Heroes, Villains, and Victims: Character Narratives in the WPS Agenda of the UNSC

Hannah Steinbach **Imge Yüzüncüoglu** **Raluca Rilla** **Manfred Stede**
University of Potsdam
Department of Linguistics
Potsdam, Germany
{hannahsteinbach0312, i.yuezuencue, ralucarilla}@gmail.com
stede@uni-potsdam.de

Abstract

We investigate how heroes, victims, and villains are constructed in debates on the United Nations Security Council's Women, Peace, and Security (WPS) agenda. Drawing from 2,566 speeches delivered between 2000 and 2019, we examine how (gendered) entities are framed within diplomatic discourse using topic modeling, clustering, and supervised learning. To assess the potential of automated character role identification, we manually annotate 54 speeches with character role labels and evaluate a fine-tuned RoBERTa classifier alongside two chat-optimized Large Language Models (DeepSeek-R1, Llama3.3 70B). Our findings reveal substantial variation in model performance, with RoBERTa demonstrating best overall performance. Our analysis shows that women are framed as both empowered agents and vulnerable subjects, while perpetrators often remain unnamed. The UNSC often casts itself as a hero by emphasizing its contributions to the WPS agenda. All code and annotated datasets are publicly available to facilitate further research on narrative framing and role attribution in this domain.

1 Introduction

According to Gehring and Grigoletto (2023), "narratives are a crucial group-based mechanism that influences human decision-making" (p. 1). The *Narrative Policy Framework (NPF)* (Jones and McBeth, 2010) defines narratives through distinct components including context, storyline, moral of the story, and characters. Building on the drama triangle (Karpman, 1968), the character roles *hero*, *victim*, and *villain* are central elements of narrative construction. While prior research has applied these frameworks extensively in areas such as climate change communication (e.g., Wolters et al. 2021; Fremann et al. 2023; Gehring and Grigoletto 2023; Grasso et al. 2025), less attention has been given to their role in diplomatic settings.

In multilateral settings like the United Nations (UN), the strategic use of narratives allows states to construct legitimacy and influence international norms. By portraying different entities as heroes, victims, or villains, speakers subtly shape perceptions of agency and responsibility. These narrative strategies play a crucial role in the Women, Peace, and Security (WPS) agenda of the UN Security Council (UNSC). Initiated by Resolution 1325 (2000), the agenda seeks to prevent violence against women and girls in (post-)conflict contexts, promote gender equality, increase women's participation in security processes, and integrate a gender perspective (UNDP, 2023). Given these objectives, one can anticipate that women are central figures in these debates, yet they are also framed in specific roles. They often appear as victims, needing protection, yet also as heroes whose (potential) participation in UN peace and security efforts is framed as beneficial. At the same time, the very call to increase their involvement implies that they are still largely excluded from these processes, reinforcing their victimhood. The UN itself is also cast as a hero, both as protector and enabler of women's involvement. Since the objectives of the UNSC are a key component of the agenda, the sessions themselves can be expected to involve similar narratives.

This paper investigates how heroes, victims, and villains are constructed in WPS open debates, and evaluates the ability of Large Language Models (LLMs) to automatically identify these roles. We begin with a meta-analysis of gender representation among UNSC speakers, followed by topic modeling using clustering techniques and BERTopic (Grootendorst, 2022) to explore dominant themes and the framing of gendered entities. To assess the potential for automated classification of narrative roles, we frame the task as a supervised learning problem and fine-tune a RoBERTa model. We then compare its performance to that of two chat-optimized LLMs: Llama and DeepSeek. The manually annotated data serves as a benchmark. Our contributions include:

1. A meta-analysis of gender representation in UNSC WPS debates.
2. An analysis of the most frequent topics in the WPS agenda using clustering methods, combined with a comparison of how gendered entities are discussed across these topics.
3. A new corpus of 54 speeches annotated for hero, victim, and villain roles, with guidelines.
4. Evaluation of automatic role classification using RoBERTa, Llama, and DeepSeek.

