

Language Models for the Detection of Manipulative Discourse and Disinformation in Text

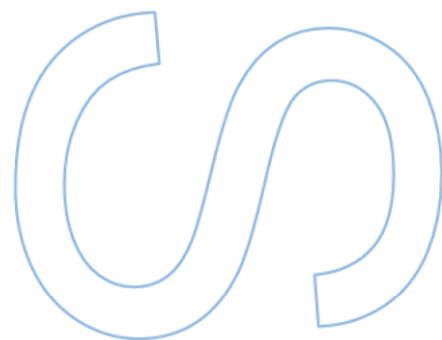
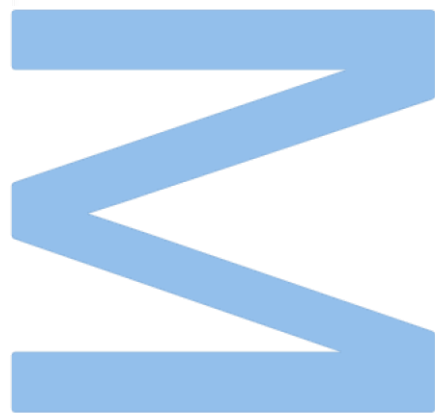
André Campos Vieira

Master in Data Science

Department of Computer Science

Faculty of Sciences of the University of Porto

2026



Language Models for the Detection of Manipulative Discourse and Disinformation in Text

André Campos Vieira

Dissertation carried out as part of the Master in Data Science
Department of Computer Science
2026

Supervisor

Nuno Guimarães, Invited Assistant Professor, Faculty of Sciences
of the University of Porto

Co-supervisor

Alípio Jorge, Full Professor, Faculty of Sciences of the University of
Porto



Acknowledgements

I would like to express my sincere gratitude to my supervisors for their guidance, support, and valuable feedback throughout the development of this dissertation. Their insights and encouragement were essential to the completion of this work.

I also thank my colleagues and friends for their support and for the stimulating discussions that helped shape many of the ideas explored in this research.

Finally, I would like to thank my family for their continuous encouragement and support throughout my academic journey.

Resumo

O discurso manipulativo nos meios de comunicação digitais nem sempre depende da falsidade. Grande parte opera através de estratégias retóricas identificáveis, como linguagem emocionalmente carregada, apelos ao medo ou a apresentação de questões complexas como escolhas binárias, que podem coexistir com conteúdo factualmente correto. Detetar estas estratégias de forma automática e localizar onde ocorrem num texto é um problema que permanece largamente inexplorado para línguas que não o inglês.

A presente dissertação aborda a deteção de 23 técnicas de persuasão em artigos informativos em português europeu, num cenário em que não existem dados de treino anotados em português. Enquadrado na CLEF-2024 CheckThat! Lab Task 3, o sistema aprende a partir de dados em seis outras línguas e transfere esse conhecimento para o português através de representações interlinguísticas e dados traduzidos automaticamente.

A abordagem proposta substitui a camada de saída single-label convencional por uma arquitetura sigmoid multi-label, permitindo detetar múltiplas técnicas co-ocorrentes no mesmo fragmento textual. Uma análise de proximidade baseada em embeddings revelou que o italiano é a língua-fonte mais próxima do português, apesar de ter um corpus menor do que o inglês. Com base neste resultado, foi desenvolvido um pipeline em duas fases: a primeira treina com as seis línguas-fonte, enquanto a segunda ajusta o modelo exclusivamente com dados traduzidos do italiano.

O sistema final alcança um F1-micro de 0.134, superando o vencedor da competição e o sistema dos organizadores. Os resultados sugerem que a escolha da arquitetura de saída tem mais impacto do que a escolha do encoder, e que a proximidade linguística supera o volume de dados como preditor da qualidade da transferência interlinguística.

Palavras-chave: Deteção de Desinformação, Discurso Manipulativo, Técnicas de Propaganda, Processamento de Linguagem Natural, Modelos Transformer, Transferência Interlinguística, Classificação ao Nível do Span

Abstract

Manipulative discourse in online media does not always depend on falsehood. Much of it operates through identifiable rhetorical strategies, such as emotionally charged language, appeals to fear, or the framing of complex issues as binary choices, that can coexist with factually accurate content. Detecting these strategies automatically and locating where they appear in a text is a problem that remains largely unexplored for non-English languages.

This dissertation tackles the detection of 23 persuasion techniques in European Portuguese news articles, in a setting where no annotated Portuguese training data is available. Working within the CLEF-2024 CheckThat! Lab Task 3 framework, the system learns entirely from data in six other languages and transfers that knowledge to Portuguese through cross-lingual representations and machine-translated data.

The proposed approach replaces the standard single-label output head with a multi-label sigmoid architecture, allowing the model to detect multiple co-occurring techniques within the same text fragment. An embedding-based proximity analysis revealed that Italian is the closest source language to Portuguese, despite having a smaller training corpus than English. Based on this finding, a two-stage training pipeline was developed: the first stage trains on all six source languages, while the second fine-tunes exclusively on Italian-translated data.

The resulting system achieves an F1-micro of 0.134, outperforming both the competition winner and the organisers' own post-competition system. The findings suggest that the choice of output architecture matters more than the choice of encoder, and that linguistic proximity outweighs data volume as a predictor of cross-lingual transfer quality.

Keywords: Disinformation Detection, Manipulation Discourse, Propaganda Techniques, Natural Language Processing, Transformer Models, Cross-linguistic Transfer, Span-Level Classification

Table of Contents

List of Figures.....	vii
List of Tables.....	viii
1. Introduction.....	1
1.1. Motivation.....	2
1.2. Objectives	3
1.3. Contributions.....	4
1.4. Document Structure.....	5
2. State of the Art	6
2.1. Manipulative Discourse and Persuasion Techniques	6
2.1.1. Propaganda, Disinformation and Misinformation	7
2.1.2. Rhetorical Techniques and Span-Level Granularity.....	8
2.1.3. Linguistic Challenges in Automatic Identification	9
2.2. Propaganda Detection Datasets and Task Description	10
2.2.1. The Propaganda Techniques Corpus (PTC).....	10
2.2.2. Evaluation Metrics.....	11
2.2.3. Class Imbalance Problem.....	12
2.3. Transformer-Based Models for Propaganda Detection	12
2.3.1. From Traditional Machine Learning to Deep Learning	12
2.3.2. Bidirectional Encoder Representations from Transformers (BERT)	14
2.3.3. Fine-Tuning Strategies for Span-Level Detection	15
2.3.4. Generative Language Models vs Encoder-Based Approaches	15
2.4. Multilingual and Cross-Lingual Propaganda Detection	16
2.4.1. Challenges in Multilingual NLP.....	16
2.4.2. Multilingual Transformer Models.....	17
2.4.3. Transfer Learning Across Languages	18
2.4.4. Low-Resource Language Scenarios	18
2.5. Research Gaps and Motivation for This Work	19
2.5.1. Limitations of Current Span-Level Detection Systems.....	19
2.5.2. Challenges in Multilingual Generalisation.....	20
2.5.3. Motivation and Objectives of This Thesis	20
3. Methods and Data	22
3.1. Problem Formulation	23
3.2. Dataset.....	23
3.2.1. CLEF-2024 CheckThat! Lab Task 3.....	23

3.2.2. Persuasion Technique Taxonomy	24
3.2.3. Document Collection	26
3.2.4. Annotation Process and Quality	26
3.2.5. Training and Development Data	27
3.2.6. Portuguese Test Dataset	27
3.3. Exploratory Data Analysis	28
3.3.1. Data Preparation	28
3.3.2. Span Length Analysis	28
3.3.3. Technique Frequency Distributions	29
3.3.4. Topic Distribution Across Languages	32
3.3.5. Class Imbalance	33
3.4. Embedding-Based Language Proximity Analysis	33
3.4.1. Methodology	33
3.5. Baseline Systems	34
3.5.1. mBERT Baseline	34
3.5.2. XLM-R Baseline	35
3.5.3. Generative Baseline: Gemini 2.5 Flash	35
3.5.4. Limitations of the BIO Formulation	35
3.6. Proposed Architecture	35
3.6.1. Encoder: XLM-RoBERTa	36
3.6.2. Multi-Label Sigmoid Output Head	36
3.6.3. Per-Output Positive Weighting	37
3.6.4. Sliding Window	37
3.7. Training Pipeline	37
3.7.1. Machine-Translated Data	38
3.7.2. Model 1	38
3.7.3. Model 2: Final System	39
3.7.4. Stage 2 Language Selection Experiments	40
3.8. Inference and Post-Processing	40
3.9. Evaluation Protocol	40
4. Results and Discussion	42
4.1. Embedding-Based Language Proximity Analysis	42
4.2. Training Dynamics	45
4.2.1. Baseline Training Curves	46
4.2.2. Proposed Model Training Curves	47
4.3. Stage 2 Language Selection	47
4.4. Overall System Comparison	49
4.5. Per-Technique Analysis	50

4.6. Threshold Analysis	52
4.7. Ablation Analysis	53
4.8. Discussion.....	54
4.8.1. Contribution of Each Design Decision	54
5. Conclusions and Future Work.....	56
5.1. Summary of the Work	56
5.2. Main Findings.....	57
5.3. Limitations	58
5.4. Future Work	59
5.5. Closing Remarks	61
Bibliography	62
A. Conference Posters.....	65

List of Figures

3.1. Data exploration and selection pipeline. The CLEF-2024 dataset is progressively filtered and transformed to determine the final training configuration.	22
3.2. Top 15 most frequent techniques in the English training set. Loaded Language and Name Calling together account for over 50% of all spans.....	30
3.3. Top 15 most frequent techniques in the Portuguese gold test set. Questioning_the_Reputation dominates, a category entirely absent from English training data.....	30
3.4. Percentage distribution of techniques in the English training set.	31
3.5. Percentage distribution of techniques in the Portuguese gold test set.	31
3.6. Heatmap of topic distribution (%) across language partitions.	32
3.7. Grouped bar chart of topic distribution by language.	33
3.8. Architecture of Model 2. Indigo boxes represent the key components: the multilingual encoder and the multi-label sigmoid head.	36
4.1. Distribution of training languages among the top-20 nearest neighbours for each of the 104 Portuguese gold documents. Italian dominates at 39.5% despite not being the largest dataset.	43
4.2. Number of Portuguese gold documents for which each training language is the dominant nearest neighbour (n=104).	43
4.3. Per-technique language distribution of nearest training neighbours (%). Italian dominates consistently across all 23 techniques, but notable variation in secondary languages motivates including all source languages in Stage 1.	44
4.4. UMAP 2D projection of document embeddings. Portuguese gold documents (stars) cluster predominantly with Italian and French. English forms a distinct and more distant cluster.	45
4.5. Training loss curves for the BIO baseline systems. Left: Stage 1 on English data. Right: Stage 2 fine-tuning on IT→PT.	46
4.6. Training loss curves for Model 1 and Model 2. Model 2 Stage 1 converges over a longer trajectory due to the larger multi-source dataset. Both Stage 2 runs are brief and stable.	47
4.7. Training loss curves for the three Stage 2 configurations, all initialised from the same Model 2 Stage 1 checkpoint. The IT→PT + FR→PT + RU→PT run requires substantially more steps and converges to a higher final loss, consistent with its lower F1 score.	49
4.8. F1-micro across the threshold range for all four systems. Model 2 and Model 1 improve monotonically up to 0.70. Both BIO baselines peak early and degrade sharply.	53
A.1. Poster presented at the FCUP Master's Day, 2026.	65
A.2. Poster presented at IJUP, University of Porto, 2026.	66

List of Tables

3.1. Dataset statistics across all splits in the CLEF-2024 CheckThat! Lab Task 3 [6]. Portuguese, the target language of this work, appears only in the test split.	24
3.2. Two-tier persuasion technique taxonomy used in the CLEF-2024 CheckThat! Lab Task 3 [5, 6]. Each technique is listed with an illustrative example.	25
3.3. Span length statistics per language partition and Euclidean distance to the Portuguese gold set. Distance is computed from the mean length profile (characters, words, tokens). Languages are ordered by distance to Portuguese.	29
3.4. Topic distribution (%) per language partition.	32
3.5. Comparison of training pipeline configurations.	39
4.1. F1 scores on the Portuguese test set for the three Stage 2 configurations tested with the Model 2 Stage 1 checkpoint.	47
4.2. Overall results on the Portuguese test set. PersuasionMultiSpan* is the organisers' post-competition system. UniBO was the CLEF-2024 competition winner.	49
4.3. Per-technique F1 scores on the Portuguese test set, sorted by gold span frequency. Bold values indicate the best result per technique.	51
4.4. Cumulative F1 gains from each key design decision on the Portuguese test set.	54

List of Abbreviations

- BERT** Bidirectional Encoder Representations from Transformers. iv, 14, 15, 17
- BIO** Beginning, Inside, Outside. 3, 14, 20, 23, 34, 35, 46, 50, 54
- CLEF** Conference and Labs of the Evaluation Forum. viii, 2–4, 10, 11, 14, 16, 18, 19, 21, 23–25, 40
- GPT** Generative Pre-trained Transformer. 15
- LLM** Large Language Model. 4, 5, 15, 16, 35
- mBERT** Multilingual Bidirectional Encoder Representations from Transformers. 3, 17, 18, 34, 36
- NLP** Natural Language Processing. iv, 3, 7, 10, 14, 16, 19, 20
- PTC** Propaganda Techniques Corpus. 8–11, 14
- SemEval** Semantic Evaluation. 3, 7, 10, 11, 14, 18, 24
- UMAP** Uniform Manifold Approximation and Projection. vii, 34, 45
- XLNet** Cross-lingual Language Model RoBERTa. 3, 15, 17, 18, 35, 37

1. Introduction

The rapid expansion of digital media has fundamentally transformed the way information is produced, distributed, and consumed. Online news platforms, social media, and instant messaging services have enabled the dissemination of content to large audiences at unprecedented speed, creating an information environment in which manipulative discourse and disinformation can spread widely before being identified or corrected. This phenomenon has attracted growing concern from researchers, policymakers, and civil society organisations, as the strategic use of rhetorical techniques to shape public opinion poses significant challenges to democratic processes and informed decision-making.

At the core of this challenge lies the concept of **manipulative discourse**: a form of communication designed not merely to inform, but to influence the beliefs, attitudes, and behaviours of an audience through strategic linguistic and rhetorical choices [1]. Unlike neutral communication, manipulative discourse is intentionally constructed to guide interpretation, frame events in a particular way, or trigger emotional reactions. It often relies on implicit assumptions, selective emphasis, and rhetorical framing rather than explicit factual distortion. This broader understanding highlights that manipulation does not require falsehood; it can coexist with factually accurate content, making its automatic detection substantially more complex than fact verification.

A particularly important manifestation of manipulative discourse is propaganda, which operates through a structured set of identifiable rhetorical strategies known as **persuasion techniques**. These include strategies such as Loaded Language, which exploits emotionally charged wording to bypass rational evaluation; Appeal to Fear, which invokes anxiety or threat to discourage critical thinking; Name Calling, which attacks individuals rather than their arguments; and False Dilemma, which presents a complex situation as having only two possible outcomes. A comprehensive taxonomy of 23 such techniques, organised into six coarse-grained categories, was introduced by Da San Martino et al. [2] and subsequently adopted as the standard framework for computational propaganda detection. This taxonomy is central to the present work: throughout this dissertation, the detection of manipulative discourse is operationalised as the automatic identification of these 23 persuasion techniques within text.

Unlike disinformation, which is typically characterised by the deliberate creation and dissemination of false information with the intent to deceive [3], persuasion-based manipulation does not depend on factual inaccuracy. A news article that is entirely truthful can still be deeply manipulative if it systematically employs Loaded Language to frame one side of a debate, uses Appeal to Authority to lend false legitimacy to a claim, or relies on Exaggeration to amplify fear. This distinction has motivated a shift in the computational research community from fact-checking tasks toward the identification of persuasion techniques and rhetorical manipulation [4], and it is this direction that the present work pursues.

The automatic identification of persuasion techniques in text has been formalised as a span-level sequence labelling task by Da San Martino et al. [2], in which systems must locate the precise character-level boundaries of manipulative text segments and assign the correct technique label to each. This formulation is considerably more challenging than document-level propaganda classification, as it requires models to capture both local linguistic signals and broader contextual dependencies in order to identify exactly where a technique begins and ends within a sentence or paragraph.

The multilingual dimension of this problem adds a further layer of complexity. While substantial progress has been made for English, most languages remain under-resourced for annotated propaganda-detection data. This scarcity motivates the use of cross-lingual transfer strategies, in which models trained on available source languages are adapted to perform detection in a target language for which no native annotations exist. The quality of this transfer depends critically on the choice of source language, the architecture of the model, and the training strategy employed, which are factors whose interactions are not yet fully understood in the propaganda detection literature [5, 6].

1.1. Motivation

This dissertation is motivated by two converging observations. First, Portuguese-language digital media has not been systematically addressed in the propaganda detection literature, despite the language being spoken by over 250 million people across multiple continents [7] and being present in online environments where coordinated disinformation campaigns have been documented [3]. The CLEF-2024 CheckThat! Lab Task 3 [6] provided the first standardised evaluation benchmark for persuasion technique detection in European Portuguese, creating an opportunity to investigate cross-lingual approaches tailored to this language.

Second, existing systems for span-level propaganda detection exhibit persistent limitations that have not been fully addressed. The standard Beginning, Inside, Outside (BIO) sequence labelling formulation, combined with a softmax output layer, is structurally incompatible with annotations where multiple persuasion techniques co-occur within the same text span, a characteristic present in a non-negligible fraction of gold annotations [2]. Systems following this design, including the top-performing entries in Semantic Evaluation (SemEval)-2023 Task 3 and the CLEF-2024 CheckThat! Lab Task 3 [5, 6], are therefore subject to a hard ceiling on recall that cannot be overcome by improving the encoder alone. Furthermore, while larger training datasets have generally been assumed to provide stronger transfer signal, recent work in low-resource NLP has suggested that linguistic relatedness between source and target languages can be a more reliable predictor of transfer quality than raw data volume [8], a hypothesis that has not been systematically tested in the context of span-level propaganda detection.

These observations motivate an investigation into alternative model architectures, training strategies, and source language selection methods for cross-lingual propaganda detection in Portuguese.

1.2. Objectives

The primary objective of this dissertation is to develop and evaluate a system for the automatic detection of persuasion techniques in Portuguese-language news articles, in the absence of native annotated training data. To this end, the following specific objectives are pursued:

- To investigate the suitability of multilingual Transformer architectures, specifically Multilingual Bidirectional Encoder Representations from Transformers (mBERT) and Cross-lingual Language Model RoBERTa (XLM-R), for cross-lingual span-level propaganda detection in Portuguese.
- To evaluate the impact of output architecture design on detection performance, comparing the standard BIO softmax formulation against a multi-label sigmoid architecture capable of handling overlapping technique annotations.
- To develop an embedding-based methodology for assessing linguistic proximity between source and target language training data, and to use this analysis to inform source language selection for cross-lingual transfer.

- To compare single-source and multi-source training strategies, examining whether combining data from multiple languages yields better coverage of the target-language technique distribution than relying on a single high-resource source language.
- To contextualise the performance of fine-tuned encoder models against a zero-shot Large Language Model (LLM) baseline, providing insight into the relative strengths of supervised and generative approaches for structured span-level detection.

