Assessing XLM-R for Metaphor Detection in Climate Technology Discourse

Urtė Jakubauskaitė
The Network Institute
Vrije Universiteit Amsterdam
urte.jakubauskaite@gmail.com

Abstract

Metaphor detection systems have achieved strong performance on general-domain texts, but their effectiveness in specialized domains remains underexplored. This study investigates the performance of a pre-trained metaphor detection model (XLM-RoBERTa) on climate-technology-related articles. While the model demonstrates robust results on general-domain data, its performance drops significantly when applied to domain-specific texts. I examine whether continued pre-training on unlabelled climate-technology-related data can enhance performance in this context. Results show that many domain-specific metaphors remain undetected. Moreover, there are only marginal improvements in precision, recall, and F1-score, with considerable variability across articles. The findings emphasize the need for more training data, refined annotation strategies, and domain-aware approaches to improve metaphor detection in expert domains.

1 Introduction

Metaphors play a vital role in how people comprehend complex and abstract concepts by linking unfamiliar ideas to familiar experiences. This linguistic phenomenon is especially influential in shaping public understanding of scientific innovations and social issues. Among such domains, emerging climate technologies stand out as an important area where effective communication can drive awareness and policy action.

Despite the importance of metaphors in climate technology discourse, detecting metaphorical language remains a challenging task. Advances in natural language processing (NLP) have shown promise for metaphor detection across multiple languages and domains. However, the effectiveness of these models in specialized contexts like climate technology has yet to be explored.

This paper investigates the capabilities of a metaphor detection model based on XLM-RoBERTa (XLM-R) within the climate technology domain. Specifically, it examines whether continued pre-training on domain-specific texts enhances model performance. By addressing this question, the study aims to contribute to improved tools for analyzing metaphorical language in climate technology communication.

2 Related Work

This study builds upon code provided by Wachowiak et al. (2022), who developed a systematic and replicable pipeline to detect image-schematic conceptual metaphors (ISCMs). These metaphors function as cognitive mechanisms that transfer knowledge from one domain to another by mapping sensorimotor experiences to abstract concepts.

Traditionally, the identification of such metaphors relied heavily on introspection or manual corpus analysis, making it time-consuming, subjective, and hard to replicate. To address these issues, Wachowiak et al. focused on the *SUPPORT* image schema in COVID-19-related texts, integrating steps such as automatic metaphor detection, dependency parsing, clustering, and frame annotation using BERT-based models.

In their automatic metaphor detection component, they trained a word-level classifier on the VU Amsterdam Metaphor Corpus, utilizing the multilingual pre-trained language model XLM-R. This model labels each word in a sentence as either literal or metaphorical and achieves a F1-Score of 0.76 for the metaphorical expressions.

However, the study does not extend such automated metaphor detection techniques to specialized domains. To address this gap, I employ their metaphor detection code to conduct a domain-specific analysis in the context of climate technology.

3 Data

To prepare and evaluate metaphor detection performance in climate technology discourse, I used several datasets, described in detail below.

3.1 Training Data

The original Vrije Universiteit Amsterdam (VUA) Metaphor Corpus was used as the primary training resource. This dataset consists of English sentences annotated for metaphor at the word level, following the MIPVU Metaphor Identification Procedure (Steen et al., 2010). Although the corpus consists of a wide range of genres, it is not domain-specific. In the original study by Wachowiak et al. (2022), the dataset is split into training, validation, and test subsets.

The training set comprises 181,488 tokens (referred to as *words* in the dataset) across 10,912 sentences, with 19,177 metaphorically used tokens (10.6%). The vocabulary size is 17,290, and the average sentence length is 16.63 tokens, with an average of 3.18 metaphors per sentence. Although I did not train the model myself, I used a pre-trained model based on this dataset as the foundation for metaphor detection in this study.

3.2 Continued Pre-training Data

To enable domain adaptation, I used a large collection of climate-technology-related articles for continued pre-training. This dataset consists of 1,462 articles, totaling 2,026,996 tokens. The average article length is 1,386.45 tokens.

The articles were selected from LexisNexis (2009) by Femke van Bruggen based on keywords focusing on texts related to *carbon capture*. However, the topics covered extend beyond carbon capture to broader related areas such as green energy, innovation policy, and environmental regulation.

