Towards robust propaganda detection

Addressing Distribution Shifts via Multi-task Learning

submitted in partial fulfillment for the degree of master of science

Alexander Hepburn
13175017

Master Information Studies
data science
Faculty of Science
University of Amsterdam

Submitted on 26-06-2025

	UvA Supervisor
Title, Name	prof. dr. Paul Groth
Affiliation	UvA, INDElab, IvI
Email	p.t.groth@uva.nl

ABSTRACT

The limitations of Transformer-based natural language processing (NLP) systems, coupled with the fragmented nature of computational propaganda research, continue to hinder system generalisability. The diverse contexts and domains that define the field pose challenges for NLP methodologies, which remain highly susceptible to distribution shifts. While these limitations are widely acknowledged, they are rarely studied systematically. This thesis explores whether multi-task learning (MTL) can improve model robustness through shared representations across persuasion detection tasks, which are central to propaganda research. Through a controlled, comparative evaluation of single-task and multi-task architectures, we examine generalisation under domain and label shifts. We find that multi-task models can recover performance degradation under data scarcity through indirect data augmentation, and can improve performance under extreme class imbalance. While we cannot definitively conclude that MTL results in domain-agnostic models, we validate the applicability and benefits of cross-task interaction for low-resource propaganda detection at scale.

KEYWORDS

natural language processing, computational propaganda, distribution shifts, transfer learning, multi-task learning

GITHUB REPOSITORY

https://github.com/aahepburn/Multi-Task-Propaganda-Detection

1 INTRODUCTION

Computational propaganda refers to the intentional use of automated and algorithmic techniques to manipulate public opinion by amplifying persuasive messages, often aimed at disrupting democratic processes or advancing specific agendas [4]. A closely related concept is disinformation, defined as false or misleading information deliberately disseminated to deceive and cause harm, often blending truthful elements with fabricated ones to craft plausible yet manipulative narratives [22, 28]. At their core, these strategies are underpinned by persuasive techniques that leverage psycholinguistic and cognitive mechanisms to influence individuals' beliefs and behaviours [8, 11, 28]. As digital media amplifies these techniques with unprecedented speed and scale, they present a growing challenge to democratic processes and healthy societal discourse.

Despite the growing body of research on computational propaganda and persuasion detection, which underpins it, developing robust natural language processing (NLP) models capable of generalising across domains remains a challenge [19, 25, 28]. Early symbolic approaches, such as those by Rashkin et al., relied on handcrafted linguistic features to distinguish between propaganda, fake news, and factual statements by leveraging rhetorical markers such as hedging and subjectivity [28]. While theoretically well-grounded, these methods struggle with adaptability and require extensive expert-driven annotation. The rise of Transformer-based architectures has since led to state-of-the-art performance in persuasion detection, but these models remain inherently limited by their known tendency to rely on spurious correlations [10].

This reliance makes them prone to performance degradation under distribution shifts—including domain shifts, label shifts, and changes in linguistic patterns [15, 19]. Furthermore, the need for expert annotation due to the complexity of these tasks means that the field remains highly fragmented in terms of its data usage—often relying on study- and domain-specific datasets [19]. Although data augmentation, multi-modal systems, and cross-lingual approaches have shown partial success in both symbolic and neural NLP methods, out-of-domain generalisation remains an obstacle [25, 33].

This thesis examines whether multi-task learning (MTL) can enhance model robustness in persuasion detection via shared linguistic representations across tasks. MTL is a type of model architecture where multiple tasks are trained concurrently using the same set of parameters, which can function as a form of inductive bias [34]. It has been explored in NLP for improving generalisation in structured prediction tasks such as named entity recognition and syntactic parsing [6, 40]. However, its application to persuasion detection is under-explored due to challenges in defining task-relatedness and the lack of aligned multi-task datasets. Given that propagandistic language exhibits recurring rhetorical strategies across domains—from political disinformation to climate change denial—MTL could be an opportunity to model shared linguistic features instead of relying on domain-specific ones [7, 11].

The SemEval 2025 shared task on "Multilingual Characterisation and Extraction of Narratives from Online News"[1] provides a suitable experimental setting for this study, as it includes two complementary subtasks: *entity framing* and *narrative classification*. These tasks are hypothesised to capture distinct but interrelated aspects of persuasive communication—entity framing examines how subjects are positioned within narratives, while narrative classification focuses on overarching discourse structures.

By evaluating whether MTL limits performance degradation under distribution shifts, this study is grounded in the line of research that has stressed the need for novel, robust NLP methods given the limitations of neural systems [15, 19].

Thus, the central research question guiding this thesis is:

- *To what extent can shared linguistic representations of persuasive strategies mitigate performance degradation caused by distribution shifts in persuasion detection NLP tasks?*

The sub-questions are as follows:

- **RQ1:** *To what extent does multi-task learning (MTL) improve generalisation compared to single-task learning (STL) in unseen domains, as measured by classification performance?*
  - **RQ1.1:** How does MTL perform compared to STL in entity framing classification under domain shifts?
  - **RQ1.2:** How does MTL perform compared to STL in narrative classification under domain shifts?
- **RQ2:** *To what extent can MTL adapt to label distribution shifts across domains more effectively than STL?*
  - **RQ2.1:** How does MTL react to label distribution shifts in entity framing, compared to STL?
  - **RQ2.2:** How does MTL react to label distribution shifts in narrative classification, compared to STL?

[1] SemEval 2025, Task 10 https://propaganda.math.unipd.it/semeval2025task10/

- **RQ3:** *To what extent does MTL learn domain-invariant representations of persuasive language, and how does this compare to STL?*
  - **RQ3.1:** Do MTL models capture transferable linguistic features across entity framing and narrative classification, as measured through SHAP analysis, and how does this differ from STL?

Together, these questions establish an empirical framework for theorising whether multi-task learning architectures can capture transferable linguistic representations of persuasion techniques, thereby improving out-of-domain generalisation.

2 RELATED WORK

Transformer-based NLP systems in computational propaganda detection remain highly susceptible to distribution shifts, limiting their generalisability and robustness. While existing research acknowledges the need for adaptation across domains, existing approaches remain largely domain-specific, with little systematic study of how to mitigate these shifts [15, 19, 31]. This section reviews the state of the art in persuasion detection, explores multi-task learning as a means of improving generalisation, and discusses interpretability techniques for assessing model robustness.

2.1 Single-Task Learning Architectures for Persuasion Detection

The state of the art in single-task learning (STL) architectures for persuasion and propaganda detection has been largely driven by transformer-based models fine-tuned on domain-specific data [7, 19, 26, 31, 33]. Recent installations of the SemEval shared tasks have served as benchmarks for propaganda and persuasion detection, with models such as BERT, RoBERTa, XLM-R, and DeBERTa demonstrating strong performance across multiple subtasks [18, 26]. For instance, in SemEval-2020 Task 11, the best-performing system achieved an F1-score of 0.62 for sentence-level propaganda detection using RoBERTa-large fine-tuned with domain adaptation techniques [18]. Similarly, in SemEval-2023 Task 3, XLM-R-large fine-tuned on persuasion technique classification achieved an F1-score of 0.71, confirming the relative effectiveness of transformer-based models when trained on task-specific datasets [26].

A key challenge in STL approaches is handling fine-grained classification tasks, such as detecting specific persuasion techniques. The NLP4IF-2019 shared task on Fine-Grained Propaganda Detection revealed that models struggled to distinguish among 18 persuasion techniques in a multi-label setting [8]. The best-performing systems incorporated RoBERTa-large with linguistic feature extraction, achieving an F1-score of 0.55, but still faced issues with class imbalance and domain shifts. Furthermore, STL models rely heavily on supervised learning, which presents a major bottleneck due to the limited availability of high-quality annotated data, particularly for under-represented persuasion techniques, exacerbating model bias towards high-frequency labels [25].

Data augmentation, multi-modality, and cross-lingual models have been explored to mitigate these limitations and are currently the focus of the field [19, 31]. Piskorski et al. experimented with data augmentation strategies to improve classification robustness for climate change denial detection, showing that synonym replacement and back-translation improved cross-domain F1-score by 3-5% on out-of-distribution test sets [25]. Additionally, cross-lingual STL models, such as those tested in SemEval-2021 Task 6, demonstrated the difficulty of transferring learned representations across languages, with BERT multilingual models seeing an F1-score drop of up to 18% on unseen languages [9]. Sharma et al. presented VECTOR, a multi-modal meme classification approach, which showed limited improvement over unimodal systems, suggesting there are benefits to combining complementary visual and text signals [33].

2.2 Multi-Task Learning in NLP

Multi-task learning (MTL) is an approach in machine learning that, under certain conditions, can improve generalisation through shared representations across related tasks [3, 6, 13]. In NLP, MTL is particularly effective in low-resource scenarios, as it enables models to transfer knowledge between tasks, mitigating overfitting compared to traditional approaches [6, 40].

Transformer-based architectures such as BERT, T5, and GPT-style models have been adapted for MTL, with T5-MTL achieving a 5-10% increase in F1-score across multiple NLP benchmarks compared to STL fine-tuning [6, 40]. MTL has also been effective in addressing distribution shifts, with MTL-trained BERT models experiencing up to 20% lower accuracy drop when tested on out-of-distribution datasets compared to STL models [38]. However, ensuring effective task-relatedness remains a key challenge, as negative transfer can degrade performance when unrelated tasks are jointly trained [38, 40].

Recent research has also explored the integration of MTL throughout the entire machine learning lifecycle, from data preprocessing to model deployment [39]. Torbarina et al. highlight the practical challenges of implementing transformer-based MTL models in real-world NLP applications. They note the trade-offs between generalisation and specialisation, showing how MTL strategies must balance task-specific and cross-task knowledge [35]. These findings align with work on parameter-sharing techniques, such as hard versus soft parameter sharing, which has implications for training efficiency and balancing task-specific versus more general representations [6].

Techniques like multi-modal alignment prompts (MmAP) and task-conditioned adapters have been developed, allowing models to optimise multiple tasks while reducing the number of trainable parameters [39]. MmAP improves task transferability by 12% on domain-adaptation benchmarks while cutting computational costs by 40% compared to full fine-tuning [39]. Similarly, adversarial filtering and contrastive learning techniques help counteract task imbalance, preventing models from overfitting to high-resource tasks [38, 40]. There has been ongoing debate around how best to structure adapter-based transfer learning: whether task-specific modules should be trained jointly to encourage cross-task interaction [34], or trained independently and later composed non-destructively [23]. Pfeiffer et al. argue that joint multi-task training with adapters is difficult to balance and prone to interference, especially when tasks differ in scale or objective [23]. They also highlight the inefficiency of full-model fine-tuning, which requires updating and storing a complete set of Transformer parameters for each task,

resulting in specialised and non-reusable models. Their AdapterFusion approach addresses both concerns by decoupling learning and composition: adapters are trained separately, and a fusion layer learns how to combine them at inference time.

2.3 Interpretability in NLP Models

The interpretability of Transformer-based NLP models has been widely studied, particularly regarding the role of attention mechanisms [37]. While some studies argue that attention weights can provide meaningful insights into model decision-making, particularly in tasks like syntax and co-reference resolution [37], others challenge their reliability, showing that attention distributions can be altered without significantly affecting predictions [12, 32].

Beyond attention-based analyses, post-hoc attribution methods such as SHAP and Integrated Gradients have been applied to assess whether Transformers capture generalisable linguistic patterns or rely on dataset-specific ones [20]. Layer-wise relevance propagation and gradient-based perturbation analyses further examine how representations evolve across Transformer layers, offering alternative insights into model behaviour [5].

