Mapping Weaponised Victimhood: A Machine Learning Approach

Samantha Butcher[a] and Beatriz De La Iglesia[b]
Department of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, U.K.

Keywords: Political Discourse, Named Entity Recognition, BERT, Entity Framing, Multi-Task Learning, Natural Language Processing.

Abstract: Political discourse frequently leverages group identity and moral alignment, with weaponised victimhood (WV) standing out as a powerful rhetorical strategy. Dominant actors employ WV to frame themselves or their allies as victims, thereby justifying exclusionary or retaliatory political actions. Despite advancements in Natural Language Processing (NLP), existing computational approaches struggle to capture such subtle rhetorical framing at scale, especially when alignment is implied rather than explicitly stated. This paper introduces a dual-task framework designed to address this gap by linking Named Entity Recognition (NER) with a nuanced rhetorical positioning classification (positive, negative, or neutral - POSIT). By treating rhetorical alignment as a structured classification task tied to entity references, our approach moves beyond sentiment-based heuristics to yield a more interpretable and fine-grained analysis of political discourse. We train and compare transformer-based models (BERT, DistilBERT, RoBERTa) across Single-Task, Multi-Task, and Task-Conditioned Multi-Task Learning architectures. Our findings demonstrate that NER consistently outperformed rhetorical positioning, achieving higher F1-scores and distinct loss dynamics. While single-task learning showed wide loss disparities (e.g., BERT NER 0.45 vs POSIT 0.99), multi-task setups fostered more balanced learning, with losses converging across tasks. Multi-token rhetorical spans proved challenging but showed modest F1 gains in integrated setups. Neutral positioning remained the weakest category, though targeted improvements were observed. Models displayed greater sensitivity to polarised language (e.g., RoBERTa TC-MTL reaching 0.55 F1 on negative spans). Ultimately, entity-level F1 scores converged (NER: 0.60–0.61; POSIT: 0.50–0.52), suggesting increasingly generalisable learning and reinforcing multi-task modelling as a promising approach for decoding complex rhetorical strategies in real-world political language.

1 INTRODUCTION

Political discourse frequently leverages group identity and moral alignment, with weaponised victimhood (WV) standing out as a powerful rhetorical strategy. Dominant actors employ WV to frame themselves or their allies as victims, thereby justifying exclusionary or retaliatory political actions. Despite advancements in Natural Language Processing (NLP), existing computational approaches struggle to capture such subtle rhetorical framing at scale, especially when alignment is implied rather than explicitly stated.

This paper introduces a novel dual-task framework designed to address this gap by linking Named Entity Recognition (NER) with a nuanced rhetorical positioning classification (positive, negative, or neutral). By conceptualising rhetorical alignment as a structured classification problem directly tied to entity references, our approach moves beyond simplistic sentiment-based heuristics in order to yield a more interpretable and fine-grained understanding of complex political discourse. We train and compare transformer-based models (BERT, DistilBERT, RoBERTa) across Single-Task, Multi-Task, and Task-Conditioned Multi-Task Learning architectures to evaluate their effectiveness.

Our findings reveal a promising dynamic: multi-task learning setups, particularly standard MTL, offer a robust framework for jointly addressing entity recognition and rhetorical positioning. While immediate gains in positioning F1 scores were modest or mixed (e.g., DistilBERT dropped slightly from 0.51 to 0.48), MTL consistently promoted more stable and efficient shared learning, evidenced by converging loss values across tasks, unlike the divergence seen in STL

[a] https://orcid.org/0009-0000-0041-6768
[b] https://orcid.org/0000-0003-2675-5826

Butcher, S. and De La Iglesia, B.
Mapping Weaponised Victimhood: A Machine Learning Approach.
DOI: 10.5220/0013673600004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 216-223

(e.g., BERT's NER loss of 0.45 vs POSIT loss of 0.99). This convergence suggests the model is internalising both tasks in a more unified way, laying essential groundwork for future refinements in rhetorical classification, particularly in contexts requiring nuanced understanding of identity and alignment.

