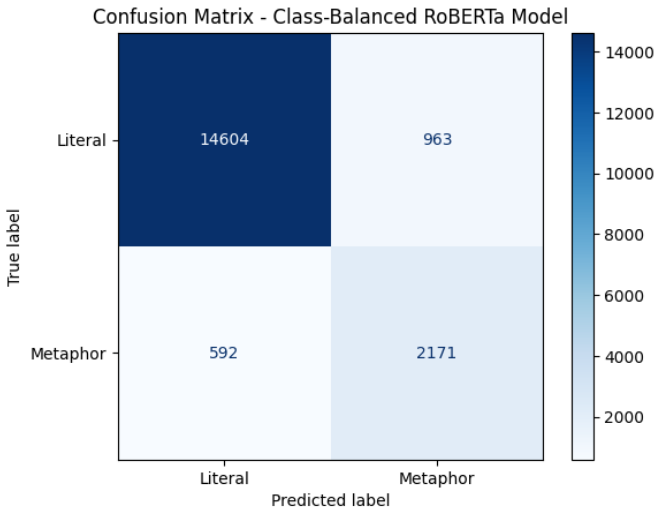

Fig. 2: Confusion Matrix

4 Concluding remarks

As observed from the evaluation metrics above, the model is much better at detecting metaphors in verbs and "other" POS than in adjectives. For the latter, recall is moderate but precision is very low, meaning the model predicts many false positives for metaphors in these categories. This could be because metaphors in these categories are more subtle or context-dependent, making them harder for the model to spot. For verbs and "other", recall is high and precision is better, especially for "other". This is a common pattern in metaphor detection observed in previous research [4][5].

4.1 Error Analysis and Example Insights

Looking beyond aggregate metrics, examples of individual predictions reveal the kinds of metaphors the model handles well and where it fails (see Table 1).

4.1.1 Success cases:

Words like "approach" (noun) and "drop" (noun) were correctly identified as metaphors. These examples suggest that the model is capable of detecting when technical or abstract terms shift meaning, e.g., "drop" in physics rather than its literal sense of something falling.

Similarly, phrases such as "world of fashion" (noun metaphor) and "little holiday" (adjective metaphor) show that the model captures common metaphorical extensions in everyday discourse.

4.1.2 Challenging cases:

Function words like "in", "to", and "that" were sometimes predicted as metaphorical. While these are rarely metaphors in themselves, they appear in metaphor-heavy contexts (e.g., "perpendicular both to the direction of motion . . . "), which may cause the model to overgeneralize. This suggests the token-level classifier is sometimes "dragged" by surrounding context rather than focusing on the word itself.

The word "virtue" illustrates another limitation: the model predicted Literal while the label was Metaphor. This reflects the difficulty of borderline cases where even humans might debate whether a phrase like "by virtue of" is metaphorical or a lexicalized idiom.

Overall, these examples confirm what the quantitative analysis already hinted:

The model is strongest with content words (especially verbs and technical nouns) where meaning shifts are clearer. It is weakest with function words and highly conventionalized expressions, where context or pragmatics play a larger role.

These findings emphasize the importance of error analysis in metaphor detection, understanding which linguistic categories the model is sensitive to, and which it misclassifies systematically is crucial, too.

Table 2: Examples with Word, POS, Labels, and Sentences

Word	POS	True Label	Pred Label	Sentence
in	other	Metaphor	Metaphor	Then his example lured his elder son Bertrand to Tripoli **in** 1112, and his younger son Alphonse Jourdain there in 1147.
approach	noun	Metaphor	Metaphor	Latest corporate unbundler reveals laid-back **approach**: Roland Franklin, who is leading a 697m pound break-up bid for DRG, talks to Frank Kane.
drop	noun	Metaphor	Metaphor	Not really, because (1) as current flows there is some potential **drop** in the electrode itself, (2) zero conductivity for part of the space cannot be electrostatically modelled since there are no dielectrics with $E_r = 0$, and (3) when current flows through two materials of different conductivity there is generally a surface charge at the boundary.

Continued on next page

Word	POS	True Label	Pred Label	Sentence
world	noun	Metaphor	Metaphor	Her favourite was the coffee shop in Fenwicks in Bond Street for this was the haunt and the meeting place of all those from the **world** of fashion.
little	adj	Metaphor	Metaphor	And it'll be really rather dolly, we'll have a **little** holiday you see together and we sort of go teaching together and...
virtue	noun	Metaphor	Literal	So the electrons want to move inwards but cannot because the ions hold them back by **virtue** of their electrostatic attraction.
to	other	Metaphor	Metaphor	The magnetic force is perpendicular both **to** the direction of motion (+z axis) and the direction of magnetic field (formula direction).
that	other	Metaphor	Metaphor	Just turn them off, and **that**'s it. Just normal conversations, the words that people use in common in different areas of the country, with accents and dialects, and one thing and another — it's for the Oxford English Dictionary, the next edition.