1.3. Contributions

The main contributions of this dissertation are as follows:

- A multi-label sigmoid architecture for span-level propaganda detection that handles overlapping technique annotations, trained with per-output positive weighting to address the severe class imbalance inherent in the task.
- An embedding-based language proximity analysis using a multilingual sentence encoder, which provides empirical evidence that linguistic relatedness between source and target languages is a stronger predictor of transfer quality than dataset size in this setting.
- A two-stage cross-lingual training pipeline combining multilingual pre-training across six source languages with Portuguese-focused fine-tuning on machine-translated data, achieving performance that surpasses the CLEF-2024 CheckThat! Lab Task 3 competition winner.
- A systematic evaluation of threshold selection strategies and their interaction with the class imbalance problem in span-level detection.
- Preliminary results of this work were presented as posters at two academic events: the FCUP Master's Day (Dia dos Mestrados, FCUP, 2026) and the Young Researchers Meeting at the University of Porto (IJUP, 2026). Both posters are reproduced in Appendix A.
- All code is openly available in the repository of this thesis*, including a streamlit based web demo application. In addition, the trained model is publicly available on Hugging Face†.

*<https://github.com/andrevieira1203/MastersThesis>

†<https://huggingface.co/AndreCVieira/persuasion-lens-pt>

1.4. Document Structure

The remainder of this dissertation is organised as follows. Chapter 2 reviews the relevant literature, covering the definition and computational study of manipulative discourse, existing corpora and task formulations, Transformer-based approaches to propaganda detection, multilingual and cross-lingual transfer strategies, and the role of LLMs in persuasion detection. Chapter 3 describes the methodology, including the dataset, model architecture, training procedure, and evaluation protocol. Chapter 4 presents and analyses the experimental results. Chapter 5 concludes the dissertation with a summary of findings and directions for future work.

2. State of the Art

This chapter reviews the state of the art at the level of detail relevant to the design decisions made in this work. Rather than giving an exhaustive technical description of every system, it focuses on the conceptual distinctions, dataset characteristics, and cross-lingual transfer strategies that directly motivate the architecture and training choices developed in Chapters 3 and 4.

2.1. Manipulative Discourse and Persuasion Techniques

Manipulative discourse refers to communicative practices designed not merely to inform, but to influence the perceptions, beliefs, and behaviours of a target audience through strategic linguistic and rhetorical choices [1]. Unlike neutral or factual communication, manipulative discourse is intentionally constructed to guide interpretation, frame events in a particular way, or trigger emotional reactions. It often relies on implicit assumptions, selective emphasis, and persuasive framing rather than explicit factual distortion [4]. In contemporary digital environments, where information spreads rapidly through social media and online platforms, manipulative communication has become increasingly visible and socially impactful, motivating substantial research efforts in its computational detection [9].

While manipulative discourse is a broad phenomenon that can manifest in many forms, its computational study requires a concrete operationalisation. In practice, manipulative discourse operates through a structured set of identifiable rhetorical strategies known as **persuasion techniques**: specific patterns of language use that aim to influence how readers interpret information. A comprehensive taxonomy of such techniques, comprising the 23 fine-grained persuasion techniques used throughout this work, was introduced by Da San Martino et al. [2] and has since been adopted as the standard framework for computational propaganda detection [5, 6].

The relationship between manipulative discourse and persuasion techniques is therefore one of concept and operationalisation: manipulative discourse is the broader phenomenon, and persuasion techniques are its measurable, annotatable manifestations in text. This distinction is central to the present work. The title of this dissertation, which refers to the detection of manipulative discourse and disinformation, reflects the broader

goal of identifying rhetorically manipulative content. The methodology, however, operationalises this goal through the automatic detection of the 23 fine-grained persuasion techniques defined in the taxonomy above. Detecting these techniques at the span level is, in effect, detecting manipulative discourse at the level of its constituent rhetorical strategies.

This operationalisation has important practical implications beyond the academic setting. Identifying the specific techniques used in a text, rather than simply classifying the text as manipulative or not, provides actionable evidence that can support media literacy education, fact-checking workflows, and content moderation systems. A reader who is told that a paragraph contains Appeal to Fear or Exaggeration gains concrete insight into how the text is attempting to influence them, whereas a binary label of “manipulative” offers no such transparency. The shift toward fine-grained, span-level detection is therefore motivated not only by scientific interest but also by practical utility.

An important distinction in computational propaganda detection lies between document-level classification and span-level identification. Earlier approaches typically focused on classifying entire articles as propagandistic or not. A natural intermediate formulation is document-level multi-label classification, which identifies which techniques a text contains but not where they occur. Span-level identification goes one step further, and this is the level the practical use case requires: the actionable evidence described above depends on locating each technique in the text, not merely on knowing that it is present somewhere in the article. However, more recent shared tasks, such as SemEval-2023 Task 3 [5], emphasize the need to detect the exact textual fragments in which specific propaganda techniques occur.

From a NLP perspective, propaganda detection thus represents a fine-grained information extraction problem situated at the intersection of discourse analysis, rhetorical modelling, and contextual language understanding. The complexity of rhetorical strategies and their dependence on pragmatic and semantic nuance make this task particularly challenging, especially in multilingual and low-resource settings.

2.1.1 Propaganda, Disinformation and Misinformation

Although propaganda, disinformation and misinformation are often used interchangeably in public discourse, they represent distinct phenomena that require different detection approaches. Disinformation refers to false information that is deliberately created and

disseminated with the intent to deceive [3]. Its defining characteristic is the combination of factual inaccuracy with intentional deception: the content is wrong, and the author knows it is wrong. Misinformation, by contrast, also involves false content but without deliberate intent to mislead; a person who unknowingly shares an inaccurate claim is spreading misinformation, not disinformation [3]. Both concepts are fundamentally tied to the truth value of the content itself.

Propaganda operates on an entirely different axis. Rather than depending on falsehood, it works through strategic framing, selective emphasis, emotional appeals, and rhetorical techniques designed to influence how an audience interprets information [1]. A propagandistic message may be entirely factually accurate and still be deeply manipulative. Consider a concrete example: reporting on a military operation, one outlet might write “Terrorist killed in US strike”, while another covers the same event as “Thousands dead in US attack that killed hidden terrorist”. Both headlines may be factually correct, but the framing, the choice of emphasis, and the emotional weight of each formulation are radically different. The first foregrounds the elimination of a threat; the second foregrounds civilian casualties. Neither is necessarily false, but both employ rhetorical strategies, specifically Loaded Language and selective framing, to guide the reader toward a particular interpretation of the same underlying facts.

This distinction is particularly relevant for computational detection. Disinformation can, in principle, be addressed through fact-checking: comparing claims against verified sources. Propaganda, however, cannot be detected by verifying facts alone, because the facts may be correct. What must be detected instead is the rhetorical structure through which those facts are presented. Accordingly, rather than asking whether a text is true or false, the present work asks which rhetorical strategies are being used to influence the reader, and where exactly in the text those strategies appear [4, 5].

2.1.2 Rhetorical Techniques and Span-Level Granularity

Propaganda operates through a set of identifiable rhetorical techniques that aim to influence interpretation rather than merely transmit information. These techniques include strategies such as loaded language, name-calling, appeal to authority, exaggeration, and black-and-white fallacy, among others.

The introduction of the Propaganda Techniques Corpus (PTC) marked a significant methodological shift by defining a taxonomy of propaganda techniques and framing the task as a

span-level sequence labeling problem. In this formulation, systems must identify the exact boundaries of text fragments where a technique occurs and assign the correct label. This differs substantially from document-level classification, as it requires precise localization and contextual understanding.

Span-level granularity increases the complexity of the task in several ways. First, rhetorical techniques may be expressed implicitly, without clear lexical markers. Second, the boundaries of a persuasive span are not always linguistically obvious, leading to annotation challenges and potential ambiguity. Third, multiple techniques may overlap or co-occur within the same sentence. These factors transform propaganda detection into a fine-grained information extraction task that demands contextual modelling and robust sequence labeling capabilities.

2.1.3 Linguistic Challenges in Automatic Identification

The automatic identification of propaganda techniques presents substantial linguistic and computational challenges that extend beyond surface-level pattern matching. Unlike tasks that rely on explicit lexical cues, propaganda often operates through implicit meaning, pragmatic inference, and discourse-level framing. Techniques such as exaggeration, appeal to emotions, or sarcasm may be indistinguishable from neutral language when analysed at the token level, requiring models to capture broader semantic and contextual dependencies that span multiple sentences or even entire documents.

A further source of difficulty lies in the variability of rhetorical expression. The same persuasive technique may manifest through widely different lexical choices, syntactic constructions, and stylistic devices depending on the author, the target audience, and the cultural context. This variability severely limits the effectiveness of rule-based systems and motivates the use of contextualised language models capable of learning distributed semantic representations that generalise across surface forms. As discussed in broad surveys of computational propaganda detection [4], achieving robust performance requires models that move beyond surface features and capture the discourse-level patterns through which rhetorical manipulation is expressed.

As highlighted in the PTC paper [2], even human annotators disagree on span boundaries and technique assignments in a non-negligible fraction of cases, illustrating the inherent ambiguity of the task. This annotation uncertainty, combined with the context-sensitivity of rhetorical strategies, makes fine-grained propaganda detection particularly challenging

and justifies the adoption of Transformer-based architectures capable of modelling long-range contextual dependencies, especially in multilingual and low-resource settings.

2.2. Propaganda Detection Datasets and Task Description

The formulation of propaganda detection as a NLP task has been strongly shaped by the availability of annotated corpora specifically designed to capture rhetorical manipulation strategies. Early computational approaches typically addressed propaganda detection as a document-level classification problem, in which entire articles were labelled as propagandistic or non-propagandistic[9]. While this formulation enabled initial progress in automatic detection, it proved insufficient for capturing the fine-grained linguistic structure through which persuasion techniques are expressed.

More recent research has therefore shifted toward span-level detection, where models are required to identify both the location and the category of rhetorical techniques within text. This transition was enabled by the development of datasets providing token-level annotations aligned with predefined persuasion taxonomies. As a consequence, propaganda detection is now commonly modeled as a sequence labeling task that requires contextual understanding across multiple linguistic levels.

Shared evaluation campaigns such as SemEval-2023 Task 3 [5] and the CLEF Check-That! Lab Task 3 [6, 10] have further standardized this formulation by introducing multilingual datasets and benchmarking protocols. These initiatives have played a central role in promoting reproducibility and enabling systematic comparison between modelling approaches, framing propaganda detection as a multilingual span-level information extraction task of increasing practical relevance.

2.2.1 The Propaganda Techniques Corpus (PTC)

The PTC, introduced by Da San Martino et al., represents one of the most influential resources in the computational study of manipulative discourse [2]. The dataset provides span-level annotations of rhetorical strategies within news articles and defines a structured taxonomy of persuasion techniques that has become a reference framework for subsequent research in propaganda detection [5, 6, 10].

This fine-grained annotation scheme supports modelling propaganda detection as a structured sequence labelling task requiring simultaneous boundary detection and technique

classification, and established the foundation for subsequent multilingual shared tasks including SemEval-2023 Task 3 and the CLEF CheckThat! Lab tasks [5, 6, 10].

2.2.2 Evaluation Metrics

Evaluation in classification and sequence labelling tasks relies on comparing the system's predictions against the gold annotations. For each class, four outcomes are possible: a **true positive** (TP) occurs when the system correctly predicts a positive instance; a **false positive** (FP) occurs when the system predicts a positive instance that is not present in the gold labels; a **false negative** (FN) occurs when the system fails to predict a positive instance that exists in the gold labels; and a **true negative** (TN) occurs when both the system and the gold labels agree on the absence of a positive instance.

From these four quantities, two complementary metrics are derived. **Precision** measures the proportion of predicted positives that are correct, while **Recall** measures the proportion of actual positives that the system successfully identifies:

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN} \quad (2.1)$$

A system with high precision but low recall produces few predictions but tends to be correct when it does predict; a system with high recall but low precision captures most positive instances but at the cost of many false alarms. Since both properties are desirable, the **F1-score** provides a single metric that balances precision and recall through their harmonic mean:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.2)$$

In multi-class settings such as persuasion technique detection, F1 scores can be aggregated in two ways. **Micro-averaged F1** computes precision and recall by summing TP, FP, and FN across all classes before computing the ratio, which means that frequent classes contribute more to the final score. **Macro-averaged F1** computes F1 independently for each class and then averages the per-class scores, giving equal weight to all classes regardless of frequency.

The distinction between micro and macro averaging is particularly important in datasets with skewed label distributions, such as the PTC and the CLEF-2024 CheckThat! Lab

corpora, where a small number of techniques account for the majority of annotated spans. In these datasets, a system that performs well only on the most frequent categories can achieve a reasonable micro-F1 while producing a very low macro-F1. For this reason, recent benchmarking campaigns report both metrics as complementary measures [5, 6], and both are used throughout this work.

2.2.3 Class Imbalance Problem

As previously mentioned, a major challenge in propaganda detection datasets lies in the uneven distribution of manipulative techniques across annotated corpora.

This imbalance often leads to inflated micro-F1 scores that do not reflect consistent performance across the full technique taxonomy. Large discrepancies between micro-F1 and macro-F1 scores frequently indicate limited robustness on minority classes, a phenomenon widely observed in recent shared-task benchmarks [5, 6]. Common mitigation strategies include class weighting during training, oversampling of minority classes, and data augmentation techniques designed to increase the representation of rare rhetorical categories [2, 6].

Despite these efforts, the gap between micro-F1 and macro-F1 performance persists in state-of-the-art systems, indicating that current mitigation approaches remain insufficient for achieving reliable detection across the complete persuasion taxonomy. This limitation is particularly pronounced in multilingual and cross-lingual settings, where the distribution of techniques varies across languages and rare categories may receive no training signal at all from certain source languages [6]. The persistence of this performance gap motivates further investigation into more effective loss formulations and training strategies, which are examined in the context of this work in the following chapters.

2.3. Transformer-Based Models for Propaganda Detection

2.3.1 From Traditional Machine Learning to Deep Learning

Early computational approaches to propaganda and manipulative discourse detection relied on two distinct types of input representation. The first was **hand-crafted features** derived from linguistic analysis, such as lexical patterns, part-of-speech distributions, syntactic constructions, and sentiment indicators [9]. These features were manually designed to capture properties that researchers believed to be indicative of manipulative language,

and were typically fed into classical classifiers such as Support Vector Machines or Logistic Regression models.

The second type of representation involved statistical models of text that required less manual engineering. **Bag-of-Words** (BoW) representations encode a document as a vector of word frequencies, discarding word order but capturing vocabulary usage patterns. **TF-IDF** (Term Frequency–Inverse Document Frequency) refines this by weighting terms according to their discriminative power across a corpus, reducing the influence of common words and amplifying terms that are distinctive to specific documents. While both BoW and TF-IDF proved useful for document-level classification tasks, they are fundamentally limited by their inability to capture word order, syntactic structure, or semantic relationships between terms.

A significant step forward came with distributed word representations such as **Word2Vec** [11], which learn dense vector embeddings by training on word co-occurrence patterns in large unlabelled corpora. Unlike sparse BoW and TF-IDF vectors, Word2Vec embeddings capture semantic similarity between words, placing related terms close together in the vector space. However, Word2Vec assigns a single static embedding to each word regardless of context, which means that polysemous words receive the same representation in all contexts, a substantial limitation for tasks where meaning is context-dependent.

Both hand-crafted features and these text representation methods proved capable of capturing surface-level signals of manipulative language in document-level classification settings [9]. However, they faced fundamental limitations when applied to fine-grained span-level detection, where the contextual interpretation of individual tokens and their surrounding discourse plays a critical role.

The transition to deep learning introduced recurrent architectures such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU), which enabled sequential modelling of text and provided richer contextual representations. These models process text token by token while maintaining an internal state that captures information from previous tokens, allowing them to model dependencies across a sentence. However, they were constrained by their difficulty in capturing long-range dependencies within lengthy documents and their limited ability to generalise across languages without dedicated multilingual training data. These limitations motivated the shift toward attention-based Transformer architectures, which address both the parallelism and the long-range dependency challenges of previous approaches and have since become the dominant

paradigm for propaganda detection [2].

2.3.2 Bidirectional Encoder Representations from Transformers (BERT)

BERT, introduced by Devlin et al. [12], marked a decisive turning point in NLP by demonstrating that deeply bidirectional Transformer encoders pre-trained on large unlabelled corpora could be fine-tuned to achieve state-of-the-art performance across a wide range of downstream tasks. Unlike earlier pre-training objectives based on unidirectional language modelling, BERT employs a Masked Language Modelling (MLM) objective that predicts randomly masked tokens conditioned on their full bidirectional context, enabling rich contextual token representations that capture both local and long-range linguistic dependencies. In the original paper, Devlin et al. demonstrated that this approach achieved state-of-the-art results on eleven NLP benchmarks, including named entity recognition and question answering, establishing pre-trained Transformer encoders as the dominant paradigm for structured prediction tasks [12].

The impact of BERT on span-level propaganda detection was substantial. The standard approach consists of appending a token classification head to the encoder output and framing the task using BIO sequence labelling, in which each token is assigned a label indicating whether it begins a persuasive span (B), falls inside one (I), or belongs to no annotated segment (O). Da San Martino et al. showed that this design substantially outperformed classical and recurrent baselines on the PTC [2], and subsequent shared-task campaigns confirmed these findings at scale: in SemEval-2023 Task 3, the majority of competitive submissions used BERT-family encoders with BIO classification heads as their core architecture [5]. Similar patterns were observed in the CLEF-2024 Check-That! Lab Task 3, where Transformer-based encoder systems dominated the leaderboard across all evaluated languages [6].

A known limitation of the BIO softmax formulation, however, is that it assigns a single label per token. This is structurally incompatible with annotations where multiple persuasion techniques overlap within the same span, a property observed in approximately 20% of gold annotations in the PTC [2]. Single-label systems systematically fail to capture this multi-label structure, producing a hard upper bound on recall that no improvement to the encoder can overcome. Despite this limitation, the vast majority of shared-task systems have continued to adopt single-label BIO formulations [5, 6], leaving multi-label alternatives largely unexplored in the propaganda detection literature.

2.3.3 Fine-Tuning Strategies for Span-Level Detection

Fine-tuning pre-trained Transformer models for span-level propaganda detection requires careful design choices regarding output architecture, loss function, and inference procedure. While effective for simple single-label spans, this formulation has been shown to be insufficient when annotations contain overlapping technique labels, as the softmax output layer can only assign one class per token.

To address class imbalance, an inherent property of propaganda datasets where positive spans constitute a small fraction of all tokens, several strategies have been explored. Weighted loss functions that penalise misclassification of minority classes more heavily than majority classes have been shown to improve recall on rare techniques, although excessively high weights can destabilise training [2]. Oversampling of minority classes and data augmentation techniques have also been proposed as complementary approaches[4].

For documents exceeding the 512-token context limit associated with the classic Transformer encoders, sliding window strategies with overlapping segments are commonly employed during both training and inference. Probability averaging across overlapping windows provides a natural aggregation mechanism that reduces boundary artefacts and improves prediction consistency [5]. These practical considerations are central to the design of competitive systems on propaganda detection benchmarks, where news articles frequently exceed the encoder's context limit.

2.3.4 Generative Language Models vs Encoder-Based Approaches

The emergence of large autoregressive language models such as Generative Pre-trained Transformer (GPT)-3 [13] and their instruction-tuned successors have introduced an alternative paradigm for propaganda detection. Unlike encoder-based models such as BERT and XLM-R, which are fine-tuned on labelled data with a task-specific classification head, generative LLMs are prompted in natural language to identify and classify persuasive spans without any task-specific training.

