Master Thesis

Staying Relevant: Metaphor Detection and Domain Relevance Classification in Immunotherapy Texts

Melina Paxinou

Supervisor Pia Sommerauer, Gudrun Reijnierse
2^{nd} reader Luis Morgado da Costa

a thesis submitted in fulfillment of the requirements for the degree of

MA Linguistics
(Text Mining)

Vrije Universiteit Amsterdam

Computational Linguistics and Text-Mining Lab
Department of Language and Communication
Faculty of Humanities

Date July 02, 2025
Student number 2854344
Word count 17122

Abstract

The purpose of this thesis is to evaluate whether models trained for metaphor detection can be outperformed by models with domain-specific pre-training, with a particular focus on the field of immunotherapy. The first task, Metaphor Detection, is a binary classification problem at a token level, where the goal is to predict whether a given metaphorical token is related to immunotherapy, using both a general-domain and a domain-adapted RoBERTa model. The second task, Domain Relevance Classification, focuses on identifying metaphors directly related to the domain of immunotherapy using a BERT model. Both tasks use data derived from the Vrije Universiteit Amsterdam (VUA) Metaphor Corpus and the Immunotherapy Metaphor Dataset compiled by Bos et al. (2025) from scientific publications and news articles. This setup provides a concrete test of whether the knowledge captured from biomedical texts during pre-training improves metaphor detection in a specialized domain.

The results show that while BioMed-RoBERTa is more sensitive to metaphorical language, its increased detection comes with greater noise, which reduces relevance precision. In contrast, XLM-RoBERTa paired with the BERT classifier achieves better overall pipeline performance.

This work shows the promise of current metaphor detection approaches for science communication and reveals their weaknesses. By improving metaphor detection and domain relevance classification, the proposed pipeline aims to support automated analyses of how metaphor shapes science communication, with implications for public understanding and expert discourse around complex scientific concepts.

Declaration of Authorship

I, Melina Paxinou, declare that this thesis, titled *Staying Relevant: Metaphor Detection and Domain Relevance Classification in Immunotherapy Texts* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a Master's degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 02-07-2025

Signed: 

Acknowledgments

First, I want to thank my supervisor, Pia Sommerauer, for her guidance, support, and patience throughout this thesis, as well as for her appreciation of my memes.

I am deeply grateful to my co-supervisor, Gudrun Reijnierse, for her valuable contribution and encouragement along the way.

I am also grateful to all the CLTL professors who offered their knowledge and advice during this journey.

Many thanks to Urtė Jakubauskaitė and Bastiaan Sizoo, the student assistants and fellow classmates who supported the annotation process and kindly dedicated their time and effort.

To my friends Elisabetta, Xin, Matt, Shenglin, Shutao, Ning, Victoria, and Manya, thank you for all the study sessions, pep talks, and perfectly timed distractions when I needed them the most. Martha, even though your path looks a little different from the rest of us, thank you for being there to support me.

To my friends back home, Alexandra, Fotini, and Theofanis, who proofread my drafts and kept me company during my all-nighters, I am truly thankful.

To my mom, your love and understanding were among the reasons I was able to keep going.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Metaphors play a crucial role in medical science communication, shaping how complex biomedical ideas are conceptualized and shared between different audiences. In cancer immunotherapy discourse, in particular, metaphors such as war, journey, or economic frames frequently appear in scientific publications and journalistic reporting, influencing how treatments are perceived and discussed by researchers, clinicians, and patients alike. Despite this prevalence, metaphor detection models developed on general-purpose datasets are rarely evaluated in highly specialized domains such as immunotherapy. This thesis addresses this gap by investigating how well existing metaphor detection models perform when applied to immunotherapy-specific texts and exploring methods to improve their performance where necessary.

Improving the detection and interpretation of metaphors in immunotherapy texts has practical implications to improve patient understanding, support clearer clinician-patient communication, and enable more accurate automated text analysis in science communication applications. By addressing this gap, this thesis contributes to both the methodological development of domain-specific metaphor detection and the broader goal of facilitating informed public understanding of medical research.

1.1 Problem Definition

A general metaphor detection model typically labels metaphorical expressions without distinguishing whether they are directly relevant to the domain context in question. As such, there is a risk that models trained on general corpora may overlook or misclassify domain-specific figurative language.

1.2 Research Questions

Based on a review of relevant literature, this thesis defines the following research questions:

Main research question:

How well do existing metaphor detection models (e.g., those trained on the VU Amsterdam Metaphor Corpus) perform on immunotherapy-related metaphors?

Sub-questions:

1. How does a general RoBERTa model perform in the task of metaphor detection on texts related to immunotherapy compared to a RoBERTa model trained with medical data?

2. Since existing models that are fine-tuned on general data adequately identify the domain-specific metaphors, how can immunotherapy-specific data and annotations be used to identify which metaphors are actually related to immunotherapy?

1.3 Approach

To address this, the first step of this research is to evaluate the effectiveness of a general-domain model on immunotherapy-related texts. Alongside this, a RoBERTa model pre-trained on biomedical data is fine-tuned and tested as a comparison point to examine whether domain-adapted pre-training provides measurable benefits for metaphor detection in science communication discourse. Furthermore, the study explores the viability of a supervised learning approach that determines whether the identified metaphors belong to the domain of immunotherapy. Understanding the behavior of the system is essential to draw reliable conclusions from automatic analyses, especially when such findings might inform communication strategies or public understanding of complex biomedical concepts.

This thesis builds upon a dataset of immunotherapy-related texts annotated for explicit metaphors, compiled from the study *'Mapping' Knowledge Dissemination: What Metaphors Reveal About the Conceptualisation of Immunotherapy in Scientific and Journalistic Communication*, published in the *Electronic Journal of Health Communication* (EJHC) in January 2025 (Bos et al., 2025). The corpus comprises both scientific articles and media texts to ensure a diverse representation of metaphorical language in this medical domain. Explicit metaphors, following the UCREL Semantic Analysis System (USAS) (Piao et al., 2015), are categorized into thematic groups such as war, journey, and business metaphors. In addition, new annotations for implicit metaphors are produced following the MIPVU procedure (Steen et al., 2010), with inter-annotator agreement assessed jointly with my supervisor and two student assistants to ensure reliability.

An existing state-of-the-art metaphor detection model that has been trained on the VU Amsterdam Metaphor Corpus is evaluated on this immunotherapy dataset to determine its ability to identify relevant metaphors in domain-specific contexts. A comparative RoBERTa model, pre-trained on biomedical texts, is also tested to examine the effect of domain adaptation on metaphor detection. The feasibility of integrating supervised learning is evaluated to enable a more precise detection of immunotherapy-specific metaphors.

Subsequently, a BERT model is trained to identify domain relevance to immunotherapy. This step allows for a two-fold pipeline: first, detecting metaphorical language, and second, filtering those metaphors for relevance to the immunotherapy domain. The final evaluation assesses this end-to-end setup on an immunotherapy dataset, assessing its potential utility for science communication research. This pipeline aims to support more targeted analyses of metaphor use in science communication, ultimately informing how complex medical topics, such as immunotherapy, are framed and understood.

1.4 Summary of Results

Both XLM-RoBERTa and BioMed-RoBERTa detect literal language well, but metaphor detection has been proven to be quite hard. BioMed-RoBERTa has a slight advantage in identifying more metaphors in immunotherapy texts. The BERT classifier performs consistently well in relevance classification.

However, in the full pipeline, XLM-RoBERTa combined with BERT outperforms BioMed-RoBERTa, suggesting that greater sensitivity in detection can sometimes lead to more noise and lower relevance accuracy.

Applying the best-performing model to analyze metaphors in immunotherapy texts would enable communication scholars to systematically identify metaphorical language that shapes public and professional discourse around immunotherapy. They could uncover prevalent metaphorical themes, track how complex biomedical ideas are framed, and potentially reveal how science communication through metaphors influences patient understanding. In addition, the metaphor identification and categorization process could potentially be automated, allowing experts to focus more on analyzing the content and impact of metaphors rather than manual annotation.

At the same time, limitations remain, as some metaphors might be missed due to the ambiguity and subtlety of figurative language. Consequently, scholars should interpret automated results as a starting point for exploration rather than definitive conclusions. Understanding these boundaries is crucial; while the models offer valuable insights, they do not replace the annotations and analysis that human expertise provides when drawing conclusions about metaphor use in science communication.

Chapter 2

Related Work

Metaphor detection is an established NLP task that has not been widely explored in medical contexts. This thesis investigates whether domain-specific language models, such as BioMed-RoBERTa (Gururangan et al., 2020), can improve the performance of metaphor detection in specialized medical texts compared to a general-domain model such as XLM-RoBERTa (Conneau et al., 2020). The dataset used is derived from the paper *'Mapping' Knowledge Dissemination: What Metaphors Reveal About the Conceptualisation of Immunotherapy in Scientific and Journalistic Communication* (Bos et al., 2025). By focusing on immunotherapy discourse, an underexplored area for figurative language research in NLP, this study explores metaphor detection, domain adaptation, and biomedical NLP.

2.1 Metaphor Use in Science Communication

Metaphors are a central part of how people understand and communicate complex ideas. They allow abstract or technical concepts to be described in more familiar terms, often drawing inspiration from everyday experiences. In health communication, for example, metaphors shape how patients and the general public think about diseases, treatments, and scientific advances. This means that the metaphors chosen by journalists or scientists can influence how a topic is perceived, what risks or hopes people associate with it, and even what decisions they make about their own health.

The dataset used in this thesis was created by Bos et al. (2025) in their paper *'Mapping' Knowledge Dissemination; What Metaphors Reveal About the Conceptualisation of Immunotherapy in Scientific and Journalistic Communication.* The purpose of their paper, which led to the creation of this dataset, was to map the use of metaphors in immunotherapy-related texts, to conceptualize immunotherapy from a medical context to the broader public. They emphasized that public awareness of a topic is as important as immunotherapy, which directly correlates with people's perception of health and treatment and can be explained metaphorically in a variety of ways. By constructing a metaphor to explain immunotherapy, we can decide which types of information are provided to the broader public and how they are presented. This way of presentation, called framing, in addition to impacting people's perception of health information, also affects their decision-making process. Depending on how a concept is framed, that is, what metaphor theme is used to describe it, people can have positive or negative opinions about new treatments.

Metaphor framing has also been extensively analyzed in the context of the COVID-

19 pandemic, where political leaders and media commonly used war metaphors to describe the public health crisis. For example, Truc (2024) conducted a critical metaphor analysis of American media discourse, highlighting how COVID-19 was framed as an enemy to be defeated in a wartime scenario. This framing was instrumental in mobilizing public urgency and compliance with strict policies, but also risked amplifying fear and legitimizing extraordinary political measures. Such studies demonstrate that metaphors can powerfully construct social realities, extending far beyond scientific accuracy to shape collective emotional and political responses.

Similarly, metaphor use plays a crucial role in climate change communication, where political leaders use figurative language to influence public attitudes and policy support. Wang and Habil (2024) analyzed speeches from COP28 and identified multiple dominant metaphor types, including war, journey, and building metaphors. These conceptual frames serve diverse persuasive functions, for example, portraying climate action as a collective battle or a shared journey toward sustainability. From an ecolinguistic perspective, they argue that carefully crafted metaphors can cultivate ecofriendly mindsets and reinforce constructive engagement with complex environmental challenges.

However, existing NLP systems for metaphor detection were not designed with this kind of targeted analysis in mind. Most metaphor detection models focus on general-purpose language and aim to find all metaphorical expressions in a text, regardless of topic. They are typically trained on general corpora, such as the VU Amsterdam Metaphor Corpus (Steen et al., 2010) and TOEFL Native Language Identification Corpus (Beigman Klebanov et al., 2018). Although these resources have resulted in great progress in automated metaphor detection, they do not address the specific needs of science communication research, where the goal is to identify metaphors relevant to a particular subject, such as immunotherapy.

This means that for such domains, researchers rely heavily on manual annotations to find and interpret relevant metaphors. General-domain models may overlook subtle figurative language in specialized texts, or flag metaphors that are linguistically valid but irrelevant to the target topic. This thesis bridges this gap by combining domain-specific language models and datasets to test whether the adaptation of metaphor detection to a specialized medical domain improves performance and usefulness for downstream science communication research.

2.2 Metaphor Detection in NLP: From Rule-based Methods to Neural Models

Over the past two decades, metaphor detection has advanced from rule-based and lexicon-driven approaches to neural and transformer-based systems. These examples are included here not to provide an exhaustive history, but to illustrate that metaphor detection is an established NLP task with diverse methodological approaches and to clarify why it is relevant to apply, test, and adapt such models for specialized domains, such as immunotherapy discourse.

For example, Turney et al. (2011) present a feature-based approach that uses abstractness as a core feature: for each sentence, they create a vector of five features: (1) the average abstractness ratings of all nouns (excluding proper nouns), (2) the average abstractness of all proper nouns, (3) the average abstractness of all verbs excluding the target verb, (4) the average abstractness of all adjectives, and (5) the average abstract-

ness of all adverbs. This example demonstrates how features such as word concreteness can capture the subtle mismatches in meaning that are characteristic of metaphorical language.

Tsvetkov et al. (2014) demonstrate a cross-lingual metaphor detection system that relies on common semantic features for SVO (subject–verb–object) relations. Specifically, the S and O components include (1) the semantic category of the word, (2) its degree of abstractness, and (3) named entity types; the V (verb) component includes only the semantic category and abstractness. This design shows how syntactic roles and lexical semantics can be combined to model the metaphorical use of verbs in different languages.

A major breakthrough came with deep learning and contextual embeddings: models such as BERT, RoBERTa, and systems such as DeepMet (Su et al., 2020) and Go Figure! (Chen et al., 2020) now achieve state-of-the-art performance by learning to detect metaphors directly from context, without explicit feature engineering. These public transformer-based models have become strong baselines and are referenced in this thesis as examples of high performing automated metaphor detection.

The 2018 VUA Metaphor Detection Shared Task (Leong et al., 2018), which was the first shared task on metaphor detection and was held at the NAACL Workshop on Figurative Language Processing, aimed to benchmark metaphor detection systems using the VU Amsterdam Metaphor Corpus. Participants were challenged to perform classification in two categories: (1) all verbs and (2) all content words, i.e., nouns, verbs, adjectives, and adverbs. All but one system used a neural network architecture.