Although this corpus is not annotated for metaphor, it allows the model to be exposed to domain-specific vocabulary and linguistic structures prior to evaluation.

3.3 Testing Data

To assess model performance on in-domain data, I used six datasets. One of these is the original VUA test dataset, followed by five manually annotated articles related to climate technology. Four of these articles were selected from the LexisNexis database (2009) and annotated for metaphor using the modified MIPVU (Steen et al., 2010) guidelines, consistent with those used in the VUA corpus. The fifth article was provided by Femke van Bruggen, a PhD candidate at Vrije Universiteit Amsterdam, who researches metaphors in climate communication. All test data, except the VUA test dataset, provide a focused environment for domain-specific evaluation.

3.3.1 VUA Benchmark Test Data

To test the model, I first ran it on the original VUA test data as a sanity check. The test set contains 58,359 tokens across 3,814 sentences, with 6,819 metaphorical tokens (11.7%) and a vocabulary size of 7,763. The average sentence length is 15.30 tokens, with an average of 3.18 metaphors per sentence.

Importantly, in both the train and test VUA datasets, the ten most frequent metaphorical words are *in*, *to*, *on*, *with*, *that*, *this*, *from*, *at*, *about*, and *have* (not necessarily in this order), indicating that function words are the most common metaphorical tokens. While our focus lies on content words, these results show that the test set well represents the whole VUA dataset.

3.3.2 Climate Technology Articles

Table 1 specifies some details about each of the four climate-technology-related articles used for model evaluation.

The four articles used for testing were randomly selected from van Bruggen's dataset of climate-technology-related articles collected from LexisNexis (2009). Manual annotation was performed independently by two annotators, Bastiaan Sizoo and Urtė Jakubauskaitė, following the principles of MIPVU (Steen et al., 2010) with slight modifications. The annotation process involved the following steps:

- Each annotator first skimmed each article individually to grasp its overall meaning,
- Starting from the first token, annotators verified the accuracy of auto-assigned part-of-speech (POS) tags,
- Only content words (nouns, verbs, adjectives, and adverbs) were considered; function words were skipped as irrelevant for this metaphor analysis,
- Tokens forming polywords (multiword expressions) were skipped entirely,

Article	Tokens	Sentences	Metaphors	Vocabulary Size	Average Sentence Length	Average Metaphors per Sentence	Function Words
1 (**4acb1**)	487	19	48	236	25.63	2.53	36.76%
2 (**7976c**)	219	10	34	145	21.90	3.40	40.64%
3 (**58717**)	304	16	43	175	19.00	2.69	38.82%
4 (**fe9d0**)	220	13	34	133	16.92	2.62	34.09%
Total	1,230	58	159	—	—	—	—

Table 1: Statistics of four annotated climate technology articles from LexisNexis.

- For each candidate token, the contextual meaning was interpreted and compared against dictionary definitions from the Longman Dictionary (2015), considering only definitions matching the word's POS and excluding technical or business-specific senses,
- A token was marked as metaphorical if there was a contrast between the contextual meaning and a more concrete, physical, or body-related basic meaning, provided the two senses were comparable,
- Tokens were also marked metaphorical if both senses were concrete but clearly different and comparable,
- For every metaphorical token, a brief motivation was recorded to document the reasoning,
- In cases of uncertainty, tokens were included as metaphorical to maintain consistency,
- Additional observations were noted in a comment column,
- After individual annotation, reconciliation sessions were held to compare results and reach consensus, with final decisions marked in the `FINAL` column of the dataset.

The annotated test data are provided together with this paper. Although the documents remain in their working state with inconsistent comments and formatting, they all share the same columns, namely `sent_id`, `token_id`, `token_text`, `pos`, and `FINAL`.

3.3.3 Femke van Bruggen's Dataset

The original dataset, provided by Femke van Bruggen, was adapted to suit this study. Femke focused specifically on carbon-removal–related signalled metaphors, each marked with a flag for identification. For example, in the sentence *The greenhouse gas acts as a kind of magic dust for concrete, making it up to 40 per cent stronger*, the metaphorical phrase *magic dust* was flagged because of the marker *as*.

Instead of using the full dataset, I manually selected only those sentences that included sufficient context. Isolated metaphorical phrases such as *buffer pool* were excluded. Additionally, I limited the selection to examples containing metaphors consisting of one or two words, omitting longer metaphorical expressions.