2 RELATED RESEARCH

WV draws on a broad set of populist rhetorical techniques, including identity framing, emotive grievance, blame attribution, and the inversion of power hierarchies. Though not always labelled explicitly as WV, such strategies have been examined across diverse political and ideological contexts, from US narratives of cultural loss and status anxiety (Bebout, 2022, 2019) to conservative and incel discourses grounded in affective grievance and perceived disempowerment (Barton Hronešová and Kreiss, 2024; Homolar and Löfflmann, 2022; Kelly et al., 2024). These appeals typically reduce complexity into binaries of victim and villain, legitimising reactionary responses through moral positioning (Johnson, 2017; Zembylas, 2021; Pascale, 2019). While WV as a cohesive phenomenon remains underexplored in NLP, its components, such as emotional tone, stance, and identity targeting, have been approached via sentiment analysis, stance detection, and entity tagging (Teso et al., 2018; Warin and Stojkov, 2023), often using lexicons or simple classifiers to surface rhetorical dynamics.

SRL has also been used to support structured analysis of rhetorical meaning, identifying roles such as actor, affected, or instrument within a sentence. While initially developed for formal text, SRL has been adapted to conversational data like tweets (Liu and Li, 2011; Xu et al., 2021), making it suitable for political discourse. However, such contexts often involve complex references, such as shifting pronouns, compound identity phrases like *the American people*, or ideologically marked groups like *the radical left*, that go beyond standard named entity boundaries. To capture these spans, researchers frequently use BIO tagging, a scheme that assigns "B-" to the beginning of an entity, "I-" to subsequent tokens, and "O" to non-entity tokens. For instance, Zhou et al. (2023) used BIO tagging to extract hate speech targets and associated framing.

Our study addresses the gap between existing component-level analyses and a more integrated modelling of rhetorical strategies like WV. While prior work has tackled sentiment, stance, and entities separately, few approaches link identity references to rhetorical alignment in a structured, scalable way. We combine these elements to model how entities are framed morally or politically, supporting future detection of WV and similar discursive strategies.

3 METHODOLOGY

Our approach consisted of three main stages: (1) identifying key rhetorical features of WV through discourse analysis and SRL; (2) constructing and annotating a training corpus drawn from a high-density source of WV rhetoric; and (3) experimenting with transformer-based architectures to evaluate model performance on rhetorical framing tasks.

3.1 Discourse and Feature Design

Discourse analysis enables examination of how language is used to construct identity, moral alignment, and power. SRL complements this by identifying who is acting, who is affected, and what the action is, revealing how agency and blame are distributed in WV. This pairing supports structured feature identification in rhetorical positioning.

A defining feature of WV is the construction of ingroups and outgroups. Ingroup references often appear via first-person plural pronouns (e.g., *we*, *us*) or identity-based phrases (e.g., *American workers*, *our public health professionals*). Outgroups are frequently vague (e.g., *they*, *these people*), inviting ideological projection. WV also commonly involves a speaker positioning themselves as protector of a threatened ingroup (Bebout, 2019). In this paper, we focus specifically on these identity references—how groups are invoked, labelled, and morally positioned within political rhetoric. By modelling both the linguistic form (namely pronouns, group identifiers and identity-based phrases) and the rhetorical stance attached to them (positive, negative, or neutral), we aim to capture the alignment strategies central to WV discourse. This targeted approach offers a scalable foundation for analysing how speakers construct legitimacy through appeals to shared identity and grievance.

3.2 Corpus Construction and Annotation

We draw on political speech corpora (USA Political Speeches Dataset, 2022; Donald Trump's Rallies Dataset, 2020), totalling 595 speeches between 2015–2024. All were attributed to a speaker known for frequent WV rhetoric. Annotation proceeded in

two stages: first, entities were identified and tagged across the corpus. These included pronouns (for example, *us*, *you*, *them*), social groups and institutions (*Americans*, *the Senate*), and also abstract ideas (like *the American Dream*). Abstract references were included because rhetorical positioning often involves praise or blame directed at concepts rather than specific agents - for example, speakers may attack ideas such as *liberalism* or defend notions like *freedom* or *our country* without attributing them to a particular group or individual. These abstract references still carry alignment or hostility and are thus critical to understanding how identity and blame are constructed.

Once entities were identified, each was assigned a rhetorical positioning label (`POSITIVE`, `NEUTRAL`, or `NEGATIVE`) based on surrounding context. This labelling was done at the entity level rather than the sentence level, as multiple entities within the same sentence could be framed differently.