The 2020 Metaphor Detection Shared Task (Leong et al., 2020), part of the ACL Workshop on Figurative Language Processing, had a similar focus on identifying metaphorical content words in English texts. It featured both the VU Amsterdam Metaphor Corpus and a TOEFL corpus annotated for metaphors. Similarly to the previous shared task, participants could compete in tracks for verbs and all content words and aimed to improve metaphor detection in terms of performance and generalizability. The systems in the 2020 submissions consisted mainly of BERT-based architectures and can be directly compared to the best systems of the 2018 shared task, as one of the datasets was used in both. The best performing system of 2018 had an F1-score of 0.651, while the best performing system of 2020 had an F1-score of 0.769. This shows how the field has advanced between the two shared tasks and that the best results of the 2020 shared task are the new state-of-the-art for both the VU Amsterdam Metaphor Corpus and TOEFL corpora.

2.3 Annotated Resources for Metaphor Detection

The VU Amsterdam Metaphor Corpus (VUA) is the first large-scale, systematically annotated corpus for metaphor detection in English (Steen et al., 2010). It comprises more than 200,000 words drawn from four genres: academic texts, news reports, fiction, and conversations. VUA's annotations follow the Metaphor Identification Procedure VU (MIPVU). The original MIP (Crisp et al., 2007) introduced an explicit step-by-step method for identifying metaphorically used words in context based on their contrast with more basic meanings. MIPVU refined this procedure to improve consistency and applicability for large corpora. The VUA Metaphor Corpus has now become the benchmark dataset for training and evaluating metaphor detection models and has been the main corpus for the 2018 and 2020 metaphor detection shared tasks.

The TOEFL corpus provides an additional annotated resource in the domain of English as a Second Language (ESL) (Beigman Klebanov et al., 2018). It consists of essays written by non-native speakers and has been annotated for metaphorical language following the same general principles as the VUA Metaphor Corpus. The TOEFL corpus brings in a mix of writing styles and learner language, which adds diversity to the data. This makes it especially useful for testing how well metaphor detection models perform on different types of language use. It was included as a second dataset in the 2020 VUA Metaphor Detection Shared Task, allowing participants to compare system performance on both native and non-native English writing.

Together, these annotated resources serve as a benchmark for developing and testing metaphor detection models in the shared tasks, as well as in independent projects. By providing consistent, high-quality metaphor annotations for a variety of domains, they enable researchers to systematically evaluate how well models recognize metaphorical language in various contexts and whether their performance remains when applied to new domains.

However, applying metaphor detection to medical texts presents unique challenges. Unlike general-domain language, medical discourse often combines highly technical terminology with figurative expressions that are context-dependent and specific to medical terminology. For example, phrases such as 'the immune system fights cancer' utilize war metaphors to provide a scientific explanation. In addition, most existing metaphor detection research focuses on texts such as news or fiction, leaving metaphor use in doctor-patient communication and scientific articles relatively underexplored. This scarcity of domain-specific resources and the complexity of health-related language and communication mean that general models might miss contextually important metaphors or fail to distinguish them from literal technical descriptions. As a result, detecting metaphors in medical texts requires training data tailored to handle specialized vocabulary and figurative framing effectively.

2.4 Health-related Metaphor Detection

In recent years, several studies have started combining metaphor detection with insights from medicine and psychology, pushing metaphor-relevant NLP toward more creative and real-world uses, such as mental health assessment and clinical support. Gutiérrez et al. (2013) introduced an innovative approach to mental health diagnosis with metaphor identification and sentiment analysis algorithms. Their methodology involved extracting metaphorical and emotional features from patient narratives and using a three-layer multilayer perceptron architecture (MLP) to predict the onset of schizophrenia. This study marked a pioneering effort to apply automated metaphor detection to psychiatric diagnostics.

David and Matlock (2018) utilized the MetaNet system of Dodge et al. (2015) to analyze metaphorical expressions related to poverty and cancer in English and Spanish corpora. MetaNet uses a frame-based approach, associating lexical items with conceptual metaphors through semantic frames. This system facilitates the identification of metaphoric expressions by mapping them to a network of conceptual metaphors, allowing for cross-linguistic and cross-domain metaphor detection.

The MAM framework, as detailed in the study by Li et al. (2019), integrates metaphor feature extraction with a CNN-RNN architecture to classify social media texts for mental illness detection. By focusing on metaphorical expressions that often convey

implicit emotions, MAM effectively captures the subtle ways people express mental health struggles through words. This approach demonstrated high recall and F1-scores in detecting depression, anorexia, and suicidal tendencies.

Panicheva et al. (2023) addressed the challenges of applying metaphor detection models to texts drafted in a psychological experiment setting by annotating the Met-Personality dataset with conceptual metaphors. They proposed a novel annotation procedure to boost inter-annotator agreement and trained state-of-the-art metaphor detection models on this dataset. The study explored correlations between metaphor usage and psychological traits, stressing the potential of metaphor analysis in psychological research.

Although these studies demonstrate the promise of automated metaphor detection for several health-related topics, they are mainly based on general-domain language models or custom architectures without domain-specific pre-training. To my knowledge, no prior research has systematically investigated whether domain-adapted transformers, such as BioMed-RoBERTa, can improve metaphor detection in specialized medical texts. Given how subtle and context-dependent metaphors can be in medical texts, looking into domain adaptation is a crucial but still overlooked step toward making metaphor detection more reliable in biomedical NLP.

2.5 Models and Domain Adaptation Techniques

This thesis builds on an existing transformer-based metaphor detection pipeline to test the impact of domain-specific pre-training for identifying metaphors in immunotherapy-related texts. The starting point is the metaphor detection system developed by Wachowiak et al. (2022), which includes the fine-tuning of the XLM-R multilingual transformer on the VU Amsterdam Metaphor Corpus. This serves as a general model capable of detecting metaphorical language in English.

To investigate whether domain-specific knowledge can improve metaphor detection in biomedical contexts, Wachowiak's training pipeline is adapted by replacing the general XLM-R model with BioMed-RoBERTa, a version of RoBERTa pre-trained on large biomedical corpora. By fine-tuning BioMed-RoBERTa on the same metaphor annotations and evaluating it on immunotherapy texts, this setup enables a direct comparison by isolating the effect of domain adaptation while preserving all other training conditions.

This approach aims to examine whether a language model with biomedical domain expertise can more accurately capture metaphorical expressions specific to immunotherapy discourse, compared to a general multilingual model such as XLM-R. The contrast between these two models reads into the role of domain-adapted transformers for figurative language tasks in specialized scientific fields.

2.5.1 Domain Adaptation Impact in NLP tasks

Domain-adapted models have been shown to consistently improve performance on NLP tasks. Named Entity Recognition (NER) and Relation Extraction often achieve several points of improvement in F1-scores compared to general models, as demonstrated by BioMed-RoBERTa evaluations (Gu et al. (2021)). These models benefit from pre-training on large-scale biomedical corpora, which helps them capture domain-specific terminology and contextual nuances that general models typically miss.

However, despite these advances, the intersection of metaphor detection and domain adaptation remains underexplored. Existing metaphor detection systems primarily focus on general-domain language, aiming to identify all metaphors regardless of topic. In contrast, research on domain-specific metaphors, such as those in biomedical texts, often lacks automated solutions.

For example, the COVID-19 metaphor detection approach by Wachowiak et al. (2022) applies a semi-automatic pipeline to a domain-specific corpus, but relies mainly on XLM-R, a language model trained on generic language. They do not explicitly investigate the impact of domain adaptation, nor adjust their models to a specific subdomain.

Moreover, current approaches to target metaphors within specific domains typically do not focus on identifying metaphors that are directly relevant to a particular target domain, in our case, immunotherapy. This creates a gap, as the identification of domain-relevant metaphors is key for applications in science communication and scientific discourse analysis.

In summary, while domain adaptation has generally proven valuable in biomedical NLP and metaphor detection has advanced in general domains, their overlap, especially for detecting metaphors tied to specific medical domains, remains an open research challenge. This thesis aims to address this gap by evaluating domain-adapted transformer models on the task of detecting immunotherapy-related metaphors, thus contributing to this research gap.

2.5.2 Drum Up SUPPORT by Wachowiak et al. (2022)

Wachowiak et al. (2022) develop a systematic and reproducible pipeline for the semi-automatic detection and analysis of image-schematic conceptual metaphors (ISCMs) within domain-specific natural language corpora, exemplified through the context of COVID-19 discourse. By integrating methods such as neural metaphor detection, dependency parsing, clustering, and frame annotation, the study aims to uncover how specific image schemas, focusing particularly on the schema SUPPORT, are semantically realized and employed metaphorically in contemporary language use. Additionally, the paper seeks to contribute to the field of cognitive linguistics and computational metaphor analysis by proposing a methodology that enables exploration of their constructional patterns and underlying cognitive structures. Ultimately, this work aims to make metaphor detection smarter by building tools that can handle a vast amount of data without relying on subjective guesses. The goal is to uncover new metaphorical connections that might otherwise go unnoticed, helping us better understand how metaphors shape the way we talk about complex and abstract ideas.

The approach is designed to be flexible, allowing for the analysis of various schemas by compiling relevant seed words relevant to different domains. For example, they mention that applying the same pipeline to other schemas, such as CONTAINMENT, would involve selecting an appropriate set of seed words.

The use of a generic pre-trained model, such as XLM-R, suggests that they rely on a domain-general language model enabled to process domain-specific data effectively. Although they do not explicitly compare a domain-specific versus a generic approach in their results, the methodology implies that their semi-automatic pipeline, powered by such models, can be applied to specific domains once suitable seed words and domain-relevant corpora are provided. This indicates a relatively domain-adaptable framework

that benefits from the strengths of pre-trained language models but requires some domain-specific seed input for best performance.

XLM-R, used by Wachowiak et al. (2022) to train their metaphor detection classifier, was developed by Conneau et al. (2020). It is a transformer-based multilingual masked language model trained on over two terabytes of filtered CommonCrawl data spanning 100 languages. The training objective follows the masked language modeling (MLM) paradigm, where the model predicts masked tokens within input sequences. The model architecture uses 12 transformer layers with multi-head attention, a hidden size of 768, and 12 attention heads.

Training data was curated through a filtering process that combined internal language identification models and the fastText classifier. The authors obtained a significantly larger and cleaner dataset compared to prior efforts, such as Wikipedia-based corpora, with the CommonCrawl datasets offering substantially increased monolingual data, especially for low-resource languages.

Evaluations were performed on multiple benchmarks, including Cross-lingual Natural Language Inference (XNLI), Machine Reading Comprehension (MLQA), Named Entity Recognition (NER) and GLUE tasks. The training strategy utilized the 'translate-train-all' approach, which includes labeled data from multiple languages to improve cross-lingual transferability. The results demonstrated that XLM-R outperforms previous multilingual models such as mBERT, achieving a new state-of-the-art average accuracy on XNLI of 83.6%.

Finally, XLM-R was evaluated on monolingual benchmarks such as GLUE, where it retained competitive performance relative to monolingual models such as RoBERTa, showing that large-scale multilingual training does not necessarily diminish language-specific capabilities. This analysis demonstrates how crucial it is to have large and well-curated datasets, as well as flexible model architectures, in order to achieve robust cross-lingual understanding in a variety of languages.

2.5.3 Bridging Domain Gaps via Adaptation Techniques

Domain adaptation refers to techniques aimed at improving a model's performance when it is applied to data from a different domain than the data on which it was trained. For example, in the unsupervised multi-source setting discussed in the paper by Wright and Augenstein (2020), the goal is to enable a model trained on labeled data from multiple source domains to accurately predict on a target domain for which no labeled data has been seen. The paper reviews several approaches, such as inducing domain-invariant representations through adversarial training and data selection strategies. The aim is to make the models robust to changes in data distribution across domains.

The linguistic characteristics of domain-specific texts, in this case immunotherapy-related news articles and scientific publications, are characterized by traits such as specialized vocabulary and medical terminology. Consequently, these characteristics differ significantly from those of everyday language. These domain shifts can lead to lower performance when using general-domain language models on specialized texts. Domain adaptation, in other words, refers to the process of retraining or fine-tuning a language model on domain-specific corpora to bridge this gap.

2.5.4 BioMed-RoBERTa

General pre-trained models, such as BERT and RoBERTa, may exhibit lower performance in biomedical tasks due to differences in linguistic characteristics in their pre-training data. To address this issue, continued pre-training on domain-specific text has emerged as a highly effective strategy. For example, the paper by Gururangan et al. (2020) examines several methodologies for domain adaptation of the RoBERTa model. The focus here is on BioMed-RoBERTa, due to the nature of the metaphor detection and domain relevance classification tasks on immunotherapy texts in this thesis. The methods mentioned in their paper describe both large-scale domain-specific pre-training and task-specific strategies, designed to improve the model's performance in biomedical classification tasks by a targeted adaptation of its representations.

Domain-Adaptive Pre-training (DAPT) involves continuing RoBERTa pre-training on a vast corpus of unlabeled biomedical texts. The corpora used include biomedical articles from sources such as PubMed. This process utilizes the objective of Masked Language Modeling (MLM), consistent with the original RoBERTa training procedure, and involves approximately 12,500 steps, equivalent to a single pass over the target corpora. The purpose of DAPT is to provide the model with domain-specific lexical and structural knowledge to improve downstream task performance.

In addition to DAPT, the authors explore Task-Adaptive Pre-training (TAPT), which involves pre-training the model further on unlabeled data directly relevant to individual biomedical tasks, such as the RCT dataset. TAPT uses smaller task-specific corpora, thus being computationally more efficient compared to DAPT. The methodology involves additional MLM training on this curated data, refining the model's understanding of task-specific characteristics. The results demonstrate that TAPT often approaches or surpasses the DAPT performance, particularly in low-resource settings.

The authors also investigate the combination of DAPT and TAPT through sequential application: first, DAPT is performed on a wide range of biomedical corpora, followed by TAPT on non-labeled task-specific datasets. This strategy combines the extensive domain expertise gained through DAPT and the precise task adaptation of TAPT, resulting in better performance in various biomedical classification tasks.

In scenarios where large domain-specific corpora are unavailable or limited, the paper proposes the use of data selection strategies. These methods involve automatically identifying and sampling texts that are most relevant to the target task or domain, thus optimizing the utility of limited data for effective pre-training.

The paper discusses the usefulness of BioMed-RoBERTa primarily in the context of biomedical text classification tasks. It stresses that RoBERTa's pre-training on biomedical domain data, through DAPT and TAPT, significantly improves performance on various biomedical NLP tasks, such as document classification and clinical note analysis. The authors emphasize that such domain-adapted models can better understand the domain-specific language, terminology, and structures typical in biomedical texts, thus enabling more accurate and reliable information extraction, classification, and question-answering in biomedical domains.

Chapter 3

Dataset

This chapter outlines the data resources developed and utilized to investigate topic-specific metaphor detection within a specialized domain, given the constraints of limited annotated data. At the beginning of this research, the primary available resources were the VU Amsterdam Metaphor Corpus (Steen et al., 2010), created using the MIPVU Metaphor Identification Procedure, and a small set of immunotherapy-related signaled metaphor examples previously compiled by Bos et al. (2025). The MIPVU framework offers a clear linguistic protocol for identifying metaphorically used words, while the signaled metaphor examples provide a domain-specific starting point for exploring characteristic metaphor patterns within the target corpus.