For data preparation, I used Python code to process and format the dataset consistently with the other test sets used in this project. The workflow included the following steps:

- The original Excel file was flattened into a single-column list of sentences, with empty cells and whitespace removed,
- After manual filtering (as described above), the resulting file was tokenized using spaCy, assigning sentence and token IDs, POS tags, and placeholders for metaphor labels,
- Following manual annotation of metaphorical expressions (they were marked in **bold** in the original dataset), token IDs were corrected to be continuous across the full dataset (rather than restarting for each sentence),
- Empty cells in the `FINAL` column were filled with zeros to mark non-metaphorical tokens.

The final dataset contains 223 sentences and 6,712 tokens. Importantly, this test set should be used solely for evaluating recall and is unsuitable for evaluating precision. Only climate-technology-related metaphors were annotated, while more general metaphorical expressions were deliberately left unmarked (even though the model may still identify them).

All the datasets, used in this study, are publicly available at Github[1].

[1] https://github.com/urtuteja/Metaphor-Detection-XLMR-Climate-Technology

4 Methodology

This study builds upon the publicly available metaphor detection model `lwachowiak/Metaphor-Detection-XLMR`, hosted on Hugging Face[2]. The model is based on XLM-R, a transformer-based multilingual model developed by Facebook AI. This particular variant was fine-tuned for word-level binary classification (metaphorical vs. literal) using the VU Amsterdam Metaphor Corpus (VUA) in English (Steen et al., 2010). While the model demonstrates strong performance on the general-domain VUA dataset, its ability to detect metaphors in specialized domains remains uncertain.

The main goal of this study is to assess the model's suitability for domain-specific metaphor detection in the context of climate technology. This is achieved by applying the model to several domain-specific test datasets (described in Section 3) and further adapting it through continued pre-training.

4.1 Baseline Evaluation

As an initial step, the original Hugging Face model is evaluated on all available test datasets: the original VUA test set, four climate-technology-related datasets, and Femke van Bruggen's dataset. This baseline evaluation serves both as a sanity check and as a reference point to assess how well the pre-trained model generalizes to previously unseen, specialized domains without any further adaptation.

4.2 Domain Adaptation via Continued Pre-training

To enhance the model's domain awareness, I apply continued pre-training (also known as domain-adaptive pre-training) on a large collection of unlabelled climate-related texts. Here, I used a portion of code provided in the blog article by the user Hey Amit (2023). Specifically, 1,462 articles comprising 2,026,996 tokens are used to further train the base XLM-R model using a masked language modeling (MLM) objective. This step helps the model internalize domain-specific vocabulary, syntax, and discourse patterns that are characteristic of climate-technology-related writing.

4.3 Supervised Fine-Tuning

Following continued pre-training, the model is fine-tuned on the original VUA training set to preserve its metaphor detection capabilities. This supervised learning step ensures that domain adaptation does not diminish the model's ability to distinguish between metaphorical and literal expressions.

4.4 Evaluation

Finally, the adapted model is evaluated on all test datasets once again. Evaluation metrics include precision, recall, and F1-score. Performance is reported separately for each dataset.

The complete code for data pre-processing, model training, and evaluation is publicly available at Github[3].

5 Experiments and Results

This section presents the results for two main stages of evaluation, namely evaluation **without** domain adaptation and evaluation **with** domain adaptation. All experiments were conducted using the test datasets described in Section 3.

5.1 Evaluation *Without* Domain Adaptation

In the baseline experiment, the original Hugging Face model (`lwachowiak/Metaphor-Detection-XLMR`) was evaluated without any modifications. The goal was to assess its performance on previously unseen, domain-specific data. The model was applied to the original VUA test set (general domain), four climate-technology-related test sets (including manually annotated metaphors), and Femke van Bruggen's dataset.

5.1.1 VUA Benchmark Test Data

To verify the performance of the original model, I first evaluated it on the VUA test set. As expected, the results match those reported in the model's documentation. The model achieves high precision (0.824) and a reasonably strong recall (0.709) for metaphorical tokens, resulting in an F1-score of 0.762.