An initial broad pass of the corpus was used to annotate entities and their rhetorical positioning, generating a large pool of examples reflecting how various entity types—pronouns, groups, institutions, and abstract concepts—were framed in context. From this, a smaller, balanced subset was curated for training, ensuring diversity in entity–position combinations while avoiding over-representation of repeated phrases or named references. This variation supports both WV detection and more robust generalisation overall.

Each speech was preprocessed by stripping timestamps and non-verbal metadata, then segmented into *context windows* averaging 130–160 characters. This segmentation strategy balances semantic coherence with model efficiency, and aligns with the study's long-term goal of applying models to social media discourse (namely Reddit), where comments are similarly brief and often fragmented. Smaller windows also help isolate rhetorical structures, particularly when multiple group references appear in close proximity.

For example, the line:

> *'They are attacking our families and destroying our country."*

contains multiple references, namely *they*, *our families*, and *our country*, each annotated independently.

Clearly, this annotation schema does not assign fixed ingroup or outgroup labels. Instead, it focuses on how each entity is rhetorically positioned within the context (POSITIVE, NEUTRAL, or NEGATIVE). This choice reflects how speakers may refer not only to allies and adversaries, but also to adjacent groups, institutions, or abstract concepts whose alignment is context-dependent. This is particularly valuable when the speaker's stance is subtle, implied, or shifts across discourse. Even for human annotators, determining group alignment often requires rereading and interpretation. By foregrounding how entities are positioned rather than what they are, this schema supports more flexible and accurate modelling of identity-related rhetoric.

3.3 Dataset Summary

The final dataset contains 5,103 labelled examples drawn from 3,325 unique windows. More examples appear than content windows because, as highlighted, a window may contain multiple examples of entities. Table 1 provides a breakdown by span type and positioning label.

Table 1: Breakdown of span types and rhetorical positioning labels.

Tag Type	**Count**
PRONOUN	2,465
IDENTITY_MARKER	2,637
Total Examples	**5,103**
Positioning Label	**Count**
POSITIVE	1,811
NEUTRAL	1,717
NEGATIVE	1,574

Although small, the dataset was carefully constructed and consistently annotated to test whether fine-tuned transformer models could learn patterns of rhetorical positioning from limited but high-quality input.

3.4 Model Selection

Transformer-based models such as BERT have become central to NLP tasks requiring contextual interpretation, including entity recognition and rhetorical classification (Aldera et al., 2021; Botella-Gil et al., 2024; Chaudhari and Pawar, 2022). Their capacity to model relational and semantic nuance makes them particularly suited to discourse-level tasks involving alignment and framing.

This study evaluates three BERT-based variants: **BERT**, **RoBERTa**, and **DistilBERT**. Each presents a different trade-off between performance and efficiency. Table 2 summarises their comparative strengths.

3.5 Process Flow and Model Architecture

All models follow a consistent preprocessing pipeline. First, labelled span data is tokenised

Table 2: Overview of selected BERT variants with strengths and limitations.

Model	Strengths	Limitations
BERT	Strong general model; good with context.	High computational cost; not task-specific.
RoBERTa	Trained on more data than BERT; often higher accuracy.	Resource-heavy; slower to train.
DistilBERT	Smaller and faster; retains 95% of BERT's performance.	Slightly lower accuracy on complex tasks.

and converted into BIO tags to delineate entity boundaries. Token alignment checks are performed to ensure that annotated spans map cleanly onto subword tokens. The processed inputs are then encoded into the format expected by the transformer model, including input IDs and attention masks. The architecture diverges at the training stage, depending on how the NER and Positioning tasks are handled:

- **Single-Task Learning (STL):** Each task is trained independently using a separate model. There is no parameter sharing or interaction between tasks.
- **Multi-Task Learning (MTL):** A shared model is trained to perform both tasks jointly. A single encoder processes the input, and two parallel classification heads are applied: one for NER, one for Positioning. The model computes separate losses for each task, which are then averaged to guide weight updates. While this approach allows the model to learn shared representations, it treats the tasks as independent in output.
- **Task-Conditioned Multi-Task Learning (TC-MTL):** This variant introduces directed task interaction. The model first predicts NER spans, which are passed through a fusion layer to produce entity-aware features used by the Positioning head. This design reflects how human annotators might work; first identifying an entity, then assessing its rhetorical stance, potentially reducing ambiguity by letting the model focus on positioning only after entity boundaries are known.