Although valuable as methodological foundations, these resources alone were insufficient to train and evaluate an automatic detection system tailored to the field of immunotherapy. Consequently, the principal aim of the methodological work was to develop an approach that could achieve reliable topic-specific metaphor detection despite the scarcity of annotated data. To this end, the research pursued two complementary strategies:

1. Incrementally extending the available data through targeted manual annotation, focusing on examples relevant to the chosen domain and topic;

2. Designing and evaluating automatic detection models that make optimal use of both the original and newly annotated texts.

The datasets described in this chapter were therefore constructed with distinct but interrelated purposes: each supports a specific aspect of developing and validating an automatic, domain-adapted metaphor detection system under resource constraints. The following sections detail the composition, annotation principles, and intended roles of each dataset within the broader experimental framework and illustrate how general-purpose linguistic protocols can be adapted and extended for effective use in a specialized biomedical context.

To accommodate these goals, data preparation followed four key steps: (1) splitting the manually annotated corpus into train and test sets for metaphor detection in a medical context; (2) preprocessing the original signaled metaphors into a token-level version suitable for NLP; (3) compiling a focused sentence-level dataset combining immunotherapy and VUA metaphors to test domain relevance; and (4) collecting automatically predicted metaphorical tokens for error analysis and qualitative insights.

3.1 Immunotherapy Metaphor Dataset

The dataset used in this thesis, referred to as the Immunotherapy Metaphor Dataset, was compiled as part of the study by Bos et al. (2025), introduced in the previous section. The study compiled two datasets consisting of 1,425 scientific articles and 2,650 newspaper articles about immunotherapy. These texts were reduced to a total of 358 scientific publications and news articles after the metaphor identification process. In these texts, 510 text fragments containing signaled metaphors were identified. These metaphors represent 210 different metaphorical words from 23 metaphorical source domains. The metaphors were categorized according to the aspects of immunotherapy they described, such as the workings of immunotherapy and its role or function.

Looking at how immunotherapy metaphors are used in both scientific and non-scientific contexts was achieved by sampling news articles and scientific publications and by identifying explicit (signaled) metaphors, such as 'tumor cells that can serve **as** *targets*'. Each text is accompanied by annotations identifying words that are used metaphorically and specifications about their source domains (e.g., war, journey). These annotations were manually curated by the authors using AntConc to locate the signal words. To determine whether a word was used metaphorically, the annotators compared its contextual meaning with its most basic concrete sense using dictionary entries from the Macmillan Dictionary, Merriam-Webster, and the Longman Dictionary of Contemporary English Online.

During this process, 510 metaphorical words related to immunotherapy were identified and marked. Of these metaphors, 209 were found in news articles and 301 in scientific publications. The most commonly used frames were WAR, PERSON, JOURNEY, BUSINESS & FINANCE, and PEOPLE'S ACTIONS, STATES & PROCESSES. The least used frames were FOOD & DRINKS, SPACE, COMPUTER, DISTANCE, and EDUCATION. These frames were identified by taking signaled metaphors and determining their relevance to immunotherapy and were decided based on the UCREL semantic analysis system (Piao et al., 2015). The annotators also categorized which aspects of immunotherapy the metaphors described, for example, the role or function of immunotherapy and medical condition.

3.2 MIPVU

The MIPVU (Metaphor Identification Procedure Vrije Universiteit) framework, as described by Steen et al. (2010), extends the original MIP (Crisp et al., 2007) into a systematic, replicable protocol for annotating metaphorically used words in authentic discourse. MIPVU provides detailed operational guidelines for determining the contextual and more basic meanings of lexical units, thereby ensuring that metaphor identification is grounded in explicit linguistic analysis rather than impressionistic judgments.

The original MIPVU corpus consists of a subset of the British National Corpus (BNC), covering various text genres such as news reports, academic prose, fiction, and conversational data. Within this corpus, metaphorical language is identified at the word level: each lexical unit is examined to ascertain whether its contextual meaning contrasts with a more basic, concrete meaning and whether this contrast can be understood by comparison, thus signaling metaphoricity.

The annotation procedure for the VUA Metaphor Corpus adhered to the MIPVU protocol, ensuring systematic and replicable identification of metaphorically used words

across a wide range of text genres. The annotators first segmented the texts into lexical units, applying part-of-speech tagging consistent with the protocol guidelines. For each lexical unit, the annotators determined its contextual meaning within the discourse and identified a more basic and concrete meaning, firstly based on the Macmillan Dictionary and, in cases where definitions were unavailable, the Longman Dictionary of Contemporary English to ensure consistency in sense differentiation.

A lexical unit was annotated as metaphorical if its contextual meaning differed from its more basic meaning but could be understood via a conceptual comparison between the two meanings, in accordance with the theoretical premises of the MIPVU framework.

Through this procedure, the VUA Metaphor Corpus provides an extensive and carefully annotated resource covering multiple registers, including news discourse, academic writing, fiction, and conversational data. Its systematic methodology has made it an indispensable benchmark for the development of automatic metaphor detection systems.

Building on this foundational resource, the present study used MIPVU both as an annotation guideline and as a reference point for developing domain-specific metaphor detection. Given that the target domain differs from the general English represented in the BNC, additional annotations were carried out on a selection of domain-relevant texts.

The following types of metaphor are defined in the MIPVU and used in the annotation of the VUA Metaphor Corpus.

Indirect Metaphor

These are the most frequent types in the corpus. An indirect metaphor occurs when a word or phrase is used in a way that contrasts with its more concrete or basic meaning, but without an explicit metaphor marker.

'The treatment *attacks* cancer cells.'

Here, *attacks* has a basic military meaning (e.g., 'soldiers attacking'), but in this context it is used figuratively to describe how a therapy works on cancer cells.

Implicit Metaphor

An implicit metaphor relies on an implied comparison that is only clear when its connection to a previous part of the discourse is understood. There is no overt metaphorical word, but the idea is sustained through context.

'Naturally, beginning such a *fight* does not necessarily mean that you'll see *it* through to remission.'

Here, *it* inherits the metaphorical framing of *fight*, but it is not directly stated. It is inferred from earlier discourse.

Direct Metaphor

In direct or explicit metaphors, the figurative comparison is made clearly, often marked with words such as *like*, or in quotation marks.

'Immunotherapy was like a long *journey*.'

This is a direct metaphor, as *journey* is explicitly used to compare immunotherapy to a trip.

3.3 Annotations

As the Immunotherapy Metaphor Dataset contained annotations of only explicit (signaled) metaphors, a small sample of the data was selected for a complete metaphor identification process following the MIPVU procedure. A fully annotated sample of the data was needed to conduct an initial evaluation of the selected metaphor detection models, in order to determine whether their performance was adequate or if further fine-tuning was necessary.

Firstly, after reviewing the full texts in a raw txt format, two news articles and three scientific publications were selected. For diversity purposes, only the abstracts of the scientific texts were included in the fully annotated dataset, so that a larger number of texts could be added to the test set. Five texts were selected to reflect a range of subjects and vocabulary, to the extent allowed by the immunotherapy context. In particular, the two news articles were chosen for their more narrative-driven style, which was expected to provide a richer variety of metaphorical words. The three scientific articles were not expected to be as rich in metaphors as the news articles, especially since they developed scientific presentations of findings, which is another reason why the abstracts were the only parts that were annotated.

The annotation process followed the guidelines of the MIPVU procedure. It was performed on a token level and only targeted context words. Subsequently, the following spaCy parts of speech were excluded from the checks: adposition, auxiliary, coordinating conjunction, determiner, numeral, particle, pronoun, proper name, punctuation, subordinating conjunction, space, symbol, and other.

Annotations were performed by me in all texts, with help from two student assistants for inter-annotator agreement purposes. The two student assistants are Master's students at VU Amsterdam, with training and experience with the MIPVU procedure and annotation guidelines. Their studies and background are Linguistics, specifically Text Mining and Forensic Linguistics.

The preprocessing included tokenization of the raw texts with spaCy information that was written in Excel format to facilitate annotations. The texts were divided between the two student assistants, who annotated a similar number of tokens. Annotation tasks included news articles and scientific publication abstracts; to optimize time, the workload was divided among annotators.

Text Title	**Tokens**	**Identified Metaphors**	**Relevant Metaphors**	**Annotators**
'The grief counsellor walked into my family's life like Mary Poppins'	1,520	106	18	3
A perspective on the impact of radiation therapy on the immune rheostat	217	20	10	2
A SKIN CANCER WONDER CURE; Survival rates up to 94% in drug trial	562	33	8	3
Science, medicine, and the future: Lung cancer	107	5	1	2
Setting the scene: a future 'epidemic' of immune-related adverse events in association with checkpoint inhibitor therapy	92	12	5	2
Total	**2,498**	**176**	**42**	**3**

Table 3.1: Annotated texts with token counts, identified metaphors, relevant metaphors, and number of annotators.

Text Title	Annotator-1 vs Annotator-3	Annotator-1 vs Annotator-2	Annotator-3 vs Annotator-2	Average Cohen's Kappa	Krippendorff's Alpha
'The grief counsellor walked into my family's life like Mary Poppins'	0.583	Not Applicable	0.629	0.606	0.606
A perspective on the impact of radiation therapy on the immune rheostat	0.597	Not Applicable	Not Applicable	0.597	0.595
A SKIN CANCER WONDER CURE; Survival rates up to 94% in drug trial	0.606	0.628	0.647	0.627	0.621
Science, medicine, and the future: Lung cancer	Not Applicable	Not Applicable	0.692	0.692	0.692
Setting the scene: a future 'epidemic' of immune- related adverse events in association with checkpoint inhibitor therapy	Not Applicable	Not Applicable	0.654	0.654	0.654

Table 3.2: Inter-annotator agreement scores (Cohen's Kappa and Krippendorff's Alpha) for metaphor annotation across texts.

Each annotator worked independently on the files and annotated on a token-by-token basis. Initially, each token's automatically assigned part of speech was checked, and if it fell on the previously mentioned part of speech categories, it was immediately categorized as non-metaphorical. For all other words, the MIPVU procedure was followed by identifying the meaning of the word in the specific context of the text at hand and comparing the concreteness of that meaning with the most basic meaning available in the dictionary. The Longman Dictionary of Contemporary English Online was used for this process (Longman, 2014).

All annotators were required to provide a justification based on the dictionary and their personal judgment as to why they marked a word as metaphorical. This helped defend their decision during the weekly meetings that were dedicated to the immunotherapy annotations in case a disagreement occurred.

The first stage of annotation consisted of the complete annotation of all context words. The relation of words with the immunotherapy context was not taken into account. Inter-annotator agreement was measured using both pairwise Cohen's Kappa and Krippendorff's Alpha. The results, as seen in Table 3.2 indicate moderate to substantial agreement, though slightly below the commonly accepted threshold of 0.67 for Krippendorff's Alpha. Although the given metaphor annotation framework significantly limits disagreements, the subjective nature of metaphor identification and the specific domain of the texts made the annotation process quite challenging.

3.4 Challenges

Applying the MIPVU protocol to a new specialized domain revealed several interesting challenges that illustrate why metaphor detection in topic-specific contexts is far from trivial. Although MIPVU provides a clear procedure for identifying metaphorically used words, using it outside its original, general-purpose setting raised questions about how strictly it should be followed and how much flexibility annotators should have

when interpreting meanings.

This makes the task harder because, in specialized domains, figurative patterns tend to be subtle and reused, which can blur the line between literal and metaphorical use. In the case of health and medical reporting, common metaphors frequently draw on conceptual frames such as miracles, warfare, and journeys, which can be subtle and conventional, but still important for understanding how the topic is framed for the reader.

Several notable cases emerged during the annotation process. One particularly difficult example was the word *wonder* in the headline *'A SKIN CANCER WONDER CURE.'* This instance came up at the very beginning of the annotation work and sparked a debate about how it should be treated based on available dictionary definitions. According to the Longman Dictionary of Contemporary English, the only entry for *wonder* was 'something that makes you feel surprise and admiration.' However, other dictionaries, such as Merriam-Webster, list an additional sense, 'miracle,' which would qualify as metaphorical according to MIPVU. This created controversy over whether such examples should be included. In the end, it was decided that using one consistent dictionary, in this case Longman, would ensure a clear workflow and minimize disagreements among annotators.

An important insight from this process is that the choice of dictionary can directly influence which words are annotated as metaphorical and which are not. When different dictionaries list different senses for the same word, annotators may reach different conclusions about whether a meaning contrast exists and whether a word meets the MIPVU criteria. This means that even when using the same procedure, the final annotations, and consequently any automatic systems trained on them, can vary depending on which lexical resource is used as the point of reference. For this reason, establishing and consistently following a single dictionary was essential in this study to ensure coherence and comparability across annotations.

Other challenges involved words that were quickly dismissed by more than one annotator because they seemed too basic to be worth considering, but which actually did have multiple senses in the dictionary and could be marked as metaphorical. For example, the word *'thing'* is one of the most common words in English, but when used in the sense of 'idea' rather than 'object,' it arguably functions metaphorically because its contextual meaning extends beyond its more basic physical sense.

These examples show that when the MIPVU procedure is applied in a new domain, metaphor identification can become more subtle and open to interpretation than it might be in general language. In this specific domain, metaphors often revolve around concepts such as miracles, cures, battles, and journeys, which means that annotators need to find balance between following the protocol step by step and recognizing how certain figurative patterns commonly appear in this type of discourse.

In general, these challenges highlight that even with a clear protocol, the detection of domain-specific metaphors is inherently complex and requires careful decisions about consistency, dictionary use, and the level of interpretive freedom allowed during annotation.

3.5 Immunotherapy-related Metaphor Annotation

In order to proceed with immunotherapy-related metaphor identification, fully annotated texts, which include explicit and implicit metaphors of all domains and topics, are

then carefully considered to establish a preliminary framework for determining which metaphors cover topics that are relevant to immunotherapy.

The first attempt at annotating this part of the dataset was intuition-based, that is, questions such as whether this word describes a disease, a treatment, or a biological concept, whether such medical concepts are explicitly mentioned directly before and after this word, and whether this word could easily be used in a variety of concepts without having a strong tie to one of them, were raised. For example, words such as '**fight** cancer', '**aggressive forms** of the disease', and '**victims**' were easily marked as domain-specific. They clearly talk about cancer, the way it is presented, how it can be treated, and the losses that it causes, therefore they were marked as immunotherapy-related without a doubt. On the contrary, '**further** results', 'He **added**', and 'that's an important **thing** to say' were marked as unrelated to immunotherapy, as they can be used in any context and do not present specific ties to the topic. They also do not affect the public's perception on cancer or its potential treatment, therefore they are considered as too generic to be marked.