Despite that, there is no clear consensus on whether interpretability methods reliably explain Transformer decision-making. The challenge remains in distinguishing causal mechanisms from correlational artefacts in learned representations.

Existing approaches to persuasion detection lack a systematic focus on robustness and interpretability. While effective in controlled settings, they struggle with distribution shifts and provide limited insight into whether models capture generalisable linguistic patterns or rely on dataset-specific biases. This thesis addresses these limitations by integrating insights from multi-task learning and deep learning interpretability to improve both generalisation and model transparency.

3 METHODOLOGY

The study adopts a comparative design that evaluates model robustness under two types of distribution shift: (1) input shift, involving lexical and topical variation, and (2) label shift, involving class imbalance and potentially semantic shift, in the case of entity framing. These shifts are instantiated across two domains— the Ukraine conflict (UA) and Climate Change (CC). Each research question is operationalised as follows:

- **RQ1 (robustness to input shift)** and **RQ2 (robustness to label shift)** are addressed through cross-domain evaluations (Ukraine and Climate Change), as well as ablations on label granularity and data augmentation to isolate the effects of distributional and label-based variation.
- **RQ3 (emergence of transferable representations)** is explored through comparisons between STL, MTL, and MTL-PAL setups, supported by ablation studies and SHAP values.

3.1 Dataset

The dataset used in this study originates from the SemEval 2025 shared task on multilingual persuasion detection[2]. It comprises 1,782 training and 178 test articles in five languages: English, Bulgarian, Hindi, Portuguese, and Russian. Each article is annotated for two subtasks: *Entity Framing* and *Narrative Classification*. The two tasks are described in more detail in subsection 3.2. Table 1 summarises the amount of labelled data used in this study.

Table 1: Dataset Summary

Task	UA Domain	CC Domain
Entity Framing (Train)	4,881 entities	579 entities
Narrative Classification (Train)	1,423 articles	491 articles
Entity Framing (Test)	357 entities	91 entities
Narrative Classification (Test)	108 articles	70 articles

Each task relies on a hierarchical label taxonomy. For narratives, labels include main narratives (e.g., "The West is responsible") and more specific sub-narratives (e.g., "NATO provokes instability in Ukraine"). Analogously, for entities, roles range from main roles (e.g., Protagonist, Antagonist) to fine-grained ones (e.g., Spy, Instigator, Guardian). See Table 2 for more details.

To standardise input and simplify downstream modelling, all non-English documents were translated into English using GPT-4o [3]. An earlier pipeline based on NLLB-200 introduced sentence-level artefacts and formatting inconsistencies. GPT-4o provided cleaner, more semantically faithful translations while maintaining alignment with pre-annotated entity spans. The use of machine translation for normalising multilingual data is standard practice in low-resource NLP and multilingual propaganda research [14]. While multilingual Transformer models such as XLM-RoBERTa have been used in related studies [26], I opted to translate the data to limit the scope of the project. This was computationally expensive, but greatly simplified the experimental setup and also allowed the use of distilled models, unsuitable for multi-lingual data. Label handling differed slightly between tasks. For narrative classification, no label translation was needed, as taxonomies were provided in English. In contrast, entity framing required careful alignment of token offsets post-translation to preserve annotation integrity.

To ensure consistent evaluation across seeds and domains, the data was split into training and validation sets using a stratified 80/20 split. Stratification was performed independently for each task, preserving the relative distribution of labels across all levels of the flattened hierarchy. This strategy prevents label dilution in the validation set, especially for rare classes.

3.2 Tasks

The two classification problems in this study—narrative classification and entity framing—can be viewed as standard document and text span classification tasks, respectively. However, their complexity increases significantly due to two shared properties: both are hierarchical and multi-label in nature. These characteristics place them at the more challenging end of the NLP spectrum [2].

A common practice in this field is to address this complexity by simplifying the problem, depending on the paper's primary goal. For example, Coan et al. restrict classification to a single level

[2] https://propaganda.math.unipd.it/semeval2025task10/

[3] Please refer to the GitHub repository for the translation pipeline, including the prompts—*machine_translation.py*.

of their taxonomy, reducing the task to a multi-class, rather than multi-label, problem [7]. Similarly, Li et al., despite achieving state-of-the-art results, treat the task as a flat multi-label classification problem, discarding the hierarchical structure entirely [14]. This tendency to simplify is common, as also noted in the review by Tsoumakas and Katakis [36].

In contrast, my approach maintains the full multi-label nature of the task, flattening the hierarchy but preserving all labels across both coarse and fine levels. In the taxonomy of Tsoumakas and Katakis, this corresponds to the Binary Relevance (PT4) problem transformation [36]. Importantly, I do not attempt to balance the class distribution across levels; instead, I preserve the natural label imbalance, allowing the frequency distribution to reflect the implicit structure of the hierarchy. This stands in contrast to approaches that focus on only one hierarchy level, which allows them to use, e.g. class-weighting techniques. In this way, the models implement the hierarchy implicitly through label distribution, rather than through additional structural constraints.

Table 2: Overview of the two persuasion detection tasks. Task 1 is performed at the entity level, while Task 2 is performed at the document level.

(a) Task 1 - Entity Framing. Given pre-annotated entities and the full news article as input, the model assigns multi-label role classifications to each entity span.

Article ID	Entity	Start	End	Label
EN_UA_024321	Yulia Navalnaya	235	249	Protagonist, Rebel
EN_CC_200141	King Charles III	1322	1337	Antagonist, Corrupt

(b) Task 2 - Narrative Classification. The model processes the tokenised news article and assigns one or more narrative and subnarrative labels.

Article ID	Narrative	Subnarrative
BG_613	Distrust towards Media	Western media is an instrument of propaganda
A9_BG_2566	Russia is the Victim	The West is russophobic

3.3 Distribution Shift Analysis

Central to the thesis is isolating the impact of input (UA/CC) and label shifts in a controlled environment to understand whether MTL systems are more resilient to such shifts. Firstly, we need to verify whether the two domains observed in the corpus are indeed sufficiently different. A selection of tests and visualisations in this subsection confirms this central assumption.

To quantify distribution shifts, we followed the approach by Li and Groth [15], applying:

- **Chi-square tests** on token and label distributions.
- **Cosine distance** on sentence embeddings.

Figure 1 illustrates the label shift across the domains; Figure 2 focuses on input shift as observed through PCA and t-SNE plots. In addition to PCA and t-SNE, we compute quantitative shift metrics. Lexical shift is measured using Jensen-Shannon Divergence ($JSD = 0.4067$), while semantic drift is captured via cosine distance between domain-level mean sentence embeddings (0.6382). Maximum Mean Discrepancy (MMD) further confirms the embedding distribution divergence ($MMD = 0.2002$), validating the experimental framing of domain shift. These results justify the comparative

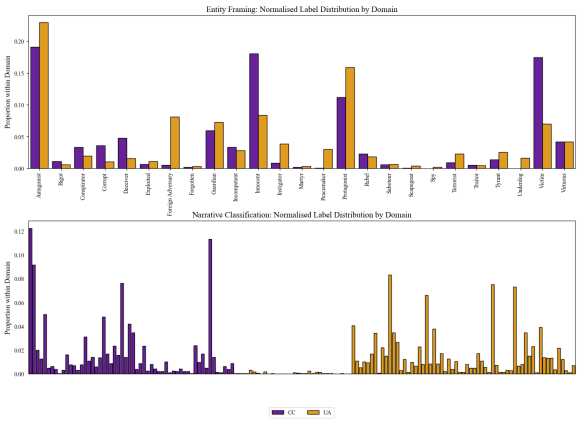


Figure 1: Normalised label distribution for tasks S1 and S2 across domains. Top: S1 task labels show broader overlap, though domain trends persist. Bar height represents proportional frequency within each domain. Bottom: S2 task labels differ radically per domain

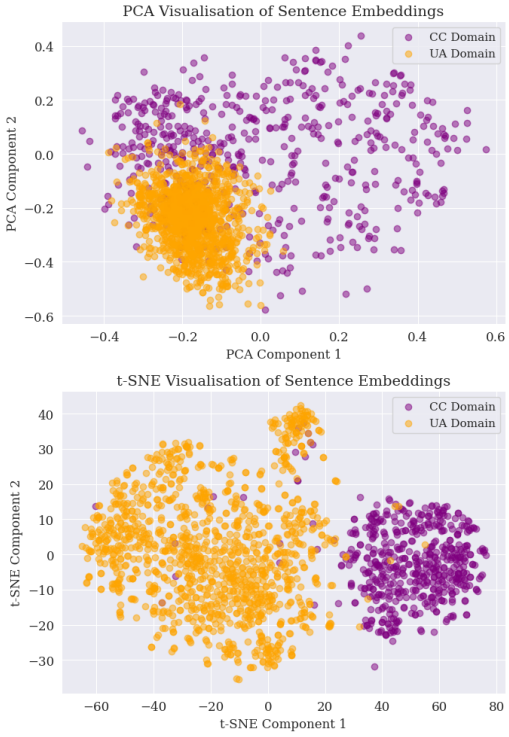


Figure 2: PCA captures linear variation in the embeddings and shows overall structure and how much variance is explained by the top components. There is moderate overlap but a noticeable central shift. t-SNE captures non-linear relationships. The domains are clearly separated, with most CC and UA points occupying different regions of the space.

design and the need for models capable of generalising across both lexical and conceptual space. Distributional differences extend to

Table 3: Distributional Shift Statistics. The results confirmed both lexical and semantic drift between domains and label skew across tasks.

Aspect	Test	Value	p-value
Input (tokens)	Chi-Square	110,926	< 0.0001
Input (semantics)	Cosine Distance	0.64	–
Entity Roles	Chi-Square	764.07	$< 10^{-140}$
Narratives	Chi-Square	1,531.45	$< 10^{-75}$

the labels themselves: the distribution of entity labels shifts across the domains; the distribution of narrative labels shifts entirely and does not overlap across the two domains.

3.4 Model Architecture

The experiments compare single-task and multi-task setups across multiple transformer backbones [4].

3.4.1 Single-Task Learning (STL). To assess the relative performance of the proposed MTL and MTL-PAL architectures, I implemented a set of STL baselines. For the majority of tasks, I use a simple fine-tuned transformer model without extensive architectural or data-level enhancements. This allows for broad comparisons across tasks under controlled and consistent conditions, and helps isolate the core mechanics of multi-task learning by reducing confounding factors introduced by more complex modelling.

3.4.2 Multi-Task Learning (MTL). Two MTL variants are explored to investigate the trade-offs between representation sharing, task specialisation, and architectural modularity:

Hard Parameter Sharing (MTL):. A single encoder is jointly fine-tuned across tasks, with separate task-specific heads. The model is trained end-to-end, and the total loss is computed as the unweighted sum of the individual task losses [6, 34]. This architecture is parameter-efficient and widely used in MTL research due to its simplicity and strong inductive bias toward learning shared representations.

Hybrid Modular Architecture (MTL-A):. Task-specific adapter modules are integrated into a shared encoder to support soft task specialisation while preserving joint learning [24, 27]. Unlike modular MTL approaches that isolate task adapters and defer integration to inference time (e.g., AdapterFusion) [23], this setup allows tasks to interact during training, encouraging the emergence of shared representations. This design is especially appropriate when tasks are semantically related but not fully aligned, as it helps balance shared linguistic structure with task-specific signal. It also improves parameter efficiency compared to full fine-tuning, making it suitable for low-resource or multi-domain scenarios. If persuasion strategies are encoded at a generalisable level, they are hypothesised to become visible precisely when the model is made to reconcile variation across related tasks.