All code and annotated datasets are publicly available and encourage further research on narrative framing and role attribution in this domain: GitHub.

2 Related Work

2.1 Narrative Framing in the United Nations

Gibbings (2011) states that "operating at the UN is akin to acquiring a second language" (p. 533), reflecting the cultural norms that shape diplomatic discourse. One of these norms is the expectation to frame issues positively: "UN speech styles encourage positive visions and utopian dreams" (Gibbings, 2011, p. 534). Based on four months of fieldwork on Resolution 1325, Gibbings notes that women are often framed as peacemakers and knowledge-bearers, while framings that deviate from this narrative are discouraged. She illustrates this with the case of two Iraqi women who, at a 2003 UN meeting, criticized the US and UK invasion and the UN's lack of support. Though they were barred from formally speaking, their critical remarks at the informal session led to them being labeled "angry." This incident reveals not only the discursive etiquette within the UN but also that when women display agency, they are expected to do so as peaceful agents.

While the UN frequently depicts women as (peaceful) "agents of change," this framing is often accompanied by portrayals of women as victims, particularly in contexts emphasizing vulnerability and protection. The UN's tendency to depict women primarily through these dual lenses has been widely criticized. Başer (2024), for instance, warns that these framings reinforce gender norms by failing to acknowledge women also as perpetrators of violence and by erasing male or LGBTQ+ victims. Similarly, Carpenter (2005) finds that the UN's civilian protection discourse heavily relies on the phrase "women and children," sidelining men. This framing emphasizes perceived vulnerability and helps garner donor support (Carpenter, 2005).

De la Rosa and Lázaro (2019) analyze metaphorical framings in UNSC resolutions from 2000 to 2015 and find a shift from portraying women solely as victims to depicting them as empowered agents in the fight against sexual violence. This suggests a change in the portrayal of women, at least within the resolutions.

Open debates, however, differ from resolutions. While resolutions are negotiated and largely agreed upon, debates reflect a broader range of Member State perspectives and are less curated. Although True and Wiener (2019) do not explicitly employ the victim, villain, and hero framework, their analysis offers valuable insights consistent with these characterizations. In their study of the 15th-anniversary open debate on WPS in 2015, they find that portrayals of women remain largely victim-focused, particularly in relation to sexual violence, thereby risking the reinforcement of stereotypes. They also identify states positioning themselves as heroes: either as champions of the agenda (e.g., Global North states) or as implementers under constraint (e.g., conflict-affected states). Others, like Russia and China, resist expansion of the agenda, emphasizing sovereignty. Perpetrator framing remains politically delicate. While non-state actors like ISIS are widely condemned, direct criticism of fellow states is rare, since it is easier to call out "breaches of the norm by non-state actors [...] than to name and shame breaches made by states present" (True and Wiener, 2019, p. 563).

True and Wiener (2019)'s analysis of an open debate offers important insights into character framings, highlighting the victimization of women (and, in some cases, states), self-portrayals of heroism by Member States, and the targeting of non-state actors as perpetrators, while notably avoiding direct criticism of present states. Given the UN's tendency to maintain a positive tone, we anticipate an emphasis on achievements and successes—thus, on heroes. Building on these findings, our next step is to assess whether these insights persist across a broader set of open debates and to explore the potential of automated approaches to process large volumes of text more efficiently. The following section reviews existing approaches to automatic character framing analysis

2.2 Automated Character Role Identification in NLP

The automated identification of character roles such as heroes, victims, and villains has long been relevant in NLP, particularly in framing analysis. Gomez-Zara et al. (2018) developed a browser extension that identifies role framings in news articles based on named entity recognition, sentiment analysis, and lexical similarity to prototypical roles. While the implementation has not undergone a formal evaluation yet, the authors acknowledge that the browser extension may struggle with complex sentence structures (Gomez-Zara et al., 2018).