Hasanain et al. provide the most direct empirical comparison to date, evaluating GPT-4 against fine-tuned encoder models on span-level propaganda detection across multiple languages [14]. Their results demonstrate that fine-tuned models consistently outperform GPT-4 in zero-shot settings, and that GPT-4's performance degrades substantially as the task moves from coarse document-level classification toward fine-grained span-level

technique detection. The authors further report that GPT-4 struggles across languages, suggesting that the limitations of generative models on this task are not language-specific but rather structural and arise from the fundamental mismatch between autoregressive text generation and the token-level precision required for span boundary detection. These findings are consistent with results from the CLEF-2024 CheckThat! Lab Task 3, where zero-shot LLM baselines achieved lower F1 scores than fine-tuned multilingual encoder systems [6]. The primary sources of failure identified in the paper include imprecise span boundaries produced by free-form generation, inconsistent technique assignment across runs, and the high computational cost of inference compared to fine-tuned encoders. Instruction tuning [15] partially mitigates some of these issues by improving prompt-following behaviour, but does not close the performance gap with fine-tuned encoders on structured span-level tasks. Taken together, these results motivate the adoption of encoder-based architectures for the present work, while retaining a generative model as a zero-shot baseline for comparative purposes.

2.4. Multilingual and Cross-Lingual Propaganda Detection

2.4.1 Challenges in Multilingual NLP

Multilingual NLP faces a set of challenges that are particularly pronounced in fine-grained tasks such as propaganda detection. A fundamental difficulty lies in the uneven distribution of annotated resources across languages: while high-resource languages such as English benefit from large labelled corpora, most languages remain under-resourced, with limited span-level annotations and fewer dedicated language models available for fine-tuning[6, 8].

Even with multilingual pre-trained encoders, performance on low-resource languages tends to lag behind high-resource ones due to imbalances in pre-training data. This phenomenon, sometimes referred to as the curse of multilinguality, reflects the tension between model capacity and the number of languages represented during pre-training [8]. Adding more languages to a fixed-capacity model tends to reduce per-language performance, particularly for languages with limited coverage in the pre-training corpus.

A further challenge specific to propaganda detection is that the distribution of rhetorical techniques varies across languages and cultural contexts. Annotated datasets for different languages may exhibit substantially different technique distributions, meaning that models trained on a source language may be systematically exposed to a different subset

of rhetorical strategies than those present in the target evaluation data [6]. This distributional mismatch limits the effectiveness of naive single-source transfer approaches and motivates the use of multi-source training strategies that provide broader technique coverage.

This is not merely a theoretical concern: the exploratory analysis in Section 3.3 confirms it directly on the CLEF-2024 data, where four techniques absent from the English training set account for 33.6% of the spans in the Portuguese test set.

2.4.2 Multilingual Transformer Models

The challenges described above have motivated the development of Transformer-based models that are pre-trained on data from many languages simultaneously, producing a shared representation space in which text from different languages can be compared and processed by the same model. Two such models have become particularly relevant for cross-lingual propaganda detection: mBERT, which extended the original BERT architecture to a multilingual setting, and XLM-R, which improved upon it with a larger and more balanced pre-training corpus. Both are described below.

The extension of BERT to multilingual settings led to mBERT, a single model pre-trained on the concatenated Wikipedia corpora of over one hundred languages using the same masked language modelling objective [12]. Despite having no explicit cross-lingual training signal, mBERT demonstrated substantial zero-shot cross-lingual transfer capabilities, enabling models fine-tuned on one language to generalise to others without requiring parallel data or language-specific architectures.

XLM-R, introduced by Conneau et al. [8], extended this approach by training on a significantly larger and more balanced multilingual corpus using the RoBERTa training methodology, which removes the Next Sentence Prediction objective and employs larger batch sizes and longer sequences. Pre-trained on 2.5 terabytes of filtered CommonCrawl data covering 100 languages, XLM-R yields substantially stronger cross-lingual representations than mBERT, particularly for low-resource languages.

Within the propaganda detection literature, both mBERT and XLM-R have been adopted as backbone encoders in shared-task systems. XLM-R has generally shown superior performance, especially in scenarios involving transfer from high-resource to low-resource languages. Conneau et al. demonstrated consistent improvements over mBERT across structured prediction tasks in low-resource settings [8], and these gains have been

confirmed in propaganda detection specifically: in both SemEval-2023 Task 3 and the CLEF-2024 CheckThat! Lab, systems based on XLM-R outperformed their mBERT counterparts on the majority of evaluated languages [5, 6].

2.4.3 Transfer Learning Across Languages

Cross-lingual transfer learning leverages models trained on one or more source languages to improve performance on a target language with limited annotated data. Two main transfer strategies have been explored in the propaganda detection literature: zero-shot transfer, in which a model fine-tuned exclusively on a source language is evaluated directly on the target language, and translate-train transfer, in which source-language annotated data is machine-translated into the target language to construct a synthetic training set.

Zero-shot transfer with XLM-R has shown promising results in structured prediction tasks, with models fine-tuned on English exhibiting performance well above the random baseline when evaluated on other languages without any target-language supervision [5]. However, performance degrades for morphologically distant languages and when the source training data does not cover techniques present in the target evaluation set, highlighting the limits of zero shot transfer in highly imbalanced multi-class settings.

Translate-train approaches have demonstrated competitive performance when the source language is linguistically close to the target, as lexical and syntactic similarity reduces translation noise and improves span boundary preservation after translation [6]. Combining multiple translated sources has been shown to improve technique coverage compared to single-source transfer, particularly when different source languages contribute complementary annotated spans. The choice of source language is therefore a non-trivial decision that affects both the quantity and the diversity of the training signal available for the target language.

2.4.4 Low-Resource Language Scenarios

In low-resource settings where no native target-language annotations exist, the selection of source languages for cross-lingual transfer is a critical design decision. Several studies have explored the use of multilingual sentence embeddings to measure document-level similarity between target-language data and training corpora across candidate source languages, providing an empirical basis for assessing which sources are most relevant to the

target distribution [8]. Linguistic proximity, reflected by shared vocabulary, morphological patterns, and syntactic structures, has also been identified as a strong predictor of transfer quality in low-resource NLP tasks. Romance languages, for instance, tend to transfer more effectively among themselves than to or from Germanic or Slavic languages, due to their shared Latin roots and similar morphosyntactic properties [8]. This observation challenges the common assumption that larger training datasets always yield better transfer, suggesting that data quality and linguistic relevance to the target language can outweigh raw dataset size in low-resource cross-lingual scenarios.

In the context of the CLEF-2024 CheckThat! Lab Task 3, Portuguese represents a language for which no native span-level propaganda annotations are available. The task therefore requires systems to rely entirely on cross-lingual transfer from the six available training languages: English, French, German, Italian, Polish, and Russian. The diversity of available source languages provides an opportunity to study the impact of source selection and multi-source combination strategies on target-language performance.

2.5. Research Gaps and Motivation for This Work

The preceding sections reviewed the current state of computational propaganda detection, covering its conceptual foundations, available datasets, modelling architectures, and cross-lingual transfer strategies. Despite the considerable progress made in recent years, several limitations persist that directly motivate the choices made in this dissertation. This section identifies the most relevant gaps in three areas: the architectural constraints of current span-level systems, the challenges of multilingual generalisation, and the specific objectives of the present work.

2.5.1 Limitations of Current Span-Level Detection Systems

Despite significant progress enabled by multilingual Transformer models, current span-level propaganda detection systems exhibit several persistent limitations. First, performance on minority rhetorical techniques remains substantially lower than on frequent categories, with macro-averaged F1 scores consistently trailing micro-averaged scores on shared-task benchmarks [5, 6]. This disparity reflects the difficulty of learning robust representations for rare classes in severely imbalanced datasets, and suggests that existing class imbalance mitigation strategies are insufficient for achieving consistent detection across the full persuasion taxonomy.

Second, the standard BIO softmax formulation used by the majority of shared-task systems is structurally incompatible with the multi-label nature of propaganda annotations, where multiple techniques can co-occur within the same span. Systems that assign a single label per token systematically fail to detect overlapping techniques, producing a hard upper bound on recall that cannot be overcome by improvements to the encoder alone [2]. This architectural limitation has received limited attention in the literature, with most competitive systems continuing to adopt single-label BIO formulations.

2.5.2 Challenges in Multilingual Generalisation

Cross-lingual generalisation in span-level propaganda detection remains an open research challenge. The quality of transfer is sensitive to the linguistic proximity between source and target languages, the completeness of technique coverage in the source training data, and the consistency of annotation schemes across languages. For languages where certain rhetorical techniques are absent from all available source data, zero-shot and translate-train transfer approaches perform poorly in those categories.

The assumption that larger training datasets always provide better transfer signal has been questioned in recent low-resource NLP research, with evidence suggesting that linguistic relatedness between source and target can outweigh dataset size [8]. However, this hypothesis has not been systematically investigated in the context of span-level propaganda detection, and the relationship between source language selection and target-language F1 performance across individual technique categories remains underexplored in the literature.

Furthermore, systematic comparisons between multi-source and single-source transfer strategies on standardised multilingual propaganda detection benchmarks are limited, leaving open questions about the relative benefits of combining languages with different degrees of linguistic proximity to the target.

2.5.3 Motivation and Objectives of This Thesis

The present dissertation is motivated by the convergence of these research gaps and the practical importance of automatic propaganda detection in contemporary information ecosystems. The increasing prevalence of manipulative discourse in digital media,

combined with the limited availability of detection tools for non-English languages, underscores the need for robust multilingual systems capable of identifying fine-grained rhetorical strategies at the span level.

This work focuses specifically on Portuguese, a language for which no native span-level propaganda annotations exist, within the evaluation framework of the CLEF-2024 Check-That! Lab Task 3 [6]. The research investigates cross-lingual transfer strategies for span-level persuasion technique detection, examining the role of source language selection, multi-source training combination, and output architecture design in determining target-language performance. Additionally, the work explores the use of multilingual sentence embeddings as a principled method for assessing linguistic proximity between source and target language data, providing an empirical basis for source language selection in low-resource cross-lingual propaganda detection.

The findings are intended to contribute to a better understanding of the conditions under which cross-lingual transfer is effective for fine-grained rhetoric detection, and to highlight architectural and data selection decisions that improve performance in low-resource multilingual settings.

3. Methods and Data

This chapter describes both the data and the methodology developed for the automatic detection of persuasion techniques in European Portuguese news articles. The task presents two fundamental challenges that shape every decision in this work: the need to identify precise boundaries of manipulative text fragments across 23 possible rhetorical categories, and the need to do so for a language with no annotated training data. The chapter begins by formalising the task and describing the dataset in full, including how documents were collected, how annotations were produced, and what statistical properties the corpus exhibits. It then walks through the exploratory analysis that revealed distributional patterns across languages, the embedding-based proximity analysis that guided source language selection, and the complete modelling pipeline from baselines through to the final system. Figure 3.1 provides an overview of the data exploration and selection process that shaped this pipeline.

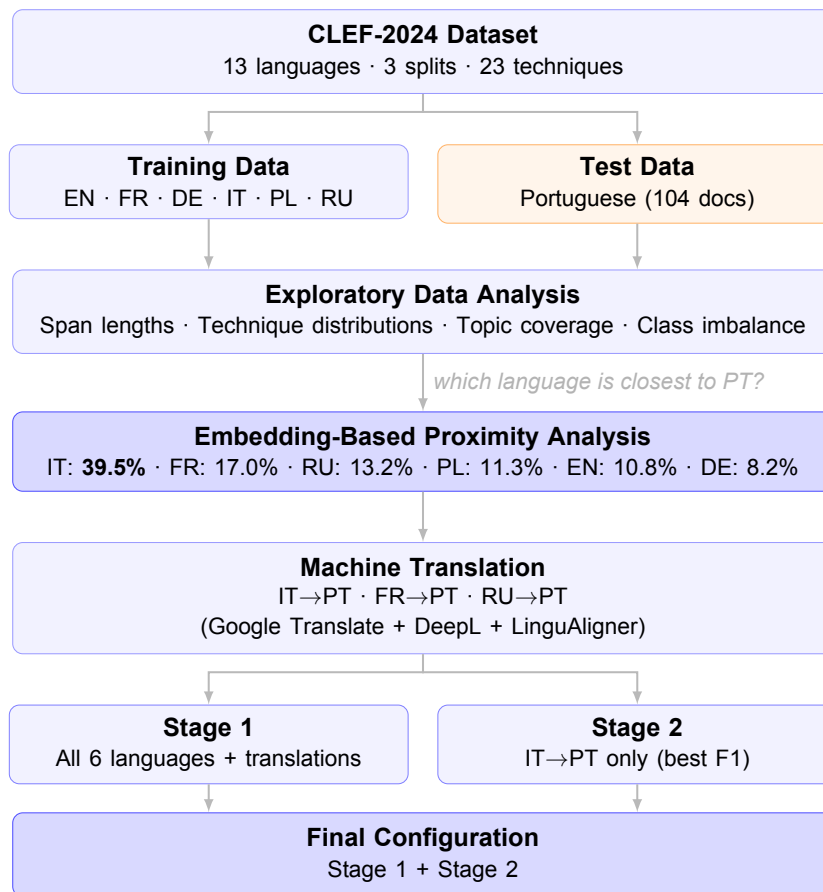


Figure 3.1: Data exploration and selection pipeline. The CLEF-2024 dataset is progressively filtered and transformed to determine the final training configuration.

3.1. Problem Formulation

The task is formulated as a span-level multi-label sequence labelling problem. Given a document D consisting of n tokens $\{t_1, t_2, \dots, t_n\}$, the system must identify all text spans $s = (t_i, \dots, t_j)$ that instantiate one or more of 23 predefined persuasion techniques and assign the correct technique label(s) to each span. This is a substantially harder problem than standard single-label sequence labelling, for two reasons that have direct implications for the architectural choices made later in this chapter.

The first is that the task is genuinely multi-label. A single token may belong to several techniques simultaneously, a property observed in approximately 20% of annotated spans in the Portuguese gold data from the CLEF-2024 test set. This rules out the standard BIO softmax formulation used in most sequence labelling tasks, since softmax normalisation can only assign one class per token by design. Any system based on this formulation will therefore have a hard upper bound on recall that cannot be overcome, regardless of encoder quality, a limitation explored empirically in Section 3.5.

The second problem is that evaluation uses partial span matching rather than exact token-level agreement. Predicted spans that overlap with gold spans receive proportional credit even if their boundaries are not perfectly aligned, which makes the metric more tolerant of boundary errors but also more complex to implement and interpret. The details of this evaluation procedure are described in Section 3.9.

Beyond these structural challenges, the challenge of the task further increases due to the complete absence of Portuguese training data. Therefore, there are no native annotations to learn from. Everything the model learns about how persuasion techniques manifest in Portuguese must be inferred from data in other languages, whether directly through cross-lingual transfer or indirectly through machine translation. This zero-resource constraint is the reason why the choice of source languages, explored through the embedding analysis in Section 3.4, turns out to be one of the most consequential decisions in the entire pipeline.

3.2. Dataset

3.2.1 CLEF-2024 CheckThat! Lab Task 3

The data used in this work was developed for the CLEF-2024 CheckThat! Lab Task 3 on persuasion technique detection [6]. The task targets the detection and classification of 23 fine-grained persuasion techniques at the text-span level in online news. As shown

in Table 3.1, the dataset covers thirteen languages distributed unevenly across three splits: six languages have both training and development data (English, French, German, Italian, Polish, and Russian), three have development data only (Georgian, Greek, and Spanish), and four appear exclusively in the test partition (Arabic, Bulgarian, Portuguese, and Slovenian). English is the only language present in all three splits. This work focuses specifically on Portuguese, the only Romance language in the test partition, for which no training or development data is available. The task builds on SemEval-2023 Task 3 [5], which used the same taxonomy but operated at the paragraph level; by moving to span-level detection, the 2024 edition substantially increases the difficulty of the problem.

Table 3.1: Dataset statistics across all splits in the CLEF-2024 CheckThat! Lab Task 3 [6]. Portuguese, the target language of this work, appears only in the test split.

Language	Training		Development		Test	
	Docs	Spans	Docs	Spans	Docs	Spans
English (EN)	536	9,002	54	1,775	98	2,599
French (FR)	211	6,831	50	1,681	–	–
German (DE)	177	5,737	50	1,904	–	–
Italian (IT)	303	7,961	61	2,351	–	–
Polish (PL)	194	3,824	47	1,491	–	–
Russian (RU)	191	4,138	72	944	–	–
Georgian (KA)	–	–	29	218	–	–
Greek (EL)	–	–	64	691	–	–
Spanish (ES)	–	–	30	546	–	–
Arabic (AR)	–	–	–	–	1,527	2,197
Bulgarian (BG)	–	–	–	–	100	1,732
Slovenian (SL)	–	–	–	–	100	4,591
Portuguese (PT)	–	–	–	–	104	1,727

3.2.2 Persuasion Technique Taxonomy

The task employs a two-tier taxonomy introduced in SemEval-2023 Task 3 [5]. At the top level, six coarse-grained categories capture the broad communicative function of each technique: Attack on Reputation, Justification, Simplification, Distraction, Call, and Manipulative Wording. These are further subdivided into 23 fine-grained techniques that constitute the classification targets. Table 3.2 presents the full taxonomy with an illustrative example for each technique, drawn from the annotation guidelines and the CLEF task description.

Table 3.2: Two-tier persuasion technique taxonomy used in the CLEF-2024 CheckThat! Lab Task 3 [5, 6]. Each technique is listed with an illustrative example.

Category	Technique	Example
Attack on Reputation	Name_Calling-Labeling	<i>"Fascist anti-vax rioters spark COVID outbreak"</i>
	Guilt_by_Association	<i>"He defends this because he associates with criminals"</i>
	Doubt	<i>"I am not sure we can trust what he says"</i>
	Appeal_to_Hypocrisy	<i>"They criticise us but do exactly the same"</i>
	Questioning_the_Reputation	<i>"What credibility can someone like that have?"</i>
Justification	Flag_Waving	<i>"This policy defends our nation and its true citizens"</i>
	Appeal_to_Authority	<i>"Since the Pope said it, we should add it to the creed"</i>
	Appeal_to_Popularity	<i>"Everybody agrees this measure is necessary"</i>
	Appeal_to_Values	<i>"This decision will protect our fundamental values"</i>
	Appeal_to_Fear-Prejudice	<i>"If we do not act now, society will be threatened"</i>
Simplification	Causal_Oversimplification	<i>"Unemployment exists because companies pay too little tax"</i>
	False_Dilemma-No_Choice	<i>"Either we accept this law or the country collapses"</i>
	Consequential_Oversimplif.	<i>"This decision will resolve all economic problems"</i>
Distraction	Straw_Man	<i>"You claim I do not care if people die without healthcare"</i>
	Red_Herring	<i>"The issue here is not the law but people's behaviour"</i>
	Whataboutism	<i>"They criticise us, but what about the other government?"</i>
Call	Slogans	<i>"Immigrants welcome, racists not!"</i>
	Appeal_to_Time	<i>"We must act immediately before it is too late"</i>
	Conversation_Killer	<i>"There is no point discussing this matter further"</i>
Manipulative Wording	Loaded_Language	<i>"This measure is completely absurd"</i>
	Repetition	<i>"They repeat lies, lies, and more lies"</i>
	Exaggeration-Minimisation	<i>"This is the worst disaster of all time"</i>
	Obfuscation-Vagueness-Conf.	<i>"Several appropriate measures have been taken"</i>

3.2.3 Document Collection

The dataset contains news articles collected from online media sources across multiple countries. For each language, the responsible annotation team manually selected articles from a curated set of news outlets, choosing material that was expected to contain a variety of persuasion strategies [6]. The selection criteria emphasised topical diversity, balanced political stances, and a mix of news genres, including both hard news reporting and opinion articles. The latter tend to contain a higher density of rhetorical techniques and were therefore deliberately overrepresented relative to their proportion in normal news consumption.