However, certain cases were more difficult to categorize. For example, '**uncovered** results' and 'the study **found**' both mention scientific methods used to study and improve cancer treatments. Although they directly concern immunotherapy, they are not particularly related to this concept. These cases showcased the need to apply the annotation criteria more strictly. Specifically, when only some of the conditions were met, such as reference to the concept of immunotherapy, its perceived impact, or a strong connection to medicine, it became necessary to develop a clearer framework and establish priorities to resolve such ambiguities.

A possible solution to this issue could be to consider the immediate context of these metaphors. Specifically, if the same sentence contains terms that are directly related to immunotherapy, such as 'cancer' or 'clinical trial', and the metaphorical word affects the perception of the public, it can be considered immunotherapy-related, as in the example '**uncovered** results'. This metaphorical word, although it can be used in a variety of contexts, refers to results that occurred in a clinical trial that aims to find ways on how to direct the body's immune system towards fighting cancer. The word itself most likely does not cause particular reactions or thoughts to the general public, however, when combined with clinical trial results that could potentially cure cancer, it could lead to feelings of hope for the readers of this particular article and falls within the margins of immunotherapy-related context.

Another possibility that occurred was to create a third category, i.e. words that fit the scientific context, but are not necessarily immunotherapy related. The borderline examples mentioned above could fit this category, as they concern subjects such as study results and findings, which, in this context, explain facts related to immunotherapy, but are not necessarily related to this domain. However, this would require additional examples that are sufficiently similar to these borderline cases in order to proceed with further training of a potential classifier that identifies all three metaphor categories.

Ultimately, since the dataset was modified for NLP purposes, the annotation process regarding the relevance of a word to immunotherapy was limited to its direct link to immunotherapy. For example, the metaphor would need to fall under the umbrella of science communication, that is, explain a medical concept in simple terms and directly impact the readers' perception of it with this particular choice of words.

3.6 Data Preparation

To accommodate the available data and the scope of this project, four steps of data preparation are needed, each serving a specific purpose in the overall experimental framework. Firstly, the manual annotations derived from the txt files available in the original data are split into train and test sets. This version is used to evaluate the classifiers' ability to detect metaphors in a medical context, based on manually verified examples.

Since the original dataset was not created for NLP purposes, preprocessing is required to convert it into a suitable format for token-level classification. In this step, the metaphorical word is located, its surrounding text is matched to the corresponding txt file in the directory, and then the text is tokenized and written to a tsv file, with an additional column indicating whether each token is metaphorical or not. This version is primarily used for measuring recall on explicit metaphors only, as only the signaled metaphors are positively marked in this format.

In addition to this full-text token-level version, a more focused version is created that contains only the sentences that include metaphors. This sentence-level version combines immunotherapy sentences with signaled metaphors, sentences extracted from the VUA Metaphor Corpus (which contains both explicit and implicit metaphors), and sentences from the fully annotated immunotherapy texts. The 120 signaled metaphors from the VUA corpus are split into train, development, and test sets and integrated with the rest of the files. This version is primarily used for measuring recall on explicit metaphors only, because it only includes positive labels for the signaled metaphors found during manual annotation, while all other tokens are treated as non-metaphorical by default. This setup allows for an assessment of how many true metaphorical tokens the model can retrieve, but it does not allow the measurement of precision, since not all possible metaphors in the text were exhaustively annotated.

Finally, a fourth version gathers the sentences with tokens that were predicted as metaphorical by the first two classifiers. This version is used to analyze the performance of the models beyond manual annotations. It provides insight into how many predicted metaphors belong to the immunotherapy-relevant class and supports a qualitative analysis of the kinds of metaphors detected automatically, as well as their usefulness in describing immunotherapy-related phenomena.

It should be noted that all training, development, and test splits followed a 60/20/20 ratio.

Chapter 4

Methods

The purpose of this thesis is to evaluate whether models trained for metaphor detection can be outperformed by models with domain-specific pre-training, with a particular focus on the field of immunotherapy. In order to approach this question systematically, the overall metaphor detection task is conceptualized in two distinct but closely associated steps. The first step involves identifying all metaphorical tokens present in the text, regardless of the topic. In the second step, those metaphorical tokens are then filtered to determine which are specifically relevant to the domain of immunotherapy.

The initial stage of evaluation includes the selection of potential models suitable for metaphor classification. These models are required to be publicly available and demonstrate performance comparable to current state-of-the-art standards for general metaphor detection. Once the models are implemented and their performance on the broad metaphor class is assessed, additional steps involve fine-tuning, where necessary, and implementing a secondary classifier to detect domain relevance within the set of metaphors extracted in the first step.

To address both parts of this pipeline, BioMed-RoBERTa is chosen as the domain-specific pre-trained language model and compared to XLM-RoBERTa for the general metaphor detection task. To evaluate domain relevance classification, a fine-tuned BERT model was used along with a Logistic Regression baseline. Although any modern transformer-based model (e.g. RoBERTa or DistilBERT) could have been suitable for this task, BERT was chosen simply because it is a reliable and well-understood model with strong performance. The method would likely generalize similarly with other transformer architectures.

This chapter provides details on the full experimental setup, including model selection, training configurations, and the two-step pipeline for detecting and categorizing metaphors. The final sections explain the evaluation strategy for each step separately, as well as for the end-to-end system, and demonstrate what this two-fold approach offers regarding the challenges of domain-specific metaphor detection under limited resource conditions.

4.1 Model Architecture

In both steps of this thesis, transformer models are used for metaphor detection and domain-relevance classification.

Specifically, the models used are:

1. XLM-RoBERTa
2. BioMed-RoBERTa-base
3. bert-base-uncased

For the second step, which involves determining whether a metaphor is relevant to immunotherapy, a Logistic Regression classifier is included as a simple baseline and is compared to a transformer-based classifier (bert-base-uncased).

4.1.1 Transformer Models

Transformer models are large language models that go beyond capturing syntactic information. Language input is represented as embeddings, that is, high-dimensional vectors, which allow them to model various aspects of natural language.

BERT and XLM-RoBERTa

BERT is a transformer model that utilizes only the encoder architecture. It is pre-trained on masked language modeling and next sentence prediction, processes that are carried out on a vast amount of text data (Devlin et al., 2019). XLM-R is a transformer-based multilingual masked language model trained on 100 languages using more than two terabytes of filtered data from the CommonCrawl corpus (Conneau et al., 2020). Inspired by developments such as RoBERTa, XLM-R is trained without the next sentence prediction task and uses extended training on a larger and more diverse dataset. Training involves significantly larger batch sizes, specifically up to 8,192, and more than 500,000 updates, which contribute to its improved performance. The model's large vocabulary and extensive multilingual data enable it to generate high-quality representations in a wide range of languages, leading to state-of-the-art results on various cross-lingual benchmarks, including XNLI, MLQA, and NER tasks.

BioMed-RoBERTa

BioMed-RoBERTa, developed by Gururangan et al. (2020), is a domain-adapted variant of the RoBERTa-base language model pre-trained for biomedical NLP tasks. It is created with continuous pre-training of the RoBERTa-base model on a large-scale collection of biomedical texts, specifically 2.68 million full-text scientific papers from the Semantic Scholar corpus. Unlike datasets that only include abstracts, this corpus contains complete articles, amounting to around 7.55 billion tokens or 47 GB of text.

The model architecture is identical to RoBERTa-base, consisting of 12 transformer layers, each with 12 attention heads and a hidden size of 768, totaling approximately 125 million parameters. Pre-training utilizes the masked language modeling task and does not include the next sentence prediction task, consistent with RoBERTa's original training strategy. This continued pre-training helps the model to learn domain-specific language patterns and vocabulary commonly found in biomedical literature.

4.1.2 Logistic Regression

Logistic Regression is a simple classification algorithm that models the probability of a given input belonging to a particular class. In the context of binary classification tasks, such as determining whether a token is related to immunotherapy, logistic regression

learns a set of weights for input features that best separate the two classes. During training, the model adjusts its weights to minimize the error between predicted and actual labels using a loss function (Jurafsky and Martin, 2025).

4.2 Experimental Design

This section outlines the intuition behind the use of the models described above, followed by details of the experimental design, parameter tuning, and evaluation metrics used.

4.2.1 Tasks Overview

The experimental design addresses two interconnected tasks. The first task, Metaphor Detection, aims to distinguish metaphorical from literal expressions within domain-specific texts, using both a general-domain and a domain-adapted RoBERTa model. The second task, Domain Relevance Classification, determines whether a given metaphorical expression is relevant to the domain of immunotherapy. This two-level setup makes it possible to develop a pipeline that first identifies metaphorical language and then assesses its domain specificity, allowing for an analysis of how domain adaptation affects overall performance.

4.2.2 Data

Both the metaphor detection and domain relevance classification tasks use data derived from two primary sources:

- the VUA Metaphor Corpus, and
- the Immunotherapy Metaphor Dataset.

Each task uses specific dataset splits for training and evaluation, structured to test performance under different conditions and to adjust to specific annotation patterns. An overview of the individual datasets used, their role in the pipeline, and how they help in the assessment process is provided below.

Both tasks make use of data derived from the two previously mentioned sources. For metaphor detection, both XLM-RoBERTa and BioMed-RoBERTa are fine-tuned and are initially evaluated only on the VUA Metaphor Corpus.

(A) Annotated Immunotherapy Subset A small, high-quality dataset consisting of five fully annotated immunotherapy-related texts, manually labeled for metaphorical and literal tokens. Of these, two are reserved for training and three for testing in the metaphor detection step.

It is initially used in the Metaphor Detection testing phase, as shown in Fig. 4.1 to evaluate how well fine-tuned XLM-R and BioMed-RoBERTa, metaphor detection models trained on out-of-domain metaphors (from VUA), transfer to domain-specific language. Its small size limits generalization; however, it allows for the complete evaluation of models in the domain of immunotherapy.

It is also used in the Domain Relevance Classification phase of the thesis, as shown in Fig. 4.2, where it is transformed into a version where only sentences with at least

one gold-labeled metaphor are incorporated. Sentences appear once for each annotated metaphor they contain, producing multiple instances when multiple metaphors are present. This version will be referred to as the **Annotated Immunotherapy Subset for Domain Relevance**.
The texts used in the training stage are the following:

1. 'The grief counsellor walked into my family's life like Mary Poppins'
2. A perspective on the impact of radiation therapy on the immune rheostat

The texts used in the evaluation stage are the following:

1. A SKIN CANCER WONDER CURE; Survival rates up to 94% in drug trial
2. Science, medicine, and the future: Lung cancer
3. Setting the scene: a future 'epidemic' of immune-related adverse events in association with checkpoint inhibitor therapy

(B) Complete Immunotherapy Corpus A large domain-specific corpus compiled from full-length immunotherapy-related texts containing only annotated signaled metaphors (i.e., metaphors explicitly indicated by words such as 'like', 'as', etc.). The texts are derived from the Immunotherapy Metaphor Dataset.

It is used in the Metaphor Detection testing phase, as shown in Fig. 4.1, to evaluate metaphor detection performance on a much broader and more varied set of real-world domain-specific texts. In this version of the dataset, only immunotherapy-related signaled metaphors are annotated, while unmarked metaphors may still be present but are treated as literal due to lack of labels. Because the annotations are incomplete, the evaluation is performed using recall alone. Precision and F1-score cannot be reliably computed, as true negatives and false positives cannot be determined due to the lack of annotations. Despite this limitation, recall is a reliable way to measure how well the system recovers metaphorical expressions from partially labeled real-world data.

(C) Combined Relevance Corpus A sentence-level dataset constructed to train and test the Domain Relevance Classifier. It includes:

- Sentences from the VUA Metaphor Corpus that contain at least one annotated metaphor (explicit or implicit). The subset of 120 explicit metaphors is split into training, development, and test sets for this task.
- sentences with explicit metaphors from the Complete Immunotherapy Corpus, and
- the fully annotated texts from the Annotated Immunotherapy Subset.

Only sentences with at least one gold-labeled metaphor are included. Sentences are repeated in the dataset so that each metaphorical token corresponds to a separate instance, allowing the model to process multiple metaphor labels per sentence independently.

It is used in the Domain Relevance Classification phase for training and testing, as shown in Fig. 4.2. It evaluates whether a model can distinguish immunotherapy-related metaphors from general-purpose ones.

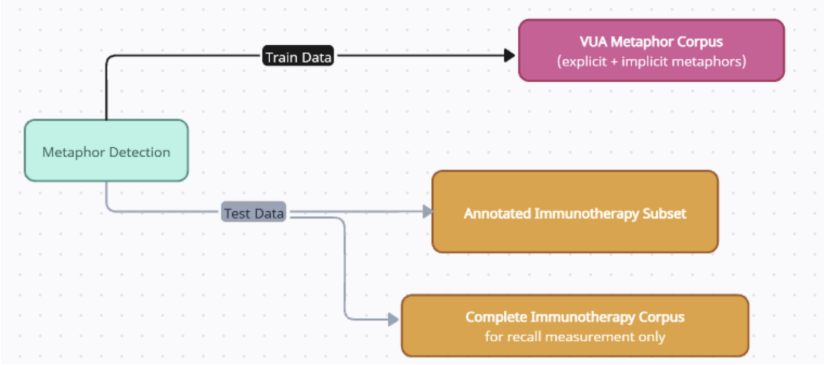


Figure 4.1: Metaphor Detection Pipeline

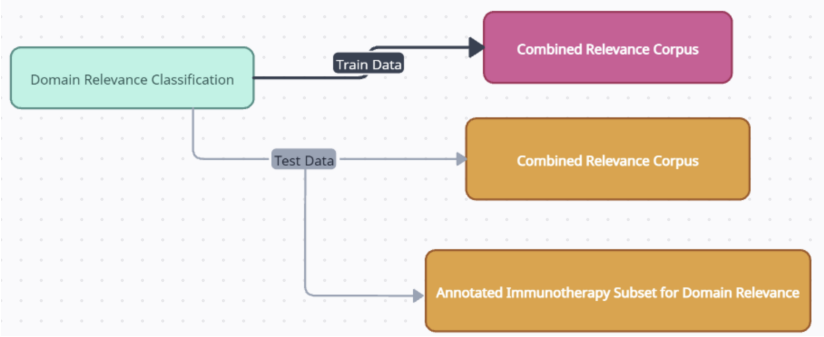


Figure 4.2: Domain Relevance Classification Pipeline

(D) Predicted Metaphor Corpus A dataset built by running the metaphor detection models (XLM-RoBERTa and BioMed-RoBERTa) over the Annotated Immunotherapy Subset and collecting sentences containing predicted metaphorical tokens.