All models are trained end-to-end using BIO-tagged supervision. Unlike standard BIO tagging, which marks only span boundaries, our approach encodes both the span structure and the entity type. We use `B-` markers (namely `B-PRONOUN`, `B-IDENTITY MARKER`) for the start of a tagged span, and corresponding `I-` markers to indicate continuation when the span is more than one token. An equivalent scheme is applied for rhetorical positioning, with tags such as `B-POSITIVE` and `I-NEGATIVE`. Each entity is therefore represented by two aligned BIO sequences: one for entity recognition and one for positioning. This structure allows models to learn from shared span boundaries while treating classification tasks independently when needed.

Table 3: Example of dual BIO-tagged tokenised span (NER and Positioning).

Token	NER BIO	POSIT BIO
They	B-PRONOUN	B-NEGATIVE
are	O	O
targeting	O	O
our	B-IDENTITY_MARKER	B-POSITIVE
veteran	I-IDENTITY_MARKER	I-POSITIVE
##s	I-IDENTITY_MARKER	I-POSITIVE
.	O	O

During training, token-level predictions are decoded into spans and compared against gold annotations, with alignment checks including both automated mismatch detection and manual review.

3.6 Training Details

All models were trained for five epochs with consistent hyperparameters (Table 4), including a batch size of 16 and learning rate of 5×10^{-5}. Early stopping was not used to ensure full convergence.

In STL and MTL, B-tags were given greater weight (e.g., B-PRONOUN = 2.0, I-PRONOUN = 1.0) to emphasise span boundaries and help the model better learn where entities begin. The `O` tag was assigned minimal weight. In MTL, a weighted joint loss (0.7 Positioning, 0.3 NER) was used to support the more complex classification task. TC-MTL introduced a warm-up phase in which the NER head was trained alone for two epochs before Positioning was added. This ensured the model had learned stable entity representations before passing them to the Positioning head. Without this, early-stage noise from untrained entity predictions could propagate, undermining Positioning accuracy. Once NER outputs had stabilised, softmax probabilities were fused with encoder hidden states to predict Positioning tags.

3.7 Evaluation Metrics

Model performance was assessed with four categories of metrics, reported separately for NER and Positioning:

Table 4: Shared hyperparameters across all models.

Hyperparameter	Value
Max sequence length	512
Epochs	5
Learning rate	5×10^{-5}
Optimiser	AdamW
Loss	CrossEntropy (ignore index = -100)
Batch size	16
Train/eval split	80/20 (seed = 42)
Random seed	42 (all frameworks)

- **Overall:** Weighted token-level accuracy, precision, recall, and F1 (includes `O` tags).
- **Entity-Level:** Macro-averaged scores across B-/I- tags only (excludes `O`).
- **Per-Label:** Precision, recall, and F1 for each specific tag (e.g., `B-PRONOUN`, `I-NEGATIVE`).
- **Loss:** Average final-epoch loss for each model.

We separate overall and entity-level metrics because overall scores include the `O` tag, which is both the most common and the easiest to predict, potentially inflating performance. Entity-level metrics exclude `O` and focus only on B-/I- tags, offering a more meaningful measure of how well the model identifies and classifies relevant spans.

4 RESULTS

Having established a consistent training setup across architectures, the following results provide an initial comparison of model performance. Results are reported separately for each model with attention to both overall trends and task-specific observations.

4.1 STL

Full results for the STL tasks can be found in Table 5 and Table 6.

The STL results show consistently strong performance across all models on the NER task, with overall F1-scores ranging from 0.89 (RoBERTa) to 0.91 (BERT and DistilBERT). Entity-level F1-scores are notably lower, peaking at 0.66 for both BERT and DistilBERT, and slightly lower for RoBERTa at 0.64. This gap highlights the increased difficulty of precise boundary detection. B- labels generally outperform I- labels, reflecting their prominence in marking span starts and their slightly higher representation in the dataset. The particularly low F1 for I-PRONOUN (e.g., 0.5181 for BERT) stems from the rarity of multi-token pronouns, making I-PRONOUN infrequent and harder to learn.