5.1.2 Climate Technology Articles

Importantly, each of the four climate-technology-related articles were evaluated seperately, because the performance varied drastically. While the

[2] https://huggingface.co/lwachowiak/Metaphor-Detection-XLMR

[3] https://github.com/urtuteja/Metaphor-Detection-XLMR-Climate-Technology

Article	Precision	Recall	F1-score
1 **(4acb1)**	0.50	0.15	0.23
2 **(7976c)**	**0.94**	0.48	0.64
3 **(58717)**	0.63	0.27	0.38
4 **(fe9d0)**	0.86	**0.55**	**0.67**

Table 2: Metaphor detection results for each article (relevant POS only) prior continued pre-training. The bold numbers indicate the highest scores.

model performs well on the VUA test set, achieving high precision and recall, its performance drops considerably on the domain-specific datasets. In particular, both the precision and recall on climate-related test sets are noticeably lower, with recall being worse and suggesting that the model fails to detect many metaphors that fall outside of its original training domain. This indicates that when the model does predict a metaphor, it is often correct, but it misses many domain-specific metaphors. Morevoer, the results, provided below, are all only for the selected, relevant POS tags (namely, nouns, verbs, adjectives and adverbs), as during the annotation process, we did not look at other parts of speech. The full results, including those for all POS, are presented in a *Metaphor-Detection-XLMR-Continued-Pretraining.html* file provided together with this paper.

1st Article: 4acb1 In the first article, the model achieved a relatively high precision (0.500) but very low recall (0.149) for metaphorical tokens, leading to a low F1-score of 0.229. This suggests that the model was conservative in labeling metaphors, resulting in many false negatives.

2nd Article: 7976c In the second article, performance improved considerably: the model reached a precision of 0.938 and a recall of 0.484 for metaphorical tokens, resulting in an F1-score of 0.638. Although recall remained modest, this article shows the second highest F1-score for metaphors among all four.

3rd Article: 58717 In the third article, the model showed moderate precision (0.625) and low recall (0.270), yielding an F1-score of 0.377 for metaphorical tokens. The pattern again reflects cautious metaphor predictions with many missed metaphorical instances.

4th Article: fe9d0 In the fourth article, the model performed slightly better, with a precision of 0.857 and recall of 0.546 for metaphorical tokens, corresponding to the highest, of 0.667, F1-score. This result suggests more balanced performance and relatively strong metaphor detection compared to other in-domain texts.

5.1.3 Femke van Bruggen's Dataset

The evaluation on Femke van Bruggen's dataset yielded a recall of 28.2% for metaphorical tokens. This test set is **not suitable for evaluating precision**, as only signaled metaphors related to climate technology were annotated. Consequently, recall is the most meaningful metric, as it indicates how well the model can retrieve the metaphors that were marked.

5.2 Evaluation *With* Domain Adaptation

To address the domain mismatch, the model was further pre-trained on unlabelled climate-related texts using masked language modeling (MLM). This continued pre-training was designed to expose the model to domain-specific vocabulary, syntactic structures, and stylistic patterns. After pre-training, the model was fine-tuned again on the original VUA training set to retain its metaphor classification capabilities.

Post-adaptation results show a marked improvement in recall across two domain-specific datasets. This suggests that continued pre-training can help the model generalize better to climate-related metaphors without sacrificing precision. Meanwhile, precision (slightly) increased for each article. Overall, the F1-scores improved significantly on one climate-domain dataset, namely, the third one, demonstrating some value of domain adaptation for metaphor detection.

5.2.1 VUA Benchmark Test Data

Following continued pre-training, the model was re-evaluated on the VUA benchmark test set. The performance remained consistent with the baseline, showing only marginal differences. The precision for metaphorical tokens slightly decreased from 0.8242 to 0.8197, while recall increased from 0.7089 to 0.7123, resulting in an unchanged F1 score of 0.7622. These results, as expected, suggest that domain-specific continued pre-training does not improve the model's ability to detect metaphors in the general domain.

5.2.2 Climate Technology Articles

1st Article: 4acb1 In the first article, the model achieved a moderate precision of 0.700 but strug-

Article	Precision	Recall	F1-score
1 **(4acb1)**	0.70	0.15	0.25
2 **(7976c)**	**1.00**	0.48	0.65
3 **(58717)**	0.67	0.32	0.44
4 **(fe9d0)**	0.87	**0.61**	**0.71**

Table 3: Metaphor detection results for each article (relevant POS only) after continued pre-training. The bold numbers indicate the highest scores.

gled with a very low recall of 0.149, resulting in a poor F1 score of 0.246. This suggests that the model was conservative in labeling metaphors - predicting them only when highly confident - which led to many metaphorical tokens being missed.