It is used to test the fully automatic pipeline, as shown in Fig. 4.3, which aims to resemble a real-world scenario where annotated metaphors would not be available and would need to be detected by a metaphor detection model before being passed to the domain relevance classifier. It allows end-to-end evaluation of how well the pipeline retrieves domain-relevant metaphors in real-world data and supports qualitative analysis of what types of metaphors are identified and how well they capture domain-specific concepts. However, as it resembles realistic conditions, errors from the metaphor detection step are propagated forward, and relevance classifications are applied to potentially noisy input.

4.2.3 Metaphor Detection

To compare the effect of domain-specific pre-training, BioMed-RoBERTa is fine-tuned for the metaphor detection task using the same training setup as Wachowiak et al. (2022). The goal is to research whether a model with pre-training on biomedical texts can outperform a general-purpose multilingual model such as XLM-RoBERTa when applied to a domain-specific setting. The original code provided by Wachowiak et al. (2022) supports RoBERTa-based architectures, which allowed the substitution of XLM-R with BioMed-RoBERTa for fine-tuning.

The XLM-RoBERTa model used in this thesis is the version fine-tuned by Wa-

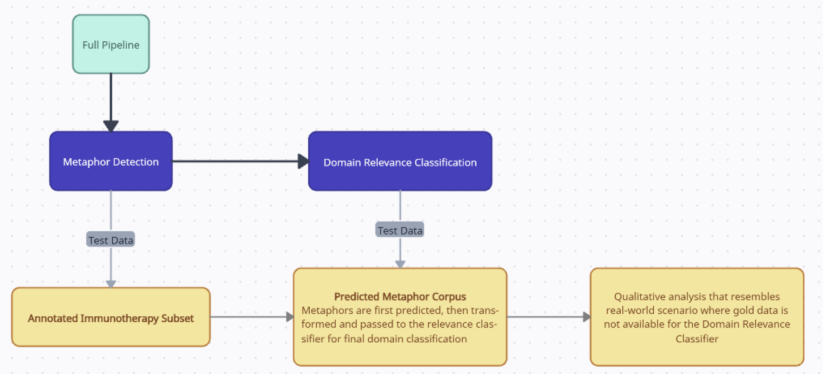


Figure 4.3: Full Pipeline: from Metaphor Detection to Domain Relevance Classification

chowiak et al. (2022), trained with a learning rate of 2e-5 for eight epochs. The best-performing model based on validation was selected using the same train-test split as in the original experiment, with 10% of the training data held out for validation. This model was chosen because of its availability and performance, which is comparable to the best performing models of the Metaphor Detection Shared Task of 2020.

4.2.4 Domain Relevance Classification

This task aims to determine whether a metaphor identified in a given sentence is relevant to the domain of immunotherapy. Given a sentence containing a metaphorical expression (as detected in the first step of the pipeline), the goal is to classify the metaphor as either domain-specific (i.e., related to immunotherapy) or general.

For example, in the sentence 'Her immune system is *like* a defense barrier,' the metaphor *'barrier'* is domain-specific: it describes a medical process and is directly related to immunotherapy mechanisms. In contrast, a sentence such as 'The study took a *step* forward in understanding the immune system' contains a metaphor (*'step'*) that is generic and not specific to immunotherapy.

This classification task is based directly on the annotation framework developed for this thesis, in which metaphors were considered domain-specific if they:

- Describe concepts such as disease, treatment, or biological mechanisms,
- Appear in contexts with explicit references to immunotherapy (e.g., 'immune system', 'cancer'),
- And/or influence how medical topics are perceived by a general audience.

This task is framed as a binary sentence-level classification problem, where each instance includes one metaphor and its surrounding sentence, and the model must determine whether that metaphor contributes to domain-specific scientific communication. The training material for both BERT and Logistic Regression (baseline) is the *Combined Relevance Corpus*.
To illustrate how domain relevance classification works, consider the following sentence, where each token is indexed starting at 0:

```
[These, drugs, are, associated, with, a, specific, mechanism, of, action,
that, has, profound, implications, for, both, immunology, and, inflammatory,
disease, .]
```

Here, the target token index is 7 ('mechanism'), which is under consideration for domain relevance classification. The corresponding label for this token is 1, indicating that it is relevant to the immunotherapy domain. This means that the classifiers should predict that the metaphorical token 'mechanism' in this sentence contributes meaningfully to the domain-specific context. Both models use the same setup and datasets.

Baseline: Logistic Regression

As domain relevance classification is a binary classification task, a Logistic Regression classifier is used as a baseline. Logistic Regression serves as a useful point of comparison for evaluating the added value of more complex models such as transformer-based architectures.

For each token marked as metaphorical, the classifier considers three simple features: (1) the token itself, (2) the directed dependency path from the token to any immunotherapy-related word in the sentence (if present), and (3) the immunotherapy-related word itself. If no path is found, 'No Path - None' is used as the path feature, and 'no relevant tokens' is assigned to the immunotherapy-related word feature. The path and immunotherapy-related word features are kept separate to avoid sparsity. In reality, combining them could lead to overly specific patterns, reducing the model's ability to generalize.

These features were selected based on the intuition that metaphors related to immunotherapy frequently involve references to immunotherapy concepts, such as 'cancer' or 'autoimmune.' The syntactic distance and dependency relations between a metaphorical token and these domain-specific terms can provide valuable cues for determining relevance. The assumption is that metaphorical expressions related to immunotherapy concepts tend to exhibit similar dependency paths.

The list of immunotherapy-related terms used in this thesis was constructed by combining two sources: (1) words that appeared frequently in the immunotherapy dataset, such as *cancer*, *tumor*, and *immunotherapy*, and (2) relevant terminology extracted from the Cancer Research Institute's ImmunoGlossary (Cancer Research Institute). To maintain generalizability and avoid sparsity, highly specific terms (e.g., rare cancer types or gene variants) were excluded. Instead, the list focuses on broader concepts such as immune cell types, treatment categories, and biological mechanisms. The list can be found in the Appendix A.

The reliance of the Logistic Regression model on traditional features limits its ability to capture deep semantic relations and context-dependent metaphorical uses. Moreover, the use of a relatively small dataset increases the risk of overfitting, especially for simpler models that lack mechanisms for capturing deeper contextual information. It is highly likely that, given a larger and more complex dataset in a real-world scenario, the performance of this model will be insufficient.

BERT

To complement the Logistic Regression baseline, a fine-tuned BERT model is used for the task of domain relevance classification. Similarly to the baseline model, BERT classifies at a token level and is fine-tuned to predict whether a given metaphorical token is relevant to the domain of immunotherapy, based on its context within the sentence.

The setup mirrors that of the baseline: only tokens previously identified as metaphorical are evaluated, and all non-metaphorical tokens in the sentence are assigned a special label of -100 to ensure that they are ignored during loss computation. This setup focuses on the task of determining domain relevance within metaphorical expressions, without the need to classify all tokens in a given sentence, thus eliminating noise in predictions and unnecessary resources.

Unlike the Logistic Regression model, which relies on hand-crafted features, BERT uses pre-trained embeddings and deep contextual information. This allows it to capture lexical and syntactic cues without explicit feature engineering. During training, BERT receives the full sentence as input and learns to classify only the metaphorical tokens based on their context, benefiting from surrounding information.

This approach addresses one of the main limitations of the baseline model, which is to generalize beyond traditional feature representations. As it does not rely on specific lists of common immunotherapy-related words, BERT is expected to be more robust to variations in sentence structure and more sensitive to cues that indicate relevance.

4.3 Evaluation

To assess the impact of domain-specific pre-training and the two-level classification pipeline, both tasks are evaluated separately and in sequence. The overall evaluation pipeline consists of two major components: (1) Metaphor Detection, where models identify metaphorical tokens in context, and (2) Domain Relevance Classification, where detected metaphors are classified as either immunotherapy-related or not. These steps are designed with a real-world use scenario in mind, forming a complete end-to-end pipeline for identifying domain-specific metaphorical language.

To quantify model performance, the standard F1-score is used, which is defined as the harmonic mean of precision and recall, treating both metrics with equal importance:

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

Precision indicates how many of the tokens predicted as metaphorical are indeed metaphorical according to the gold standard annotations. Precision for the metaphor class is calculated as the number of true positives (TP) divided by the total number of tokens predicted as metaphorical, which includes both true positives and false positives (FP):

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

In the context of binary metaphor detection, a high precision score means the model labels a token as metaphorical only when it is confident, thus minimizing false positives.

Recall, on the other hand, measures the model's ability to identify all actual metaphors present in the text. It is calculated as the number of true positives divided by the sum of true positives and false negatives (FN):

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

A high recall score means that the model captures as many metaphorical instances as possible, even at the risk of including some false positives.

Together, precision and recall provide a balanced view of how well each model detects metaphorical language. The F1-score combines them into a single metric, allowing a straightforward comparison between models and experimental conditions (Manning et al., 2008).

4.3.1 Metaphor Detection

To evaluate the extent to which domain-specific pre-training benefits metaphor detection, the performance of BioMed-RoBERTa is compared to that of the general-domain XLM-RoBERTa model on the same classification task. Both models are fine-tuned to distinguish metaphorical from non-metaphorical expressions, their only difference being their pre-training data. The use of immunotherapy texts derived from news articles and scientific publications makes this setting particularly suitable for testing whether general-domain pre-training is suitable for the task or whether domain-adapted models boost performance.

The metaphor detection models are evaluated on the *Annotated Immunotherapy Subset*, a fully annotated dataset containing both explicit and implicit metaphor labels. This dataset supports full evaluation using precision, recall, and F1-score, allowing for a complete comparison between models.

Additionally, models are evaluated on the *Complete Immunotherapy Corpus*, which contains only annotated signaled metaphors (i.e., metaphors explicitly marked with cue words such as 'like' or 'as'). Because unmarked metaphors are treated as literal due to the lack of labels, precision and F1-score cannot be reliably computed. Therefore, only recall is reported on this dataset. Using recall alone provides information on how well each model recovers metaphorical expressions in real-world immunotherapy discourse with incomplete annotations.

This follows standard practice for working with small amounts of data or incomplete annotations. Each evaluation dataset is selected to match the characteristics of the task: fully annotated texts allow for full metric calculation, while partially labeled datasets can only support recall. This separation ensures that the models are assessed fairly according to what each dataset can support.

This approach draws a well-rounded picture of how XLM-RoBERTa and BioMed-RoBERTa perform in the immunotherapy domain. Keeping the fine-tuning process the same for both models means that any differences in their results can be directly linked to the impact of domain-specific pre-training. Overall, this setup offers a clear test of whether pre-training on biomedical texts helps a model better detect metaphors in this specialized context.

According to the results of the two RoBERTa classifiers described in Chapter 5, it was determined that the metaphor class was adequately predicted by both models. Consequently, the experimental setup of this thesis follows the route of targeted metaphor detection: the creation of a classifier which, given a metaphor and its context, identifies whether the metaphor describes a domain-specific concept.

4.3.2 Relevance Classification

For Domain Relevance Classification, a baseline Logistic Regression classifier and a fine-tuned BERT model are trained using the *Combined Relevance Corpus*. Both models are evaluated at a token level, using only the subset of metaphorical tokens given

the context of the surrounding sentence. Performance is measured by precision, recall, and F1-score. In this task, positive examples are metaphorical tokens labeled as immunotherapy-relevant, while negative examples are metaphorical tokens labeled as not relevant to immunotherapy.

Evaluation is conducted on the test set of the *Combined Relevance Corpus*, which includes metaphorically annotated tokens labeled for their relevance to immunotherapy. This dataset was constructed specifically for this second-level classification task and enables full metric calculation.

As a secondary evaluation, the same metrics are applied only on the *Annotated Immunotherapy Subset for Domain Relevance*. This smaller set ensures consistency in comparison and helps clarify whether the mixed training data (VUA + immunotherapy) of the *Combined Relevance Corpus* is sufficient for models to learn the distinction between general and domain-specific metaphors in real-world biomedical texts. The inclusion of the VUA corpus ensures that the model is exposed to a wide variety of general-domain metaphors that are not relevant to immunotherapy. This contrast is essential for the models to learn to differentiate domain-specific metaphors from general ones.

The results of this stage, along with those from metaphor detection, form the basis for evaluating the ability of the entire pipeline to extract domain-relevant metaphorical expressions from real-world biomedical discourse.

4.3.3 Full pipeline: from Metaphor Detection to Domain Relevance Classification

To perform this end-to-end evaluation, the same metrics are applied to the tokens identified as metaphorical by the metaphor detection models in the three immunotherapy texts. These predicted metaphor tokens, along with their sentences, are saved to a new file and passed to the second level of classification, thus forming the *Predicted Metaphor Corpus*. Evaluating in these circumstances ensures that the practical utility of the models can be tested in settings that reflect real-world usage. As gold labels are likely to be unavailable in real-world settings, a qualitative analysis of predicted metaphor tokens, including how relevant they are to immunotherapy, is an important part of assessing the models' practical usefulness.

In this step of the evaluation, the effectiveness of the full pipeline is measured from metaphor detection to domain relevance classification. Here, the final output is compared to the gold metaphors for the entire *Annotated Immunotherapy Subset*, not just the ones predicted by the model.

To do this, each token is matched to the gold annotations and each token predicted by the pipeline (excluding predefined irrelevant parts-of-speech such as punctuation, conjunctions, determiners, etc.). Any gold metaphor that was never predicted at all is added to the false negatives, directly penalizing the pipeline for missing metaphors at the detection stage. Then, the evaluation continues to whether the predicted metaphors were correctly classified as relevant or irrelevant. This end-to-end metric reflects the real-world scenario in which a missed metaphor is equally as problematic as a misclassified relevance.

Additionally, the total number of gold metaphors (filtered by POS), the total number of metaphors predicted by the pipeline, and the number of gold metaphors that were completely missed (false negatives) are tracked throughout this process.

	END-TO-END: Metaphor Detection + Domain Relevance
TP	Correctly detected **and** relevant
FP	Predicted relevant but **not gold metaphor**
FN	**Missed gold metaphor** or misclassified as irrelevant
TN	Correctly identified as irrelevant **or not predicted**

Table 4.1: Correspondence between classifications and their meaning in the end-to-end evaluation.

Chapter 5

Results

This chapter presents a comprehensive evaluation of metaphor detection performance in the different models used in the experimentation phase of this thesis. Firstly, the results of fine-tuning and testing of XLM-RoBERTa by Wachowiak et al. (2022) and BioMed-RoBERTa on the VUA Metaphor Corpus are presented in Sub-section 5.1.1. Then, XLM-RoBERTa and BioMed-RoBERTa's performances are evaluated on the *Annotated Immunotherapy Subset* (Sub-section 5.1.2), followed by a comparison of only recall between the two models on the *Complete Immunotherapy Corpus* (Sub-section 5.1.3. The goal is to assess whether domain-specific pre-training on biomedical texts offers any advantage over a general-purpose multilingual model in detecting metaphoric language, particularly in health-related contexts.