In the POSIT task, overall performance remains strong, with BERT achieving the highest overall F1-score (0.88), closely followed by DistilBERT (0.87) and RoBERTa (0.86). However, entity-level F1-scores are more modest, ranging from 0.51 to 0.52 across models. Neutral spans proved most difficult for all models, with I-NEUTRAL F1-scores ranging from 0.43 (DistilBERT) to 0.45 (RoBERTa). RoBERTa also achieved the best performance on B-NEGATIVE (F1: 0.5638), suggesting increased sensitivity to more polarised language. Overall, STL provides stable and competitive results, though entity-level detection—particularly of internally continued spans—remains a key challenge.

4.2 MTL

Full results can be shown in Table 7 and Table 8.

The MTL results show strong NER performance across all models, with BERT and DistilBERT achieving the highest overall F1 (0.89) and RoBERTa slightly behind (0.87). Entity-level F1 remains lower, with all models performing similarly. As with STL, boundary detection proves challenging, especially for internally continued spans, with precision trailing recall. DistilBERT shows the highest sensitivity to span detection, while I-IDENTITY MARKER consistently outperforms its B- counterpart, indicating improved internal span modelling under MTL.

In the POSIT task, overall performance is comparable to STL, with F1-scores of 0.87 for BERT and DistilBERT, and 0.86 for RoBERTa. Entity-level F1 scores cluster around 0.50–0.51 across models. RoBERTa performs best on B-NEGATIVE and B-NEUTRAL, while BERT leads on I-POSITIVE. Neutral spans remain the most difficult across all models, particularly I-NEUTRAL, which shows the weakest performance. Overall, MTL supports strong rhetorical classification and internal span learning but continues to struggle with boundary precision and neutral positioning.

4.3 TC-MTL

Results for this final style of architecture can be found in Table 9 and Table 10.

TCMTL results show strong NER performance across all models, with overall F1-scores ranging from 0.88 (RoBERTa) to 0.89 (BERT). BERT and RoBERTa both achieve the highest entity-level F1 (0.61), though RoBERTa benefits from stronger recall (0.87) despite lower precision. I-IDENTITY MARKER outperforms B-IDENTITY across models, indicating improved internal span recognition.

Table 5: Overall and entity-specific performance metrics for NER and POSIT tasks (STL pipeline).

Task	Model	Eval Loss	Overall Acc.	Overall Prec.	Overall Rec.	Overall F1	Entity Prec.	Entity Rec.	Entity F1
NER	BERT	0.45	**0.90**	**0.93**	**0.90**	**0.91**	**0.53**	0.87	**0.66**
	DistilBERT	**0.34**	**0.90**	**0.93**	**0.90**	**0.91**	0.52	**0.91**	**0.66**
	RoBERTa	0.30	0.88	0.92	0.88	0.89	0.51	0.89	0.64
POSIT	BERT	0.99	**0.86**	**0.91**	**0.86**	**0.88**	**0.41**	0.71	**0.52**
	DistilBERT	0.94	0.85	**0.91**	0.85	0.87	0.40	0.71	0.51
	RoBERTa	**0.92**	0.83	**0.91**	0.83	0.86	0.40	**0.74**	**0.52**

Table 6: Per-label precision, recall, and F1-scores for NER and POSIT tasks (STL pipeline).

Task	Label	BERT			DistilBERT			RoBERTa		
		P	R	F1	P	R	F1	P	R	F1
NER	B-PRONOUN	**0.60**	0.95	**0.74**	0.57	0.97	0.72	0.59	**0.98**	**0.74**
	I-PRONOUN	0.38	0.83	0.52	**0.39**	**0.88**	**0.54**	0.38	0.79	0.51
	B-IDENTITY MARKER	**0.56**	0.87	**0.68**	0.54	**0.92**	**0.68**	0.52	0.91	0.66
	I-IDENTITY MARKER	**0.59**	0.83	0.69	**0.59**	**0.89**	**0.71**	0.54	**0.89**	0.67
POSIT	B-POSITIVE	**0.38**	0.77	0.51	0.37	0.79	0.50	**0.38**	**0.81**	**0.52**
	I-POSITIVE	**0.45**	0.79	**0.57**	0.44	**0.81**	0.57	0.40	0.77	0.53
	B-NEUTRAL	**0.43**	0.62	0.51	0.40	0.60	0.48	0.42	**0.68**	**0.52**
	I-NEUTRAL	**0.36**	0.56	0.44	0.35	0.56	0.43	0.35	**0.62**	**0.45**
	B-NEGATIVE	**0.44**	0.76	0.55	0.42	0.74	0.53	**0.44**	**0.77**	**0.56**
	I-NEGATIVE	**0.42**	**0.78**	**0.55**	0.40	0.74	0.52	**0.42**	**0.78**	**0.55**