2nd Article: 7976c The second article shows a similar trend. Here, the model reached a perfect precision of 1.000, meaning all predicted metaphorical tokens were correct. The recall remained stable at 0.484, yielding a higher, but not significantly, F1 score of 0.652. Therefore, although the model correctly identified more metaphors, it still failed to detect many.

3rd Article: 58717 In the third article, the precision increased slightly (not significantly) to 0.667, and recall significantly increased, but still remained low at 0.324, leading to an F1 score of 0.436. This performance highlights ongoing challenges with identifying a broader set of metaphorical expressions, suggesting that the model continues to struggle with generalizing to less familiar metaphor uses.

4th Article: fe9d0 The fourth article presents the best overall performance among the four. The model achieved both high precision (0.870) and improved recall (0.606), resulting in the highest F1 score (0.714). This indicates a more balanced performance, suggesting that the metaphors in this article were more aligned with the patterns learned during training and further reinforced during continued pre-training.

5.2.3 Femke van Bruggen's Dataset

On Femke's dataset, the model's performance on metaphorical tokens remained limited, with a recall of 0.3359. This result suggest that the model misses the majority of metaphorical expressions, even when they are marked.

5.3 Summary

Before continued pre-training, the model demonstrated only limited success in metaphor detection across individual articles. Precision was generally moderate (ranging from 0.500 to 0.938), but recall scores were consistently low, especially for metaphorical tokens (ranging from 0.149 to 0.546), which resulted in low F1 scores. This suggests that while the model avoided a large number of false positives, it struggled to identify the majority of metaphors. Performance across articles was highly inconsistent, likely due to domain mismatch and insufficient metaphor coverage in the training data.

After continued pre-training on metaphor-labeled data, the model's performance improved across the board. Improvements were slightly visible in the climate technology articles: for instance, in the fourth article, the F1 score for metaphorical tokens increased from 0.6667 to 0.7143. However, each article improved in either precision or recall, but not both.

Performance on Femke's dataset showed a modest increase in recall for metaphorical tokens (from 0.282 to 0.336). However, this increase is barely significant, as was also the case in the other datasets, already discussed above.

In summary, continued pre-training slightly improved the model's metaphor detection capabilities. While performance remains variable across domains, the results suggest that even limited domain-adaptive pre-training can help the model slightly better identify metaphorical language. However, due to the small size of the test data, it is difficult to draw definitive conclusions. Moreover, it is clear that continued pre-training alone is not sufficient to achieve strong performance in domain-specific metaphor detection.

6 Error Analysis

The experimental results were less promising than expected, showing significantly lower performance of climate-technology-related articles compared to benchmark test data. To better understand the limitations of this study and to identify potential avenues for improvement in domain-specific metaphor detection, an extensive error analysis was conducted. However, it was done only on four climate-technology-related articles, as they are the core of this study.

Article	Before Pre-training			After Pre-training		
	Precision	Recall	F1-score	Precision	Recall	F1-score
1 **(4acb1)**	0.50	0.15	0.23	0.70	0.15	0.25
2 **(7976c)**	0.94	0.48	0.64	1.00	0.48	0.65
3 **(58717)**	0.63	0.27	0.38	0.67	0.32	0.44
1 **(4acb1)**	0.86	0.55	0.67	0.87	0.61	0.71

Table 4: Comparison of metaphor detection results (relevant POS only) before and after continued pre-training across four climate technology articles. Results highlighted in dark yellow indicate improvements compared to those prior to continued pre-training, while results in dark green represent significant improvements ($\geq 5\%$).

6.1 Methodology

To gain deeper insights into the model's performance on domain-specific metaphor detection, I conducted an extensive error analysis. During the evaluation phase, besides reporting standard performance metrics such as precision, recall, and F1-score, I also generated confusion matrices to better understand the types of classification errors. Additionally, I extracted all misclassified tokens, including both false positives and false negatives.

To organize these errors systematically, I created a detailed spreadsheet for each article. In these sheets, all metaphorical tokens were listed and manually categorized into climate-technology-related (CTR), climate-related (CR), technology-related (TR), or non-relevant categories. The technology-related category also included terms related to research and finance. I acknowledge that this categorization is somewhat subjective but provides a useful framework for qualitative analysis.