3.4.3 Encoders. We initially considered four encoders—DeBERTa-V3-base and RoBERTa-base [16], which were selected to represent the current state of the art, as noted in Section 2.1. DistilBERT and MiniLM-L12-H384 were considered as light-weight, distilled alternatives [30]. To manage the computational demands of the ablation experiments, all ablations in this study were conducted using DistilBERT as the encoder. Although RoBERTa demonstrated superior performance in the main evaluation, DistilBERT was chosen as the default encoder to ensure practical training times across the numerous experimental conditions. The goal of these ablations is not to achieve state-of-the-art accuracy, but to isolate the effects of label granularity, class imbalance, and cross-task structure on model performance. Since distilBERT preserves the relative performance trends between STL, MTL, and MTL-PAL architectures observed in RoBERTa-based models (Table 8), it provides a suitable basis for analysing the mechanisms underlying multi-task generalisation.

[4] The detailed implementation of the models can be found in the code repository in *single_task.py* and *multi_task.py* respectively.

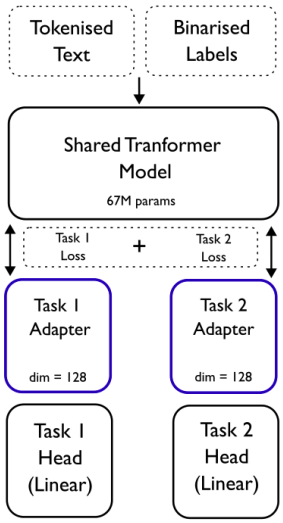


Figure 3: Multi-task learning setup and multi-task learning with task-specific adapters (blue). Both setups fine-tune the parameters of the shared encoder by optimising their summed loss. The single-task learning setup consists of only one of the tasks during training and does not include an adapter.

3.5 Evaluation Metrics

The evaluation strategy integrates standard multi-label classification metrics with post-hoc interpretability tools to address the research questions outlined in Section 1. The goal is to assess model robustness under distribution shifts (RQ1–RQ2) and to investigate whether multi-task learning fosters transferable representations (RQ3–RQ3.1).

To evaluate predictive performance across both in-domain and cross-domain settings, we report:

Micro-F1: Captures overall classification quality by aggregating true positives (TP), false positives (FP), and false negatives (FN) across all labels:

$$\text{Micro-F1} = \frac{2 \cdot \sum_{l=1}^{L} \text{TP}_l}{2 \cdot \sum_{l=1}^{L} \text{TP}_l + \sum_{l=1}^{L} \text{FP}_l + \sum_{l=1}^{L} \text{FN}_l}$$

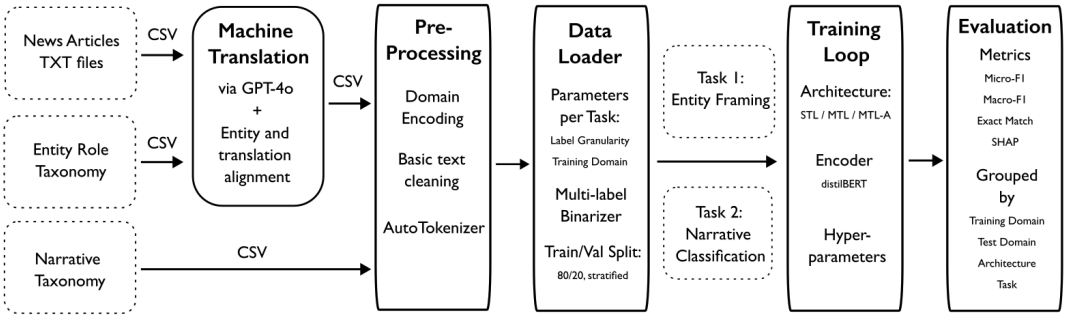


Figure 4: Visual Summary of the Data Pipeline

It is sensitive to frequent classes and provides a high-level summary of model performance.

Macro-F1: Averages F1-scores across all labels, weighting each class equally:

$$\text{Macro-F1} = \frac{1}{L}\sum_{l=1}^{L}\frac{2\cdot \text{TP}_l}{2\cdot \text{TP}_l + \text{FP}_l + \text{FN}_l}$$

This metric emphasises performance on rare or minority labels, making it important for analysing robustness under label imbalance (RQ2).

Exact Match Ratio (EMR): Measures the proportion of samples where the predicted label set exactly matches the gold label set:

$$\text{Exact Match} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[\hat{y}_i = y_i\right]$$

EMR is a stringent metric for multi-label settings, indicating how often the model gets every label correct for a given instance.

Notation:

- L: Total number of labels
- N: Number of instances
- $\text{TP}_l, \text{FP}_l, \text{FN}_l$: True/False positives/negatives for label l
- y_i, $\hat{y}_i$: Ground truth and predicted label vectors for instance i

To assess the significance of model performance, each configuration is evaluated over 20 random seeds, and 10 for each ablation study. For every metric, I report the mean and standard deviation. To determine whether observed differences between models (STL vs MTL or MTL-A) are statistically meaningful, I conduct paired hypothesis tests using the two-sided Wilcoxon signed-rank test, which is appropriate for non-normally distributed results. All comparisons are performed under controlled conditions—same encoder, training setup, task, and test domain—to ensure that significance reflects the effect of the modelling approach alone.

3.5.1 SHAP Values. To investigate how different model configurations attend to linguistic features, I use the SHAP method to perform post-hoc attribution analysis [17]. This allows for the identification of token-level contributions to individual predictions, independent of internal model components such as attention weights. The goal is to explore whether multi-task models rely on more generalisable or task-relevant features than their single-task counterparts. SHAP values are computed for each input token using a transformer-compatible explainer. For each example, I compare the token-level attributions across STL, MTL, focusing on differences in what each model considers most informative.

3.6 Experimental Setup

We perform a comprehensive evaluation of the three architectures within the limits of the available dataset. Our experiments are primarily data-centric—by varying the properties of data inputs, we evaluate the observed change in model performance across the two domains (UA and CC) and two tasks (EF and NC).

Table 4: Model training hyperparameters used across all experiments

Hyperparameter	Value
Pretrained Models	`distilbert-base-uncased`
Tokeniser	AutoTokenizer (matching model)
Max Sequence Length	512
Batch Size	8
Learning Rate	3e-5
Optimiser	AdamW
Loss Function	BCEWithLogitsLoss
Epochs	4
Random Seeds	{31, 32, 33, 34, 35, 36, 37, 38, 39, 40 42, 43, 44, 54, 55, 71, 72, 73, 74, 75}
Threshold	0.35
Validation Strategy	Exact Match threshold sweep
Hardware	GPU A100 via Google Colab

Ablation Studies. Recent work in multi-task learning (MTL) has highlighted its potential benefits, but often without disentangling the underlying mechanisms. Mulyar et al.[21] consolidate eight clinical NLP tasks into a single MTL model to reduce inference cost, reporting average task performance without probing why MTL helps or hinders certain tasks. Rodriguez et al.[29] take a step further by ablating data and label types to study generalisation in entity recognition, but their setup does not fully isolate the effects of shared inductive bias, implicit data augmentation, or regularisation.

Our ablation studies aim to go further, not just testing whether MTL helps, but under what conditions and why. We design three ablations to isolate the effects of task interaction, label distribution shifts (RQ2), and domain divergence (RQ1), while also supporting

our investigation of transferable representations (RQ3). We perform three ablations: (1) cross-domain training, (2) data duplication and (3) label granularity. They are described in more detail in the respective results figures.

4 RESULTS

In this section, we examine the results of task interaction in the MTL and MTL-A setups and compare these results with STL baseline models where no such interaction occurs. Every subsection is introduced with an observation most relevant to the thesis; each figure caption explains the experimental setup for the subsection in which it appears. For conciseness, we define "EF:UA" as entity framing (EF) trained on Ukraine data (UA). Analogously, "NC:CC" means narrative classification trained on Climate Change data (CC).

4.1 Main Results

See Table 5 for this result. Initially, we observe that data-rich models perform competitively under the STL framework. However, MTL significantly boosts performance in models marked by more extreme cases of class imbalance and label shift, at the expense of insignificant negative transfer for the stronger task. Significance tests are summarised in Appendix C: Tables 9 and 10.

4.2 Ablation: Cross-domain training

See Table 6 for this ablation. To further understand the initial patterns, we examined whether a task trained on one domain can implicitly improve the metrics of that domain in another task. This probes MTL's ability to generalise when task-specific signals are drawn from disjoint domain distributions. We find that, by exposing the encoder to UA narratives, we completely recover the performance degradation of UA entities for EF:CC. Significance tests are summarised in Appendix C Tables 11 and 12.

4.3 Ablation: Data Duplication

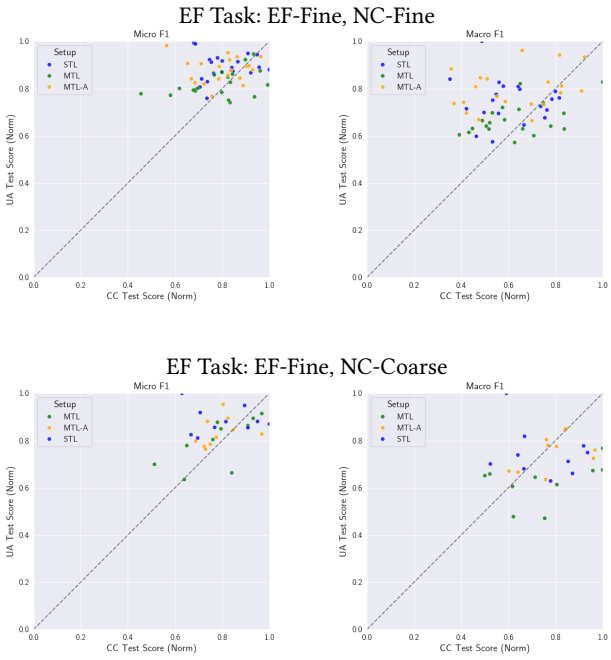
See Table 7 for this result. We duplicate the training data for each task to match the combined token count of the multi-task setup. This isolates the effect of supervision volume from that of multi-task interaction, helping to determine whether performance gains arise from task interaction or simply more exposure to labelled data. Since we are operating in low-resource and highly imbalanced settings, we examine whether the patterns we observed earlier hold when compared to simple data augmentation. Significance tests are summarised in Appendix C Tables 13 and 14.

4.4 Ablation: Label Granularity

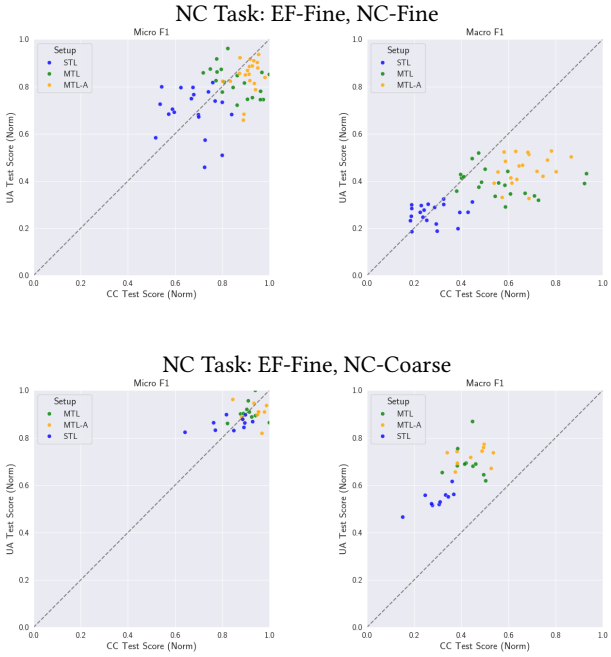
See Figure 5 for this ablation. When we reduce NC to coarse-grained labels (EF–Fine, NC–Coarse), MTL-A maintains an advantage in narrative classification, while EF remains largely unaffected. This suggests that NC continues to benefit from auxiliary supervision, though the label simplification narrows STL–MTL performance gaps.