The topics covered were chosen based on their known prevalence of persuasion techniques in online media, and included the Russian–Ukraine war, climate change, COVID-19, abortion, migration, and the Israeli–Palestinian conflict. Some language partitions additionally included elections coverage. All languages except Arabic share these core topics, enabling meaningful cross-lingual comparison of technique usage across different national media contexts.

3.2.4 Annotation Process and Quality

The annotation process followed a standardised protocol applied consistently across all languages before the data was split into training, development, and test sets [6]. For all languages except Arabic, each document was annotated independently by two annotators fluent in the language and trained in the persuasion technique taxonomy. A curator subsequently analysed the two sets of annotations to produce the final gold labels. For Arabic, three annotators per paragraph were used, with two curators consolidating the results.

The training of annotators was systematic. Each annotator received comprehensive annotation guidelines, completed two sets of flashcards of increasing complexity covering the full technique taxonomy, and participated in discussion sessions with expert annotators from previous annotation campaigns. This process was designed to ensure consistent understanding of technique definitions and span boundary conventions across language-specific teams, which is critical for cross-lingual comparability.

Several quality assurance measures were applied after initial annotation. Label distributions were compared across languages to detect systematic cross-lingual inconsistencies. Texts were sorted alphabetically to make it easier to spot similar articles that had received

conflicting annotations. Annotations were clustered based on semantic similarity to flag outliers for review. Senior annotators performed random checks across all language partitions.

The overall inter-annotator agreement before curation, measured by Krippendorff's α , is 0.404. While this falls below the commonly recommended threshold of 0.667, this figure is computed at the span level and therefore jointly penalises technique labelling and exact boundary placement. The organisers report that, on inspection, annotators most often agreed on the technique while the spans differed only slightly [6]. The relatively low α therefore reflects the inherent difficulty of exact span-level annotation rather than poor annotation quality, and it is precisely this observation that motivates the partial span matching evaluation adopted in Section 3.9. The curation process was specifically designed to resolve these boundary disagreements.

3.2.5 Training and Development Data

As Table 3.1 shows, the training and development partitions cover six languages with substantial size differences: English has the most annotated spans (9,002) while Russian has the fewest (4,138). No Portuguese data exists in either split.

A critical detail that only becomes apparent when examining the per-technique breakdown is that four categories are entirely absent from the English training data: `Questioning_the_Reputation`, `Appeal_to_Values`, `Consequential_Oversimplification`, and `Appeal_to_Time`. As the exploratory analysis in Section 3.3 reveals, these four techniques collectively account for 33.6% of all annotated spans in the Portuguese test set. This observation has direct implications for the training strategy, and it was one of the primary reasons for adopting a multi-source approach rather than relying on the largest available dataset.

3.2.6 Portuguese Test Dataset

The Portuguese test dataset was developed specifically for the CLEF-2024 task and constitutes the first application of the persuasion technique annotation framework to European Portuguese [6]. It consists of 104 news articles manually selected from 28 European Portuguese news media sources, combining both news and opinion pieces to ensure annotation diversity. The corpus covers seven topics: the Israeli–Palestinian conflict (24 articles), the Russia–Ukraine war (19), climate change (13), elections (13), abortion (12),

COVID-19 (11), and migration (11). Elections were included because the Portuguese parliamentary election cycle during the collection period produced a large volume of opinion articles rich in persuasion techniques.

The test set contains 1,727 annotated spans across 1,501 paragraphs, covering all 23 technique categories. The inter-annotator agreement for Portuguese is $\alpha = 0.587$, the highest of all five test languages, indicating relatively strong annotation consistency. This is likely because the Portuguese annotation team included researchers directly involved in the design of the task, who brought a deeper familiarity with the annotation conventions.

3.3. Exploratory Data Analysis

Before any modelling decisions were made, an exploratory analysis was conducted on the full multilingual corpus. The goal was to understand the statistical properties of the data well enough to anticipate potential problems and make informed design choices. All analysis was implemented in Python, with spans loaded from the raw task files into a unified JSONL format and processed with `pandas` and `matplotlib`.

3.3.1 Data Preparation

The raw data was provided in a format where article texts and span annotations were stored in separate directories, one annotation file per article. A preprocessing pipeline was developed to merge these into a unified JSONL representation, where each record contains the article identifier, the full article text, and all annotated spans with their character-level offsets and technique labels. Separate JSON files were constructed for each language and split combination, enabling independent loading and filtering during training and evaluation.

3.3.2 Span Length Analysis

One of the first things to examine in a span-level detection task is how long the spans actually are, since this directly affects decisions about window size, post-processing filters, and expected model behaviour. Table 3.3 presents summary statistics for span length across all language partitions.

Portuguese stands out immediately: it has the longest spans of any partition, with a mean of 88.2 characters, 14.5 words, and 21.1 tokens. English, the largest training set, has the shortest spans at 46.3 characters on average, nearly half the Portuguese mean. This

Table 3.3: Span length statistics per language partition and Euclidean distance to the Portuguese gold set. Distance is computed from the mean length profile (characters, words, tokens). Languages are ordered by distance to Portuguese.

Partition	Spans	Mean length			Median length			Dist. to PT
		Char	W	Tok	Char	W	Tok	
PT gold	1,727	88.2	14.5	21.1	63	11	15	0.0
RU train	4,138	78.1	11.3	20.3	54	8	14	10.7
PL train	3,824	73.2	10.7	19.9	52	8	14	15.5
DE train	5,737	66.9	9.4	16.1	39	5	10	22.5
FR train	7,181	64.9	10.6	17.1	41	7	11	24.0
IT train	7,961	57.4	9.1	14.1	40	6	10	32.0
EN train	9,002	46.3	7.7	11.1	24	4	6	43.7

difference likely reflects stylistic conventions in Portuguese journalism, where persuasive constructions tend to be expressed through longer syntactic structures. From a practical standpoint, this means that the length filtering thresholds used in post-processing (Section 3.8) need to be calibrated to Portuguese rather than to the training languages.

The last column of Table 3.3 shows the Euclidean distance between each language’s mean span length profile and Portuguese. Russian and Polish are structurally closest, while English is the most distant. This ordering is notably different from the linguistic proximity ranking established through the embedding analysis in Section 3.4, where Italian and French dominate, suggesting that span length patterns and linguistic family membership capture different dimensions of cross-lingual similarity.

3.3.3 Technique Frequency Distributions

The distribution of techniques across the corpus is where the most actionable insights emerge. Figures 3.2 and 3.3 compare the English training set and the Portuguese gold test set, while Figures 3.4 and 3.5 show the corresponding percentage breakdowns.

The contrast is stark. In the English training data, Loaded Language (33.5%) and Name Calling (17.1%) together make up over half the spans. In Portuguese, the hierarchy inverts: Questioning_the_Reputation leads at 19.1%, while Loaded Language drops to third at 15.1%. A model trained only on English therefore learns to predict Loaded Language aggressively while having no signal for the most prevalent Portuguese technique. This mismatch is perhaps the single strongest argument for the multi-source training strategy adopted in this work.

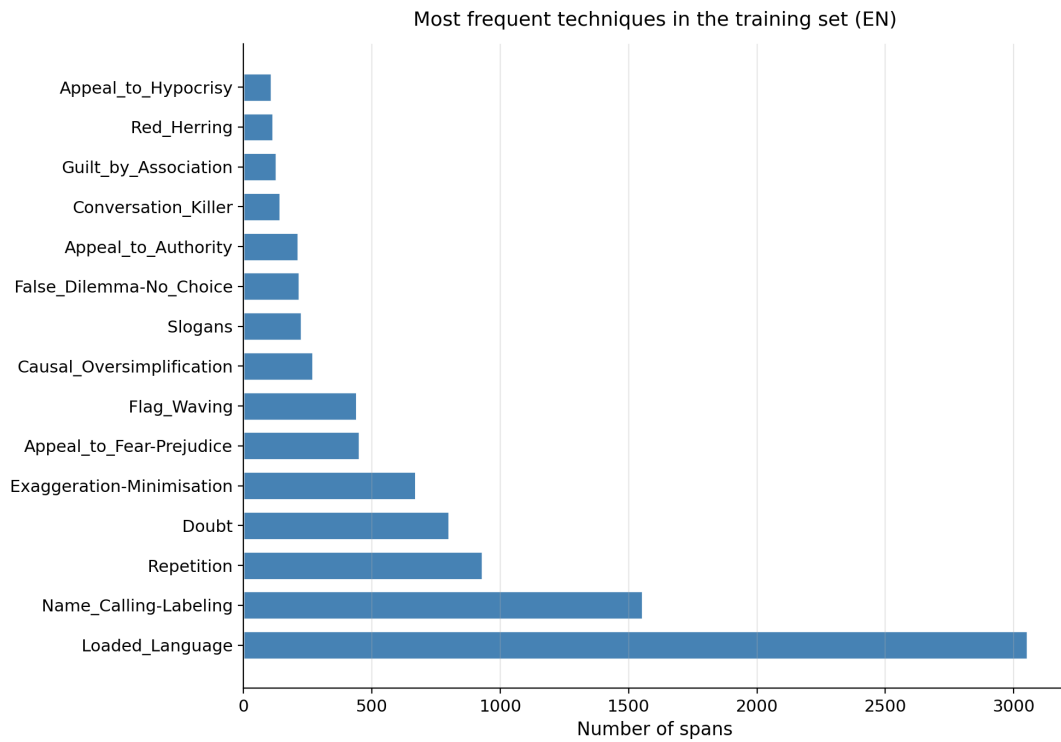


Figure 3.2: Top 15 most frequent techniques in the English training set. Loaded Language and Name Calling together account for over 50% of all spans.

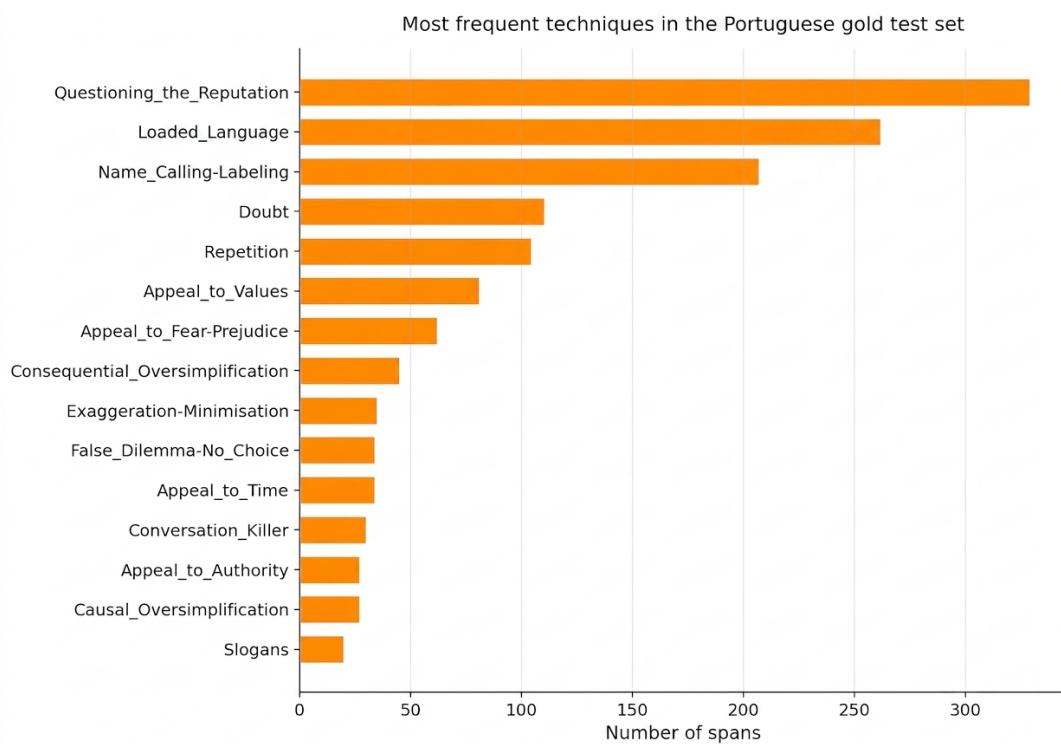


Figure 3.3: Top 15 most frequent techniques in the Portuguese gold test set. Questioning_the_Reputation dominates, a category entirely absent from English training data.

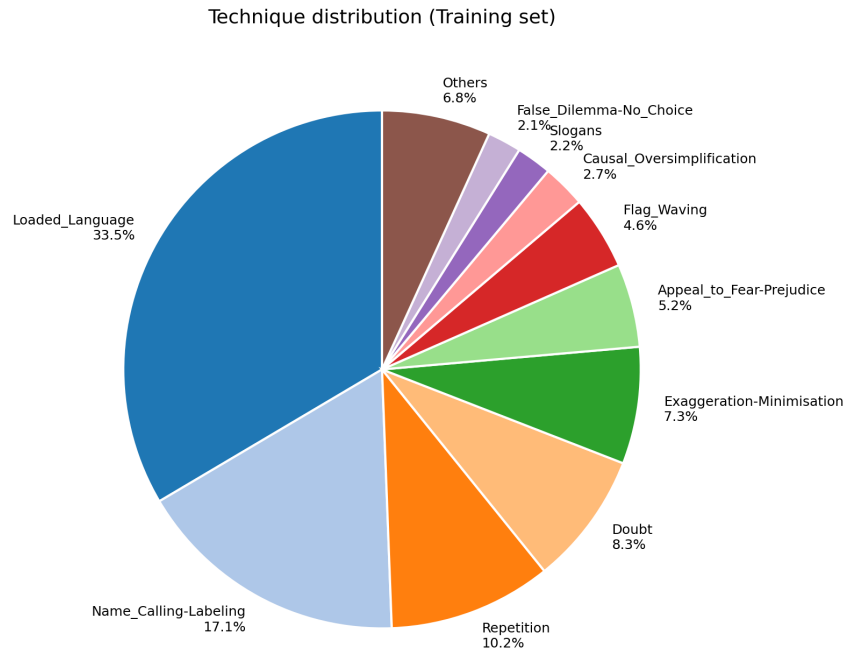


Figure 3.4: Percentage distribution of techniques in the English training set.

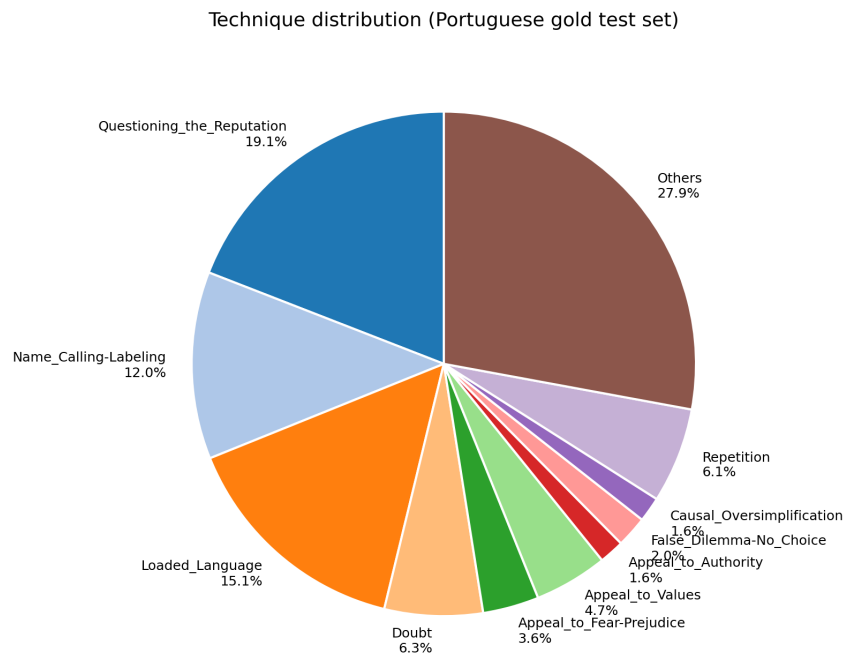


Figure 3.5: Percentage distribution of techniques in the Portuguese gold test set.

3.3.4 Topic Distribution Across Languages

The technique frequency analysis above focuses on what is being detected. An equally important question is what the articles are about, since persuasion techniques tend to cluster around specific topics. Table 3.4 and Figures 3.6 and 3.7 present the topic distribution.

Table 3.4: Topic distribution (%) per language partition.

Language	Abort.	Climate	Covid	Elect.	Israel-P.	Migrat.	Ukraine-R.
German (DE)	0.5	13.7	18.5	8.1	0.9	10.4	47.9
English (EN)	1.4	2.1	5.4	39.9	7.9	21.0	22.4
French (FR)	4.5	12.4	14.9	9.9	0.8	13.6	43.8
Italian (IT)	0.9	9.6	16.4	9.1	0.3	10.8	52.9
Polish (PL)	7.0	17.5	17.5	10.0	0.0	5.0	43.0
Portuguese (PT)	11.5	10.6	11.5	13.5	23.1	10.6	19.2
Russian (RU)	5.4	2.1	7.9	5.0	2.5	0.4	76.8

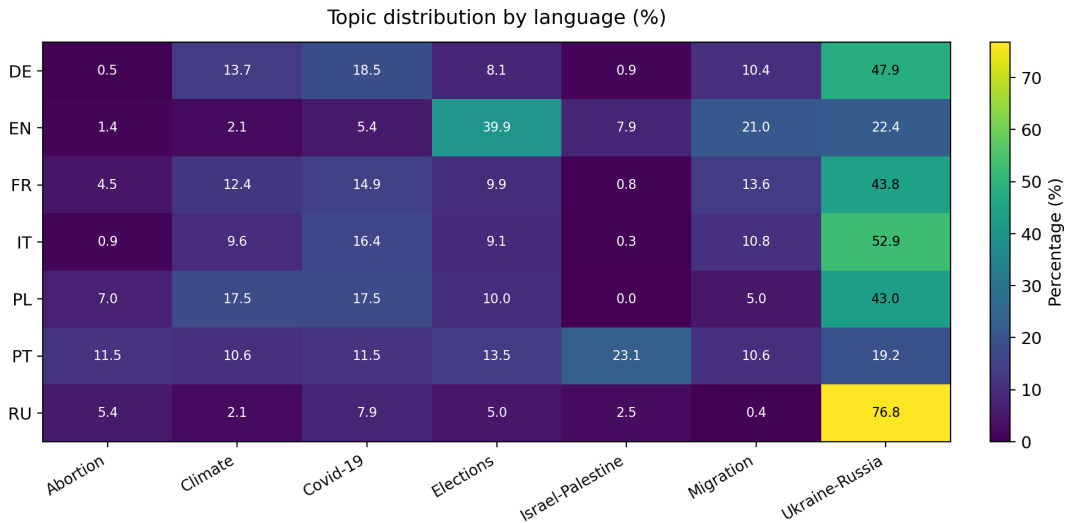


Figure 3.6: Heatmap of topic distribution (%) across language partitions.