The spreadsheet highlights metaphorical tokens with color coding to facilitate quick visual inspection: tokens metaphorical only in combination, namely, phrasal verbs such as *look forward*, are marked in **blue**; lines with tokens that should be ignored due to irrelevant part-of-speech tags, such as *down* (the second part of the phrasal verb *breaking down*), are in **black**; and lines with tokens missed by the model are shaded in **gray**. Moreover, I marked cases where the model's predictions changed after continued pre-training as *DIFFERENCE* in white font on a **dark red** background to highlight the impact of this training phase.

This method allowed me to analyze both the quantitative and qualitative aspects of the model's errors, identifying systematic challenges and potential areas for future improvement in domain-specific metaphor detection.

6.2 Quantitative Error Analysis

Table 5 provides a comparative overview of the confusion matrix values across four domain-specific articles, before and after continued pre-training (CP). Overall, the improvements after continued pre-training are modest but consistent. For instance, in the first article, the number of false positives decreased from 7 to 3 while true positives remained stable, indicating an increase in precision. Similarly, in the second article, false positives were completely eliminated (from 1 to 0), while true positives and false negatives remained unchanged. The third article saw a small improvement, with true positives increasing from 10 to 12 and false negatives decreasing accordingly, although false positives remained the same. Finally, the fourth article demonstrated the clearest benefit of continued pre-training, with true positives increasing from 18 to 20 and false negatives dropping from 15 to 13. These results suggest that continued pre-training can provide small, but meaningful improvements in metaphor detection performance across multiple article contexts.

Article	Model	TP	FP	FN	TN
1 **(4acb1)**	Without CP	7	7	40	168
	With CP	7	3	40	172
2 **(7976c)**	Without CP	15	1	16	69
	With CP	15	0	16	70
3 **(58717)**	Without CP	10	6	27	96
	With CP	12	6	25	96
4 **(fe9d0)**	Without CP	18	3	15	55
	With CP	20	3	13	55

Table 5: Confusion matrix results without and with continued pre-training across four climate-technology-related articles (only for relevant POS).

6.3 Qualitative Error Analysis

6.3.1 1st Article: 4acb1

For the first article, only 7 out of 47 metaphorical tokens with the relevant part of speech (POS) (48 in total) were correctly identified, reflecting the model's recall. Among those 47 tokens, 4 were climate-technology-related (CTR), and only

1 of these was correctly identified. Furthermore, out of 22 climate-related (CR) and technology-related (TR) tokens, only 2 were correctly identified. Among 21 non-relevant metaphorical tokens with the correct POS, 4 were correctly identified. Finally, it is important to note that two tokens, *part* and *play*, from the CR group were correctly identified by the model after continued pre-training. However, this improved version also introduced two new errors: it failed to identify *decarbonise* (from the CTR group) and *release* (CR), both of which were correctly identified by the initial model.

Therefore, while the model struggled the most with CR and TR tokens, the CTR group was too small to draw firm conclusions. Overall, it seems the model had difficulty with all metaphorical tokens, not just those related to climate technologies.

However, for the first article, the model did reduce the number of false positives from 7 to 3. After continued pre-training, the model was able to correctly identify tokens such as *tree*, *planting*, *absorbed*, and the adverb *in* (which is not relevant for my study).

6.3.2 2nd Article: 7976c

For the second article, there were 34 metaphorical tokens in total, of which 31 had a relevant POS tag. Out of those 31 tokens, 15 were correctly identified. One token out of the 31 was climate-technology-related (CTR), but it was not identified by the model. Out of 16 climate-related (CR) and technology-related (TR) tokens, 9 were correctly identified by the model. Out of 14 non-relevant metaphorical tokens, 6 were correctly identified. Finally, the model after continued pre-training managed to correct a mistake made by the previous model by correctly identifying *board* (from the non-relevant token group). However, it failed to identify *value* (also from the not-relevant group), which had been correctly recognized by the previous model.

Similarly to the first article, the groups were too small to draw definite conclusions, and the model struggled across all token groups. However, it appears that the CR and TR groups were less problematic this time.

Finally, before continued pre-training, the model had one false negative, namely, *infrastructure*, which was correctly identified after continued pre-training.