Next, the performance of a Logistic Regression classifier and a BERT classifier are evaluated based on the ability to classify metaphors in relation to whether they are used to describe an immunotherapy-related concept. The evaluation process is performed on the *Combined Relevance Corpus* (Sub-section 5.2.1, as well as on the *Annotated Immunotherapy Subset for Domain Relevance* (Sub-section 5.2.2. This two-fold evaluation aims to determine whether a model could identify relevance to a domain-specific concept.

Finally, the results of an end-to-end evaluation of the complete pipeline are presented in Sub-section 5.2.3. The predictions of both metaphor detection models (XLM-RoBERTa and BioMed-RoBERTa) are passed to the domain relevance classifier (BERT), so that a real-world scenario is simulated with the *Predicted Metaphor Corpus*, where annotations of metaphoricity and domain relevance are unavailable.

5.1 Metaphor Detection

5.1.1 VUA Metaphor Corpus

Table 5.1 presents the performance of XLM-RoBERTa, as reported by Wachowiak et al. (2022), alongside BioMed-RoBERTa trained as part of this thesis. Both models were evaluated on the VUA Metaphor Corpus under the same training conditions, with the pre-training domain being the only differentiating factor. The results show that both models perform very strongly on *Literal* tokens, with identical F1-scores of 0.97, and only small differences in precision and recall. However, on *Metaphorical* tokens, there are more notable differences: XLM-RoBERTa achieves slightly higher precision (0.82 vs. 0.80), while BioMed-RoBERTa achieves better recall (0.75 vs. 0.71), resulting in a

	Precision XLM-R / BioMed	**Recall** XLM-R / BioMed	**F1-Score** XLM-R / BioMed	**Support**
Literal	0.96 / 0.97	0.98 / 0.98	0.97 / 0.97	51,540
Metaphorical	0.82 / 0.80	0.71 / 0.75	0.76 / 0.78	6,819

Table 5.1: Classification report comparing XLM-RoBERTa (XLM-R) and BioMed-RoBERTa (BioMed) on metaphor detection on the VUA test set. Metrics are presented as XLM-R / BioMed.

small F1-score advantage for BioMed (0.78 vs. 0.76).

These results suggest that while domain-specific pre-training does not drastically outperform general-domain pre-training in metaphor detection, BioMed-RoBERTa shows slightly more balanced performance, especially in identifying a wider range of metaphors (higher recall). This could indicate a slight benefit from its exposure to scientific language, even in a general-domain metaphor task.

However, it remains unclear whether the performance of BioMed-RoBERTa is better due to its domain-specific pre-training on biomedical texts or differences in architecture that make it more efficient on metaphor detection. Additional experiments that isolate these factors would be required to identify the source of improvement.

5.1.2 Annotated Immunotherapy Subset

Table 5.2 presents a side-by-side comparison of XLM-RoBERTa and BioMed-RoBERTa in the metaphor detection task.

Both models perform very well on the *Literal* class, with high precision and recall values. XLM-RoBERTa achieves an F1-score of 0.911, while BioMed-RoBERTa achieves 0.910, although it slightly trades off recall for higher precision.

For the *Metaphorical* class, performance is lower in both models, which is typical for minority classes in imbalanced datasets. XLM-RoBERTa reaches a precision of 0.383 and recall of 0.360 (F1-score: 0.371), while BioMed-RoBERTa offers a slightly more balanced trade-off with precision 0.392 and higher recall at 0.400, leading to an F1-score of 0.396.

BioMed-RoBERTa achieves slightly higher macro and weighted average F1-scores (macro: 0.641 vs 0.653, weighted: 0.842 vs. 0.844), suggesting it may better handle the minority class without significantly sacrificing overall performance.

	Precision XLM-R / BioMed	**Recall** XLM-R / BioMed	**F1-Score** XLM-R / BioMed	**Support**
Literal	0.907 / 0.912	0.915 / 0.909	0.911 / 0.910	340
Metaphorical	0.383 / 0.392	0.360 / 0.400	0.371 / 0.396	50
Macro Avg	0.645 / 0.652	0.637 / 0.654	0.641 / 0.653	390
Weighted Avg	0.840 / 0.845	0.844 / 0.844	0.842 / 0.844	390

Table 5.2: Classification report comparing XLM-RoBERTa (XLM-R) and BioMed-RoBERTa (BioMed) on metaphor detection. Metrics are presented as XLM-R / BioMed.

5.1.3 Complete Immunotherapy Corpus

Model	Recall
XLM-RoBERTa	0.475
BioMed-RoBERTa	0.581

Table 5.3: Recall scores for XLM-RoBERTa and BioMed-RoBERTa on the ***Complete Immunotherapy Corpus***, which contains only *explicit* metaphor annotations. Since implicit metaphors are unlabeled, only recall can be calculated.

Testing on this dataset based on recall produced results of 0.475 (XLM-RoBERTa) and 0.581 (BioMed-RoBERTa), as shown in Table 5.3. The difference in recall between the two models reflects the influence of pre-training data on metaphor detection performance. XLM-RoBERTa, designed as a general-purpose model, lacks specialized exposure to medical language. In contrast, BioMed-RoBERTa is pre-trained on domain-specific corpora, thus may be more capable to determine whether a token is used metaphorically to describe a medical concept. This likely contributes to its better recall, indicating better sensitivity to metaphorical cues within medical contexts.

5.2 Domain Relevance Classification

5.2.1 Combined Relevance Corpus

A comparison of Logistic Regression and BERT on the domain relevance classification task shows differences in their performance. As shown in Table 5.4, Logistic Regression achieves a respectable performance, with a strong ability to predict irrelevant tokens (precision = 0.832, recall = 0.912, F1 = 0.870) and slightly lower performance on relevant ones (precision = 0.880, recall = 0.779, F1 = 0.827). These results suggest that the model is fairly capable at identifying immunotherapy-related metaphors in a mixed dataset.

As expected, BERT outperforms Logistic Regression across all metrics. It achieves more balanced precision and recall across both classes. For irrelevant instances, BERT achieves a precision of 0.975 and a recall of 0.928 (F1 = 0.951), while for relevant instances, the model reaches a precision of 0.918 and an even higher recall of 0.971 (F1 = 0.944). These results indicate that BERT not only maintains high precision in detecting relevant metaphors but also exhibits a superior ability to retrieve nearly all true positive cases, minimizing false negatives.

The macro and weighted average scores further demonstrate BERT's advantage. The macro F1-score reaches 0.947 and the weighted F1-score reaches 0.948, compared to Logistic Regression's macro and weighted F1-scores of 0.848 and 0.850, respectively. This performance gap reflects the deeper contextual understanding of BERT, which appears to be especially beneficial in capturing the linguistic cues that distinguish metaphorical expressions relevant to the domain. However, it should be noted that at this stage of evaluation, the noise introduced by the composition of the dataset (a combination of the VUA Metaphor Corpus and the Immunotherapy Metaphor Dataset) may contribute to the increased performance of both models. This can be explained by the fact that, in this round of testing, the model mainly needs to distinguish metaphors

that are present in medically relevant texts from metaphors that are found in texts of other domains.

	Precision LogReg / BERT	**Recall** LogReg / BERT	**F1-Score** LogReg / BERT	**Support**
Irrelevant	0.832 / 0.975	0.912 / 0.928	0.870 / 0.951	125
Relevant	0.880 / 0.918	0.779 / 0.971	0.827 / 0.944	104
Macro Avg	0.856 / 0.946	0.845 / 0.950	0.848 / 0.947	229
Weighted Avg	0.854 / 0.949	0.852 / 0.948	0.850 / 0.948	229

Table 5.4: Classification report for Logistic Regression and BERT models on domain relevance classification. Metrics are shown for the Combined Relevance Corpus.

	Precision LogReg / BERT	**Recall** LogReg / BERT	**F1-Score** LogReg / BERT	**Support**
Irrelevant	0.923 / 0.950	0.375 / 0.594	0.533 / 0.731	32
Relevant	0.375 / 0.480	0.923 / 0.923	0.533 / 0.632	13
Macro Avg	0.649 / 0.715	0.649 / 0.758	0.533 / 0.681	45
Weighted Avg	0.765 / 0.814	0.533 / 0.689	0.533 / 0.702	45

Table 5.5: Classification report for Logistic Regression and BERT models on immunotherapy relevance detection. Metrics are shown for the Annotated Immunotherapy Subset for Domain Relevance.

5.2.2 Annotated Immunotherapy Subset for Domain Relevance

The classification results comparing Logistic Regression and BERT on the *Annotated Immunotherapy Subset for Domain Relevance* (Table 5.5) reveal significant differences in their performance on identifying *Irrelevant* and *Relevant* classes compared to the previous evaluation layer. Both models achieved high precision for the *Irrelevant* class, with BERT reaching 0.950 compared to Logistic Regression's 0.923. BERT also demonstrated better recall (0.594) compared to Logistic Regression (0.375) for this class, suggesting that BERT was able to identify a greater proportion of truly irrelevant instances. This implies that BERT has a better capacity to capture most irrelevant samples without missing as many.

For the *Relevant* class, both models achieved high recall (0.923 for Logistic Regression and 0.923 for BERT), which means they successfully identified nearly all relevant instances in the dataset. However, BERT showed higher precision (0.480) than Logistic Regression (0.375) for the relevant category, indicating that BERT made fewer false positive errors in predicting relevance.

When considering the F1-score, BERT consistently outperformed Logistic Regression for both classes. The F1-score for the *Irrelevant* class was 0.731 for BERT versus 0.533 for Logistic Regression, and for the *Relevant* class, BERT scored 0.632 compared to Logistic Regression's 0.533. These differences demonstrate that BERT provides a more balanced and reliable classification in both categories.

The macro-average and weighted-average scores further show BERT's advantage. The macro-average treats both classes equally and shows improvements in precision (0.715 vs. 0.649), recall (0.758 vs. 0.649), and F1-score (0.681 vs. 0.533) for BERT, reflecting improved overall performance regardless of class size.

In general, BERT's better recall and balanced F1-scores indicate a stronger ability to generalize and capture both relevant and irrelevant instances effectively. However, it is also evident that performance has declined compared to the results obtained on the *Combined Relevance Corpus*. This difference can be attributed to the increased complexity of the task, as the models are required to classify metaphors as relevant or irrelevant in a very limited dataset that consists only of immunotherapy-related sentences. Whereas the original classification involved the identification of relevant metaphors in a diverse range of text types, that is, those of the VUA Metaphor Corpus, the current task focuses exclusively on the distinction of metaphors within immunotherapy-related texts. These texts contain a mix of both relevant and irrelevant metaphorical language, making the distinction more challenging. This increased difficulty likely originates from the more subtle semantic distinctions the models are required to navigate within a specific domain, which complicates the identification relevant metaphors and contributes to the performance decline.

5.2.3 Predicted Metaphor Corpus

Metric	**Precision**	**Recall**	**F1-score**	**Support**
End-to-End: Detection + Relevance				
Irrelevant	0.830	0.698	0.759	63
Relevant	0.208	0.357	0.263	14
Gold metaphors (filtered): 49		Predicted metaphors: 47; Missed: 30		

Table 5.6: Relevance classification and end-to-end results for XLM -RoBERTa metaphor detection classifier with BERT domain relevance classifier.

Metric	**Precision**	**Recall**	**F1-score**	**Support**
End-to-End: Detection + Relevance				
Irrelevant	0.739	0.523	0.613	65
Relevant	0.088	0.200	0.122	15
Gold metaphors (filtered): 49		Predicted metaphors: 50; Missed: 29		

Table 5.7: Relevance classification and end-to-end results for BioMed metaphor detection classifier with BERT domain relevance classifier.

When comparing the two pipelines (Tables 5.6 and 5.7), XLM-RoBERTa combined with the BERT relevance classifier achieved better performance than the BioMed-RoBERTa with the same relevance classifier.

When assessing the complete pipeline[1], which accounts for both detection and relevance classification (including missed metaphors as false negatives), the XLM+BERT

[1]Line charts for score comparison of the full pipeline can be found in Appendix B.

combination again outperformed BioMed+BERT across all metrics. The XLM-BERT pair achieved a higher precision (0.208 vs. 0.088) and recall (0.357 vs. 0.200) for *Relevant* metaphors, reaching an F1-score of 0.263 as opposed to 0.122. This indicates that the BioMed model failed to detect a larger proportion of actual metaphors and produced more *Irrelevant* results.

However, since this task is as much qualitative as it is quantitative, and the goal is not only to identify domain-specific metaphors but also to interpret and analyze them, a thorough error analysis is crucial to better understand these results. As discussed earlier in this thesis, metaphor detection faces challenges, such as inconsistencies in lexical resources, variation in annotation guidelines, and the subjective nature of figurative language. Consequently, these issues can propagate and affect downstream tasks as well. Therefore, it is important to evaluate not only whether the models achieve good quantitative performance but also whether the metaphors identified as relevant genuinely describe aspects of immunotherapy in ways that carry meaning compared to those found by a general-purpose model.

Chapter 6

Error Analysis

Following the quantitative results, an error analysis is performed on the *Annotated Immunotherapy Subset*, focusing first on the metaphor detection task. Here, common sources of confusion are studied, including metaphorical tokens that resemble literal expressions, and vice versa. The goal of this error analysis is to establish whether a model with domain-specific pre-training provides more meaningful results compared to a model with general pre-training.

The error analysis progresses in stages, beginning with a general examination of the most common errors in the metaphor detection task in Sub-section 6.1.1. This includes examples where literal and metaphorical uses are difficult to distinguish and cases where contextual information is insufficient for accurate disambiguation.

Next, a comparative analysis of the XLM-RoBERTa and BioMed-RoBERTa models is presented in Sub-section 6.1.2, highlighting differences in false positives and false negatives, and examining whether domain-specific pre-training produces more meaningful or domain-adapted predictions. Special attention is given to tokens that are uniquely misclassified by each model to establish whether the difference in pre-training results in finding tokens which describe immunotherapy-related concepts, even if their use is not metaphorical by the annotation standards used in this thesis.

Additional analysis is dedicated to the relevance classification stage in Section 6.2, where the goal is to identify which detected metaphors of the *Annotated Immunotherapy Subset for Domain Relevance* are genuinely relevant to the biomedical context.

In Section 6.3, the interaction between metaphor detection errors and relevance misclassification is unpacked by illustrating how errors in the first stage propagate to the second, specifically in the *Predicted Metaphor Corpus*.