In the fully coarse (EF–Coarse, NC–Coarse) setup, STL outperforms both MTL and MTL-A across almost all metrics, reinforcing the view that hard parameter sharing is only advantageous in imbalanced settings. Full results are available in Appendix C, Table 7.

Figure 5: Label Granularity Ablation. NC is trained on coarse-grained labels only. All models are trained on UA-CC.



(a) Entity Framing. EF metrics remain stable across NC granularity conditions.



(b) Narrative Classification. Coarse labels reduce STL–MTL gaps, especially in micro-F1; macro-F1 suggests stronger overfitting.

Table 5: Main Results. The y-axis of the table represents the three model architectures (Model) grouped by their training data (Data). Each row is a model. Each of these nine models is evaluated across two tasks: entity framing (EF) and narrative classification (NC)—see the x-axis. For each of these tasks, we report three metrics: Micro-F1, Macro-F1 and Exact Match. These metrics are reported separately for UA and CC test sets. The models presented here use the full fine-grained label taxonomy.

Model	Data	EF Task Metrics (UA)			EF Task Metrics: (CC)			NC Task Metrics: (UA)			NC Task Metrics: (CC)		
		Micro-F1	Macro-F1	Exact Match	Micro-F1	Macro-F1	EM	Micro-F1	Macro-F1	EM	Micro-F1	Macro-F1	EM
STL	UA	**0.474** ± 0.015	**0.186** ± 0.016	**0.200** ± 0.022	**0.754** ± 0.038	**0.152** ± 0.014	**0.598** ± 0.060	0.302 ± 0.041	0.056 ± 0.008	0.024 ± 0.014	0.000 ± 0.000	0.000 ± 0.000	0.091 ± 0.061
MTL	UA	0.448 ± 0.014	0.171 ± 0.013	0.196 ± 0.028	0.729 ± 0.035	0.142 ± 0.014	0.579 ± 0.044	0.375 ± 0.029	0.099 ± 0.013	0.038 ± 0.016	0.244 ± 0.101	0.017 ± 0.010	0.041 ± 0.031
MTL-A	UA	0.454 ± 0.015	0.185 ± 0.018	0.192 ± 0.020	0.735 ± 0.031	0.141 ± 0.012	0.584 ± 0.064	**0.386** ± 0.030	**0.122** ± 0.019	**0.044** ± 0.011	0.347 ± 0.068	0.026 ± 0.006	0.071 ± 0.039
STL	UA-CC	**0.469** ± 0.015	0.188 ± 0.018	0.201 ± 0.025	**0.764** ± 0.027	0.154 ± 0.011	0.631 ± 0.035	0.305 ± 0.043	0.044 ± 0.007	0.011 ± 0.010	0.373 ± 0.052	0.027 ± 0.008	0.071 ± 0.015
MTL	UA-CC	0.450 ± 0.015	0.173 ± 0.013	**0.203** ± 0.020	0.757 ± 0.039	0.156 ± 0.013	0.619 ± 0.041	0.360 ± 0.028	0.066 ± 0.010	0.030 ± 0.014	0.473 ± 0.047	0.055 ± 0.015	0.151 ± 0.039
MTL-A	UA-CC	0.465 ± 0.014	**0.198** ± 0.017	0.202 ± 0.027	0.761 ± 0.030	**0.156** ± 0.015	**0.634** ± 0.034	**0.371** ± 0.032	**0.076** ± 0.010	**0.037** ± 0.013	**0.501** ± 0.023	**0.064** ± 0.009	**0.165** ± 0.033
STL	CC	0.332 ± 0.044	0.066 ± 0.008	0.044 ± 0.022	0.676 ± 0.027	0.125 ± 0.012	0.551 ± 0.053	0.000 ± 0.000	0.000 ± 0.000	0.112 ± 0.081	0.401 ± 0.019	0.053 ± 0.008	0.113 ± 0.021
MTL	CC	**0.348** ± 0.041	**0.079** ± 0.009	0.070 ± 0.027	**0.694** ± 0.048	**0.144** ± 0.013	0.548 ± 0.069	0.000 ± 0.000	0.000 ± 0.000	0.068 ± 0.052	0.401 ± 0.011	0.055 ± 0.008	0.111 ± 0.023
MTL-A	CC	0.343 ± 0.025	0.075 ± 0.006	**0.083** ± 0.033	0.679 ± 0.040	0.138 ± 0.011	**0.558** ± 0.055	0.000 ± 0.000	0.000 ± 0.000	0.097 ± 0.058	**0.440** ± 0.023	**0.077** ± 0.010	**0.151** ± 0.024

Note: Average scores across 20 random seeds. DistilBERT. EF = Entity Framing, NC = Narrative Classification, UA = Ukraine, CC = Climate Change.

Entity Framing.
Pattern 1: MTL generally fails to significantly outperform STL. MTL results in a significant negative transfer of 2-3% in micro-F1 scores.
Pattern 2: For EF:CC, both domains experience a drop in performance, potentially due to less training data, but UA macro scores drop notably more (-10-12%) than CC scores (-2-3%).
Pattern 3: For the data-rich EF:UA and EF:UA-CC, MTL-PAL shows a consistent pattern of recovering some of the negative transfer, but this is statistically insignificant.

Narrative Classification.
Pattern 1: NC:UA and NC:UA-CC perform significantly better under MTL with gains of 6-10% for in-domain micro-F1.
Pattern 2: MTL achieves significant gains in NC at the expense of insignificant performance degradation in EF.
Pattern 3: NC:UA tested on CC and NC:CC tested on UA is a case of zero-shot learning because the narratives do not overlap across the two domains. Zero-shot learning is outside the scope of this thesis; therefore, we do not reflect on these results.

Table 6: Ablation: Cross-domain training. We train one task on data from just one domain, e.g., entity framing (EF) on UA labels and narrative classification (NC) on CC labels.

Model	Train	EF Metrics (UA)			EF Metrics (CC)			Train	NC Metrics (UA)			NC Metrics (CC)		
		Micro	Macro	EM	Micro	Macro	EM		Micro	Macro	EM	Micro	Macro	EM
STL	UA	**0.472** ± 0.015	0.183 ± 0.014	0.193 ± 0.013	0.745 ± 0.045	0.147 ± 0.018	0.597 ± 0.063	CC	0.000 ± 0.000	0.000 ± 0.000	0.125 ± 0.099	0.398 ± 0.019	0.053 ± 0.009	0.114 ± 0.023
MTL	UA	0.445 ± 0.012	0.173 ± 0.016	0.184 ± 0.016	0.728 ± 0.066	0.145 ± 0.014	0.585 ± 0.093	CC	0.381 ± 0.028	0.073 ± 0.013	0.036 ± 0.014	0.500 ± 0.048	0.061 ± 0.012	0.149 ± 0.044
MTL-A	UA	0.465 ± 0.010	**0.202** ± 0.019	**0.205** ± 0.016	**0.752** ± 0.036	**0.148** ± 0.011	**0.638** ± 0.046	CC	0.379 ± 0.024	0.081 ± 0.006	0.040 ± 0.014	**0.507** ± 0.030	**0.065** ± 0.019	**0.170** ± 0.030
STL	CC	0.336 ± 0.039	0.069 ± 0.007	0.043 ± 0.015	0.674 ± 0.037	0.128 ± 0.012	0.543 ± 0.054	UA	0.312 ± 0.027	0.058 ± 0.005	0.023 ± 0.011	0.000 ± 0.000	0.000 ± 0.000	0.083 ± 0.062
MTL	CC	0.448 ± 0.016	0.176 ± 0.017	**0.202** ± 0.025	0.739 ± 0.047	0.145 ± 0.018	0.597 ± 0.052	UA	0.373 ± 0.026	0.070 ± 0.009	0.033 ± 0.013	0.487 ± 0.032	0.055 ± 0.012	**0.160** ± 0.040
MTL-A	CC	**0.459** ± 0.026	**0.198** ± 0.021	0.197 ± 0.025	**0.746** ± 0.037	**0.149** ± 0.016	**0.623** ± 0.045	UA	**0.378** ± 0.032	**0.079** ± 0.011	**0.036** ± 0.020	**0.498** ± 0.038	**0.073** ± 0.012	0.154 ± 0.049

Note: Average scores across 10 random seeds. DistilBERT. EF = Entity Framing, NC = Narrative Classification, UA = Ukraine, CC = Climate Change.

Entity Framing.
Pattern 4: EF:UA does not benefit from the NC:CC task signal.
Pattern 5: For EF:CC, MTL significantly outperforms STL by 10% on F1 scores, and almost 16% on Exact Match. The NC:UA task signal helped to improve the performance of EF:CC on UA metrics.

Narrative Classification.
Pattern 4: NC:UA benefits less from EF:CC than it does from EF:UA in Pattern 1 (Main Results). We note symmetric results for NC:CC.

Table 7: Ablation: Data augmentation (duplication). STL-aug and MTL-aug were trained on duplicated datasets.

Model	Train	EF: UA			EF: CC			NC: UA			NC: CC		
		Micro	Macro	EM	Micro	Macro	EM	Micro	Macro	EM	Micro	Macro	EM
STL	UA	**0.472** ± 0.015	0.183 ± 0.014	0.193 ± 0.013	**0.745** ± 0.045	0.147 ± 0.018	**0.597** ± 0.063	0.312 ± 0.027	0.058 ± 0.005	0.023 ± 0.011	0.000 ± 0.000	0.000 ± 0.000	0.083 ± 0.062
STL-aug	UA	0.462 ± 0.021	**0.236** ± 0.020	**0.225** ± 0.022	0.730 ± 0.034	**0.155** ± 0.008	0.582 ± 0.044	**0.408** ± 0.012	0.114 ± 0.009	0.044 ± 0.014	0.258 ± 0.130	0.016 ± 0.008	0.064 ± 0.030
MTL	UA	0.445 ± 0.015	0.168 ± 0.013	0.186 ± 0.033	0.727 ± 0.036	0.143 ± 0.015	0.571 ± 0.049	0.365 ± 0.025	0.096 ± 0.014	0.040 ± 0.015	0.261 ± 0.120	0.019 ± 0.012	0.044 ± 0.034
MTL-aug	UA	0.448 ± 0.017	0.226 ± 0.011	0.217 ± 0.019	0.728 ± 0.043	0.147 ± 0.023	0.574 ± 0.040	0.403 ± 0.012	**0.145** ± 0.008	**0.052** ± 0.011	0.397 ± 0.046	0.031 ± 0.004	0.104 ± 0.018
STL	UA-CC	**0.469** ± 0.013	0.190 ± 0.019	0.193 ± 0.018	**0.770** ± 0.031	0.156 ± 0.011	0.630 ± 0.033	0.308 ± 0.048	0.046 ± 0.008	0.012 ± 0.011	0.376 ± 0.058	0.028 ± 0.008	0.069 ± 0.016
STL-aug	UA-CC	0.459 ± 0.013	**0.233** ± 0.009	**0.229** ± 0.015	0.764 ± 0.025	**0.164** ± 0.027	**0.658** ± 0.032	0.377 ± 0.024	0.079 ± 0.010	**0.041** ± 0.010	0.491 ± 0.027	0.061 ± 0.007	0.184 ± 0.030
MTL	UA-CC	0.451 ± 0.014	0.173 ± 0.014	0.199 ± 0.018	0.769 ± 0.032	0.161 ± 0.013	0.627 ± 0.033	0.367 ± 0.022	0.070 ± 0.007	0.031 ± 0.012	0.473 ± 0.058	0.055 ± 0.020	0.151 ± 0.044
MTL-aug	UA-CC	0.449 ± 0.012	0.222 ± 0.014	0.221 ± 0.012	0.739 ± 0.053	0.163 ± 0.022	0.607 ± 0.061	**0.409** ± 0.021	**0.109** ± 0.012	0.037 ± 0.015	**0.522** ± 0.029	**0.086** ± 0.015	**0.194** ± 0.050
STL	CC	0.336 ± 0.039	0.069 ± 0.007	0.043 ± 0.015	0.674 ± 0.037	0.128 ± 0.012	0.543 ± 0.054	0.000 ± 0.000	0.000 ± 0.000	0.125 ± 0.099	0.398 ± 0.019	0.053 ± 0.009	0.114 ± 0.023
STL-aug	CC	0.344 ± 0.031	0.087 ± 0.007	**0.090** ± 0.029	0.704 ± 0.023	0.149 ± 0.010	**0.599** ± 0.031	0.000 ± 0.000	0.000 ± 0.000	0.027 ± 0.022	**0.498** ± 0.035	**0.122** ± 0.028	**0.210** ± 0.030
MTL	CC	**0.352** ± 0.038	0.078 ± 0.010	0.066 ± 0.026	0.687 ± 0.042	0.139 ± 0.011	0.535 ± 0.073	0.000 ± 0.000	0.000 ± 0.000	0.079 ± 0.059	0.400 ± 0.014	0.054 ± 0.009	0.111 ± 0.027
MTL-aug	CC	0.327 ± 0.021	**0.089** ± 0.007	0.081 ± 0.015	**0.716** ± 0.034	**0.156** ± 0.011	0.597 ± 0.049	0.000 ± 0.000	0.000 ± 0.000	0.041 ± 0.029	0.479 ± 0.016	0.097 ± 0.011	0.206 ± 0.031