Table 7: Overall and entity-specific performance metrics for NER and POSIT tasks (MTL pipeline).

Task	Model	Eval Loss	Overall Acc.	Overall Prec.	Overall Rec.	Overall F1	Entity Prec.	Entity Rec.	Entity F1
NER	BERT	**0.73**	**0.87**	**0.93**	**0.87**	**0.89**	**0.47**	0.88	**0.61**
	DistilBERT	0.90	**0.87**	**0.93**	**0.87**	**0.89**	**0.47**	**0.92**	**0.61**
	RoBERTa	0.74	0.86	0.92	0.86	0.87	**0.47**	0.86	**0.61**
POSIT	BERT	**0.73**	**0.85**	**0.91**	**0.85**	**0.87**	**0.42**	0.71	**0.51**
	DistilBERT	0.90	0.84	**0.91**	0.84	**0.87**	0.37	0.68	0.48
	RoBERTa	0.74	0.83	**0.91**	0.83	0.86	0.39	**0.73**	**0.51**

Table 8: Per-label precision, recall, and F1-scores for NER and POSIT tasks (MTL pipeline).

Task	Label	BERT			DistilBERT			RoBERTa		
		P	R	F1	P	R	F1	P	R	F1
NER	B-PRONOUN	0.54	0.96	0.69	0.54	0.96	0.69	**0.57**	**0.97**	**0.72**
	I-PRONOUN	0.31	0.73	0.44	0.32	**0.85**	**0.47**	**0.37**	0.60	0.46
	B-IDENTITY MARKER	**0.48**	0.93	**0.64**	0.47	**0.94**	0.63	0.46	0.91	0.61
	I-IDENTITY MARKER	**0.53**	0.91	**0.67**	0.52	0.93	**0.67**	0.47	**0.95**	0.63
POSIT	B-POSITIVE	0.36	**0.82**	0.50	**0.41**	0.59	0.48	0.40	0.73	**0.52**
	I-POSITIVE	0.41	**0.81**	**0.55**	**0.45**	0.65	0.53	0.41	0.79	0.54
	B-NEUTRAL	**0.48**	0.59	**0.53**	0.31	0.73	0.43	0.38	**0.76**	0.51
	I-NEUTRAL	**0.46**	0.40	0.43	0.28	**0.70**	0.40	0.32	0.69	**0.44**
	B-NEGATIVE	0.40	**0.78**	0.53	0.41	0.72	0.52	**0.45**	0.71	**0.55**
	I-NEGATIVE	**0.41**	**0.83**	**0.54**	0.38	0.69	0.49	0.39	0.69	0.50

Table 9: Overall and entity-specific performance metrics for NER and POSIT tasks (TCMTL pipeline).

Task	Model	Eval Loss	Overall Acc.	Overall Prec.	Overall Rec.	Overall F1	Entity Prec.	Entity Rec.	Entity F1
NER	BERT	0.89	**0.88**	**0.93**	**0.88**	**0.89**	**0.47**	0.89	**0.61**
	DistilBERT	0.78	**0.88**	**0.93**	**0.88**	**0.89**	0.46	**0.91**	0.60
	RoBERTa	**0.67**	0.86	0.92	0.86	0.88	**0.47**	0.87	**0.61**
POSIT	BERT	0.89	**0.86**	**0.91**	**0.86**	**0.88**	**0.41**	0.69	0.51
	DistilBERT	0.78	0.85	**0.91**	0.85	0.87	0.40	0.69	0.50
	RoBERTa	**0.67**	0.83	**0.91**	0.83	0.86	**0.41**	**0.73**	**0.52**

Table 10: Per-label precision, recall, and F1-scores for NER and POSIT tasks (TCMTL pipeline).