Sap et al. (2017) examine how stereotypes are formed through language by extending the connotation frame lexicon of Rashkin et al. (2016) with verbs that capture "power" and "agency:" "power" reflects the level of authority conveyed by a verb, while "agency" indicates how much a character is “powerful, decisive, and capable of pushing forward their own storyline” (Sap et al., 2017, p.4). To test this empirically, they analyze gender bias in modern films by comparing how male and female characters differ in attributed power and agency. Their approach offers a useful tool for studying narrative structure and power dynamics.

Stammbach et al. (2022) reformulate character role labeling as a reading comprehension task. Prompting GPT-3 with context and role-specific questions (e.g., “Who is the hero?”), they evaluate performance across diverse domains: news articles on fracking, Disney film summaries, and US State of the Union addresses. GPT-3 outperforms rule-based baselines (Gomez-Zara et al., 2018), and proves to be especially effective on narrative text.

Character framing has been frequently studied in the context of ecology and climate change. The developers of the *Character-Role Narrative Framework* (Gehring and Grigoletto, 2023), which builds on the drama triangle (Karpman, 1968), applied it to study US climate policy on Twitter as an example of large-scale climate change discourse, revealing how narrative roles influence the virality and public reception of environmental messages (Gehring and Grigoletto, 2023). For their analysis, they used RoBERTa (Liu et al., 2019) combined with XGBoost (Chen and Guestrin, 2016), concluding that integrating LLMs with a limited training dataset is a "very efficient approach to measuring narratives in large text data sets" (Gehring and Grigoletto, 2023, p. 32).

Following this approach, Grasso et al. (2025) evaluate four LLMs (GPT-4o, GPT-4, GPT-4-turbo, and LLaMA-3.1-8B) for character role detection (adding the class "beneficiary") and categorization on tweets from the *Ecoverse* dataset and editorial paragraphs from *Nature & Science*. They then adapt the best performing model, GPT-4, to create two new annotated datasets of tweets and editorial paragraphs, followed by a subsequent error analysis. They conclude that LLMs are in fact suitable for large-scale character-role analysis in the environmental domain, though they advise that human validation remains necessary.

While previous research has demonstrated the potential of LLMs for role detection in narrative, journalistic texts, and social media posts such as tweets, their applicability to formal diplomatic discourse remains underexplored.

3 Dataset and Task

3.1 Data

Using the dataset of UNSC debates from 1997–2019 by Schoenfeld et al. (2019), we identified 68 sessions (4613 speeches) addressing the WPS agenda here. We removed presidential speeches (which are mainly formalities), the duplicate of one debate, and the proceedings of S/PV.6180 from 2009 under the meeting of the year 2001, after observing that there was no WPS agenda meeting in 2001. The final dataset consists of a total of 2566 speeches, for which we each extracted metadata including year, session, speaker affiliation, gender (based on honorifics), and language of the speech using regular expressions.

3.2 Manual Annotation of Dataset

We manually annotated 54 speeches for the roles of "hero," "victim," and "villain," taking into account both commonly held understandings of these roles and the definitions provided by Klapp (1954), Cohen (2011), Gomez-Zara et al. (2018), Bergstrand and Jasper (2018) and Jasper et al. (2018). The full character definitions can be found in Section Appendix C. The speeches were sampled to roughly reflect their distribution across different years. Based on four pilot annotations, we developed corpus-specific guidelines. In addition to excluding abstract entities (e.g. "sexual violence"), we expanded the definitions of the roles to adapt them

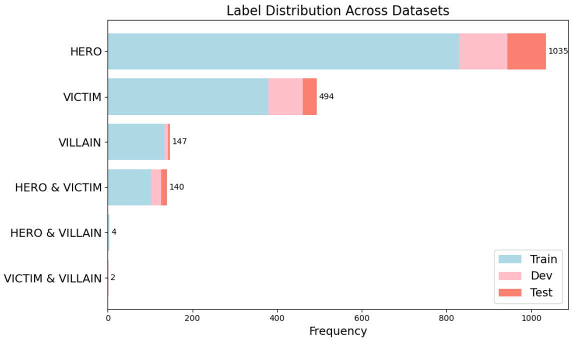


Figure 1: Character Label Distribution.

to the context of the data. Two important additions were made: the "hero" category was extended to encompass actors as having potential to bring about positive change but facing structural oppression, and the "villain" category to include actors portrayed as obstructing equal rights and justice for victims. Following Fremann et al. (2023), who criticizes existing datasets and frameworks for assuming that entities can only be assigned a single role rather than recognizing that they can be associated with multiple roles simultaneously, we allowed overlapping roles and nested annotations.