Note: Average scores across 10 random seeds. DistilBERT. EF = Entity Framing, NC = Narrative Classification, UA = Ukraine, CC = Climate Change.

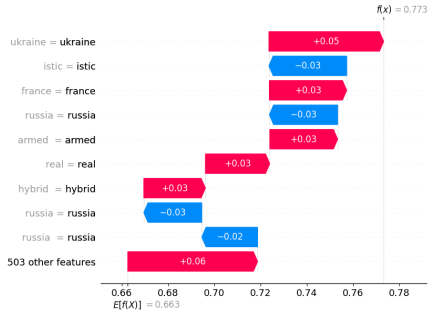
Entity Framing.
Pattern 6: STL-aug yields a macro-F1 improvement of 3-4% for EF:UA-CC, making it comparable to MTL.

Narrative Classification.
Pattern 5: MTL models consistently outperform STL and STL-aug models, which further confirms that the NC benefits more from the MTL setup.

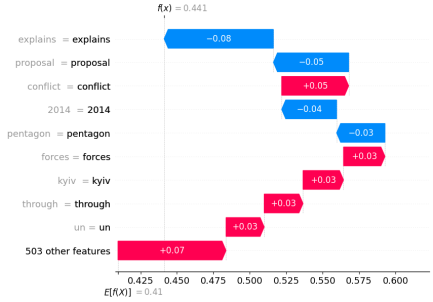
4.5 Post-Hoc Analysis

Having approached the questions quantitatively, we also present several interpretability results—namely, SHAP values. These contribute to model explainability but should be considered in light of

the ongoing academic debate regarding their effectiveness (Section 2.3).



(a) Sample 4 – STL - *Antagonist*



(b) Sample 4 – MTL - *Antagonist*

Figure 6: SHAP visualisations. Each waterfall plot illustrates the additive contributions of individual input tokens to the predicted probability of the specified label—*Antagonist* in this case. Pink bars indicate positive contributions, while blue bars show negative contributions. $E[f(x)]$ is the expected model output, and $f(x)$ is the actual prediction for the class. This allows for limited insight into the model but should be interpreted with care (Section 2.3).

5 DISCUSSION

Consistent with recent findings in multitask learning research, our results show that single-task learning (STL) performs competitively in stable, data-rich settings, such as Entity Framing (EF) in the Ukraine or mixed domains, where label distributions are well-covered and task-specific signals are strong. In such cases, the specialisation of STL models may outweigh the regularisation benefits of multitask learning (MTL) [6]. While Stickland and Murray demonstrate the promise of task-specific adapters for efficient MTL, they also note the risk of negative task interference [34], which was observed in e.g., EF:UA (Table 5). Martino et al. [19] highlight that most current propaganda systems still rely on STL setups—our findings suggest that this is a viable approach in less extreme cases of class imbalance or data scarcity.

More complex persuasion detection approaches already exist. Where Sharma et al. explore generalisation through modality—aligning image and text features to promote abstraction across multimodal inputs—this work examines whether a similar effect can emerge from task-level supervision alone [33]. While our results do not support a uniform benefit of MTL across tasks, we observe that narrative classification (NC) tends to benefit more from MTL than EF. This is likely due to NC's higher label imbalance and greater distributional shift in taxonomy across domains. In such settings, MTL appears to provide a modest buffer against overfitting to domain-specific cues, though we do not claim that this is due to broader or more abstract internal representations without definitive interpretability evidence.

MTL does not always improve robustness. In EF, performance gains are inconsistent and sometimes reversed, aligning with findings by Torbarina et al., who identify task interference as a central challenge in multitask setups [35]. This is especially relevant when tasks differ in granularity, structure, or domain alignment. Our results echo other findings in recent MTL research: Mulyar et al. and Rodriguez et al. similarly observed that while MTL can improve generalisation across tasks in biomedical NLP, the gains were often uneven and contingent on latent task compatibility [21, 29]. Rodriguez et al., for example, report a negative transfer of 2–3% in F1-score for some tasks. Our results are directionally similar—MTL and MTL-A configurations outperformed STL models when performance is averaged across tasks, but not necessarily for all tasks individually. Whether such trade-offs are desirable ultimately depends on the application's goals. The adapter-based MTL variant (MTL-A) attempts to address this trade-off by allowing task specialisation, and has shown more stable performance than naïve MTL in several cases. In particular, MTL-A insignificantly outperforms standard MTL for the EF task in the cross-domain ablation (Table 6), suggesting that allowing for modularisation, when combined with hard parameter sharing, can be a viable hybrid approach.

One alternative explanation for the results is that, during MTL training, the encoder essentially benefits from far greater supervision than during STL. Since the data setup is scarce and imbalanced, this needed to be addressed before making any claims regarding MTL's role as inductive bias, and was done through the data augmentation ablation (See Table 7). We find that simple data duplication does improve EF and NC performance, though MTL configurations, such as indirect (cross-task) out-of-domain augmentation, achieve higher generalisation gains (see Table 6).

5.1 Limitations

It remains unclear why entity framing trained on CC performs worse on UA evaluation metrics than the model trained on UA data performs on CC metrics. One possible explanation is that the binary operationalisation of domain shift—Ukraine versus Climate Change—fails to capture meaningful distinctions in linguistic, rhetorical, or topical features, thus somewhat contradicting the distribution shift analysis (Figure 1). Alternatively, the asymmetry may reflect differences in domain informativeness: the UA domain may contain structures or cues that generalise to CC, whereas the reverse may not hold. While a deeper textual analysis could, in theory, explain these differences, establishing causal relationships between such properties and model behaviour will be difficult, particularly given known challenges in the interpretability of Transformers (Section 2.3). A more tractable approach would be to expand the domain set and to vary it systematically in terms of lexical and semantic diversity. This will require a directed effort in terms of

data collection and/or annotation, since assembling a dataset from existing publicly available data is difficult due to the fragmented data practices mentioned in the Introduction 1.

A second limitation stems from architectural rigidity. The models assume robustness arises from shared encoder training across tasks, yet this comes at the cost of adaptability. Once the encoder is co-trained, adapting it to new tasks or domains risks catastrophic forgetting. This rigidity limits practical deployment of such systems unless they are paired with continual learning strategies or dynamic adaptation methods such as adapter fusion [23]. The robustness gains observed here may not generalise to modular or evolving real-world systems.

The final limitation is epistemological and pertains not just to this thesis but to the field more broadly. Much of computational propaganda research inherits assumptions from adjacent areas like fact-checking, a field grounded in factual claims and relatively stable and "objective" annotation schemes. In contrast, concepts like “antagonist” or “narrative strategy” are socially constructed, culturally contingent, and often contested. From a media theory or post-truth perspective, it is unclear to what extent these concepts can be treated as ground truth in the same way. From a conventional machine learning computational perspective, this is not a typical limitation to include. However, if we view these systems as embedded in a media landscape that has tangible implications in the real world, a closer consultation with media researchers could greatly benefit the design of such systems in the future, and is another area where this work is lacking.

6 CONCLUSION

This thesis examined whether shared linguistic representations, learned through multi-task learning (MTL), can serve as an inductive bias to improve model robustness in the face of distribution shifts within persuasion detection tasks. Addressing the broader challenge of generalisation in low-resource NLP, we evaluated MTL against single-task learning (STL) across two tasks—entity framing and narrative classification—and two domains: Ukraine conflict and Climate Change.

In response to RQ1, we find that performance degradation in out-of-domain settings cannot be attributed solely to lexical or semantic shifts, as evidenced in Figure 1. Rather, the asymmetric drop in performance—particularly the stronger generalisation of UA-trained models to CC data, compared to the reverse—suggests that class imbalance, data scarcity, and label shift exert a greater influence than surface-level variation. This asymmetry hints at the UA domain being more informative in the context of these two tasks, which appears to act as a stronger source of inductive bias. Notably, MTL proves most effective in out-of-domain settings when the auxiliary task provides complementary signal from a semantically richer or more diverse source. These results indicate that successful cross-domain transfer hinges more on the depth and breadth of training signal than on input similarity alone.

In response to RQ2, MTL shows clear benefits in scenarios marked by label shift and extreme class imbalance, especially for narrative classification. This task, being more distributionally skewed, benefits from the inductive regularisation offered by multi-task training, which stabilises performance without requiring direct (task-specific) data augmentation. These benefits are either less pronounced or absent under coarse-grained, more balanced label settings.

In response to RQ3, the evidence suggests that MTL contributes to more generalisable representations, but not necessarily domain-agnostic ones. The most substantial performance recovery occurs not when training and test domains align directly, but when the auxiliary task provides stronger domain-specific signals (e.g., narrative classification trained on Ukraine data improving entity framing evaluated on Ukraine data). This points toward domain-aware, task-mediated transfer, rather than domain-agnostic models. Ultimately, MTL likely helps in part due to the additional supervision of the encoder; nevertheless, the observed instances of task interaction suggest that it can offer benefits that single-task or task-specific data augmentation cannot. Although SHAP analyses hint at less reliance on domain-specific tokens under MTL, such interpretability tools remain indirect and limited in diagnostic value.

This thesis provides a systematic investigation of MTL as a robustness-enhancing strategy for NLP in computational propaganda research. Our findings validate MTL as a practical method to mitigate performance degradation under label imbalance and low-resource constraints. While it does not unequivocally produce domain-invariant representations, it enables more stable performance across tasks and domains, assuming a sufficient degree of task complementarity.