Task	Label	BERT			DistilBERT			RoBERTa		
		P	R	F1	P	R	F1	P	R	F1
NER	B-PRONOUN	0.54	**0.97**	0.69	0.53	**0.97**	0.69	**0.55**	**0.97**	**0.70**
	I-PRONOUN	0.32	0.79	**0.45**	0.27	**0.83**	0.40	**0.35**	0.62	0.44
	B-IDENTITY MARKER	**0.51**	0.91	**0.65**	0.50	0.92	**0.65**	0.47	**0.93**	0.62
	I-IDENTITY MARKER	0.53	0.89	0.67	**0.54**	0.92	**0.68**	0.49	**0.94**	0.65
POSIT	B-POSITIVE	**0.40**	0.74	**0.52**	0.39	0.71	0.50	0.35	**0.82**	0.49
	I-POSITIVE	0.45	0.74	**0.56**	**0.46**	0.68	0.55	0.39	**0.83**	0.53
	B-NEUTRAL	0.42	**0.67**	0.52	0.43	0.56	0.49	**0.47**	0.61	**0.53**
	I-NEUTRAL	0.33	0.59	0.42	0.39	**0.63**	**0.48**	**0.40**	0.51	0.45
	B-NEGATIVE	0.42	0.70	0.53	0.36	**0.80**	0.50	**0.43**	0.79	**0.55**
	I-NEGATIVE	**0.41**	0.68	0.51	0.37	0.79	0.50	0.40	**0.82**	**0.54**

In the POSIT task, overall F1 remains high across models, BERT (0.88), DistilBERT (0.87), and RoBERTa (0.86), with entity-level F1 tightly clustered around 0.51–0.52. Precision remains low (sitting at around 0.40), with recall helping offset performance gaps. I-NEUTRAL continues to be the most difficult label, though DistilBERT performs slightly better than others. BERT achieves the highest I-POSITIVE F1 (0.56), while RoBERTa leads on B-NEGATIVE (0.55). TCMTL improves span consistency but leaves challenges in neutral classification and boundary precision.

5 DISCUSSION

NER consistently outperformed POSIT across all architectures, achieving higher entity-level F1-scores and, in the STL configuration, significantly lower evaluation losses. For instance, BERT recorded a loss of 0.45 on the NER task compared to 0.99 on POSIT. However, in MTL and TC-MTL setups, loss values were often identical across tasks within a given model, suggesting that loss alone may not reliably capture relative task complexity in multi-task configurations.

Rhetorical positioning spans proved more difficult to model. Entity-level precision for POSIT remained low across models, typically around 0.40 to 0.42, and span fragmentation was a frequent error. Models would often correctly tag salient identity tokens such as “American” but fail to include the full expression “the American people,” leading to incomplete representations of rhetorical intent. Despite label weighting, I-tags such as I-POSITIVE consistently achieved higher F1-scores than their corresponding B-tags, indicating stronger modelling of internal span content. However, this pattern was less consistent for negative spans, where I-NEGATIVE scores were often similar to or slightly lower than B-NEGATIVE.

RoBERTa showed consistently strong recall, such as a score of 0.87 on the NER task in TC-MTL, but underperformed in span precision. It frequently omitted key contextual modifiers, such as possessives like “our” in phrases like “our public health professionals.” In these cases, the model successfully identified the core entity (“public health professionals”) but failed to capture the full rhetorical framing, diminishing its ability to model speaker alignment or affiliation.

MTL showed no clear gains in POSIT performance. Entity-level F1-scores remained flat or declined compared to STL, for example, DistilBERT dropped from 0.51 to 0.48, while TC-MTL produced no consistent improvements across tasks. These results suggest that task conditioning may require more data, architectural adjustment, or strategies such as curriculum learning to realise its full benefits.

At the label level, BERT achieved the strongest results on I-POSITIVE (F1: 0.56 in TC-MTL), while RoBERTa led on B-NEGATIVE (F1: 0.55). However, all models struggled with I-NEUTRAL, which consistently had the lowest F1-scores across settings, underlining the persistent difficulty of detecting subtle or non-polar rhetorical positioning.