Three patterns stand out. First, Russian is overwhelmingly dominated by the Ukraine-Russia topic (76.8%), which means that despite having reasonable span-length proximity to Portuguese, Russian data is topically very narrow. Second, the English partition is heavily skewed toward Elections (39.9%), reflecting the US political context during data collection, while Portuguese Elections coverage (13.5%) reflects a Portuguese parliamentary cycle. Third, and perhaps most importantly, Portuguese has the highest Israel-Palestine coverage of any language at 23.1%, while most training languages have near-zero representation on this topic. No single source language captures the full thematic breadth of the Portuguese test set.

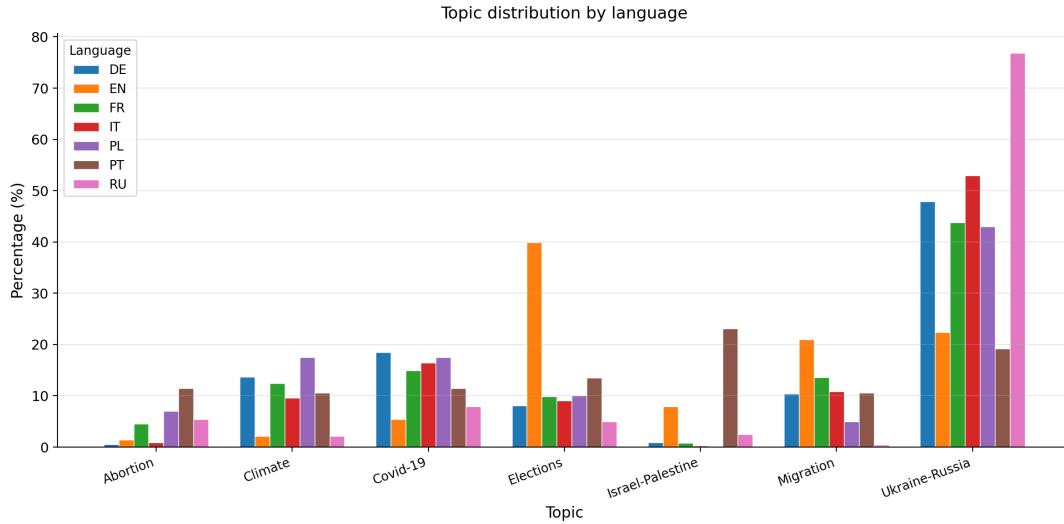


Figure 3.7: Grouped bar chart of topic distribution by language.

3.3.5 Class Imbalance

The exploratory analysis confirms that class imbalance in this dataset is severe and operates at multiple levels simultaneously. At the token level, only 0.7% of token-label pairs are positive, meaning that 72% of all predictions should be O (outside). At the technique level, the top three categories account for over half of all annotated spans, while the five rarest categories together represent less than 4%.

These imbalances, compounded by the cross-lingual variation documented above, motivate both the weighted loss function described in Section 3.6 and the multi-source training strategy in Section 3.7.

3.4. Embedding-Based Language Proximity Analysis

The exploratory analysis revealed that the Portuguese test set has a fundamentally different technique distribution from the English training data, and that no single training language provides complete thematic coverage. But it did not directly answer the question of which language, or combination of languages, is most similar to Portuguese at the document level. To address this, an embedding-based proximity analysis was conducted before any model was trained (Figure 3.1, middle level).

3.4.1 Methodology

Document embeddings were generated using the `paraphrase-multilingual-MiniLM-L12-v2` model from the `sentence-transformers` library [16], which projects texts from all languages into a shared vector space where cosine similarity can be computed directly

across languages. For each document, a combined feature vector was created by concatenating the textual embedding with the document’s normalised technique distribution (the 23-dimensional vector of its per-technique span counts, divided by the total number of spans in the document so that the entries sum to one), with the technique block scaled by a factor of 5.0 so that it contributes meaningfully alongside the higher-dimensional embedding. This design ensures that similarity captures both topical overlap and rhetorical strategy alignment, which is particularly relevant for propaganda detection.

For each of the 104 Portuguese gold documents, the 20 nearest training documents were identified by cosine similarity across all six source languages, yielding 2,080 nearest-neighbour pairs. The language distribution of these neighbours was then analysed globally, per document, and per technique. A 2D UMAP projection was also produced for visual inspection.

The full quantitative results are reported in Chapter 4; here it suffices to note that the analysis was designed to inform two specific decisions: whether Stage 1 should include all six languages or only a subset, and which machine-translated dataset to prioritise for Stage 2.

3.5. Baseline Systems

Three baseline systems were developed, each designed to test a different hypothesis about what matters most for this task. The two encoder-based baselines test whether the choice of encoder is more important than the output architecture, while the generative baseline tests whether a state-of-the-art LLM can match fine-tuned performance through prompting alone.

3.5.1 mBERT Baseline

The simplest reasonable approach is to take an off-the-shelf multilingual encoder and train it on the available data using the standard BIO formulation. The first baseline does exactly this, using `bert-base-multilingual-cased` (mBERT) with a softmax head over the 47-class BIO label set.

Training follows two stages: Stage 1 on English data, Stage 2 fine-tuning on IT→PT translated data. Italian was chosen over French for Stage 2 based on preliminary experiments showing consistently higher performance, consistent with the embedding analysis findings.

3.5.2 XLM-R Baseline

The second baseline replaces the encoder with `xlm-roberta-base` (XLM-R) while keeping the same BIO softmax head and training procedure. If XLM-R substantially outperforms mBERT under these conditions, it would suggest that encoder quality is the main bottleneck. If not, the architecture is the constraint. The results in Chapter 4 point clearly to the latter.

3.5.3 Generative Baseline: Gemini 2.5 Flash

Gemini 2.5 Flash, a recent instruction-tuned generative LLM from Google DeepMind, was evaluated on the Portuguese test articles in a few-shot setting with no fine-tuning. The model was accessed via the `langextract` library and received a structured prompt specifying all 23 technique categories, with explicit instructions to extract only exact text spans and assign each to one predefined label. Twenty-three few-shot examples in Portuguese were included (one per technique). Due to API rate limits, articles were processed sequentially with a 30-second pause between requests.

3.5.4 Limitations of the BIO Formulation

The two encoder baselines share structural limitations that no amount of training data or encoder improvement can resolve. The softmax head assigns at most one label per token, creating a hard recall ceiling for the co-occurring technique annotations described in Section 3.1. Meanwhile, the extreme class imbalance documented in Section 3.3 pushes the model toward predicting nothing, resulting in very low recall. These are the specific problems that the proposed architecture, described next, is designed to solve.

3.6. Proposed Architecture

The architecture of the proposed system, shown in Figure 3.8, is a direct response to the three problems identified through the baselines and the exploratory analysis: the need for cross-lingual transfer, the multi-label nature of the annotations, and the extreme class imbalance.

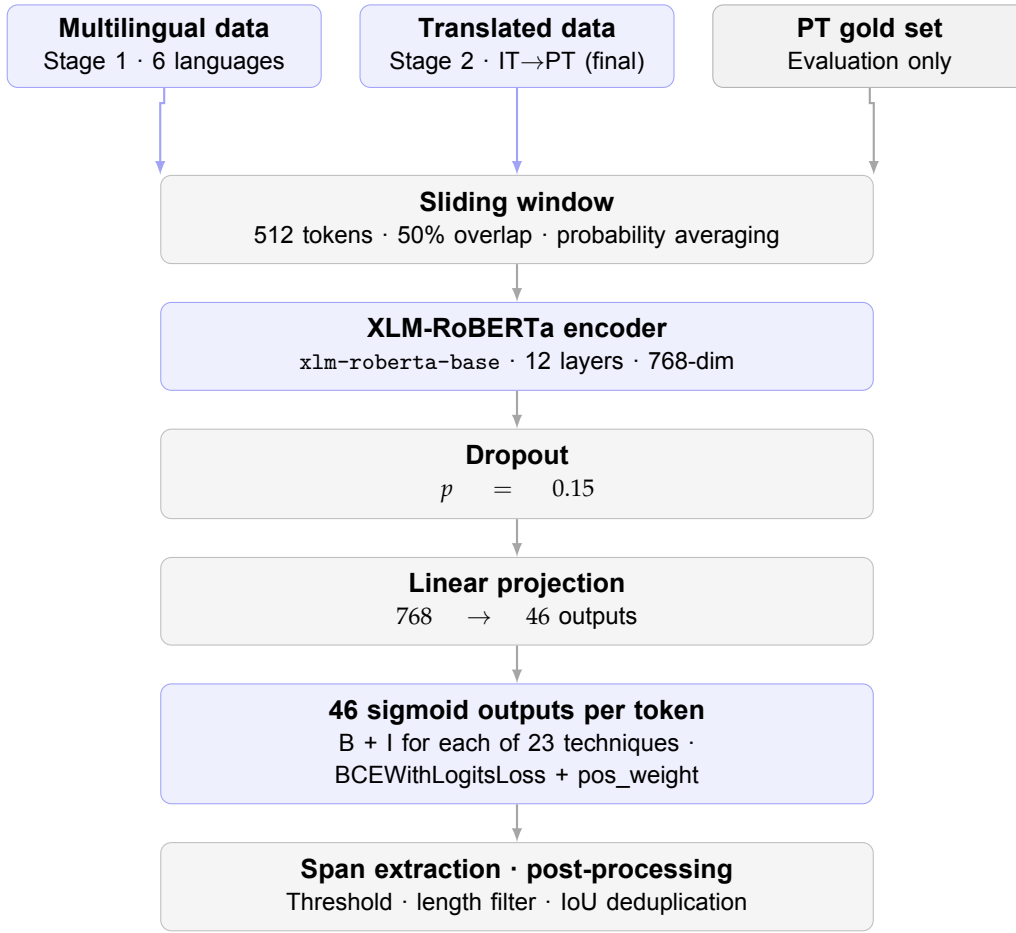


Figure 3.8: Architecture of Model 2. Indigo boxes represent the key components: the multilingual encoder and the multi-label sigmoid head.

3.6.1 Encoder: XLM-RoBERTa

The encoder is `xlm-roberta-base`, chosen for its superior cross-lingual transfer capabilities compared to `mBERT`, as established in Section 2.4.2 and confirmed by Conneau et al. [8]. For a zero-resource target language, the quality of the shared multilingual representation space is the single most important property of the encoder, and XLM-R provides the best available starting point.

3.6.2 Multi-Label Sigmoid Output Head

The key architectural innovation is replacing the softmax head with 46 independent sigmoid outputs: one B (Beginning) and one I (Inside) output for each of the 23 techniques. Because each sigmoid is independent, the model can predict any combination of techniques for a given token, matching the multi-label structure of the gold annotations. The head consists of a Dropout layer ($p = 0.15$) followed by a linear projection from 768 dimensions to 46 scalar outputs.

3.6.3 Per-Output Positive Weighting

To counteract the extreme class imbalance identified in Section 3.3, BCEWithLogitsLoss is used with a per-output `pos_weight`: each of the 46 binary outputs receives an individual weight computed from the ratio of negative to positive instances for that specific output on the Stage 1 training data, capped between 10 and 100. These weights are derived only from the training data and never from the Portuguese test set, so they introduce no test-set leakage. In practice, all 23 begin outputs reach the ceiling of 100, as do the inside outputs of the rarer techniques; only the inside outputs of the most frequent techniques fall below it, for example Doubt at 15.3, Questioning the Reputation at 24.7, and Loaded Language at 32.4. In total, 36 of the 46 outputs sit at the ceiling and none at the floor of 10, giving a median weight of 100.0 and a mean of 91.3. This calibration prevents the all-zeros degenerate solution that standard BCE produces under extreme imbalance.

3.6.4 Sliding Window

Since 94% of articles exceed the 512-token context window of XLM-R, each document is split into overlapping windows of 512 tokens with a stride of 256 (50% overlap), following the task organisers' approach [6]. During inference, sigmoid probabilities for tokens appearing in multiple windows are averaged, which reduces boundary artefacts at window edges.

3.7. Training Pipeline

Training uses a two-stage approach: **Stage 1** builds broad multilingual representations by training on multiple source languages, while **Stage 2** fine-tunes specifically for Portuguese using machine-translated data. This procedural structure is shared by two distinct versions of the pipeline. The first, **Model 1**, served as a proof of concept for the multi-label sigmoid architecture. In contrast, **Model 2** incorporated the findings from the embedding analysis to expand the training data and refine the loss weighting.

While the data and specific configurations within each stage distinguish the two models, both rely on the same machine-translation process for Stage 2 fine-tuning. Therefore, before detailing each pipeline's configuration, summarised in the lower levels of Figure 3.1, it is worth describing how this translated data was produced.

3.7.1 Machine-Translated Data

To bring training data closer to the target language, machine-translated versions of three source language corpora were produced. The Italian and French training sets were translated into Portuguese using Google Translate while the Russian training set was translated using DeepL with the PT-PT (European Portuguese) target variant. In all three cases, translation was performed at the article level, splitting long texts into paragraph-level chunks to respect API limits.

The more challenging step was transferring the span boundary annotations from the source text to the translated text. For this, LinguAligner was used, an annotation alignment tool that combines multiple alignment strategies in a configurable pipeline. The configuration used in this work consisted of four sequential stages: lemma matching, word-level alignment using `bert-base-multilingual-uncased`, Gestalt pattern matching, and Levenshtein distance matching. For each annotated span in the source language, the aligner identified the corresponding text fragment in the Portuguese translation and returned its character-level offsets. Spans for which the alignment failed or produced invalid offsets ($\text{start} \geq \text{end}$, or end beyond the document length) were discarded.

Across the three translation methods, alignment discarded a substantial fraction of the source spans: 3,226 of the 7,961 Italian spans (40.5%), 3,015 of 7,181 for French (42.0%), and 1,042 of 4,138 for Russian (25.2%). Most of these discards correspond to spans whose boundaries could not be reliably projected onto the translated text, which is the main source of the annotation noise discussed below.

This process produced three translated datasets: IT→PT (302 documents, 4,735 spans), FR→PT (208 documents, 4,166 spans), and RU→PT (190 documents, 3,096 spans). While the multi-stage alignment pipeline accurately preserves most span boundaries, it inevitably introduces some noise, particularly for techniques that depend on word order or subtle lexical choices that do not translate cleanly. Despite this limitation, the translated data proved to be an effective fine-tuning resource, as the Stage 2 results in Chapter 4 demonstrate.

3.7.2 Model 1

Model 1 was the first system to adopt the multi-label sigmoid architecture and established that the approach was viable. **Stage 1** trained on English (536 documents) combined with IT→PT (302 documents), producing approximately 3,294 sliding window instances.

At this point, the embedding analysis had not yet been conducted, so the data selection was based on the assumption that English, as the largest dataset, combined with Italian-translated data, as the most linguistically plausible Romance language proxy, would provide a reasonable starting point. A global `pos_weight` of 80 was applied uniformly to all 46 outputs. Hyperparameters: learning rate 3×10^{-5} , batch size 8 with gradient accumulation 2 (effective batch size 16), 10 epochs with early stopping (patience 3), warmup ratio 6%, weight decay 0.01, fp16 mixed precision. **Stage 2** fine-tuned on IT→PT with encoder embeddings frozen, learning rate reduced to 1×10^{-5} , up to 15 epochs with early stopping (patience 3).

3.7.3 Model 2: Final System

Model 2 was built on Model 1 by incorporating two changes motivated by the embedding analysis and the class imbalance observations from the EDA. First, Stage 1 was expanded to include all six source languages rather than just English, ensuring that the model had exposure to every technique category present in the Portuguese test set. Second, the global `pos_weight` was replaced by per-output weights computed from the actual class distribution, allowing rare techniques to receive a stronger gradient signal without destabilising training on frequent categories.

Stage 1 included IT (303 documents), FR (211), EN (536), DE (177), PL (194), RU (191), IT→PT (302), FR→PT (208), and development sets for FR, DE, IT, PL, and RU, totalling approximately 10,000 sliding window instances. All other hyperparameters remained identical to Model 1. **Stage 2** fine-tuned on IT→PT only, following the language selection experiments described in Section 3.7.4. Table 3.5 summarises the key differences between the two configurations.

Table 3.5: Comparison of training pipeline configurations.

Component	Model 1	Model 2
Stage 1 data	EN + IT→PT (~3,294 windows)	6 languages + translations + dev sets (~10,000 windows)
Stage 2 data	IT→PT only (~919 windows)	IT→PT only (~919 windows)
<code>pos_weight</code>	Global: 80	Per-output: 30–100
Threshold	Global	Global

3.7.4 Stage 2 Language Selection Experiments

After Stage 1, three Stage 2 configurations were tested using the same Model 2 checkpoint:

- **IT→PT only:** Italian-translated data (302 documents)
- **IT→PT + FR→PT:** Italian plus French (510 documents)
- **IT→PT + FR→PT + RU→PT:** adding Russian to the above

The results, presented in Chapter 4, show that Italian alone produces the best performance, with each additional language degrading F1. This outcome, consistent with the embedding analysis, led to IT→PT being selected for the final system.

3.8. Inference and Post-Processing

During inference, documents are processed with the same sliding window strategy used in training, and sigmoid probabilities are averaged across overlapping windows. Consecutive tokens above the global threshold are grouped into technique-specific spans. Three post-processing filters are applied: **length filtering** removes spans shorter than 2 or longer than 600 characters (calibrated to the Portuguese gold distribution described in Section 3.3); **punctuation removal** discards punctuation-only spans; and **deduplication** retains only the higher-scoring span when two spans of the same technique overlap with IoU above 0.5.

The global threshold is determined by grid search over 0.10–0.70 on the Italian development set, maximising F1-micro. Per-technique thresholds were explored but discarded as they constitute data leakage: in the competition setting, participants had no access to gold test labels for threshold tuning.

3.9. Evaluation Protocol

Evaluation uses the official CLEF-2024 partial span matching metric [6]. For each predicted span p and gold span g of the same technique:

$$S(p, g) = L(p, g) \times I(p, g) \quad (3.1)$$

where $L(p, g)$ penalises predictions longer than twice the gold span, and $I(p, g)$ measures the proportion of the gold span covered. Predictions covering at least 50% of a gold

span with acceptable length receive full credit. Matching is greedy by maximum score, and both micro-averaged and macro-averaged F1 are computed. Given the severe class imbalance, both metrics are necessary: micro-F1 reflects overall performance dominated by frequent techniques, while macro-F1, which weights all 23 categories equally, reveals whether the system generalises beyond the head of the distribution.

With the methodology, data, and evaluation framework now fully described, Chapter 4 presents the results of applying this pipeline to the Portuguese test set.

4. Results and Discussion

The results presented in this chapter cover four distinct but interconnected aspects of the experimental work. The chapter begins with an embedding-based language-proximity analysis, whose findings shaped the design of the training pipeline before any model was trained. It then examines how the systems behaved during training, followed by a comprehensive evaluation of all systems on the Portuguese test set at both the global and per-technique levels. The chapter closes with a threshold sensitivity analysis and a discussion that draws together the main findings.

4.1. Embedding-Based Language Proximity Analysis

The first and perhaps most consequential finding of this work concerns not the models themselves, but the data used to train them. Before any fine-tuning was performed, the embedding-based similarity analysis described in Section 3.4 was conducted to determine which of the six available source languages was most relevant to the Portuguese target domain.