4.2 Limitations and Future Work

4.2.1 Alternative Labeling Strategy

As was mentioned earlier, metaphors were annotated based on "metaphor-related words" (MRWs), which include direct, indirect, and implicit metaphors. While this provides broad coverage, it may not align well with what a token-level classifier can realistically capture. For instance, some implicit metaphors depend more on sentence-level semantics than on individual words, which explains why function words ("in", "to", "that") were sometimes incorrectly tagged as metaphors. Future work could explore alternative labeling strategies, perhaps distinguishing between metaphor types, or shifting to a phrase-level (e.g. adjective-noun pair) rather than word-level task.

4.2.2 "Other" POS Category

Another limitation of the current analysis that is closely connected to the previous point, lies in the treatment of the "other" part-of-speech (POS) category. This group includes all tokens that are not nouns, verbs, or adjectives, such as prepositions, determiners, conjunctions, and adverbs. While some of these words can occasionally be used metaphorically (e.g., down in "prices went down"), they are extremely heterogeneous in function and frequency. As a result, the model struggles to learn a consistent pattern for this class, and evaluation results for "other vs. literal" are difficult to interpret. In some cases, the classifier shows high recall for this category, but this is likely due to noisy predictions rather than a genuine ability to detect metaphoricity.

From a theoretical perspective, most metaphor research emphasizes nouns, verbs, and adjectives as the primary carriers of metaphorical meaning. By contrast, function words are less central to the phenomenon, even if they occasionally participate in metaphorical expressions. For this reason, future work could focus solely on the more semantically meaningful classes.

Another promising direction would be to incorporate syntactic or dependency information, which may help capture metaphorical uses of function words that depend heavily on context. For example, prepositions like *over* or *beyond* might reveal metaphorical usage more clearly when analyzed in relation to their head words.

Overall, removing or restructuring the "other" category could lead to cleaner, more interpretable results and better alignment with linguistic theory, while future experiments could investigate whether subcategories of function words play a more meaningful role in metaphor detection.

4.2.3 POS imbalance

Out of the 2763 metaphor tokens in the test set, the distribution is highly skewed:

Nouns: 600 (22%)
Adjectives: 255 (9%)
Verbs: 697 (25%)
Other: 1211 (44%)

Aside for last category, verbs make up the majority of metaphor labels, while adjectives are far less represented. This uneven distribution likely explains why the model performs so differently across categories. Addressing this imbalance could take several forms:

- Ensuring similar numbers of training examples per POS category (through resampling).

- Assigning different class weights per POS category, rather than applying a single balance between "literal" and "metaphor."

- Explicitly incorporating POS tags as features into the model (e.g., joint training with a POS embedding), so that the classifier has more linguistic context.

4.2.4 Model and Training Setup

Due to the time constraints, only distilroberta-base, faster and lighter version of RoBERTa-base, was tested, which is efficient but not as strong as larger variants (e.g., roberta-base, roberta-large).

Future work could compare different transformer architectures, or leverage domain-adapted pretraining for better metaphor sensitivity.

AI Usage Disclaimer

Parts of this project have been developed with the assistance of OpenAI's ChatGPT (GPT-4). AI was used to support the development of project ideas, the structuring of methodological workflows, the drafting of descriptive texts, and the identification of relevant datasets and references. Additionally, AI provided alternative explanations for technical steps and contributed ideas for structuring the methodology, such as approaches for class balancing and rapid hyperparameter search. These suggestions were critically assessed, tested, and adapted to fit the project's goals.

AI also pointed out potential methodological weaknesses, including the risks of overfitting when using small validation sets for hyperparameter tuning and the importance of avoiding data leakage. All content produced with AI assistance has been carefully reviewed, edited, and validated by me. I take full responsibility for the final content and its accuracy, relevance, and academic integrity

References

[1] Choi, M., Lee, S., Choi, E., Park, H., Lee, J., Lee, D., Lee, J.: MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1763–1773. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.naacl-main.141 . https://aclanthology.org/2021.naacl-main.141/

[2] Song, W., Zhou, S., Fu, R., Liu, T., Liu, L.: Verb metaphor detection via contextual relation learning. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4240–4251. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.327 . https://aclanthology.org/2021.acl-long.327/

[3] Steen, G.J., Dorst, A.G., Herrmann, J.B., Kaal, A.A., Krennmayr, T., Pasma, T.: A Method for Linguistic Metaphor Identification. From MIP to MIPVU. John Benjamins, Amsterdam (2010)

[4] Elzohbi, M., Zhao, R.: ContrastWSD: Enhancing Metaphor Detection with Word Sense Disambiguation Following the Metaphor Identification Procedure (2024). https://arxiv.org/abs/2309.03103

[5] Liu, J., O'Hara, N., Rubin, A., Draelos, R., Rudin, C.: Metaphor detection using contextual word embeddings from transformers. In: Klebanov, B.B., Shutova, E., Lichtenstein, P., Muresan, S., Wee, C., Feldman, A., Ghosh, D. (eds.) Proceedings of the Second Workshop on Figurative Language Processing, pp. 250–255. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.figlang-1.34 . https://aclanthology.org/2020.figlang-1.34/