6 CONCLUSION

While MTL showed the most promise for learning both entity and rhetorical positioning tasks, future work will explore how to further optimise this setup, particularly through better span boundary detection and improved handling of neutral positioning. Despite its intuitive design, TC-MTL has not yet yielded consistent gains, suggesting that the sequential dependency it models may require more sophisticated integration or richer supervision to translate into mea-

surable improvements. One particular area of focus, which should benefit all models going forward, will be an expanded training dataset which includes more diverse examples from a broader range of sources.

Beyond these models, we plan to test other transformer architectures (such as deBERTa or a BiLSTM-enhanced BERT model) and apply transfer learning to new datasets from social media and news. These domains will provide more diverse rhetorical strategies and enable evaluation of generalisation beyond the original corpus. Ultimately, we aim to scale this framework toward more robust detection of WV discourse across varied contexts.

REFERENCES

Aldera, S., Emam, A., Al-Qurishi, M., Alrubaian, M., and Alothaim, A. (2021). Exploratory data analysis and classification of a new arabic online extremism dataset. 9:161613–161626.

Barton Hronešová, J. and Kreiss, D. (2024). Strategically hijacking victimhood: A political communication strategy in the discourse of Viktor Orbán and Donald Trump. pages 1–19.

Bebout, L. (2019). Weaponizing victimhood: Discourses of oppression and the maintenance of supremacy on the right. In Nadler, A. and Bauer, A., editors, *News on the Right*, pages 64–83. Oxford University PressNew York, 1 edition.

Bebout, L. (2022). Weaponizing victimhood in u.s. political culture and the january 6, 2021, insurrection. Submission to the Select Committee to Investigate the January 6th Attack on the United States Capitol.

Botella-Gil, B., Sepúlveda-Torres, R., Bonet-Jover, A., Martínez-Barco, P., and Saquete, E. (2024). Semi-automatic dataset annotation applied to automatic violent message detection. 12:19651–19664.

Chaudhari, D. D. and Pawar, A. V. (2022). A systematic comparison of machine learning and NLP techniques to unveil propaganda in social media. Publisher: IGI Global.

Donald Trump's Rallies Dataset (2020). Donald trump's rallies. Kaggle. https://www.kaggle.com/datasets/christianlillelund/donald-trumps-rallies [Accessed July 2025].

Homolar, A. and Löfflmann, G. (2022). Weaponizing masculinity: Populism and gendered stories of victimhood. 16(2):131–148.

Johnson, P. E. (2017). The art of masculine victimhood: Donald trump's demagoguery. 40(3):229–250.

Kelly, M., Rothermel, A.-K., and Sugiura, L. (2024). Victim, violent, vulnerable: A feminist response to the incel radicalisation scale. 18(1):91–119.

Liu, P. and Li, S. (2011). A corpus-based method to improve feature-based semantic role labeling. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pages 205–208. IEEE.

Pascale, C.-M. (2019). The weaponization of language: Discourses of rising right-wing authoritarianism. 67(6):898–917.

Teso, E., Olmedilla, M., Martínez-Torres, M., and Toral, S. (2018). Application of text mining techniques to the analysis of discourse in eWOM communications from a gender perspective. *Technological Forecasting and Social Change*, 129:131–142.

USA Political Speeches Dataset (2022). Usa political speeches. Kaggle. https://www.kaggle.com/datasets/beridzeg45/usa-political-speeches [Accessed July 2025].

Warin, T. and Stojkov, A. (2023). Discursive dynamics and local contexts on Twitter: The refugee crisis in Europe. *Discourse & Communication*, 17(3):354–380.

Xu, K., Wu, H., Song, L., Zhang, H., Song, L., and Yu, D. (2021). Conversational semantic role labeling. 29:2465–2475.

Zembylas, M. (2021). Interrogating the affective politics of white victimhood and resentment in times of demagoguery: The risks for civics education. 40(6):579–594.

Zhou, L., Caines, A., Pete, I., and Hutchings, A. (2023). Automated hate speech detection and span extraction in underground hacking and extremist forums. 29(5):1247–1274.