As detailed there, document embeddings were generated using `paraphrase-multilingual-MiniLM-L12-v2` and combined with normalised technique distributions 3.4.1, and the 20 nearest training documents were identified by cosine similarity for each of the 104 Portuguese gold documents. Italian accounts for 39.5% of all nearest neighbours, followed by French (17.0%), Russian (13.2%), Polish (11.3%), English (10.8%), and German (8.2%). Together, Italian and French represent 56.5% of all neighbours, a majority accounted for by the two Romance languages alone.

The contrast between English and Italian is worth emphasising. English is the language most present in the training dataset by a substantial margin, with 536 documents and 9,002 annotated spans. Italian has 303 documents and 7,961 spans. Yet English accounts for only 10.8% of nearest neighbours to Portuguese documents, while Italian, from a smaller dataset, accounts for nearly four times as much. This result provides direct empirical evidence that **linguistic proximity outweighs dataset size** as a predictor of transfer quality in this setting.

Aggregated statistics, however, can be misleading. While Italian dominates globally, the picture at the individual document level is considerably more varied. Figure 4.2 shows the

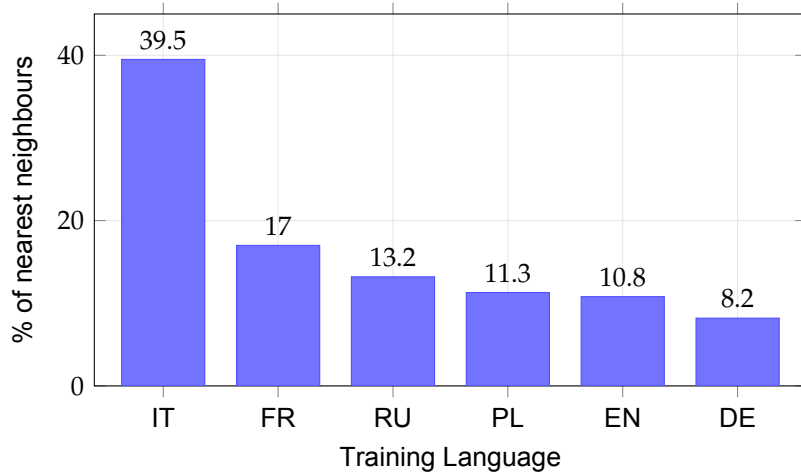


Figure 4.1: Distribution of training languages among the top-20 nearest neighbours for each of the 104 Portuguese gold documents. Italian dominates at 39.5% despite not being the largest dataset.

dominant training language for each of the 104 Portuguese documents separately. Italian is the primary language for 55 documents (52.9%), but for the remaining 49 (47.1%), another language is more relevant. French dominates for 14 documents, English for 13, Russian for 10, German for 7, and Polish for 5.

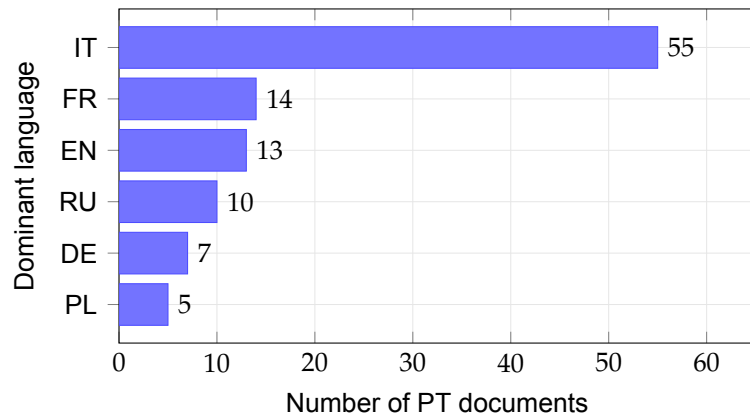


Figure 4.2: Number of Portuguese gold documents for which each training language is the dominant nearest neighbour (n=104).

This heterogeneity is the reason why restricting training to a single language, even the globally dominant Italian, would be suboptimal for Stage 1. Almost half of the Portuguese test documents are better served by a different source. This finding was the direct motivation for the multi-source Stage 1 training adopted in Model 2.

The analysis also reveals interesting patterns at the technique level. Figure 4.3 presents the language distribution broken down across all 23 persuasion techniques. For each technique, the Portuguese gold documents that contain it are collected, and the source

languages of their 20 nearest neighbours are aggregated and normalised to percentages, so that each row sums to 100% and shows where the training signal for that technique is most likely to originate. Italian remains dominant for every technique category, ranging from 37% for Appeal_to_Authority and Whataboutism to 55% for Guilt_by_Association. English contributes above-average proportions for Obfuscation-Vagueness-Confusion (23%), Repetition (16%), and Conversation_Killer (15%). French is strongest for Appeal_to_Popularity (29%), Slogans (22%), and Straw_Man (22%). These patterns suggest that different techniques have different linguistic homes across the training data, and that no single source language provides complete technique coverage for Stage 1.

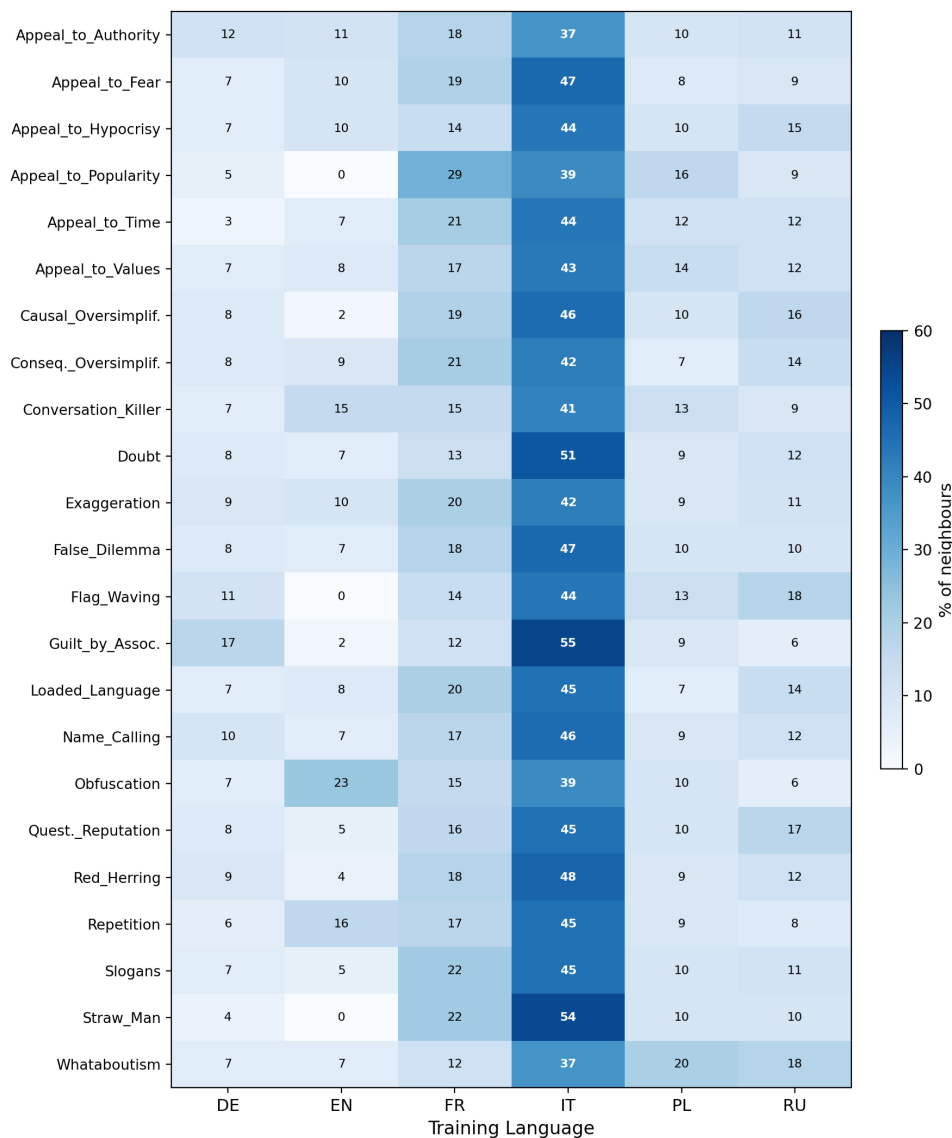


Figure 4.3: Per-technique language distribution of nearest training neighbours (%). Italian dominates consistently across all 23 techniques, but notable variation in secondary languages motivates including all source languages in Stage 1.

Figure 4.4 offers a complementary geometric perspective on these relationships. The 2D UMAP projection of the embeddings produced by `paraphrase-multilingual-MiniLM-L12-v2` (described in Section 3.4) shows that Portuguese documents occupy the central-right region of the shared multilingual space, clustering predominantly around Italian and French points. English documents form a separate cluster in the lower-left that Portuguese documents rarely approach. Russian documents are more dispersed, with some overlap near the Portuguese region, an overlap likely driven by shared topical content around the Russia-Ukraine war rather than any linguistic affinity. Taken together, the UMAP projection provides an intuitive visual confirmation of what the nearest-neighbour statistics quantify: Portuguese text is simply more similar to Romance language text than to English text, and training data selection should reflect this.

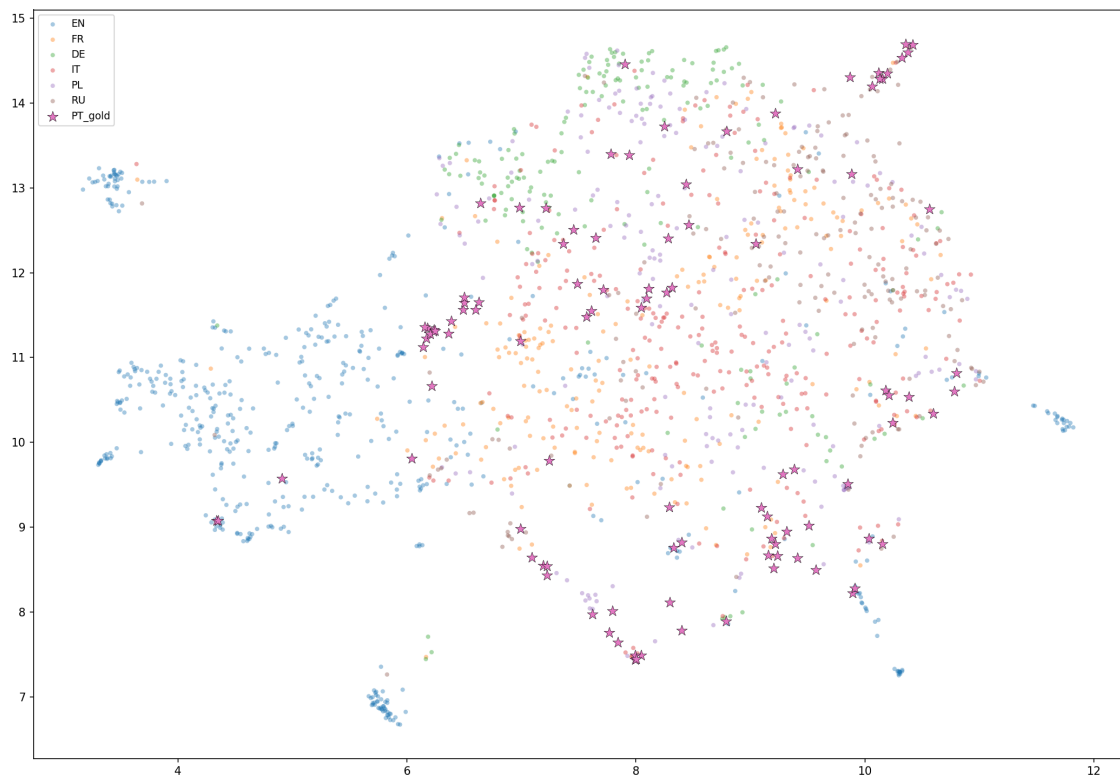


Figure 4.4: UMAP 2D projection of document embeddings. Portuguese gold documents (stars) cluster predominantly with Italian and French. English forms a distinct and more distant cluster.

4.2. Training Dynamics

All training runs were monitored using `Weights & Biases*`, which logged training loss at each step. The loss curves presented in this section (Figures 4.5, 4.6, and 4.7) are derived from these logged values.

*<https://wandb.ai/>

4.2.1 Baseline Training Curves

Figure 4.5 shows the loss curves for the mBERT and XLM-R baselines during both stages of training. During Stage 1 on English data, the two models follow remarkably similar trajectories despite having different encoder architectures, suggesting that the BIO softmax formulation is the binding constraint rather than anything specific to the encoder. The Stage 2 curves on IT→PT data tell a slightly different story: XLM-R consistently maintains a higher training loss than mBERT throughout fine-tuning, foreshadowing its weaker evaluation performance.

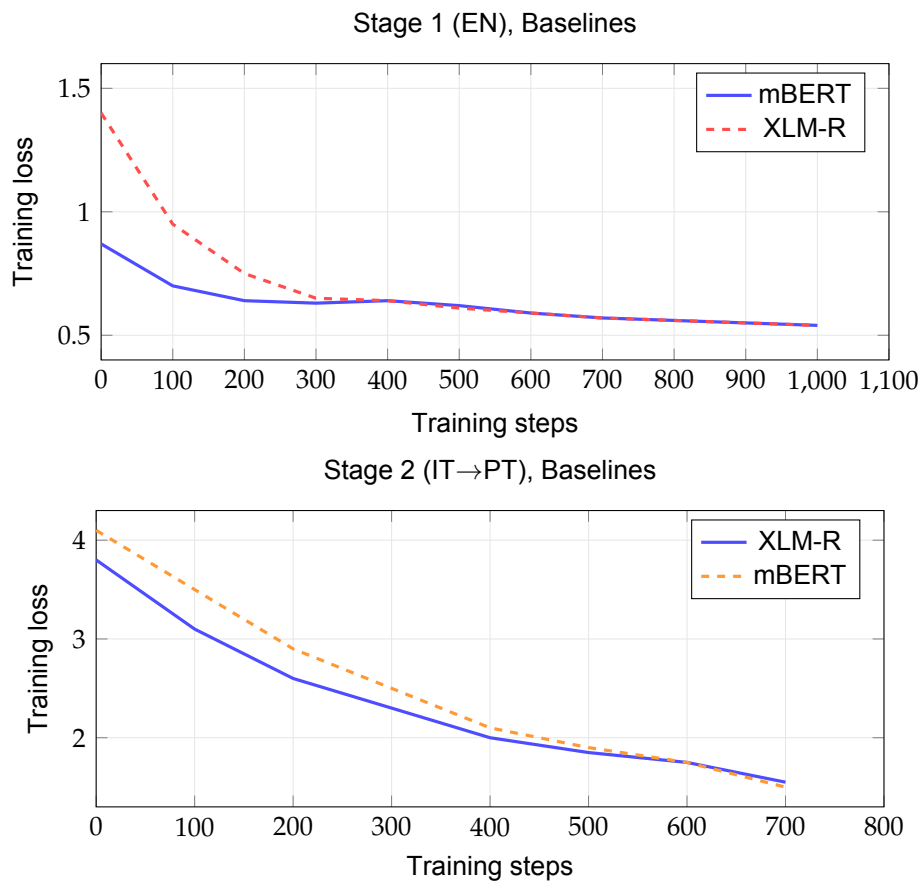


Figure 4.5: Training loss curves for the BIO baseline systems. Left: Stage 1 on English data. Right: Stage 2 fine-tuning on IT→PT.

4.2.2 Proposed Model Training Curves

Figure 4.6 shows the training dynamics for the proposed systems. Model 2 Stage 1 requires approximately 5,000 steps to converge, compared to roughly 2,000 for Model 1, reflecting the approximately three-fold increase in training instances from the expanded multi-source setup. Both Stage 2 runs are strikingly brief by comparison, converging within the first few hundred steps to stable low-loss values, reflecting successful knowledge transfer from Stage 1.

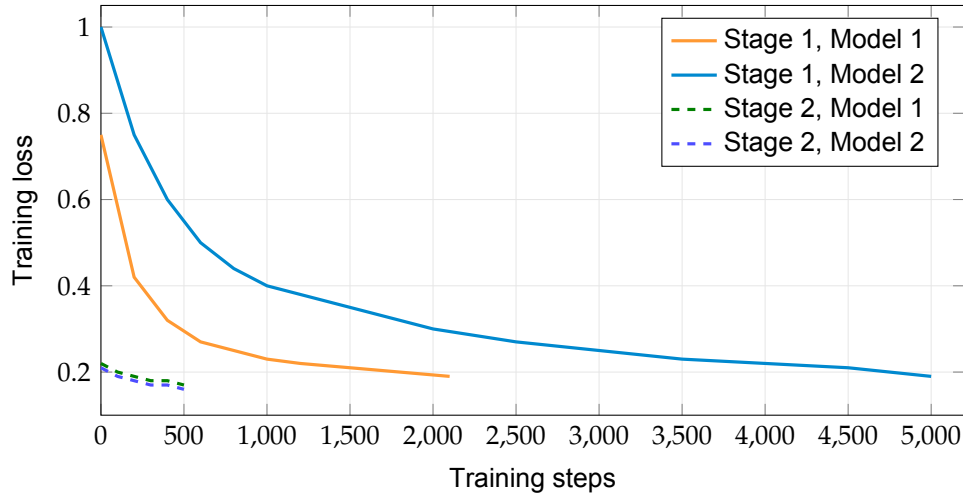


Figure 4.6: Training loss curves for Model 1 and Model 2. Model 2 Stage 1 converges over a longer trajectory due to the larger multi-source dataset. Both Stage 2 runs are brief and stable.

4.3. Stage 2 Language Selection

As described in Section 3.7.4, three configurations were evaluated for Stage 2 fine-tuning. All three start from the same point: the model obtained after completing Stage 1 training on all six source languages. The only variable is the data used for Stage 2 fine-tuning. Table 4.1 presents the results.

Table 4.1: F1 scores on the Portuguese test set for the three Stage 2 configurations tested with the Model 2 Stage 1 checkpoint.

Stage 2 data	F1-micro	F1-macro
IT→PT only	0.134	0.103
IT→PT + FR→PT	0.128	0.099
IT→PT + FR→PT + RU→PT	0.117	0.095

The results reveal a clear and somewhat counterintuitive pattern. Adding French-translated data to Italian-translated data decreases performance, and adding Russian-translated

data decreases it further. The Italian-only configuration achieves the best F1-micro (0.134) and F1-macro (0.103), and was the configuration selected for the final system.

This finding is consistent with the embedding analysis: Italian is the most linguistically proximate source by a substantial margin. The addition of French, despite being the second most similar language globally, appears to introduce noise rather than a complementary signal at the fine-tuning stage. A plausible explanation is that Stage 2 represents a short fine-tuning pass on a small dataset, and the model benefits from a focused, high-quality signal in the closest available language rather than a more diverse but less coherent mixture. The inclusion of Russian, a structurally distant language, degrades performance more sharply, confirming that linguistic distance is the dominant factor at this stage. This is a particularly noteworthy result because it shows that the intuition from Stage 1, where all six languages helped, does not carry over to Stage 2: the dynamics of broad multilingual pre-training and focused cross-lingual fine-tuning are fundamentally different.