The 54 annotated speeches were split into 44 training set, 5 development set, and 5 test set speeches. Annotations were done individually by the authors using INCEpTION (Klie et al., 2018). Inter-annotator agreement averaged at 0.38 (Krippendorff's α).[1]The relatively low initial agreement stemmed partly from differing interpretations of role definitions, which prompted a subsequent refinement of the guidelines. It was also influenced by smaller inconsistencies in span selection, such as whether to include punctuation, modifiers, or complete noun phrases. These discrepancies reflected the inherent subjectivity and complexity of role labeling task. Nevertheless, after collaboratively reviewing all speeches and discussing disagreements, we reached full consensus on the final labels.

3.3 Annotated Data Insights

The annotation of the 54 speeches offers valuable insights. The following statistics were generated only after model evaluation to avoid introducing any bias.

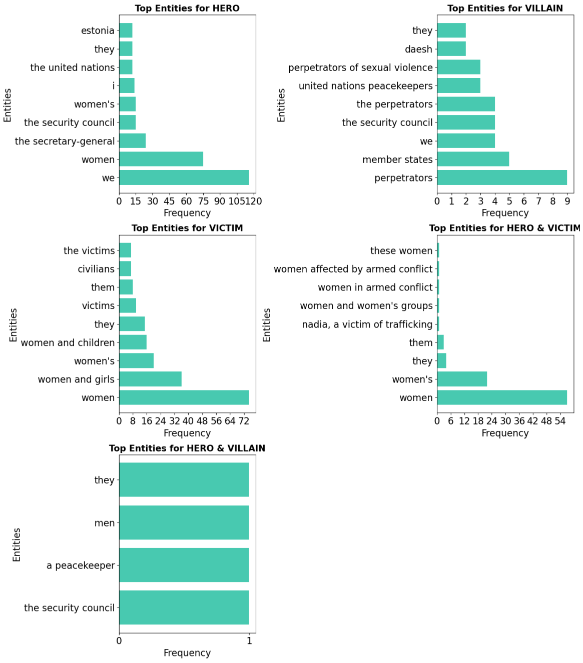


Figure 2: Top Entities per Label

Figure 1 shows the frequency of roles within the annotated dataset. The "hero" label appears most frequently—twice as often as "victim" and seven times more than "villain," aligning with previous research indicating that speeches at the UNSC are generally framed positively. Figure 2 shows the top entities per label. As expected, "women" dominates the "victim" category, alongside group mentions like "women and girls" and "women and children." For "hero," the most common entity is "we," referring to speakers, Member States, or the Council itself, supporting findings that states emphasize their own contributions (True and Wiener, 2019). Interestingly, "women" also ranks second among heroes, indicating that women are not only portrayed as victims but also credited for their agency and/or potential. This dual role is echoed in the "hero-victim" category, which includes mentions like "women," and "Nadia, a victim of trafficking."

The "villain" label appears much less frequently and is often assigned to vague actors such as "perpetrators" or "perpetrators of sexual violence." In some cases, the Council villainizes itself:

(1) "Such debates remain necessary, considering how far **we** are from full implementation."

As noted earlier, we broadened the "villain" category to include entities that impede equal rights. In this example, the Council acknowledges falling short of its own goals regarding women's equal rights. While this self-criticism is subtle, it reflects a characteristic diplomatic strategy where direct

[1] A detailed overview of pairwise agreement scores between individual annotators and our gold labels can be found in Appendix B.

blame is often avoided or softened. This example highlights that the way character roles are defined and assigned can vary across domains, which must be carefully considered when developing automated classification methods.