6.3.3 3rd Article: 58717

The third article contained 43 metaphorical tokens, 37 of which had a relevant POS tag. While none of the 3 climate-technology-related (CTR) tokens were identified by the model, 7 out of 20 climate-related (CR) or technology-related (TR) tokens were correctly identified. Moreover, 3 out of 14 non-relevant metaphorical tokens were found. Finally, the second version of the model corrected three mistakes by successfully identifying *answer* (TR), *reward* (non-relevant), and *make* (TR). However, it failed to identify *going* (non-relevant), *green* (TR), and *great* (non-relevant), all of which had been correctly identified by the previous model.

Therefore, although the group sizes were again small, it is clear that the model struggles with identifying different types of metaphors, not only those that are domain-specific.

Finally, both sets of results, before and after continued pre-training, contained 6 false positives. However, in the earlier results, the model incorrectly identified *bottom*, a mistake that was resolved after continued pre-training. At the same time, the model incorrectly identified *long* after continued pre-training, even though this error did not occur previously.

6.3.4 4th Article: fe9d0

Finally, in the fourth article, the model was presented with 34 metaphorical tokens, 33 of which had a relevant POS tag. There was one climate-technology-related (CTR) token, which the model failed to identify. Moreover, there were 18 tokens from the climate-related (CR) and technology-related (TR) group, 8 of which were correctly identified. Out of 14 non-relevant metaphorical tokens, 9 were identified by the model. Finally, two tokens, *area* (non-relevant) and *hub* (TR), were correctly identified after continued pre-training, indicating small improvements introduced by domain adaptation.

Although the group sizes are too small to draw any concrete conclusions, it appears that in this article, the model handled non-relevant metaphorical tokens better than those related to climate technologies.

Lastly, both versions of the model, before and after continued pre-training, produced 3 false positives: *ups*, *includes*, and *net*. Notably, *ups* appeared as part of the word *start-ups*, and *net* occurred in the expression *net-zero*. It is possible that if these

tokens had not been separated, they might have been identified correctly.

6.4 Detailed analysis of the tokens

6.4.1 Frequent Tokens

Table 6 presents the tokens that appeared more than once across four climate-technology-related articles. For each token, the table shows its frequency, its classification category (CTR, CR/TR, or non-relevant), and the number of times it was correctly identified by each model.

The token *green* is the most frequent, appearing four times. However, three of these instances occur in a single article, namely 58717. Only one instance is correctly identified by the model without domain adaptation, while the domain-adapted model fails to detect it entirely.

The token *removal* appears once in each of the articles 7976c, 58717, and fe9d0, but it is never identified by either model. In contrast, *look* also appears once in each of the same articles and is consistently detected by both models.

Undone appears twice in the first article (4acb1) but is never correctly identified. Similarly, *ways* appears twice in the same article and is correctly identified once. The token *saying* is also seen twice in 4acb1, but both instances are missed by the models. The same article mentions *compounds* twice, and both occurrences are missed as well.

Growth appears once in 7976c and once in fe9d0, and it is correctly identified in both cases. *Going* occurs twice in 58717, and is correctly identified once by the model without domain adaptation.

Stage appears once in 58717 and once in fe9d0, and is correctly identified in both cases. The token *social* appears twice in 58717, but both instances are missed. On the other hand, *puts* also appears once in 58717 and once in fe9d0, and both are correctly identified.

Area occurs twice in fe9d0; it is identified once by each model, with the model without domain adaptation missing one instance. The token *talent* also appears twice in fe9d0 but is missed entirely by both models.

Finally, the tokens *can* and *for* each appear twice in the dataset, but they are excluded from analysis due to having POS tags deemed irrelevant for this study.

Therefore, I can conclude that model performance varies depending on the token and its context. For example, *growth* is consistently well identified by both models, indicating that these metaphors may be easier for the models to recognize or occur in clearer contexts. In contrast, tokens like *undone*, *saying*, and *compounds* are frequently missed, suggesting challenges in detecting metaphors related to these words. Moreover, for certain tokens, such as *green*, the model without domain adaptation outperforms the domain-adapted model, while for others, for example, *puts*, both models perform well. This suggests that domain adaptation does not uniformly improve metaphor detection and may require further tuning or additional data. Furthermore, it is important to note that even frequently occurring tokens like *removal* and *talent* are often missed by both models, implying that frequency alone is insufficient for reliable metaphor detection.

6.4.2 Factors Influencing the Models' Performance

The reason why each climate-technology-related article was evaluated separately is due to the significant differences observed in their performance. Below, I present several factors that may have contributed to these variations.