Figure 4.7 presents the training loss curves for the three configurations, which help to explain this pattern from a training dynamics perspective. The IT→PT and IT→PT + FR→PT runs both converge to low loss values within a few hundred steps, consistent with their similar F1 scores. The IT→PT + FR→PT + RU→PT run behaves very differently: it starts from a substantially higher loss, requires more than three times as many steps to complete, and converges to a noticeably higher final loss despite the longer training. This is consistent with the addition of Russian-translated data, which introduces a less coherent training signal that the model has more difficulty fitting, ultimately resulting in the lowest F1 score of the three configurations.

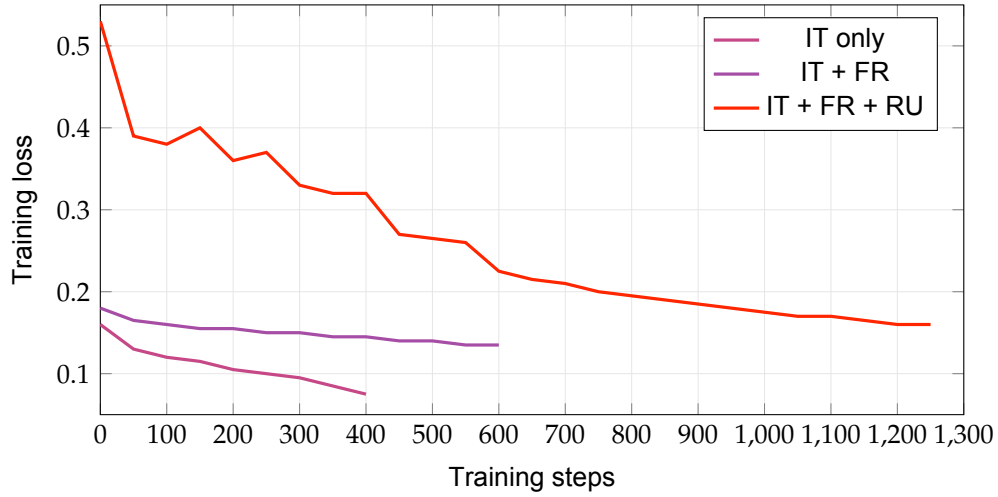


Figure 4.7: Training loss curves for the three Stage 2 configurations, all initialised from the same Model 2 Stage 1 checkpoint. The IT→PT + FR→PT + RU→PT run requires substantially more steps and converges to a higher final loss, consistent with its lower F1 score.

4.4. Overall System Comparison

Table 4.2 brings together the results for all systems evaluated on the Portuguese test set, alongside the competition reference scores. Model 2 here refers to the final configuration with IT→PT-only.

Table 4.2: Overall results on the Portuguese test set. PersuasionMultiSpan* is the organisers’ post-competition system. UniBO was the CLEF-2024 competition winner.

System	Thr.	F1-micro	F1-macro	Precision	Recall	Spans
Model 2	0.70	0.134	0.103	0.101	0.182	2,791
PersuasionMultiSpan*	–	0.132	0.120	–	–	–
Model 1	0.70	0.111	0.071	0.084	0.166	3,136
UniBO (1st place)	–	0.107	0.073	–	–	–
mBERT Baseline	0.40	0.057	0.017	0.054	0.059	1,799
XLM-R Baseline	0.30	0.043	0.005	0.033	0.060	2,873
Gemini 2.5 Flash	–	0.037	0.040	0.156	0.021	228
Baseline (zero-shot)	–	0.002	–	–	–	–

Model 2 achieves F1-micro = 0.134, which places it above the organisers’ post-competition system PersuasionMultiSpan (0.132) and 25% above the competition winner UniBO (0.107). This is a strong result for a system relying entirely on cross-lingual transfer with no native Portuguese training data, and it is the first time in this work that the organiser system is surpassed.

These absolute values are low by the standards of most NLP benchmarks, but this reflects the difficulty of the task rather than a weakness of any single system: the competition winner and the organisers’ post-competition system score in the same range (0.107 and 0.132). This low ceiling is consistent with the intrinsic challenge of the annotation itself. The inter-annotator agreement on this data, measured by Krippendorff’s α , is only 0.404 before curation, well below the 0.667 threshold usually considered reliable, and even for Portuguese, the highest-agreement test language, it reaches just 0.587 (Section 3.2.4). When trained annotators disagree this often on span boundaries and technique labels [2], the performance achievable by an automatic system is also a challenge, and the scores in Table 4.2 are best interpreted relative to this task-specific difficulty and to the competing systems rather than against high-agreement benchmarks.

The baseline systems stand out for how poorly they perform relative to their encoder quality. The XLM-R baseline (0.043) achieves lower F1-micro than the mBERT baseline (0.057) despite using XLM-R, which has been shown to produce superior cross-lingual representations due to its larger and more balanced pre-training corpus (2.5 TB of CommonCrawl data across 100 languages, compared to mBERT’s Wikipedia-only training) [8]. Its macro-F1 (0.005) is also substantially lower. This paradox makes sense once the structural limitations of the BIO formulation are considered: the softmax output collapses XLM-R’s richer representations onto an even narrower set of frequent classes than mBERT. Encoder quality matters less than architecture when the architecture is fundamentally unsuited to the task.

Gemini 2.5 Flash occupies an interesting position. Its overall F1-micro (0.037) places it below both BIO baselines, but its precision of 0.156 is the highest of any system: when Gemini does make a prediction, it tends to be correct. The problem is that it produces just 228 spans against 1,727 in the gold set, collapsing recall to 0.021. The model interprets the instruction to extract only exact spans very conservatively, opting to predict nothing rather than risk an incorrect boundary.

4.5. Per-Technique Analysis

Table 4.3 presents per-technique F1 scores for all systems, sorted by gold span frequency in the Portuguese test set. The per-technique breakdown reveals patterns that the overall F1 scores cannot capture on their own, and some of these patterns are particularly telling.

Table 4.3: Per-technique F1 scores on the Portuguese test set, sorted by gold span frequency. Bold values indicate the best result per technique.

Technique	Gold	Model 2	Model 1	mBERT	XLM-R	Gemini
Questioning_the_Reputation	329	0.214	0.167	0.000	0.000	0.012
Loaded_Language	261	0.052	0.074	0.056	0.000	0.064
Name_Calling-Labeling	207	0.119	0.117	0.174	0.042	0.022
Doubt	109	0.056	0.055	0.000	0.000	0.052
Repetition	106	0.025	0.009	0.000	0.000	0.000
Appeal_to_Values	81	0.195	0.163	0.037	0.007	0.019
Appeal_to_Fear-Prejudice	62	0.142	0.116	0.028	0.026	0.023
Consequential_Oversimplif.	45	0.138	0.129	0.000	0.001	0.030
False_Dilemma-No_Choice	34	0.124	0.133	0.000	0.002	0.000
Appeal_to_Time	34	0.201	0.044	0.000	0.000	0.035
Exaggeration-Minimisation	35	0.035	0.036	0.022	0.008	0.168
Appeal_to_Authority	27	0.086	0.034	0.030	0.000	0.043
Conversation_Killer	30	0.050	0.049	0.009	0.000	0.048
Causal_Oversimplification	27	0.064	0.057	0.000	0.000	0.070
Slogans	20	0.075	0.108	0.007	0.000	0.000
Appeal_to_Hypocrisy	11	0.051	0.055	0.000	0.000	0.100
Obfuscation-Vagueness-Conf.	13	0.024	0.009	0.000	0.000	0.211
Flag_Waving	10	0.061	0.050	0.031	0.000	0.000
Guilt_by_Association	8	0.014	0.000	0.000	0.019	0.000
Appeal_to_Popularity	3	0.043	0.024	0.000	0.000	0.028
Red_Herring	8	0.000	0.000	0.000	0.000	0.000
Straw_Man	6	0.000	0.000	0.000	0.000	0.000
Whataboutism	3	0.000	0.000	0.000	0.000	0.000
Micro F1		0.134	0.111	0.057	0.043	0.037
Macro F1		0.103	0.071	0.017	0.005	0.040

The five best-performing techniques for Model 2 are Questioning_the_Reputation (F1=0.214), Appeal_to_Time (0.201), Appeal_to_Values (0.195), Appeal_to_Fear-Prejudice (0.142), and Consequential_Oversimplification (0.138). Four of these five are entirely absent from the English training data. This is a direct reflection of the multi-source Stage 1 training: by incorporating Italian, French, German, Polish, and Russian data, Model 2 gains exposure to exactly the technique distributions that dominate the Portuguese test set but that English data cannot provide. Had training been restricted to English, performance on these categories would likely have been close to zero, as it is for both BIO baselines.

At the other extreme, three techniques resist detection entirely: Red_Herring (8 gold spans), Straw_Man (6 spans), and Whataboutism (3 spans). Every system scores zero on all three. There is no threshold or architectural change that can fix this, because the problem is upstream: there are simply too few positive examples in the training data for the model to learn any meaningful pattern. Eight spans of Red_Herring spread across six training languages gives roughly one example per language, which is not enough to learn from. Addressing these tail categories would require collecting or generating additional annotated examples specifically for them.

Comparing Model 2 and Model 1 technique by technique isolates the effect of the expanded multi-source Stage 1. Model 2 outperforms Model 1 on 17 of the 20 techniques with at least one gold span. The largest gains are on Appeal_to_Time (+0.157), Questioning_the_Reputation (+0.047), and Appeal_to_Values (+0.032), concentrated precisely on the techniques where Italian and French data contribute most strongly.

One genuinely surprising result comes from Gemini 2.5 Flash. Despite its low overall F1, Gemini achieves the best score of any system on Obfuscation-Vagueness-Confusion (0.211) and Exaggeration-Minimisation (0.168). Two factors likely explain this. First, both techniques have fairly distinctive linguistic properties that can be conveyed precisely in a natural language prompt: Obfuscation involves intentionally vague or non-committal language, and Exaggeration involves obvious overstatement. Second, these are categories where the fine-tuned encoder models themselves perform poorly (Model 2 scores just 0.024 and 0.035 respectively), partly because they are underrepresented in the training data. Gemini's advantage may therefore reflect not so much a structural strength of generative models as a relative weakness of the encoder models on these specific categories, combined with the fact that few-shot prompting does not depend on training data frequency at all.

4.6. Threshold Analysis

The results presented so far were all obtained at a single threshold value per system. However, the choice of threshold has a substantial effect on F1 scores, and the way each system responds to threshold variation reveals something about how well its output probabilities are calibrated.

Figure 4.8 traces the F1-micro of all four systems as the decision threshold varies from 0.10 to 0.70.

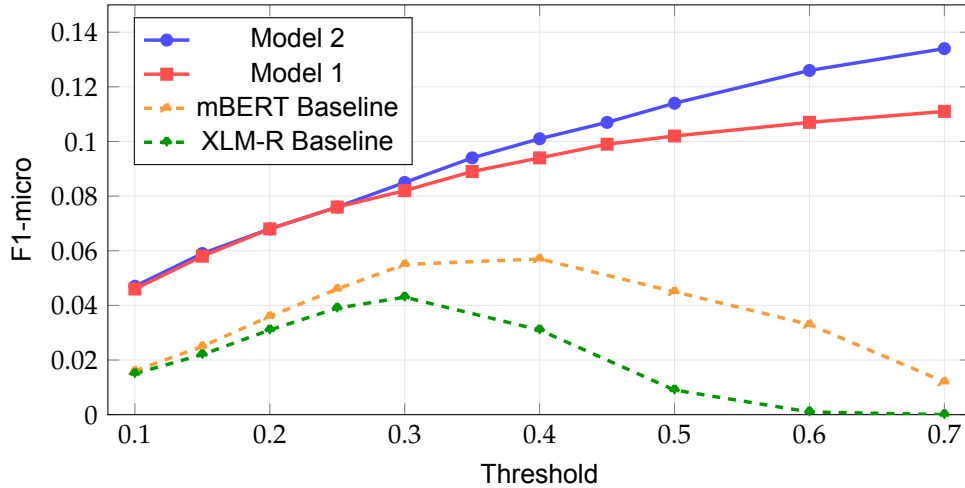


Figure 4.8: F1-micro across the threshold range for all four systems. Model 2 and Model 1 improve monotonically up to 0.70. Both BIO baselines peak early and degrade sharply.

Model 1 and Model 2 improve monotonically across the entire range, with F1-micro continuing to rise from threshold 0.10 all the way to 0.70. This is the signature of a well-calibrated model: as the threshold increases, marginal predictions are filtered out, and those that remain tend to be correct. The two BIO baselines tell the opposite story, peaking at low thresholds (mBERT at 0.40, XLM-R at 0.30) and degrading sharply afterwards. The XLM-R baseline collapses to essentially zero by threshold 0.70, reflecting the over-confidence problem of softmax normalisation: predictions that pass a high threshold are not substantially more reliable than those that pass a low one.

The global threshold of 0.70 was selected on the Italian development set, the source language closest to Portuguese, and applied once to the Portuguese test set; the value coincided with the optimum observed on the Portuguese data in Figure 4.8.

4.7. Ablation Analysis

Having examined the behaviour of each system individually, it is useful to step back and quantify the cumulative contribution of each design decision. Table 4.4 reorganises the main systems into an ablation view, tracking the cumulative impact of each architectural and data choice from the BIO baseline to the final Model 2.

The dominant contribution comes from replacing the BIO softmax head with the multi-label sigmoid architecture and global positive weighting, which alone accounts for a +0.068 gain in F1-micro. The second step, expanding to all six source languages with per-output positive weighting and IT→PT-only Stage 2, adds +0.023 in F1-micro and +0.032 in F1-macro. The macro gain (+45.1%) is proportionally much larger than the micro gain (+20.7%).

Table 4.4: Cumulative F1 gains from each key design decision on the Portuguese test set.

System / Change	F1-micro	F1-macro
XLM-R Baseline (BIO softmax)	0.043	0.005
+ Multi-label sigmoid + global pos_weight (Model 1)	0.111	0.071
+ Per-output pos_weight + all 6 languages + IT→PT Stage 2 (Model 2)	0.134	0.103
Total gain over BIO baseline	+0.091	+0.098

4.8. Discussion

4.8.1 Contribution of Each Design Decision

The ablation results in Table 4.4 confirm the central hypothesis of this work: the BIO formulation, not the encoder or the training data, is the primary bottleneck for span-level persuasion technique detection. The architectural shift from softmax to sigmoid is responsible for the majority of the overall gain. The fact that multi-source training primarily benefits minority techniques, as reflected by the disproportionately larger macro-F1 gain (+45% vs +21% in micro-F1), reinforces the argument made in Section 4.1: different source languages cover different parts of the technique taxonomy, and including all of them ensures that the model has at least some exposure to every category present in the Portuguese test set.

The Stage 2 experiments reveal an important asymmetry between the two training stages. In Stage 1, including all six languages clearly helps: the broader coverage of technique distributions and document types provides a richer initialisation. In Stage 2, the opposite is true: restricting fine-tuning to the single most linguistically proximate language produces the best results. Adding French-translated data hurts, and adding Russian hurts more.

The most plausible explanation relates to the nature of Stage 2 itself. It is a short, focused fine-tuning pass on a small dataset, designed to shift the model’s predictions from a multilingual distribution toward Portuguese specifically. In this context, signal quality matters more than quantity, and Italian-translated data provides the cleanest, most relevant signal. A valid hypothesis is that French-translated data introduces a slightly different distribution that partially conflicts with the Italian signal without adding enough unique value to compensate, though confirming this would require a more detailed analysis of the per-technique overlap between the two translated datasets. Russian-translated data is structurally distant enough to actively degrade the Portuguese-specific representations built up during Stage 2. In the context of this task and target language, the optimal language

selection strategy for Stage 2 proved to be more restrictive than for Stage 1. Whether this asymmetry generalises to other target languages and tasks remains an open question that would benefit from further investigation.

The embedding analysis in Section 4.1 already demonstrated that linguistic relatedness matters more than data volume in this setting. The Stage 2 results reinforce this conclusion from a different angle: not only does Italian dominate the nearest-neighbour distribution, but using only Italian-translated data for fine-tuning outperforms configurations that add more data from less related languages.

The limitations of this work, including assumptions about data availability and the sensitivity of generative baselines to prompt design, are discussed in Section 5.3.

5. Conclusions and Future Work

This dissertation set out to answer a practical question with a clear constraint: can persuasion techniques be automatically detected in European Portuguese news articles when no annotated Portuguese training data exists? The results presented in Chapter 4 suggest that this is feasible, though far from solved: the final system achieves modest but meaningful F1 scores that surpass existing baselines and the competition winner, while several technique categories remain entirely undetectable. The path to these results revealed findings that extend beyond the original question and that may inform future work on cross-lingual transfer for structured prediction tasks.

This chapter begins with a summary of the work conducted, followed by a discussion of the main findings and their implications. It then acknowledges the limitations of the approach and identifies directions for future research. The chapter closes with brief remarks on the broader significance of the work.

5.1. Summary of the Work

The work began with a careful analysis of the task and its inherent challenges. Span-level persuasion technique detection in a zero-resource language is hard for at least three distinct reasons: the multi-label nature of the annotations makes standard sequence labelling architectures structurally inadequate, the severe class imbalance at both the token and technique levels makes standard loss functions trivially exploitable, and the complete absence of Portuguese training data means that every design choice carries an additional risk of language mismatch. Each of these problems required a different solution, and the evolution from the BIO baselines to the final system reflects how they were addressed one by one.

The first challenge was addressed by replacing the BIO softmax output head with a multi-label sigmoid architecture producing 46 independent binary predictions per token, one Beginning and one Inside output token for each of the 23 persuasion techniques. This architectural change, combined with positive weighting to prevent the all-zeros trivial solution, formed the basis of Model 1 and produced the largest single performance gain in the entire work.

The second and third challenges were addressed jointly through the embedding-based language proximity analysis and the multi-source training strategy of Model 2. Rather than assuming that the largest available dataset would provide the best transfer signal, an empirical analysis was conducted to measure the similarity between Portuguese gold documents and the training corpora of each source language. The results showed that Italian was by far the most similar source language to Portuguese in the shared embedding space, despite having a substantially smaller training corpus than English. This finding motivated including all six source languages in Stage 1 training, ensuring broad technique coverage, while restricting Stage 2 fine-tuning to Italian-translated data, which provided the most linguistically coherent signal for the Portuguese-specific fine-tuning objective.

The final system, Model 2 with IT→PT-only, achieves $F1\text{-micro} = 0.134$ and $F1\text{-macro} = 0.103$ on the Portuguese test set, surpassing both the CLEF-2024 competition winner and the organisers' post-competition system (Table 4.2) and establishing a new state-of-the-art result for persuasion technique detection in Portuguese.

5.2. Main Findings

Several findings emerged from this work that have implications beyond the specific task and language studied. They are discussed here in order of their impact on the final system.

The most important finding is also the most actionable: the choice of output architecture matters more than the choice of encoder. The XLM-R BIO baseline, which uses the strongest available multilingual encoder, is actually outperformed by the mBERT BIO baseline because the softmax formulation causes XLM-R's richer representations to collapse onto an even narrower set of frequent techniques. Switching to the multi-label sigmoid head with the same XLM-R encoder more than doubles the $F1\text{-micro}$ of the best BIO baseline. For any future work on span-level persuasion technique detection, the output architecture deserves at least as much attention as the pre-trained model sitting beneath it.

This architectural insight, however, only gets the system so far. The second major finding concerns where the training data comes from. The embedding analysis provided direct empirical evidence that linguistic relatedness is a more reliable predictor of cross-lingual transfer quality than dataset size: Italian, despite having a smaller corpus than English, dominates the nearest-neighbour distribution for Portuguese documents across all 23 technique categories. This challenges a common assumption in multilingual NLP

and has practical implications for low-resource settings, where investing in data from a linguistically close source may yield better returns than accumulating more data in a distant high-resource language.