Together, these findings offer a clearer picture of where current models succeed, where they fall short, and what kinds of improvements might help advance metaphor detection and relevance identification in domain-specific scientific texts.

6.1 Metaphor Detection

6.1.1 XLM-RoBERTa and BioMed-RoBERTa

The results presented in Table 5.2 suggest that while BioMed-Roberta and XLM-RoBERTa perform similarly, BioMed-RoBERTa shows slightly better results. To better interpret the scores analyzed in Chapter 5, an error analysis is performed on the individual predictions.

Table 6.1 presents False Positives (FP) and False Negatives (FN) for both XLM-RoBERTa and BioMed-RoBERTa, focusing only on tokens with parts of speech that represent content words (e.g., nouns, verbs, adjectives, adverbs), with irrelevant POS categories excluded, as they were automatically assigned a *Literal* label during the annotation process. The analysis was conducted over all token instances in the *Annotated Immunotherapy Subset*, which is small enough to allow for manual inspection of the full sentence context. Since metaphorical meaning often depends heavily on surrounding text, evaluating whether a token is metaphorical requires considering its complete context rather than the token in isolation. Bolded tokens are unique to the errors of one model and do not appear in the corresponding list for the other model.

In the context of science communication, both false positives and false negatives carry significant weight. False negatives, metaphors that are present but not detected, represent missed opportunities to identify figurative framing that could shape public understanding. False positives, meanwhile, risk misrepresenting literal statements as metaphorical, which can introduce noise to the results. Since metaphor analysis is often used to simplify how complex topics are communicated, both types of error directly affect the reliability and depth of downstream qualitative analysis.

To begin with, we can immediately notice the similarities in the two FP and FN columns. This similarity in misclassifications may signify that the models struggle with similar tokens and their surrounding context. Moreover, it demonstrates the difference in annotations compared to those of the VUA Metaphor Corpus, on which the models were originally trained, and specifically the difference in the dictionary referenced. For this thesis, the Longman Dictionary of Contemporary English Online was used, whereas the VUA Metaphor Corpus annotators relied mainly on the Macmillan English Dictionary for Advanced Learners.

XLM-RoBERTa exhibits several false positives on common verbs and nouns such as *testing*, *increase*, and *standard*, suggesting a tendency to over-predict metaphoricity for general-purpose verbs that often appear in abstract or figurative contexts in scientific texts.

BioMed-RoBERTa, while sharing many of these false positives with XLM-RoBERTa, also includes additional terms such as *CURE*, *breakthrough*, and *miracle*, which can be viewed as more content-specific predictions. This indicates that the model is more attuned to terms that carry figurative or evaluative connotations within biomedical discourse. While these predictions do not meet the metaphorical criteria defined by the annotation guidelines in this thesis, they suggest that the model is sensitive to terminology that may carry figurative meaning in wider scientific discourse.

6.1.2 Notable Examples

In the following examples, representative sentences have been selected from the *Annotated Immunotherapy Subset*. They include tokens marked in bold (false positives) and italics (false negatives), which are of particular interest for error analysis.
XLM-RoBERTa:

- A skin cancer wonder cure; **Survival** rates up to 94% in a **drug** trial. **Scientists testing** wonder drugs on Irish patients **believe** they could have found the cure for skin cancer.
- These drugs are associated with a specific *mechanism* of action that has profound implications for both immunology and inflammatory disease.

XLM-R		BioMed	
FP	**FN**	**FP**	**FN**
adverse	epidemic	**future**	epidemic
therapy	events	adverse	events
Cancer	checkpoint	therapy	checkpoint
CPIs	checkpoint	Cancer	landscape
mainly	landscape	CPIs	growing
lung	growing	lung	number
prevention	number	prevention	mechanism
environmental	mechanism	**CURE**	looks
Survival	looks	Survival	covering
%	covering	%	outlining
drug	outlining	drug	developments
SCIENTISTS	developments	SCIENTISTS	contribute
testing	contribute	here	rise
believe	rise	increase	poor
here	poor	survival	examines
increase	examines	**immune**	developments
survival	developments	taking	shows
taking	shows	part	near
part	near	phase	harnessing
phase	harnessing	**breakthrough**	fight
starting	fight	tantalising	uncovered
tantalising	uncovered	bring	further
bring	further	About	dramatic
About	dramatic	**spreads**	stages
next	stages	**miracle**	**aggressive**
have	**show**	next	**victims**
standard	Highest	**aspiration**	Highest
DEREK	**released**	have	**highest**
POWER	branded	standard	branded
	start	**breakthrough**	care
	foundational		
	care		

Table 6.1: False Positives (FP) and False Negatives (FN) for XLM-R and BioMed models on metaphor detection. Tokens in bold appear only in that model's list and not in the corresponding list of the other model.

- Researchers are pioneering methods of *harnessing* the body's own immune system to *fight* cancer.
- Survival of lung cancer is still *poor*, and new treatments are urgently needed.
- There is the **tantalising** possibility of a cure with immunotherapy.

BioMed-RoBERTa:

- A skin cancer wonder **cure**; **Survival** rates up to 94% in a **drug** trial. **Scientists testing** wonder drugs on Irish patients believe they could have found the cure for skin cancer.
- Global biopharma company Bristol-Myers Squibb is producing the two **miracle** drugs being used in the trial - nivolumab and ipilimumab, *branded* as Yervob on the market.

- He said: 'This really is an amazing **breakthrough**.'

False Positives (FP)

XLM-RoBERTa exhibits a tendency to overpredict metaphoricity for domain-specific terms. For example, *Survival*, *drug*, *SCIENTISTS*, *testing*, *believe*, and *tantalising* are all marked as FP, indicating that the model incorrectly labels these literal biomedical terms as metaphorical. This suggests that the model lacks sufficient domain-specific background to distinguish technical quantitative vocabulary and discourse markers from figurative expressions. The verb *believe* is a standard reporting verb in scientific texts and is used literally in context.

BioMed-RoBERTa displays a similar pattern of false positives, with similar tokens incorrectly labeled metaphorical. However, it also positively predicts words that describe immunotherapy-related concepts, particularly in article titles and spoken examples, such as *miracle drug* and *amazing breakthrough*. Both expressions describe a drug that is said to *cure* cancer, a token also predicted as metaphorical by the biomedical model. This overextension to literal medical terms further illustrates the difficulty of the model in discriminating rhetorical emphasis from genuine metaphorical usage. For instance, *CURE* in the given headline context functions as a literal descriptor rather than an abstract figurative construct.

False Negatives (FN)

Both models exhibit false negatives predominantly for verbs and mechanisms that instantiate well-known conceptual metaphors, particularly the framing of disease treatment as warfare. In the XLM-RoBERTa outputs, tokens such as *mechanism*, *harnessing*, *fight*, and *poor* are incorrectly classified as literal, despite clearly suggesting metaphorical usage (e.g., the immune system 'fighting' cancer, or disease outcomes conceptualized in evaluative terms such as 'poor survival'). Although explicit examples for BioMed-RoBERTa do not include these cases, the error tables confirm that the same metaphorical tokens are missed by this model as well, implying an underlying common limitation in capturing such metaphors.

Model Differences

While both models display similar patterns, and despite the fact that the dataset in question is fairly limited, some differences can be observed. XLM-RoBERTa tends to produce more false positives on functional and reporting verbs (*believe*, *testing*). In contrast, BioMed-RoBERTa more frequently misclassifies literal expressions used hyperbolically, such as *CURE* and *breakthrough*, suggesting a sensitivity to rhetorical emphasis typical in journalism, but less so in technical research contexts.

6.2 Domain Relevance Classification

6.2.1 Logistic Regression and BERT

Table 6.2 presents the false positives (FP) and false negatives (FN) produced by the baseline Logistic Regression classifier and the BERT-based model on the domain relevance classification task tested on the gold data of the *Annotated Immunotherapy Corpus for Domain Relevance*. In this setting, both models operate on given metaphors

LogReg		BERT	
FP	**FN**	**FP**	**FN**
events	**mechanism**	events	**landscape**
growing		developments	
number		contribute	
looks		rise	
covering		developments	
outlining		shows	
developments		**uncovered**	
contribute		**dramatic**	
rise		found	
examines		stages	
developments		show	
shows		reach	
near		**thing**	
found		**thing**	
stages		**branded**	
show		**market**	
reach		start	
leading		head	
released		reinforce	
start			
head			
reinforce			

Table 6.2: False Positives (FP) and False Negatives (FN) for Logistic Regression and BERT on immunotherapy-specific metaphor detection. Tokens in bold appear only in that model's list and not in the corresponding list of the other model.

and are provided with context. The models then must decide whether each instance belongs specifically to the target domain.

Both models exhibit false positives for general biomedical or research-related expressions such as *events*, *developments*, and *stages*. This suggests a tendency to overextend domain relevance labels to generic scientific language that is not specific to immunotherapy. Logistic Regression, in particular, misclassifies terms such as *looks*, *covering*, and *outlining* as domain-relevant, which can easily be explained by the superficial features of the directed dependency path to an immunotherapy related word and the immunotherapy related word itself.

In terms of false negatives, the models fail to label certain domain-relevant terms that are crucial in the context in which they are used, such as *mechanism* and *landscape*. In their respective contexts, they are immunotherapy-specific, as they frame how immunotherapy functions and how the field evolves. However, as standalone terms, they can be overlooked due to their distance from the field of immunotherapy.

In general, these patterns demonstrate that while BERT outperforms a simple Logistic Regression baseline in capturing contextual clues for domain relevance, both models remain limited by the subtle boundary between general biomedical language and immunotherapy-specific content. This can also be explained by the limited size of relevant training data, as well as the nature of this data: BERT may learn to associate specific lexical items or sentence patterns with relevance labels, rather than generalizing to broader conceptual cues. To improve performance, the model would require more diverse training examples of metaphorical language tied to immunotherapy, ideally cov-

ering a range of expression styles and contexts, so it can learn the abstract concept of domain relevance rather than overfitting to specific patterns.

6.2.2 Notable Examples

In the following examples, representative sentences have been selected from the *Annotated Immunotherapy Subset for Domain Relevance*. They include tokens marked in bold (false positives) and italics (false negatives).

Logistic Regression:

- Nevertheless, only 20% of smokers develop lung cancer and while prevention is important, environmental factors are expected to **contribute** to the predicted rise in the incidence of lung cancer in the next 25 years.
- This review **examines** potential new therapeutic developments which have arisen from a greater understanding of the molecular and cellular biology of lung cancers.
- These drugs are associated with a specific *mechanism* of action that has profound implications for both immunology and inflammatory disease.

BERT:

- And they have **uncovered** further **dramatic** results in a worldwide clinical trial involving 7,000 people.
- Global biopharma company Bristol-Myers Squibb is producing the two miracle drugs being used in the trial — nivolumab and ipilimumab, branded as Yervob on the **market**.
- Cancer immune therapy with checkpoint inhibitors (CPIs) has changed the *landscape* of treatment for a growing number of indications.

Logistic Regression:

The Logistic Regression classifier mislabels tokens such as *contribute* and *examines* as domain-relevant. This behavior is to be expected due to the nature of its feature set, which is limited to syntactic and surface lexical cues. For instance, *contribute* appears in a causal phrase describing environmental factors influencing lung cancer incidence, which provides the classifier with a short dependency path to the term 'cancer'. The model is very likely to overfit on co-occurrence with high-frequency domain terms (e.g., *lung cancer*) without the capacity to determine whether simple co-occurrence suffices for a positive classification. Similarly, *examines* occurs in a sentence reporting potential new therapeutic developments in lung cancer; here, the use of the verb in reporting scientific results and its proximity to 'cancer' appears sufficient for the classifier, despite its generic meaning.

In contrast, logistic regression fails to flag *mechanism* as domain-relevant, even though the sentence explicitly connects the mechanism of action of the drug to the immunological and inflammatory processes central to immunotherapy. This omission reflects the limited depth of the model: it lacks any means to generalize the semantics of *mechanism of action* as a domain-defining concept without explicit keyword matches in the feature set.

	XLM-RoBERTa	BioMed-RoBERTa
True Positives (TP)	5	3
True Negatives (TN)	32	19
False Positives (FP)	3	16
False Negatives (FN)	9	11
Total gold domain-relevant metaphors	14	14

Table 6.3: True and False Positives and Negatives occurring from the end-to-end pipeline.

BERT:

BERT shows similar patterns, but with improved sensitivity to context in general. In these examples, BERT displays false positives for adjectives and verbs such as *uncovered* and *dramatic*, which often coincide with newsworthy medical findings, but do not necessarily indicate immunotherapy-specific relevance. Similarly, a generic commercial noun such as *market* is incorrectly flagged, simply because it appears alongside immunotherapy pharmaceutical products in the same sentence.

On the false negative side, BERT overlooks *landscape* in the phrase *changed the landscape of treatment for a growing number of indications*. In this context, the landscape change is directly attributed to checkpoint inhibitor therapy, which is a key element of immunotherapy, but the model fails to link this abstract frame to the domain label, possibly due to the broader use of the token in other contexts and the limited amount of training data.

Model Differences

In general, both models are characterized by shared challenges: differentiating between general language, in biomedical or other texts, and the narrower immunotherapy domain when common tokens appear alongside domain-relevant cues. Logistic Regression's dependency on direct lexical features produces more systematic false positives, while BERT's embeddings allow a more precise contextual interpretation but still struggle with abstract metaphors or domain-specific figurative constructs that do not co-occur frequently in the training data.

6.3 Error Propagation in the Full Pipeline

Table 6.3 summarizes the performance of the complete pipeline, combining metaphor detection and domain relevance classification. This evaluation reflects the realistic use case: the system must first detect whether a token is used metaphorically and then determine if it is specific to the immunotherapy domain.

A core source of false negatives in the pipeline stems directly from the initial metaphor detection task. If a metaphorical expression is not recognized in the first stage, it never reaches the domain classifier, regardless of its context. This issue primarily affects domain-specific figurative language that the models interpret literally. For example, terms such as *fight* and *harnessing*, as shown in Tables 6.4 and 6.5, which often represent processes in immunotherapy-related discourse, are overlooked by both models in this context.

XLM-R+BERT		
FP	**FN**	**TP**
reach	epidemic	**checkpoint**
market	**checkpoint**	**aggressive**
head	landscape	forms
	harnessing	**victims**
	mechanism	harnessing
	fight	
	poor	
	foundational	
	care	

Table 6.4: Error analysis for XLM-RoBERTa+BERT. Tokens in bold appear only in that model's list and not in the corresponding list of the other model.