- **Contextual Clarity:** Tokens like *growth* often appear in clear and consistent metaphorical contexts, such as technology-related ones, making them easier for models to detect,
- **Token Polysemy and Ambiguity:** Highly polysemous tokens, such as *going*, pose challenges due to their multiple literal and metaphorical meanings,
- **Frequency and Training Data:** Frequent tokens like *removal* and *talent* may still be missed if their metaphorical uses are diverse or underrepresented in the training data,
- **Domain-Specific Adaptation:** Domain adaptation can improve detection for some tokens but may reduce performance if the adapted data is limited or biased,
- **Syntactic and Semantic Roles:** Tokens appearing in complex or rare syntactic constructions, such as *undone*, are harder for models to identify metaphorically,
- **Annotation Quality:** Model success depends heavily on high-quality, consistent annotations. Inconsistent labeling leads to detection

Token	Frequency	CTR	CR/TR	Not-relevant	Correct Without CP	Correct With CP
green	4	0	4	0	1	0
team	3	0	3	0	0	0
removal	3	3	0	0	0	0
look	3	0	0	3	3	3
undone	2	0	2	0	0	0
ways	2	0	0	2	1	1
saying	2	0	0	2	0	0
compounds	2	0	2	0	0	0
growth	2	0	1	1	2	2
~~can~~	2					
going	2	0	0	2	1	0
stage	2	0	2	0	2	2
~~for~~	2					
social	2	0	0	2	0	0
puts	2	0	1	1	2	2
area	2	0	1	1	1	2
talent	2	0	2	0	0	0

Table 6: Common metaphorical tokens across four climate-technology-related articles, sorted by frequency, with classification groups and model output columns. Rows containing tokens with non-relevant POS tags are crossed out.

difficulties. This factor may be especially important in our study, as the two annotators preparing the four climate-technology-related articles were still in training and often disagreed on the labels.

7 Conclusions and Recommendations

8 Conclusion

This study evaluated the performance of a pre-trained metaphor detection model (XLM-R) on domain-specific, climate-technology-related articles. While the model performed well on general-domain data, its effectiveness dropped considerably on domain-specific texts. Continued pre-training with unlabelled climate-related data led to occasional, slight improvements in precision, recall, and F1-score. However, performance remained highly variable across articles, and many domain-specific metaphors were still missed. These findings highlight the limitations of current models in specialized domains and emphasize the need for more targeted training data and refined annotation strategies.

8.1 Limitations

This study has several limitations that should be addressed in future work:

- **Test Data Size:** The amount of climate-technology-related test data was limited (only 158 metaphorical tokens in total), which constrains the robustness of the evaluation. Future studies should include larger and more diverse test sets to better assess model generalization.
- **Annotation Consistency:** Training annotators and achieving agreement proved challenging. Annotation consistency benefits from having a clear and specific goal from the outset. In this study, annotation guidelines evolved over time, which may have introduced inconsistencies.
- **POS Tagging Accuracy:** POS tagging errors were not rigorously addressed. The guidelines specified that if the auto-assigned part-of-speech tag was incorrect, annotators should add a separate column with the correct tag. However, this rule was not followed. Future efforts should focus on better handling and correcting incorrect POS tags to improve model performance.
- **Tokenization and Compound Words:** Currently, some multi-word expressions, such as *start-up*, are treated as separate tokens. Future work should explore combining such compounds into single tokens. Tools like spaCy or lexical resources like WordNet may assist with this task.
- **Contextual Analysis of Tokens:** It is advisable to analyze corpus tokens more closely to understand the contexts in which they appear. This can improve the interpretation and modeling of metaphorical language. For example, with more time, it would be beneficial to examine the contexts in which each token appears in the VUA Metaphor Corpus.

References

2009. LexisNexis Academic Database. Last Accessed: 30th July, 2025.

Hey Amit. 2023. Fine-Tuning vs Continued Pretraining. Last Accessed: 2nd August, 2025.

Pearson. 2015. Longman Dictionary of Contemporary English. Last Accessed: 2nd August, 2025.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

Lennart Wachowiak, Dagmar Gromann, and Chao Xu. 2022. Drum Up SUPPORT: Systematic Analysis of Image-Schematic Conceptual Metaphors. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 44–53, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.