The relationship between data selection and training stage turned out to be more nuanced than initially expected. While Stage 1 clearly benefits from including all six source languages, providing broader technique coverage and more diverse document representations, Stage 2 behaves differently. The best results come from using only Italian-translated data, and each additional language degrades performance. This asymmetry suggests that when the fine-tuning dataset is small and the objective is language-specific, signal quality matters more than signal diversity, at least in the context of this task and target language.

The comparison with Gemini 2.5 Flash adds another dimension to the picture. The high-precision, low-recall profile of the generative baseline confirms that instruction-tuned LLMs are not yet competitive with fine-tuned encoders for this task. The model produces dramatically fewer spans than are present in the gold set, reflecting a fundamental mismatch between autoregressive generation and the token-level precision that span boundary detection requires. That said, Gemini’s strong performance on Obfuscation-Vagueness-Confusion and Exaggeration-Minimisation suggests that LLMs may retain an advantage on techniques with distinctive, easily describable linguistic markers, an asymmetry that could potentially be exploited in future hybrid systems.

Finally, three techniques, Red_Herring, Straw_Man, and Whataboutism, achieved F1 = 0 across every system evaluated, including the final model. With only 8, 6, and 3 gold spans respectively, no amount of loss weighting or multi-source training can compensate for having so few examples spread across six languages. The model simply has nothing to learn from. Addressing these tail categories remains an open problem for the field and would likely require dedicated data collection or augmentation efforts.

5.3. Limitations

Several limitations should be acknowledged when interpreting the results of this work. The Portuguese test set, while sufficient for reliable overall F1 estimates, is relatively small at 104 documents, and per-technique F1 scores for rare categories are subject to considerable variance. A different sample of Portuguese articles could plausibly shift the scores for low-frequency techniques by several percentage points.

The embedding-based language proximity analysis assumes access to the unannotated Portuguese test documents prior to training. The analysis combined document text embeddings with the normalised technique distribution of each document, which for the Portuguese documents drew on the gold labels; however, these labels only informed source-language selection and were never used during training or model selection, and repeating the analysis with text embeddings alone leaves the central finding unchanged, with Italian remaining the dominant nearest language. This assumption introduces a practical constraint: in a real-world deployment scenario where a new target language emerges and no documents are yet available, the same proximity analysis could not be conducted. The source language selection strategy proposed in this work therefore depends on having at least a small corpus of unannotated target-language text, which may not always be available and limits the scalability of this approach to truly zero-resource settings where not even raw documents exist.

The evaluation of Gemini 2.5 Flash as a generative baseline should be interpreted with the caveat that the results are sensitive to prompt engineering. The specific prompt structure, the choice and number of few-shot examples, the instruction phrasing, and even the output format constraints can all substantially influence the model’s behaviour. A different prompt design could plausibly produce higher recall at the cost of precision, or improve performance on specific technique categories. The results reported here therefore reflect one particular prompting configuration rather than an upper bound on what generative models can achieve on this task.

The training loss curves reported in Chapter 4 were logged using Weights & Biases during the actual training runs, but the plotted data points were sampled at regular step intervals rather than at every individual step. Finally, the three undetectable technique categories identified in Section 5.2 represent a hard limit of the current approach that the directions proposed in Section 5.4 aim to address.

Despite these limitations, the results are robust enough to support the main conclusions of this work and to motivate the directions for future research described next.

5.4. Future Work

The findings of this dissertation open several directions for future investigation, ranging from immediate extensions of the current system to broader research questions.

The most natural starting point concerns the zero-resource assumption itself. This work demonstrates that competitive performance is achievable without any Portuguese training data, but it does not quantify how much that performance would improve if even a small amount of native annotation became available. Would annotating as few as ten or twenty Portuguese documents and incorporating them into Stage 2 produce a meaningful improvement over the Italian-translated data, or does the current system already capture most of the achievable signal? Understanding where that inflection point lies would help determine when investing in native annotation is actually worthwhile, and when the cross-lingual approach developed here is sufficient on its own.

A related direction concerns the quality of the translated training data. The current system uses machine-translated data with character-level span alignment, and while this is effective, the translation process inevitably introduces noise, particularly for techniques that rely on subtle lexical or pragmatic cues. Future work could investigate whether filtering out low-confidence span alignments before training improves the signal-to-noise ratio, whether the quality of translation correlates with downstream performance for specific technique categories, and whether back-translation strategies could generate additional augmented examples with more natural Portuguese phrasing.

The embedding-based proximity analysis could also be pushed further. In this work it was applied at the language level, to determine which source languages to include in training. A natural extension would be to apply it at the individual document level: ranking each source document by its textual similarity to the unannotated Portuguese documents and retaining only the most similar ones for training. This would rely exclusively on document embeddings without requiring any knowledge of Portuguese annotations, and could reduce the noise introduced by training documents that are topically or stylistically distant from the Portuguese domain.

On the modelling side, the per-technique results suggest an interesting direction for hybrid systems. As discussed in Section 5.2, the generative baseline showed a selective advantage on techniques with distinctive linguistic markers. A system that uses the fine-tuned encoder as its primary detector but routes specific techniques to a generative specialist could potentially combine the strong recall of the encoder-based approach with the precision that LLMs show on these particular categories. Designing an effective routing mechanism, whether based on confidence scores, technique category, or some combination of both, would be a concrete and testable research direction.

The problem of extremely rare techniques also deserves dedicated attention. Targeted data augmentation using generative models to synthesise additional training examples, few-shot learning approaches that leverage the textual definitions from the annotation guidelines, or hierarchical classification strategies that first predict at the coarse-grained category level before refining to the fine-grained label are all plausible avenues that could be explored.

Beyond Portuguese, the methodology developed here is not language-specific. The embedding-based proximity analysis, the multi-source training strategy, and the Stage 2 language selection procedure could all be applied to Bulgarian and Slovenian, the other low-resource test languages in the CLEF-2024 dataset. Validating the finding that linguistic proximity outweighs dataset size across a wider range of language pairs would contribute additional evidence to the growing body of work on cross-lingual transfer [8]. The classification architecture is not tied to Portuguese either. Since the multi-label sigmoid head helps wherever spans carry more than one technique, testing it on the other CLEF-2024 languages would show whether the gains seen for Portuguese also appear elsewhere.

The current evaluation is also limited to a single test set covering specific topics such as the Israeli-Palestinian conflict and the Russia-Ukraine war. How well the model generalises to new topics, different news sources, or different time periods remains an open question, and investigating its sensitivity to topic distribution shifts would be important for any real-world deployment in a media monitoring context.

Finally, the outputs of the current system, while accurate enough to be useful, are not inherently interpretable. A system that not only identifies a span as Appeal to Fear but also explains why, in terms of the specific linguistic features that triggered the prediction, would be substantially more useful for journalists, fact-checkers, and media literacy educators. Integrating attention-based or gradient-based explanation methods with the multi-label sigmoid architecture, or generating natural language explanations conditioned on the encoder's predictions, could make the system both more transparent and more practically valuable.

5.5. Closing Remarks

This dissertation demonstrates that it is possible to build a competitive span-level persuasion technique detector for a language with no native annotated training data, provided

that the architectural choices are appropriate for the multi-label structure of the task and that the training data selection is guided by empirical evidence of linguistic proximity rather than by data volume alone. The result is a system that surpasses the state of the art on the Portuguese CLEF-2024 benchmark while contributing methodological insights that are likely to generalise to other low-resource languages and similar structured prediction tasks.

The broader lesson is perhaps the most important takeaway: in cross-lingual transfer for structured NLP tasks, the assumptions that seem most natural (more data is always better, stronger encoders always win, architecture matters less than the pre-training) turn out to be systematically wrong in this setting. Questioning those assumptions and replacing them with empirically grounded decisions is what made the difference between a competitive baseline and a system that surpasses the state of the art.

References

- [1] G. Jowett and V. O'Donnell, *Propaganda and Persuasion*. Sage, 2018. [Cited on pages 1, 6, and 8.]
- [2] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov, “Fine-grained analysis of propaganda in news articles,” in *Proceedings of EMNLP-IJCNLP*, 2019, pp. 5636–5646. [Cited on pages 1, 2, 3, 6, 9, 10, 12, 14, 15, 20, and 50.]
- [3] C. Wardle and H. Derakhshan, “Information disorder: Toward an interdisciplinary framework for research and policymaking,” Council of Europe, Tech. Rep., 2017. [Cited on pages 2 and 8.]
- [4] G. Da San Martino *et al.*, “A survey on computational propaganda detection,” *IJCAI*, 2020. [Cited on pages 2, 6, 8, 9, and 15.]
- [5] J. Piskorski, N. Stefanovitch, G. Da San Martino, and P. Nakov, “Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup,” in *SemEval*, 2023. [Cited on pages viii, 2, 3, 6, 7, 8, 10, 11, 12, 14, 15, 18, 19, 24, and 25.]
- [6] M. Hasanain *et al.*, “Overview of the clef-2024 checkthat! lab task 3 on persuasion techniques detection,” in *CLEF Working Notes*, 2024. [Cited on pages viii, 2, 3, 6, 10, 11, 12, 14, 16, 17, 18, 19, 21, 23, 24, 25, 26, 27, 37, and 40.]
- [7] D. M. Eberhard, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the World*, 26th ed. Dallas, Texas: SIL International, 2023. [Online]. Available: <https://www.ethnologue.com> [Cited on page 2.]
- [8] 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 *Proceedings of ACL*, 2020, pp. 8440–8451. [Cited on pages 3, 16, 17, 19, 20, 36, 50, and 61.]
- [9] A. Barrón-Cedeño, G. Da San Martino, I. Jaradat, and P. Nakov, “Proppy: Organizing the news based on their propaganda content,” in *IJCAI*, 2020. [Cited on pages 6, 10, 12, and 13.]

- [10] J. Piskorski *et al.*, “Overview of the clef-2023 checkthat! lab task 3: Persuasion techniques detection,” in *CLEF Working Notes*, 2023. [Cited on pages 10 and 11.]
- [11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *Proceedings of ICLR*, 2013. [Cited on page 13.]
- [12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186. [Cited on pages 14 and 17.]
- [13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal *et al.*, “Language models are few-shot learners,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020. [Cited on page 15.]
- [14] M. Hasanain, F. Ahmed, and F. Alam, “Can GPT-4 identify propaganda? Annotation and detection of propaganda spans in news articles,” in *Proceedings of LREC-COLING 2024*, 2024, pp. 2724–2744. [Cited on page 15.]
- [15] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, “Finetuned language models are zero-shot learners,” in *International Conference on Learning Representations*, 2022. [Cited on page 16.]
- [16] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using siamese BERT-networks,” in *Proceedings of EMNLP-IJCNLP*, 2019, pp. 3982–3992. [Cited on page 33.]

A. Conference Posters

FCUP Master's Day

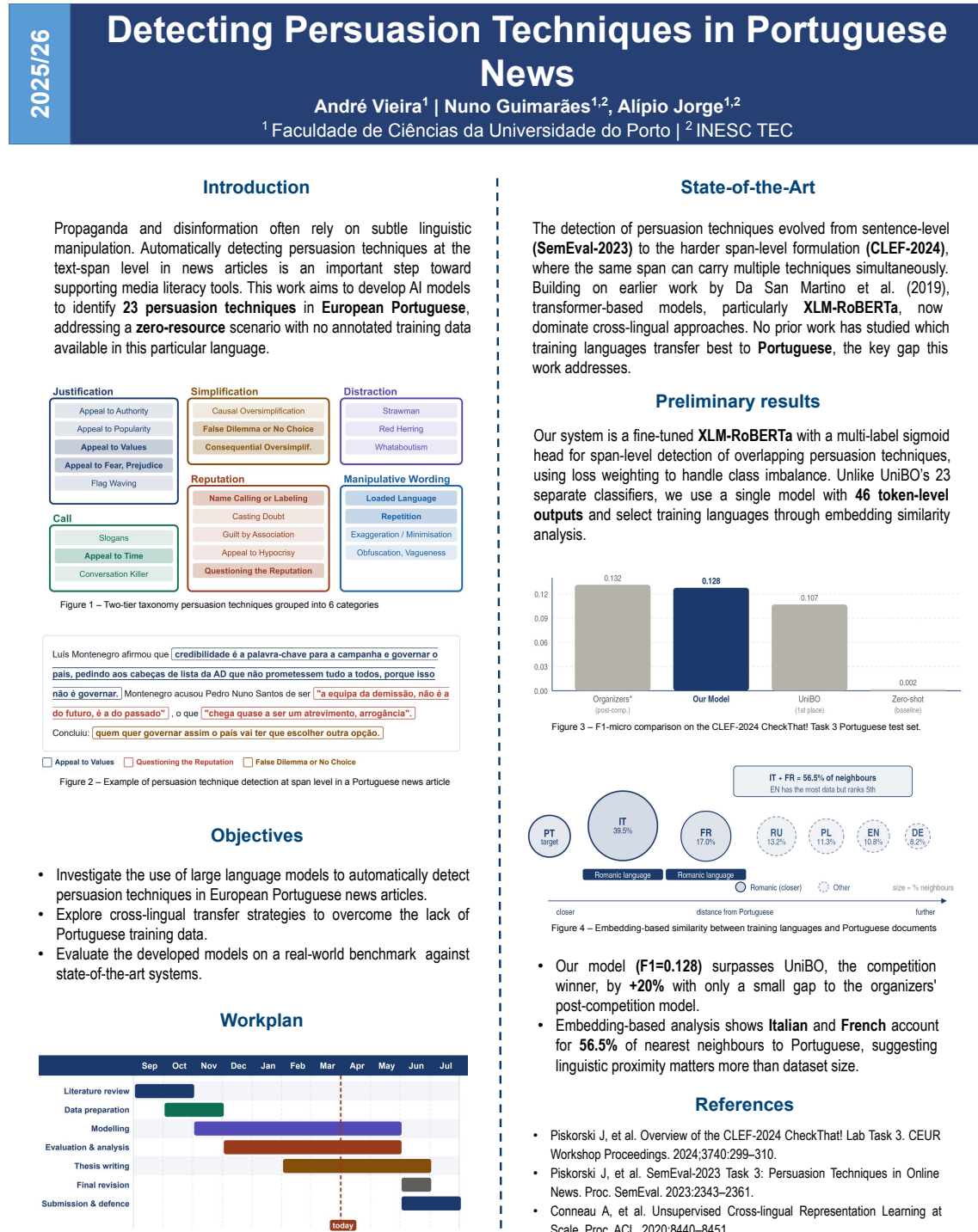


Figure A.1: Poster presented at the FCUP Master's Day, 2026.

IJUP - Young Researchers Meeting

Multilingual Language Models for Span-Level Detection of Persuasion Techniques in Disinformation Discourse

André Vieira | Nuno Guimarães | Alípio Jorge
Faculdade de Ciências da Universidade do Porto | INESC TEC
Portugal

Introduction

- Propaganda and disinformation often rely on subtle linguistic manipulation.
- Automatically detecting persuasion techniques at the text-span level in news articles supports media literacy tools.
- This work develops an AI model to identify **23 persuasion techniques** in **European Portuguese** texts, with **no annotated training data** available in this language for this task.

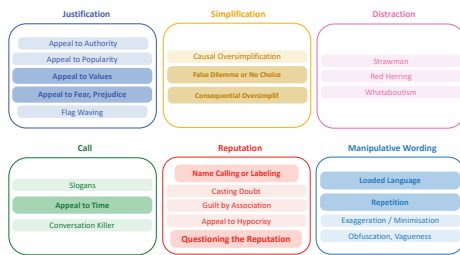


Figure 1. Two-tier taxonomy persuasion techniques grouped into 6 categories; **darker, bold** techniques appear most frequently in the dataset.

Luís Montenegro afirmou que **credibilidade é a palavra-chave para a campanha e governar o país**, pedindo aos cabeças de lista da AD que **não prometessem tudo a todos, porque isso não é governar**. Montenegro acusou Pedro Nuno Santos de ser **"a equipa da demissão, não é a do futuro, é a do passado"**, o que **"chega quase a ser um atrevimento, arrogância"**.
Concluiu: **quem quer governar assim o país vai ter que escolher outra opção.**

○ Appeal to values ○ Questioning the Reputation ○ False Dilemma or No Choice

Figure 2. Example of persuasion technique detection at span level in a Portuguese news article

Aim

- Investigate the use of language models to automatically detect persuasion techniques in European Portuguese news articles.
- Explore cross-lingual transfer strategies to overcome the lack of Portuguese training data.
- Evaluate the developed models on a real-world benchmark (CLEF-2024) against state-of-the-art systems.

Methodology

Our system is a fine-tuned **XLM-RoBERTa** with a multi-label sigmoid head producing **46 binary outputs per token** (2×23 techniques). Training language selection is guided by **embedding-based similarity** rather than data volume alone.

Year	Task	Level	Languages	Techniques
2019	NLP4IF	Span	En	18
2020	SemEval-2020	Span	En	14
2022	SemEval-2022	Paragraph	9	23
2024	CLEF-2024 ★	Span	5	23

Figure 3. Resources and Results in Persuasion Techniques Tasks

Results

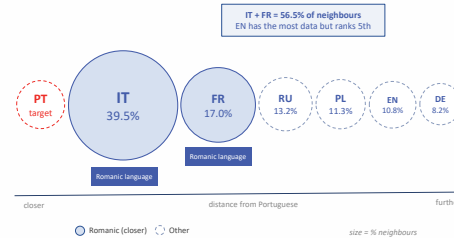


Figure 4. Embedding-based similarity between training languages and Portuguese documents

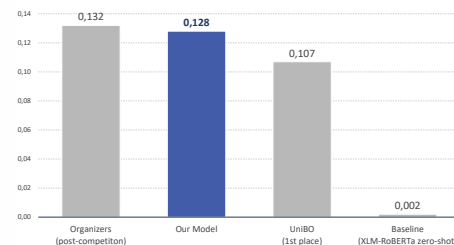


Figure 5. F1-micro comparison on the CLEF-2024 CheckThat! Task 3 Portuguese test set.

Our model (**F1=0.128**) surpasses UniBO, the competition winner, by **+20%** with only a small gap to the organizers' post-competition model.

Embedding analysis shows **Italian** and **French** account for **56.5%** of nearest neighbours to Portuguese, linguistic proximity matters more than dataset size.

Conclusion

Cross-lingual Transfer

Romance languages (IT, FR) transfer more effectively to Portuguese than English, despite English having the largest dataset.

Embedding-guided Selection

Linguistic proximity via multilingual embeddings is a reliable proxy for transfer effectiveness in zero-resource settings.

Architecture

A single multi-label model with per-output calibration outperforms 23 separate classifiers for span-level detection.

Impact

The system supports automated media monitoring in Portuguese, contributing to disinformation detection at scale.

References

- Piskorski, J. et al. Overview of the CLEF-2024 CheckThat! Lab Track 5. CLEF Workshop Proceedings, 2024(1742):290-315.
- Piskorski, J. et al. SemEval-2023 Task 3: Persuasion Techniques in Online News. Proc. SemEval, 2023:2415-2461.
- Conneau, A. et al. Unsupervised Cross-lingual Representation Learning at Scale. Proc. ACL, 2020:8440-8451.



ORGANIZAÇÃO



APOIO



Figure A.2: Poster presented at IJUP, University of Porto, 2026.