Aside from this, only two specific villain groups appear: Daesh (ISIS) and UN peacekeepers. The hero-villain dual label is rare but includes references to entities such as "the Security Council" and "a peacekeeper," indicating that while peacekeepers are sent with good intentions, they are sometimes also perpetrators of (sexual) violence. Although it is often implied that perpetrators are male, specific individuals or groups are very rarely identified.

Over time, portrayals of "women" fluctuate (Appendix, Figure 5) but do not show a clear shift from victimhood to heroism as observed in resolutions (de la Rosa and Lázaro, 2019). "Men" are mentioned far less frequently and are often referenced alongside women as perpetrators, allies, or co-victims (Appendix, Figure 6). Mentions of male victims without women or children are rare, reinforcing the framing of "women and children" as the prototypical vulnerable group (Carpenter, 2005). Mentions of male victims increase slightly after 2012, while portrayals of men as perpetrators decline (Appendix, Figure 7). This could suggest increasing recognition of male victims of armed conflict, as well as a reduction in the portrayal of men as perpetrators.

These patterns confirm our core hypotheses: women are primarily framed as victims, though also as heroes; villains remain vague or externalized; and men are largely secondary, often mentioned in relation to women and/or children.

4 Exploratory Work

4.1 Meta-Analysis

Since a core goal of the WPS agenda is to promote the inclusion of women in peace and security processes not only as subjects of protection but as active participants, we examined the representation at the top: within the UN Security Council itself. This led to our first research question: "What does the gender representation of the UNSC look like?"

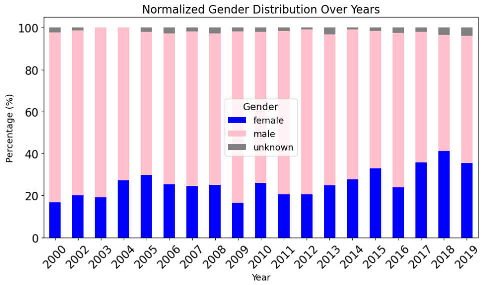


Figure 3: Number of speeches by gender across years. Speakers without gendered honorifics or titles are marked as "unknown."

Using collected meta-information, Figure 3 shows that from 2000 to 2019, women never delivered more than half of all speeches. Female representation increased from under 20% in 2000 to nearly 40% in 2019, peaking just above 40% in 2018. However, this trend is inconsistent and fluctuates across years. Among the permanent members (Appendix, Figure 8), the US stands out with over 80% of speeches delivered by women. The UK follows at just ~20%, followed by France (~10%) and Russia (~2%), while China had no female speakers at all. This imbalance raises questions about the relationship between representation and discourse; though a systematic comparison by country is beyond the scope of this study, it remains a promising direction for future work.

4.2 Topics

4.2.1 Bertopic

To analyze how gendered entities are framed, we applied BERTopic (Grootendorst, 2022) to all sentences across the corpus that mention either unambiguous female terms (“women”, “girls”) or male terms (“men”, “boys”), excluding sentences where both co-occur. BERTopic generated significantly more topics for female entities (309) than for male ones (7). The number of topics for female mentions were reduced to nine with an in-built function for interpretability and selected topics visualized using word clouds (see all word clouds in Appendix E.3).

Female entities. Some topics focused on institutional discourse, with terms like “director,” “delivered,” and “Ms.” Others centered on the WPS agenda itself, highlighting keywords such as “resolution,” “peace,” and “UN-Women”. Several topics emphasized women as victims, particularly of sexual violence (Figure 9a), human trafficking (Figure 9b), and domestic abuse (Figure 9c).

One topic links women to peacekeepers (Figure 9e), framing them both as heroes via calls for greater female participation and as victims of sexual abuse by (implicitly male) peacekeepers. The following quotes highlight this dual framing:

(2) “We are convinced of the advantages inherent in increasing the number of women Blue Helmets"

(3) "[W]e must take every possible step to root out sexual exploitation and abuse by peacekeepers"

Male entities. Topics related to male entities were fewer and less thematically coherent. One addressed sexual violence (Figure 9f), framing men as both perpetrators and victims:

(4) "[...] the role of men as perpetrators but also as victims of sexual violence in conflict."