Future work should extend this research by incorporating more diverse domains to better disentangle the contributions of input and label shift. This greater domain diversity may further strain the epistemological limitations of current annotation practices in humanities-related machine learning, which suggests that closer interdisciplinary collaboration is needed. In parallel, more adaptive MTL architectures, such as adapter fusion, could be explored. This dual focus on architecture *and* data is important for building more robust and scalable propaganda detection systems.

REFERENCES

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2623–2631. https://doi.org/10.1145/3292500.3330701

[2] Enrique Amigo and Agustín Delgado. 2022. Evaluating Extreme Hierarchical Multi-label Classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 5809–5819. https://doi.org/10.18653/v1/2022.acl-long.399

[3] Peter Bell and Steve Renals. 2015. Regularization of context-dependent deep neural networks with context-independent multi-task training. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4290–4294. https://doi.org/10.1109/ICASSP.2015.7178780 ISSN: 2379-190X.

[4] Samantha Bradshaw and Philip N. Howard. 2018. The Global Organization of Social Media Disinformation Campaigns. *Journal of International Affairs* 71, 1.5 (2018), 23–32. https://www.jstor.org/stable/26508115 Publisher: Journal of International Affairs Editorial Board.

[5] Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer Interpretability Beyond Attention Visualization. 782–791. https://openaccess.thecvf.com/content/CVPR2021/html/Chefer_Transformer_Interpretability_Beyond_Attention_Visualization_CVPR_2021_paper.html

[6] Shijie Chen, Yu Zhang, and Qiang Yang. 2024. Multi-Task Learning in Natural Language Processing: An Overview. https://doi.org/10.48550/arXiv.2109.09138 arXiv:2109.09138 [cs].

[7] Travis G. Coan, Constantine Boussalis, John Cook, and Mirjam O. Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific Reports* 11, 1 (Nov. 2021), 22320. https://doi.org/10.1038/s41598-021-01714-4

[8] Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Findings of the NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, Anna Feldman, Giovanni Da San Martino, Alberto Barrón-Cedeño, Chris Brew, Chris Leberknight, and Preslav Nakov (Eds.). Association for Computational Linguistics, Hong Kong, China, 162–170. https://doi.org/10.18653/v1/D19-5024

[9] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images. https://doi.org/10.48550/arXiv.2105.09284 arXiv:2105.09284 [cs].

[10] Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Transactions of the Association for Computational Linguistics* 10 (Oct. 2022), 1138–1158. https://doi.org/10.1162/tacl_a_00511

[11] Tiffiany Howard, Brach Poston, and April Lopez. 2024. Extremist Radicalization in the Virtual Era: Analyzing the Neurocognitive Process of Online Radicalization. *Studies in Conflict & Terrorism* 47, 8 (Aug. 2024), 862–887. https://doi.org/10.1080/1057610X.2021.2016558 Publisher: Routledge _eprint: https://doi.org/10.1080/1057610X.2021.2016558.

[12] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. https://doi.org/10.48550/arXiv.1902.10186 arXiv:1902.10186 [cs].

[13] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. 2024. Distribution Matching for Multi-Task Learning of Classification Tasks: A Large-Scale Study on Faces & Beyond. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 3 (March 2024), 2813–2821. https://doi.org/10.1609/aaai.v38i3.28061 Number: 3.

[14] Dailin Li, Chuhan Wang, Xin Zou, Junlong Wang, Peng Chen, Jian Wang, Liang Yang, and Hongfei Lin. 2024. CoT-based Data Augmentation Strategy for Persuasion Techniques Detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 1315–1321. https://doi.org/10.18653/v1/2024.semeval-1.190

[15] Xue Li and Paul Groth. 2024. How different is different? Systematically identifying distribution shifts and their impacts in NER datasets. *Language Resources and Evaluation* (July 2024). https://doi.org/10.1007/s10579-024-09754-8

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/arXiv.1907.11692 arXiv:1907.11692 [cs].

[17] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[18] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov. 2020. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. https://doi.org/10.48550/arXiv.2009.02696 arXiv:2009.02696 [cs].

[19] Giovanni Da San Martino, Stefano Cresci, Alberto Barron-Cedeno, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A Survey on Computational Propaganda Detection. https://doi.org/10.48550/arXiv.2007.08024 arXiv:2007.08024 [cs].

[20] Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-Based Explanation Methods: A Review for NLP Interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 4593–4603. https://aclanthology.org/2022.coling-1.406/

[21] Andriy Mulyar, Ozlem Uzuner, and Bridget McInnes. 2021. MT-clinical BERT: scaling clinical information extraction with multitask learning. *Journal of the American Medical Informatics Association* 28, 10 (Oct. 2021), 2108–2115. https://doi.org/10.1093/jamia/ocab126

[22] Archita Pathak, Rohini K. Srihari, and Nihit Natu. 2021. Disinformation: analysis and identification. *Computational and Mathematical Organization Theory* 27, 3 (Sept. 2021), 357–375. https://doi.org/10.1007/s10588-021-09336-x

[23] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. https://doi.org/10.48550/arXiv.2005.00247 arXiv:2005.00247 [cs].

[24] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, Online, 46–54. https://doi.org/10.18653/v1/2020.emnlp-demos.7

[25] Jakub Piskorski, Nikolaos Nikolaidis, Nicolas Stefanovitch, Bonka Kotseva, Irene Vianini, Sopho Kharazi, and Jens Linge. 2022. Exploring Data Augmentation for Classification of Climate Change Denial: Preliminary Study. *CEUR Workshop Proceedings* (2022). https://publications.jrc.ec.europa.eu/repository/handle/JRC128613

[26] Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori (Eds.). Association for Computational Linguistics, Toronto, Canada, 2343–2361. https://doi.org/10.18653/v1/2023.semeval-1.317

[27] Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A Unified Library for Parameter-Efficient and Modular Transfer Learning. https://doi.org/10.48550/arXiv.2311.11077 arXiv:2311.11077 [cs].

[28] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 2931–2937. https://doi.org/10.18653/v1/D17-1317

[29] Nicholas E. Rodriguez, Mai Nguyen, and Bridget T. McInnes. 2022. Effects of data and entity ablation on multitask learning models for biomedical entity recognition. *Journal of Biomedical Informatics* 130 (June 2022), 104062. https://doi.org/10.1016/j.jbi.2022.104062

[30] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://doi.org/10.48550/arXiv.1910.01108 arXiv:1910.01108 [cs].

[31] Estela Saquete, David Tomás, Paloma Moreda, Patricio Martínez-Barco, and Manuel Palomar. 2020. Fighting post-truth using natural language processing: A review and open challenges. *Expert Systems with Applications* 141 (March 2020), 112943. https://doi.org/10.1016/j.eswa.2019.112943

[32] Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 2931–2951. https://doi.org/10.18653/v1/P19-1282

[33] Shivam Sharma, Atharva Kulkarni, Tharun Suresh, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. Characterizing the Entities in Harmful Memes: Who is the Hero, the Villain, the Victim?. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 2149–2163. https://doi.org/10.18653/v1/2023.eacl-main.157

[34] Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 5986–5995. https://proceedings.mlr.press/v97/stickland19a.html ISSN: 2640-3498.

[35] Lovre Torbarina, Tin Ferkovic, Lukasz Roguski, Velimir Mihelcic, Bruno Sarlija, and Zeljko Kraljevic. 2023. Challenges and Opportunities of Using Transformer-Based Multi-Task Learning in NLP Through ML Lifecycle: A Survey. https://doi.org/10.48550/arXiv.2308.08234 arXiv:2308.08234 [cs].

[36] Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining (IJDWM)* 3, 3 (2007), 1–13. https://EconPapers.repec.org/RePEc:igg:jdwm00:v:3:y:2007:i:3:p:1-13 Publisher: IGI Global.

[37] Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention Interpretability Across NLP Tasks. https://doi.org/10.48550/arXiv.1909.11218 arXiv:1909.11218 [cs].

[38] Zirui Wang, Zihang Dai, Barnabas Poczos, and Jaime Carbonell. 2019. Characterizing and Avoiding Negative Transfer. 11293–11302. https://openaccess.thecvf.com/content_CVPR_2019/html/Wang_Characterizing_and_Avoiding_Negative_Transfer_CVPR_2019_paper.html

[39] Yi Xin, Junlong Du, Qiang Wang, Ke Yan, and Shouhong Ding. 2024. MmAP: Multi-Modal Alignment Prompt for Cross-Domain Multi-Task Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 14 (March 2024), 16076–16084. https://doi.org/10.1609/aaai.v38i14.29540 Number: 14.

[40] Yu Zhang and Qiang Yang. 2022. A Survey on Multi-Task Learning. *IEEE Transactions on Knowledge and Data Engineering* 34, 12 (Dec. 2022), 5586–5609. https://doi.org/10.1109/TKDE.2021.3070203 Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Appendix A HYPER-PARAMETERS

To ensure model stability and reliable comparisons, I conducted systematic hyperparameter tuning using Optuna, a Bayesian optimisation framework [1]. Each experiment was treated as an independent optimisation problem. The objective function maximised macro-averaged F1 on a validation set. All tuning experiments used consistent search spaces across models and tasks, with 3 trials per configuration. After initial sweeps, I observed that certain hyperparameter values consistently yielded strong performance across both encoders and task formulations. Specifically, learning rates in the 2e5 to 3e5 range provided stable convergence; batch sizes of 8 generally proved to be the optimum; three to four epochs sufficed in most cases, and thresholds of 0.3 to 0.4 optimised macro-F1. Rather than fine-tuning parameters per experiment, I selected the most common high-performing configuration as a shared baseline for subsequent ablations. This reflects the principle of parsimony and ensures that observed differences are due to model architecture or learning setup, not parameter idiosyncrasies.

Appendix B GENERATIVE AI USED IN THIS THESIS

My experience with GenAI tools like Copilot is that, especially with more complex pipelines like mine, they can help if and only if the user has a very precise improvement in mind. In all other cases, unless the user is designing a canonical PyTorch implementation of a known algorithm, they tend to overcomplicate things to a degree that wastes time rather than saves it. In my case, since I was programming the system from scratch independently, GenAI often acted as a "sanity check" when tracing the root cause of bugs. For example, I had a working loop for routing data according to label granularity to make a mixed training set. I then suddenly needed to do the same, but based on another filtering condition. GenAI was generally good at pointing out the problem with the new loop when shown a version of the loop that worked as intended. It was also generally very helpful for troubleshooting LaTeX, or converting my Python visualisations into LaTeX and vice versa. Prompts like “What is the equivalent of this TensorFlow function in PyTorch?” were helpful, too. Tools like Grammarly were used for grammar checking and rephrasing some sentences.

Appendix C FULL RESULTS

Table 8: This table mirrors **Main Results** and validates them by running the models using the RoBERTa-base encoder—this type of encoder would generally be used in the state-of-the-art studies identified earlier. Since the number of experiments was high in this study, we opted for a distilled model instead and report these results only for comparison.