BioMed-RoBERTa+BERT		
FP	**FN**	**TP**
found	epidemic	forms
up	**checkpoint**	harnessing
show	checkpoint	**foundational**
reach	landscape	
marks	harnessing	
added	mechanism	
thing	fight	
thing	poor	
leading	**aggressive**	
has	**victims**	
lowest	care	
released		
market		
start		
head		
reinforce		

Table 6.5: Error analysis for BioMed-RoBERTa+BERT. Tokens in bold appear only in that model's list and not in the corresponding list of the other model.

Even when metaphors are correctly identified, the domain relevance step remains imperfect. Both models misclassify generic scientific metaphors as immunotherapy-specific or fail to flag subtle immunotherapy-specific metaphors as relevant. For example, in both pipelines, *reach* in 'a cure could be within reach' is labeled as immunotherapy-specific, although this usage is general and not uniquely related to the domain. In contrast, figurative terms such as *landscape* and *epidemic*, which describe the shifting research terrain or the spread of immune-related adverse events, are often misclassified as irrelevant when, in fact, they frame key domain concepts.

The two pipelines show complementary tendencies. The XLM-RoBERTa+BERT pair takes a cautious approach. When false positives are given, they tend to be tokens that come with rich contexts. The BioMed-RoBERTa+BERT pair, on the other hand, is noticeably more permissive. It tends to over-label common tokens such as *found*, *thing*, *up*, or *show* as both metaphorical and domain relevant, which pushes up its false positive rate.

Together, these patterns reveal how errors accumulate across the pipeline. If the

metaphor detection model misses a metaphorical token, the domain classifier never gets a chance to evaluate it. Therefore, relevant cases remain unseen. At the same time, when the metaphor detection stage works correctly, the domain classifier can still get it wrong, classifying generic metaphors as immunotherapy-specific. This double source of error limits both precision and recall in real-world scenarios. Especially tricky cases, such as *landscape* as used in 'changing treatment landscape', require a deeper sense of context and more sophisticated disambiguation than the current models can reliably provide.

Chapter 7

Discussion

This thesis has presented an investigation of metaphor detection within the specialized domain of immunotherapy-related texts. The experiments combined the evaluation of an existing metaphor detection model trained on general metaphor corpora and its comparison to a domain-adapted RoBERTa model pre-trained on biomedical data, alongside an attempt to create a BERT classifier that can detect domain-relevant immunotherapy metaphors. The results were quite interesting, as they revealed significant limitations and challenges in this research area.

First, the study revealed that, while XLM-RoBERTa by Wachowiak et al. (2022), fine-tuned on general-domain data, does reasonably well in identifying metaphors, its ability to capture metaphors specific to immunotherapy is limited. This model tends to overgeneralize, labeling metaphorical expressions that may not be relevant to the immunotherapy context or missing domain-specific figurative language crucial for accurate interpretation. The RoBERTa model that was pre-trained on biomedical data showed some improvements, suggesting that domain-specific pre-training can provide valuable contextual information. However, when combined with the BERT domain relevance classifier, performance was insufficient.

One of the most important reasons for these mixed results may lie in the lack of fully annotated data for immunotherapy metaphors. Although this thesis benefited from the annotated dataset of Bos et al. (2025), as well as the complete annotations of five texts related to immunotherapy, the size and scope of the data remained limited. Due to this issue, the model's ability to learn domain-specific patterns was insufficient. Furthermore, metaphor annotation requires a number of annotators who can dedicate time and effort to the project, which was only partially feasible given the time constraints and available resources.

Another practical limitation was time. Metaphor detection is a computationally and intellectually intensive task, demanding iterative experimentation and exhaustive error analysis to produce reliable results. This study laid an important foundation by fine-tuning BioMed-RoBERTa as a domain adaptation strategy, but was inevitably limited by the time frame of a Master-level thesis.

Despite these challenges, the findings point to several promising directions for future work that could address current research gaps. Fine-tuning with larger quantities of domain-relevant data may improve performance. The immunotherapy corpus could be expanded through additional annotation efforts, perhaps by involving more annotators or utilizing domain adaptation techniques such as pseudo-labeling (Jia et al., 2019). Within this semi-supervised setup, the general-domain model generates pseudo-

labels for additional unlabeled immunotherapy texts. By adding these pseudo-labeled instances to the training data, the model iteratively refines its ability to recognize metaphors specific to the immunotherapy domain.

Domain adaptation strategies, such as continued pre-training on immunotherapy-specific corpora, could improve model generalization within this domain. In addition, exploring other architectures beyond BERT for domain relevance classification, including generative models, could be a way to achieve more targeted results in the science communication domain.

Lastly, another extension would be to expand beyond immunotherapy to other medical subdomains with BioMed-RoBERTa, or to shift to different domains altogether with language models developed with entirely different pre-training data.

Chapter 8

Conclusion

This thesis aimed to evaluate whether existing metaphor detection models adequately identify metaphors in immunotherapy-related texts and whether metaphor identification can be improved through domain-specific pre-training. To this end, two main tasks were addressed: metaphor detection at a token level and domain relevance classification. The first involved comparing a general-domain XLM-RoBERTa model with a biomedical domain-adapted BioMed-RoBERTa to assess their ability to identify metaphorical expressions. The second focused on determining whether the metaphors detected were specifically relevant to immunotherapy. Both tasks were carried out using data from the VUA Metaphor Corpus and a domain-specific dataset compiled from scientific publications and news articles on immunotherapy. These experiments provided an assessment of the impact of biomedical pre-training on metaphor detection performance in immunotherapy science communication.

In conclusion, this research confirms that metaphor detection in specialized medical discourse is a complex yet important challenge. General-purpose models serve as a baseline but are likely to underperform when tasked with identifying metaphorical expressions that hold important semantic and communicative functions within specialized domains. The modest improvements observed with domain-specific pre-training show the need for more extended data resources. Although this thesis encountered practical limitations, including restricted data availability and time constraints, it successfully demonstrates the feasibility of this approach and provides directions for future research. With further development, domain-specific metaphor detection models could substantially improve NLP applications in science communication, supporting clearer interpretation and patient understanding.

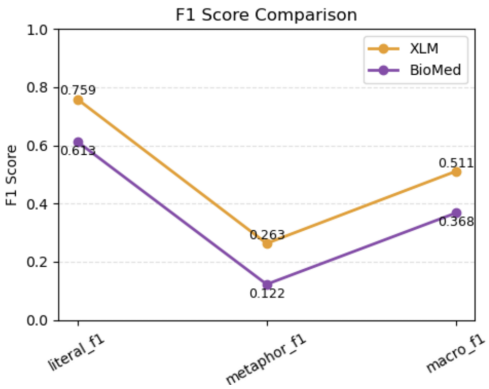
Appendix A

Immunotherapy-Related Terms

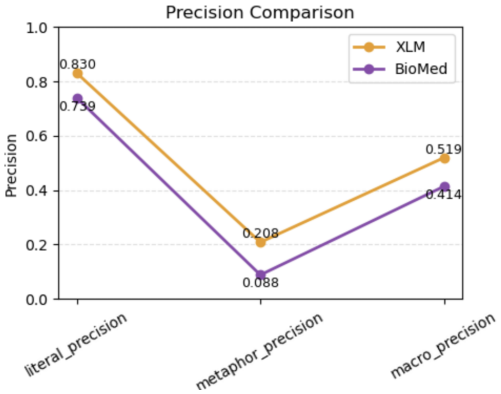
- adoptive
- adenocarcinoma
- adenovirus
- allergy
- antibody
- antigen
- antigen-presenting
- autoimmune
- b-cell
- biomarker
- biopsy
- bone marrow
- cancer
- car-t
- checkpoint
- checkpoint inhibitor
- chemotherapy
- clinical trial
- ctla-4
- cytokine
- dendritic
- effector
- endocrine
- fusion-protein
- genetic mutation
- immune
- immune activation
- immune checkpoint
- immune modulator
- immune response
- immune system
- immune-related
- immuno-oncology
- immunogenic
- immunohistochemistry
- immunomodulation
- immunosuppression
- immunotherapy
- inflammation
- insulin
- insulinoma
- interleukin
- lymph
- lymph node
- lymphocyte
- macrophage
- melanoma
- melanocytes
- monoclonal
- neoplasm
- oncology
- pd-1
- pd-l1
- perforin
- radiation
- recurrence
- regulatory
- remission
- stem cell
- t-cell
- therapy
- tolerance
- tumor
- tumour
- t-cell
- vaccine
- white blood cell

Appendix B

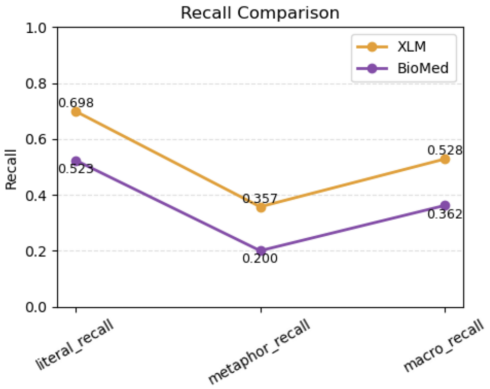
Full Pipeline Score Charts



(a) F1-score



(b) Precision



(c) Recall

Figure B.1: Comparison of F1-score, Precision, and Recall for the full pipeline.

Bibliography

B. Beigman Klebanov, C. W. B. Leong, and M. Flor. A corpus of non-native written English annotated for metaphor. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 86–91, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2014. URL https://aclanthology.org/N18-2014/.

N. Bos, L. Vandeberg, A. Oerlemans, M. Hulscher, and W. G. Reijnierse. 'mapping' knowledge dissemination: What metaphors reveal about the conceptualisation of immunotherapy in scientific and journalistic communication. *European Journal of Health Communication*, 6(1):60–82, 2025. doi: 10.47368/ejhc.2025.103.

Cancer Research Institute. Immunoglossary, 2024. URL https://www.cancerresearch.org/immunoglossary. Last accessed: 2025-06-25.

X. Chen, C. W. B. Leong, M. Flor, and B. Beigman Klebanov. Go figure! multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. In B. B. Klebanov, E. Shutova, P. Lichtenstein, S. Muresan, C. Wee, A. Feldman, and D. Ghosh, editors, *Proceedings of the Second Workshop on Figurative Language Processing*, pages 235–243, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.figlang-1.32. URL https://aclanthology.org/2020.figlang-1.32/.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

P. Crisp, R. Gibbs, A. Deignan, G. Low, G. Steen, L. Cameron, E. Semino, J. Grady, A. Cienki, Z. Kövecses, and T. Group. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22, 01 2007. doi: 10.1207/s15327868ms2201_1.

M. David and T. Matlock. Cross-linguistic automated detection of metaphors for poverty and cancer. *Language and Cognition*, 10(3):485–513, 2018.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and

T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

E. Dodge, J. Hong, and E. Stickles. Metanet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49, Denver, Colorado, 2015. Association for Computational Linguistics. URL https://aclanthology.org/W15-1405.pdf.

Y. Gu, R. Tinn, H. Cheng, M. Lucas, T. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021.

S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don't stop pretraining: Adapt language models to domains and tasks, 2020. URL https://arxiv.org/abs/2004.10964.

E. D. Gutiérrez, P. R. Corlett, C. M. Corcoran, and G. A. Cecchi. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 93–104. Association for Computational Linguistics, 2013.

C. Jia, X. Liang, and Y. Zhang. Cross-domain ner using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2464–2474. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1236. URL https://aclanthology.org/P19-1236.pdf.

D. Jurafsky and J. H. Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models, 2025. URL https://web.stanford.edu/~jurafsky/slp3. Chapter 5: Logistic Regression. Online manuscript released January 12, 2025.

C. W. Leong, B. B. Klebanov, E. Shutova, P. Lichtenstein, S. Bowman, Y. Zhang, and R. Mihalcea. A report on the VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana, July 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-0907.

C. W. Leong, B. B. Klebanov, E. Shutova, P. Lichtenstein, I. Dagan, and C.-k. Lo. A report on the 2020 vu and toefl metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.figlang-1.3v2.pdf.

Z. Li, Y. Wang, F. Wang, and Z. Fu. Mam: A metaphor-based approach for mental illness detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3932–3942. Association for Computational Linguistics, 2019.

Longman. Longman dictionary of contemporary english online, 2014. URL `https://www.ldoceonline.com/`. Accessed: 2025-06-27.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

E. Panicheva, E. Zheltonozhskii, A. Volkov, E. Pivovarova, Y. Cherepanova, and V. Malykh. Towards automatic conceptual metaphor detection for psychological tasks. *Personality and Individual Differences*, 199:111877, 2023.

S. Piao, F. Bianchi, C. Dayrell, A. D'Egidio, and P. Rayson. Development of the multilingual semantic annotation system. pages 1268–1274, 01 2015. doi: 10.3115/v1/N15-1137.

G. J. Steen, A. G. Dorst, J. B. Herrmann, A. A. Kaal, T. Krennmayr, and T. Pasma. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins, Amsterdam, 2010.

C. Su, F. Fukumoto, X. Huang, J. Li, R. Wang, and Z. Chen. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In B. B. Klebanov, E. Shutova, P. Lichtenstein, S. Muresan, C. Wee, A. Feldman, and D. Ghosh, editors, *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.figlang-1.4. URL `https://aclanthology.org/2020.figlang-1.4/`.

T. T. T. Truc. Analyzing war metaphors in the context of the covid-19: A critical metaphor analysis. *Journal of Language Teaching and Research*, 15(6):1242–1249, 2024. doi: 10.17507/jltr.1506.21. URL `https://jltr.academypublication.com/index.php/jltr/article/view/8950`.

Y. Tsvetkov, L. Boytsov, A. Gershman, E. Nyberg, and C. Dyer. Metaphor detection with cross-lingual model transfer. volume 1, 01 2014. doi: 10.3115/v1/P14-1024.

P. D. Turney, Y. Neuman, D. Assaf, and Y. Cohen. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690. Association for Computational Linguistics, 2011. URL `https://aclanthology.org/D11-1063`.

L. Wachowiak, D. Gromann, and C. Xu. Drum up SUPPORT: Systematic analysis of image-schematic conceptual metaphors. In D. Ghosh, B. Beigman Klebanov, S. Muresan, A. Feldman, S. Poria, and T. Chakrabarty, editors, *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 44–53, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.flp-1.7. URL `https://aclanthology.org/2022.flp-1.7/`.

Y. Wang and H. Habil. Critical metaphor analysis of climate change in cop28 speeches: An ecolinguistic perspective. *World Journal of English Language*, 14(5):49–61, 2024. doi: 10.5430/wjel.v14n5p49. URL `https://www.sciedupress.com/journal/index.php/wjel/article/view/25277`.

D. Wright and I. Augenstein. Transformer based multi-source domain adaptation. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pages 7963–7974, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.639. URL https://aclanthology.org/2020.emnlp-main.639/.