Other topics focused on men's engagement in gender equality and leadership (Figures 9g–9h), blending heroism, advocacy, and power.

Overall, women are frequently framed as victims but also as (potential) heroes in peacekeeping contexts. Men are mentioned less often and usually in relation to women—as perpetrators, allies, or fellow victims. Villain roles are rarely explicit and often assigned to non-state actors. This implicitness complicates role detection, especially when relying solely on gendered search terms.

4.2.2 Extracting Embeddings and Clustering

We generated 1024-dimensional embeddings for each speech using mxbai-embed-large (Li and Li, 2023), then applied UMAP (McInnes et al., 2018) for dimensionality reduction, and DBSCAN (Ester et al., 1996) for its ability to handle undefined cluster counts and outliers in high-dimensional data. After parameter tuning (eps=6.5, min_samples=5), we identified 18 distinct clusters (see Appendix, Figure 10), revealing two key patterns.

First, we found 14 country-specific clusters, suggesting both consistent national narratives across multiple sessions and standardized diplomatic templates. Beyond typical formalities, we also notice entire phrases such as "*Women are formidable negotiators, mediators, and peacebuilders*" and "*Sexual violence is not cost-free*" are reproduced across multiple speeches, raising questions about the performative versus substantive nature of these statements. Secondly, a smaller subset of clusters grouped together thematically similar speeches. One such cluster centers around sexual violence in armed conflict, with most speakers framing it as a crime against humanity and a threat to peace and security, emphasizing the continued need to take action. Nevertheless, this cluster also includes a speech by the Russian delegation in which the speaker is solely expressing concern over the expansion of UN mandate language beyond existing Security Council resolutions, highlighting that even within thematically coherent clusters, substantive differences in framing exist.

While this approach offers a scalable way to detect patterns in country discourse, it fails to capture *how* countries frame issues like sexual violence, obscuring speaker intentions and level of assertiveness. We therefore turn to a more fine-grained analysis of character-role annotations using (1) a fine-tuned RoBERTa model and (2) direct prompting of LLMs to complement our topic-modeling and clustering approaches.

4.3 RoBERTa Model

Various BERT models were developed by participants in a similar shared task on detecting whether characters in memes were glorified, vilified, or victimized (Sharma et al., 2022). Both the top-performing system (Kun et al., 2022) and the third-ranked system (Singh et al., 2022) incorporated RoBERTa (Liu et al., 2019). Inspired by these works and Gehring and Grigoletto (2023), who also used a RoBERTa model, we fine-tuned RoBERTa for our task.

4.3.1 Preprocessing steps

We adapted the NER task using the standard Beginning-inside-outside (BIO) tagging format (Tjong Kim Sang and Buchholz, 2000) (see Table 5 in the Appendix for an example). Overlapping entities were merged into single combined labels. Since such overlaps were relatively rare (see Figure 1), maintaining them as distinct tags would not have yielded reliable model performance. This resulted in a total of 13 labels[2].

4.3.2 Implementation

To fine-tune the RoBERTa model for the multi-class classification task, we encoded labels and pre-tokenized the input texts using pre-trained RoBERTa embeddings. Model selection was based on performance on the dev set. Details on hyperparameter configurations, performance (for m1-m4), and the rationale for ultimately selecting model m4 for evaluation on the test set are provided in Appendix D.2.

4.3.3 Evaluation & Interpretation

We used the default mode of `seqeval` (Nakayama, 2018), a framework designed for the evaluation of sequence labeling tasks, thus allowing some leniency in evaluation.

[2] Labels: HERO, VICTIM, VILLAIN, HERO_VICTIM, HERO_VILLAIN, VICTIM_VILLAIN, each marked with B- and I-prefixes.