Setup	Data	EF: UA			EF: CC			NC: UA			NC: CC		
		Micro-F1	Macro-F1	Exact Match	Micro-F1	Macro-F1	EM	Micro-F1	Macro-F1	EM	Micro-F1	Macro-F1	EM
STL	UA	**0.526** ± 0.006	**0.256** ± 0.019	**0.282** ± 0.019	**0.821** ± 0.006	**0.194** ± 0.032	**0.678** ± 0.013	0.342 ± 0.008	0.070 ± 0.003	0.037 ± 0.000	0.020 ± 0.034	0.001 ± 0.002	0.029 ± 0.025
MTL	UA	0.521 ± 0.007	0.246 ± 0.008	0.275 ± 0.022	0.801 ± 0.032	0.183 ± 0.013	0.612 ± 0.054	0.401 ± 0.011	0.117 ± 0.007	0.059 ± 0.011	0.294 ± 0.015	0.018 ± 0.001	0.071 ± 0.038
MTL-A	UA	0.514 ± 0.027	0.249 ± 0.027	0.249 ± 0.053	0.775 ± 0.030	0.175 ± 0.011	0.593 ± 0.050	0.404 ± 0.011	0.132 ± 0.011	0.056 ± 0.019	0.377 ± 0.049	0.028 ± 0.003	0.095 ± 0.041
STL	UA-CC	**0.531** ± 0.009	0.251 ± 0.016	0.269 ± 0.010	0.812 ± 0.008	0.186 ± 0.020	0.700 ± 0.006	0.328 ± 0.040	0.051 ± 0.006	0.028 ± 0.019	0.404 ± 0.032	0.029 ± 0.008	0.119 ± 0.046
MTL	UA-CC	0.524 ± 0.021	**0.256** ± 0.015	0.281 ± 0.031	0.829 ± 0.023	0.202 ± 0.010	0.652 ± 0.023	0.396 ± 0.022	0.097 ± 0.003	0.056 ± 0.019	0.526 ± 0.006	0.079 ± 0.005	0.181 ± 0.008
MTL-A	UA-CC	0.530 ± 0.015	0.256 ± 0.013	0.287 ± 0.011	0.823 ± 0.017	0.187 ± 0.016	0.663 ± 0.028	0.395 ± 0.013	0.096 ± 0.004	0.043 ± 0.014	0.495 ± 0.002	0.091 ± 0.013	0.138 ± 0.036
STL	CC	0.404 ± 0.022	0.100 ± 0.011	0.104 ± 0.029	0.805 ± 0.021	0.187 ± 0.005	**0.659** ± 0.033	0.000 ± 0.000	0.000 ± 0.000	0.068 ± 0.053	0.438 ± 0.028	0.070 ± 0.006	0.171 ± 0.025
MTL	CC	0.378 ± 0.020	0.101 ± 0.008	0.117 ± 0.039	0.805 ± 0.008	0.182 ± 0.013	0.656 ± 0.039	0.000 ± 0.000	0.000 ± 0.000	0.031 ± 0.019	0.433 ± 0.018	0.078 ± 0.017	0.162 ± 0.016
MTL-A	CC	**0.418** ± 0.020	**0.103** ± 0.004	**0.127** ± 0.019	**0.811** ± 0.019	**0.187** ± 0.005	0.648 ± 0.029	0.000 ± 0.000	0.000 ± 0.000	0.037 ± 0.033	0.445 ± 0.026	0.101 ± 0.008	0.152 ± 0.036

Note: Average scores across 3 random seeds. RoBERTa-base. EF = Entity Framing, NC = Narrative Classification, UA = Ukraine, CC = Climate Change.

Table 9: Main Results (trained on UA-CC) – Wilcoxon p-values. Statistically significant differences ($p < 0.05$) are marked with an asterisk.

Task	Comparison	Micro (UA)	Macro (UA)	EM (UA)	Micro (CC)	Macro (CC)	EM (CC)
EF	STL vs MTL	0.000*	0.002*	0.881	0.388	0.701	0.325
	STL vs MTL-A	0.498	0.231	0.705	0.927	0.784	0.932
	MTL vs MTL-A	0.006*	0.000*	0.896	0.927	1.000	0.397
NC	STL vs MTL	0.000*	0.000*	0.001*	0.000*	0.000*	0.000*
	STL vs MTL-A	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*
	MTL vs MTL-A	0.231	0.004*	0.047*	0.021*	0.021*	0.201

Table 10: Main Results (trained on CC) – Wilcoxon p-values. Statistically significant differences ($p < 0.05$) are marked with an asterisk.

Task	Comparison	Micro (UA)	Macro (UA)	EM (UA)	Micro (CC)	Macro (CC)	EM (CC)
EF	STL vs MTL	0.261	0.000*	0.003*	0.133	0.000*	0.984
	STL vs MTL-A	0.498	0.003*	0.002*	0.756	0.007*	0.776
	MTL vs MTL-A	0.674	0.143	0.173	0.231	0.202	0.643
NC	STL vs MTL	–	–	0.091	0.756	0.648	0.721
	STL vs MTL-A	–	–	0.546	0.000*	0.000*	0.000*
	MTL vs MTL-A	–	–	0.210	0.000*	0.000*	0.000*

Table 11: Cross-domain Ablation (EF-UA, NC-CC) Wilcoxon p-values. Statistically significant differences ($p < 0.05$) are marked with an asterisk.

Task	Comparison	Micro (UA)	Macro (UA)	EM (UA)	Micro (CC)	Macro (CC)	EM (CC)
EF	STL vs MTL	0.002*	0.002*	0.002*	0.105	0.002*	0.301
	STL vs MTL-A	0.002*	0.002*	0.002*	0.002*	0.006*	0.002*
	MTL vs MTL-A	0.014*	0.020*	0.008*	0.375	0.557	0.137
NC	STL vs MTL	0.004*	0.010*	0.062	0.002*	0.002*	0.027*
	STL vs MTL-A	0.002*	0.002*	0.020*	0.002*	0.002*	0.018*
	MTL vs MTL-A	0.695	0.160	0.648	0.770	0.695	0.406

Table 12: Cross-domain Ablation (EF-CC, NC-UA) Wilcoxon p-values. Statistically significant differences ($p < 0.05$) are marked with an asterisk.

Task	Comparison	Micro (UA)	Macro (UA)	EM (UA)	Micro (CC)	Macro (CC)	EM (CC)
EF	STL vs MTL	0.002*	0.002*	0.002*	0.014*	0.064	0.016*
	STL vs MTL-A	0.002*	0.002*	0.002*	0.002*	0.014*	0.002*
	MTL vs MTL-A	0.275	0.020*	0.795	0.770	0.625	0.277
NC	STL vs MTL	0.004*	0.004*	0.148	0.002*	0.002*	0.010*
	STL vs MTL-A	0.004*	0.002*	0.105	0.002*	0.002*	0.027*
	MTL vs MTL-A	0.770	0.027*	0.898	0.557	0.004*	0.922

Table 13: Data Augmentation Ablation (trained on UA-CC) – Wilcoxon p-values. Statistically significant differences ($p < 0.05$) are marked with an asterisk.

Task	Comparison	Micro (UA)	Macro (UA)	EM (UA)	Micro (CC)	Macro (CC)	EM (CC)
EF	STL vs STL-aug	0.049*	0.002*	0.004*	0.770	0.375	0.041*
	MTL vs MTL-aug	0.846	0.002*	0.020*	0.160	0.770	0.523
	STL-aug vs MTL	0.232	0.002*	0.002*	0.625	1.000	0.064
NC	STL vs STL-aug	0.004*	0.002*	0.002*	0.004*	0.002*	0.002*
	MTL vs MTL-aug	0.010*	0.002*	0.562	0.027*	0.004*	0.004*
	STL-aug vs MTL	0.232	0.049*	0.066	0.375	0.275	0.039*

Table 14: Data Augmentation Ablation (trained on CC) – Wilcoxon p-values. Statistically significant differences ($p < 0.05$) are marked with an asterisk.

Task	Comparison	Micro (UA)	Macro (UA)	EM (UA)	Micro (CC)	Macro (CC)	EM (CC)
EF	STL vs STL-aug	0.625	0.002*	0.002*	0.010*	0.002*	0.029*
	MTL vs MTL-aug	0.027*	0.002*	0.223	0.105	0.010*	0.039*
	STL-aug vs MTL	0.922	0.010*	0.164	0.432	0.064	0.043*
NC	STL vs STL-aug	–	–	0.008*	0.002*	0.002*	0.002*
	MTL vs MTL-aug	–	–	0.250	0.002*	0.002*	0.002*
	STL-aug vs MTL	–	–	0.035*	0.002*	0.002*	0.002*

Figure 7: Ablation: Label granularity. Average values across 10 seeds. DistilBERT. We vary the label taxonomy (fine/coarse) independently for EF and NC, resulting in four configurations (fine–fine, coarse–fine, fine–coarse, coarse–coarse). These experiments assess the role of label resolution in cross-task learning and representation sharing.

EF–Fine, NC–Fine granularity (Baseline for this experiment)

Model	Train	EF: UA			EF: CC			NC: UA			NC: CC		
		Micro	Macro	EM	Micro	Macro	EM	Micro	Macro	EM	Micro	Macro	EM
STL	UA	**0.472** ± 0.015	0.183 ± 0.014	0.193 ± 0.013	**0.745** ± 0.045	**0.147** ± 0.018	0.597 ± 0.063	0.312 ± 0.027	0.058 ± 0.005	0.023 ± 0.011	0.000 ± 0.000	0.000 ± 0.000	0.083 ± 0.062
MTL	UA	0.445 ± 0.015	0.168 ± 0.013	0.186 ± 0.033	0.727 ± 0.036	0.143 ± 0.015	0.571 ± 0.049	0.365 ± 0.025	0.096 ± 0.014	0.040 ± 0.015	0.261 ± 0.120	0.019 ± 0.012	0.044 ± 0.034
MTL-A	UA	0.458 ± 0.015	**0.191** ± 0.017	**0.198** ± 0.015	0.743 ± 0.026	0.139 ± 0.015	**0.602** ± 0.052	**0.380** ± 0.031	**0.118** ± 0.017	**0.044** ± 0.011	**0.350** ± 0.062	**0.027** ± 0.006	**0.079** ± 0.035
STL	UA-CC	**0.469** ± 0.013	0.190 ± 0.019	0.193 ± 0.018	**0.770** ± 0.031	0.156 ± 0.011	**0.630** ± 0.033	0.308 ± 0.048	0.046 ± 0.008	0.012 ± 0.011	0.376 ± 0.058	0.028 ± 0.008	0.069 ± 0.016
MTL	UA-CC	0.451 ± 0.014	0.173 ± 0.014	0.199 ± 0.018	0.769 ± 0.032	**0.161** ± 0.013	0.627 ± 0.033	0.367 ± 0.022	0.070 ± 0.007	0.031 ± 0.012	0.473 ± 0.058	0.055 ± 0.020	0.151 ± 0.044
MTL-A	UA-CC	0.464 ± 0.016	**0.194** ± 0.016	**0.201** ± 0.029	0.758 ± 0.033	0.155 ± 0.018	0.627 ± 0.040	**0.387** ± 0.016	**0.082** ± 0.007	**0.036** ± 0.013	**0.510** ± 0.019	**0.069** ± 0.009	**0.154** ± 0.035
STL	CC	0.336 ± 0.039	0.069 ± 0.007	0.043 ± 0.015	0.674 ± 0.037	0.128 ± 0.012	0.543 ± 0.054	0.000 ± 0.000	0.000 ± 0.000	0.125 ± 0.099	0.398 ± 0.019	0.053 ± 0.009	0.114 ± 0.023
MTL	CC	**0.352** ± 0.038	0.078 ± 0.010	0.066 ± 0.026	0.687 ± 0.042	0.139 ± 0.011	0.535 ± 0.073	0.000 ± 0.000	0.000 ± 0.000	0.079 ± 0.059	0.400 ± 0.014	0.054 ± 0.009	0.111 ± 0.027
MTL-A	CC	0.349 ± 0.021	**0.078** ± 0.006	**0.086** ± 0.033	**0.690** ± 0.027	**0.145** ± 0.011	**0.560** ± 0.042	0.000 ± 0.000	0.000 ± 0.000	0.077 ± 0.041	**0.447** ± 0.016	**0.081** ± 0.009	**0.154** ± 0.023

Note: This table is essentially identical to **Main Results**–the difference is in the lower number of random seeds to make it comparable to subsequent granularity ablations. Average scores across 10 random seeds. DistilBERT. EF = Entity Framing, NC = Narrative Classification, UA = Ukraine, CC = Climate Change.

EF–Fine, NC–Coarse granularity

Model	Train	EF: UA			EF: CC			NC: UA			NC: CC		
		Micro	Macro	EM	Micro	Macro	EM	Micro	Macro	EM	Micro	Macro	EM
STL	UA	**0.472** ± 0.015	**0.183** ± 0.014	0.193 ± 0.013	**0.745** ± 0.045	**0.147** ± 0.018	**0.597** ± 0.063	0.490 ± 0.030	0.180 ± 0.011	0.146 ± 0.022	0.393 ± 0.122	0.055 ± 0.017	0.169 ± 0.077
MTL	UA	0.445 ± 0.011	0.159 ± 0.010	0.188 ± 0.022	0.736 ± 0.064	0.147 ± 0.012	0.582 ± 0.098	0.483 ± 0.026	0.202 ± 0.014	0.153 ± 0.025	0.479 ± 0.053	0.074 ± 0.010	0.259 ± 0.040
MTL-A	UA	0.462 ± 0.013	0.178 ± 0.008	0.192 ± 0.017	0.735 ± 0.049	0.138 ± 0.014	0.580 ± 0.080	**0.492** ± 0.038	**0.215** ± 0.019	**0.156** ± 0.028	**0.503** ± 0.039	**0.074** ± 0.013	**0.303** ± 0.031
STL	UA-CC	**0.469** ± 0.013	0.190 ± 0.019	0.193 ± 0.018	**0.770** ± 0.031	0.156 ± 0.011	**0.630** ± 0.033	0.474 ± 0.015	0.135 ± 0.010	0.154 ± 0.026	0.575 ± 0.043	0.138 ± 0.024	0.357 ± 0.020
MTL	UA-CC	0.449 ± 0.023	0.168 ± 0.017	**0.201** ± 0.018	0.764 ± 0.035	0.157 ± 0.014	0.623 ± 0.023	**0.502** ± 0.023	0.175 ± 0.018	0.144 ± 0.028	0.613 ± 0.023	0.187 ± 0.021	**0.383** ± 0.042
MTL-A	UA-CC	0.457 ± 0.014	**0.191** ± 0.014	0.190 ± 0.022	0.765 ± 0.020	**0.160** ± 0.009	0.613 ± 0.047	0.499 ± 0.022	**0.181** ± 0.010	0.153 ± 0.023	**0.623** ± 0.023	**0.193** ± 0.026	0.381 ± 0.049
STL	CC	**0.336** ± 0.039	0.069 ± 0.007	0.043 ± 0.015	0.674 ± 0.037	0.128 ± 0.012	0.543 ± 0.054	0.000 ± 0.000	0.000 ± 0.000	0.040 ± 0.044	0.591 ± 0.035	**0.296** ± 0.055	0.376 ± 0.040
MTL	CC	0.322 ± 0.041	0.069 ± 0.007	**0.068** ± 0.031	0.705 ± 0.028	0.134 ± 0.021	0.577 ± 0.051	0.000 ± 0.000	0.000 ± 0.000	0.029 ± 0.011	**0.600** ± 0.026	0.271 ± 0.039	0.399 ± 0.024
MTL-A	CC	0.334 ± 0.041	**0.070** ± 0.010	0.052 ± 0.020	0.666 ± 0.031	0.124 ± 0.014	0.541 ± 0.063	0.000 ± 0.000	0.000 ± 0.000	0.022 ± 0.026	0.592 ± 0.040	0.279 ± 0.049	**0.400** ± 0.033

Note: We present a selected set of visualisations of these UA-CC models in the Results section. DistilBERT. EF = Entity Framing, NC = Narrative Classification, UA = Ukraine, CC = Climate Change.

EF–Coarse, NC–Fine granularity

Model	Train	EF: UA			EF: CC			NC: UA			NC: CC		
		Micro	Macro	EM	Micro	Macro	EM	Micro	Macro	EM	Micro	Macro	EM
STL	UA	**0.600** ± 0.015	**0.553** ± 0.019	**0.518** ± 0.018	**0.796** ± 0.029	**0.694** ± 0.053	**0.762** ± 0.037	0.312 ± 0.027	0.058 ± 0.005	0.023 ± 0.011	0.000 ± 0.000	0.000 ± 0.000	0.083 ± 0.062
MTL	UA	0.547 ± 0.012	0.440 ± 0.022	0.408 ± 0.037	0.639 ± 0.032	0.434 ± 0.056	0.578 ± 0.051	0.347 ± 0.137	0.102 ± 0.076	0.073 ± 0.074	0.183 ± 0.208	0.028 ± 0.031	**0.176** ± 0.069
MTL-PAL	UA	0.537 ± 0.016	0.430 ± 0.043	0.395 ± 0.030	0.633 ± 0.033	0.436 ± 0.070	0.560 ± 0.065	**0.386** ± 0.110	**0.123** ± 0.080	**0.094** ± 0.084	**0.306** ± 0.157	**0.040** ± 0.033	0.166 ± 0.113
STL	UA-CC	**0.605** ± 0.018	**0.562** ± 0.023	**0.520** ± 0.031	**0.817** ± 0.024	**0.715** ± 0.057	**0.799** ± 0.029	0.308 ± 0.048	0.046 ± 0.008	0.012 ± 0.011	0.376 ± 0.058	0.028 ± 0.008	0.069 ± 0.016
MTL	UA-CC	0.546 ± 0.021	0.449 ± 0.039	0.402 ± 0.054	0.684 ± 0.030	0.559 ± 0.048	0.610 ± 0.065	0.299 ± 0.139	0.064 ± 0.052	0.063 ± 0.067	0.488 ± 0.126	0.102 ± 0.077	**0.229** ± 0.147
MTL-A	UA-CC	0.539 ± 0.017	0.451 ± 0.033	0.386 ± 0.044	0.697 ± 0.022	0.549 ± 0.044	0.664 ± 0.044	**0.342** ± 0.104	**0.076** ± 0.050	**0.067** ± 0.070	**0.515** ± 0.100	**0.108** ± 0.070	0.223 ± 0.137
STL	CC	0.477 ± 0.044	0.365 ± 0.049	0.376 ± 0.029	**0.720** ± 0.041	**0.548** ± 0.083	**0.685** ± 0.051	0.000 ± 0.000	0.000 ± 0.000	0.125 ± 0.099	0.398 ± 0.019	0.053 ± 0.009	0.114 ± 0.023
MTL	CC	**0.489** ± 0.045	**0.385** ± 0.047	0.376 ± 0.045	0.688 ± 0.039	0.536 ± 0.066	0.620 ± 0.078	0.000 ± 0.000	0.000 ± 0.000	0.074 ± 0.070	0.480 ± 0.094	**0.140** ± 0.096	0.229 ± 0.134
MTL-A	CC	0.469 ± 0.052	0.345 ± 0.085	**0.379** ± 0.039	0.687 ± 0.047	0.534 ± 0.049	0.638 ± 0.084	0.000 ± 0.000	0.000 ± 0.000	0.075 ± 0.087	**0.486** ± 0.075	0.131 ± 0.079	**0.244** ± 0.124

Note: We note that when we switch the label granularity of EF to 'coarse' only, MTL introduces a much more substantial negative transfer than observed in fine-grained models. NC still benefits from the MTL setup, but here we see that MTL-A actually outperforms the naive MTL approach. Average scores across 10 random seeds. DistilBERT. EF = Entity Framing, NC = Narrative Classification, UA = Ukraine, CC = Climate Change.

EF–Coarse, NC–Coarse granularity

Model	Train	EF: UA			EF: CC			NC: UA			NC: CC		
		Micro	Macro	EM	Micro	Macro	EM	Micro	Macro	EM	Micro	Macro	EM
STL	UA	**0.600** ± 0.015	0.553 ± 0.019	0.518 ± 0.018	**0.796** ± 0.029	**0.694** ± 0.053	**0.762** ± 0.037	0.490 ± 0.030	0.180 ± 0.011	0.146 ± 0.022	0.393 ± 0.122	0.055 ± 0.017	0.169 ± 0.077
MTL	UA	0.546 ± 0.026	0.443 ± 0.021	0.396 ± 0.040	0.634 ± 0.031	0.421 ± 0.078	0.576 ± 0.062	0.466 ± 0.030	0.167 ± 0.013	0.141 ± 0.020	0.363 ± 0.104	0.053 ± 0.015	0.236 ± 0.043
MTL-A	UA	0.551 ± 0.008	0.432 ± 0.039	0.392 ± 0.043	0.618 ± 0.053	0.430 ± 0.059	0.570 ± 0.082	**0.494** ± 0.026	**0.203** ± 0.020	**0.154** ± 0.026	**0.449** ± 0.067	**0.069** ± 0.012	**0.267** ± 0.038
STL	UA-CC	**0.605** ± 0.018	**0.562** ± 0.023	**0.520** ± 0.031	**0.817** ± 0.024	**0.715** ± 0.057	**0.799** ± 0.029	**0.474** ± 0.015	**0.135** ± 0.010	**0.154** ± 0.026	0.575 ± 0.043	0.138 ± 0.024	**0.357** ± 0.020
MTL	UA-CC	0.549 ± 0.014	0.449 ± 0.040	0.390 ± 0.059	0.686 ± 0.052	0.550 ± 0.076	0.599 ± 0.129	0.439 ± 0.048	0.119 ± 0.013	0.121 ± 0.018	0.588 ± 0.033	0.162 ± 0.033	0.350 ± 0.038
MTL-A	UA-CC	0.541 ± 0.017	0.440 ± 0.037	0.378 ± 0.022	0.706 ± 0.015	0.559 ± 0.046	0.670 ± 0.036	0.450 ± 0.041	0.129 ± 0.017	0.131 ± 0.025	**0.603** ± 0.029	**0.178** ± 0.036	0.340 ± 0.031
STL	CC	0.477 ± 0.044	0.365 ± 0.049	0.376 ± 0.029	**0.720** ± 0.041	**0.548** ± 0.083	**0.685** ± 0.051	0.000 ± 0.000	0.000 ± 0.000	0.040 ± 0.044	**0.591** ± 0.035	**0.296** ± 0.055	**0.376** ± 0.040
MTL	CC	**0.504** ± 0.058	**0.377** ± 0.096	**0.378** ± 0.054	0.674 ± 0.040	0.527 ± 0.062	0.591 ± 0.098	0.000 ± 0.000	0.000 ± 0.000	0.049 ± 0.030	0.556 ± 0.024	0.220 ± 0.040	0.341 ± 0.035
MTL-A	CC	0.477 ± 0.057	0.332 ± 0.078	0.370 ± 0.042	0.673 ± 0.052	0.529 ± 0.065	0.615 ± 0.100	0.000 ± 0.000	0.000 ± 0.000	0.032 ± 0.039	0.565 ± 0.021	0.217 ± 0.027	0.346 ± 0.042

Note: For the final granularity configuration—coarse/coarse—we arrive at results where STL outperforms MTL more conclusively, supporting our conclusion that hard paremeter sharing is more beneficial in imbalanced and data-scarce settings. Average scores across 10 random seeds. DistilBERT. EF = Entity Framing, NC = Narrative Classification, UA = Ukraine, CC = Climate Change.