	P	R	F1	S
HERO	**0.69**	0.76	0.72	95
VICTIM	0.67	**0.81**	**0.73**	32
VILLAIN	0.43	0.75	0.55	4
HERO_VICTIM	0.36	0.33	0.35	12
micro avg	0.65	0.73	0.69	143
macro avg	0.54	0.66	0.59	143
weighted avg	0.65	0.73	0.69	143

Table 1: Fine-tuned RoBERTa results: precision (P), recall (R), F1, and support (S) by label, plus micro, macro, and weighted averages.

The performance of the RoBERTa-based classifier is summarized in Table 1. The model performs best for the HERO and VICTIM classes, achieving F1 scores of 0.72 and 0.73, respectively. For the VILLAIN class, it shows a high recall (0.75), but lower precision (0.43) and F1 (0.55) scores, indicating frequent false positives. We acknowledge that this is largely due to the fact that there are only four occurrences of label. Most likely, performance would improve with more training data for this category.

The classifier struggles most with the overlapping label HERO_VICTIM (F1: 0.35), while the remaining overlapping classes HERO_VILLAIN and VICTIM_VILLAIN are not represented in the test set. The low performance scores suggest difficulty in distinguishing overlapping roles, likely due to ambiguity in textual representations.

The identical micro and weighted F1 scores (0.69) indicate strong overall performance, especially on frequent classes like HERO and VICTIM. The lower macro-average (0.59), however, highlights class-specific performance gaps, again highlighting that the model struggles with rarer labels. Addressing this imbalance (e.g., through data augmentation, different sampling techniques, or loss function reweighting) could improve performance on underrepresented roles and therefore result in a more balanced model.

Overall, our fine-tuned RoBERTa model performed well, with its predictions closely aligning with our annotations. The consistency of its outputs suggests a high degree of reliability in capturing the intended role assignments. Interestingly, the model also identified some entities that we overlooked in our annotation process, indicating its potential to enhance annotation quality and reduce human error. Therefore, the model could serve as a valuable tool for automating or assisting in large-scale annotation efforts, provided that occasional inconsistencies are accounted for through further refinement or human verification.

4.4 LLMs

Following Stammbach et al. (2022), we evaluated the outputs of two LLMs in zero- and few-shot settings: DeepSeek-R1 (Guo et al., 2025) and Meta-Llama-3-70B (Grattafiori et al., 2024).[3]

DeepSeek, designed for reasoning, is enhanced through pure reinforcement learning, built on the fine-tuned DeepSeek-V3-Base model, and utilized with a reinforcement learning framework. Llama, a state-of-the-art open-weight transformer-based model developed by Meta, has demonstrated strong generalization capabilities, making it a suitable candidate for comparison (Grattafiori et al., 2024).

While we acknowledge that a strict comparison between the two LLMs that we utilized is not entirely fair in virtue of the immense difference in parameters, we still like to comparatively report differences in their performance.

4.4.1 Preprocessing

We prompted models to reproduce the input speech with in-line role tags (e.g., <HER>Ambassador Powers</HER>). We then converted both model outputs and gold annotations into span-based JSON files for comparison. Due to inconsistencies within the gold annotations (e.g., whitespace, newline characters), we implemented a character-level alignment process to ensure accurate offset matching between model predictions and gold spans. Using fuzzy matching (`difflib.SequenceMatcher`[4]) with a similarity threshold of 0.98 and offset mapping adjustments, we recalibrated span boundaries to enable precise comparison.

4.4.2 Implementation

The prompts were iteratively refined with ChatGPT using the dev set, ensuring none of the models chosen for testing has an advantage with prompts optimized to its understanding. Each prompt included the task description followed by the annotation guidelines (Appendix C), a defined template for the output, and a short repetition of the task to support long-text understanding. The few-shot prompt also includes two annotated speeches from the train set. To see the full prompts, please refer to our GitHub repository.

[3] For readability, we refer to DeepSeek-R1 and Meta-Llama-3-70B as "DeepSeek" and "Llama," respectively, throughout the paper.

[4] https://github.com/python/cpython/blob/3.13/Lib/difflib.py

4.4.3 Evaluation & Interpretation

Table 2 presents the performance of LLMs on the test set using partial match evaluation, to stay in line with our reporting of more lenient evaluation results from RoBERTa. Partial matches required a span overlap of $\geq 50\%$ and a one-to-one mapping between predicted and gold spans, penalizing outputs that over-segment spans.

Label	Model	P	R	F1
HERO	DeepSeekZero	0.39	**0.76**	0.52
	DeepSeekFew	**0.48**	0.72	**0.58**
	LlamaZero	0.29	0.59	0.39
	LlamaFew	0.36	0.65	0.46
VICTIM	DeepSeekZero	0.49	0.65	0.56
	DeepSeekFew	**0.74**	**0.67**	**0.71**
	LlamaZero	0.4	0.65	0.5
	LlamaFew	0.43	0.65	0.52
VILLAIN	DeepSeekZero	0.33	0.25	0.29
	DeepSeekFew	0.67	0.5	0.57
	LlamaZero	**0.75**	**0.75**	**0.75**
	LlamaFew	0.67	0.5	0.57
Micro avg	DeepSeekZero	0.42	0.69	0.52
	DeepSeekFew	**0.54**	**0.7**	**0.61**
	LlamaZero	0.33	0.61	0.43
	LlamaFew	0.38	0.65	0.48
Macro avg	DeepSeekZero	0.4	0.55	0.45
	DeepSeekFew	**0.63**	0.63	**0.62**
	LlamaZero	0.48	**0.66**	0.55
	LlamaFew	0.49	0.6	0.52
Weighted avg	DeepSeekZero	0.44	**0.69**	0.53
	DeepSeekFew	**0.58**	0.66	**0.62**
	LlamaZero	0.36	0.63	0.46
	LlamaFew	0.42	0.61	0.5

Table 2: Classification results for LLMs across role labels. **DeepSeekZero** and **LlamaZero** refer to zero-shot prompting; **DeepSeekFew** and **LlamaFew** use few-shot prompting.

Unlike RoBERTa, which used composite class labels to represent multi-label annotations, we treated each role label independently in the LLM setup. This approach preserved annotation granularity and improved evaluation fidelity by crediting individual label matches.

As shown in Table 2, DeepSeek generally outperforms Llama, particularly in the few-shot setting: For both HERO and VICTIM, DeepSeekFew achieves the highest F1 scores (0.58 and 0.71), with DeepSeekZero yielding the highest recall for HERO (0.76 vs. 0.72 for DeepSeekFew). However, for VILLAIN, LlamaZero performs best (Precision: 0.54, Recall: 0.75, F1: 0.75), though this is just based on four test instances and may not generalize. In future work, this pattern should be investigated with a more balanced dataset.

Overall, DeepSeekFew yields the highest scores across all aggregates: micro-F1 (0.61), macro-F1 (0.62), and weighted-F1 (0.62). These results suggest better generalization and handling of class imbalance compared to Llama. DeepSeekFew achieves a slightly higher macro-average F1 (0.62) than RoBERTa (0.59). This may be due to our decision to treat each role label independently in the LLM setup, avoiding multi-label combinations (RoBERTa's lowest scoring class). However, both LLMs generally underperform relative to RoBERTa. In the next section, we analyze how and why these models fail.

4.5 Error Analysis

Noun Phrases, Articles and Pronouns The tested LLMs often struggled to capture full NPs, often separating phrases such as "women and children." Moreover, articles, pronouns, and descriptive modifiers were frequently omitted (e.g., "thousands of" in "thousands of women and girls"). Additionally, some entities were overextended to include abstract concepts (e.g., "the zero tolerance policy of the secretary general" instead of "the secretary-general"). Importantly, few-shot prompting slightly improved consistency, suggesting that prompt length and structure may impact model stability (Wu et al., 2024).

Mislabeling Formal Entities DeepSeekZero regularly incorrectly labeled speakers as heroes. DeepSeekFew frequently mislabeled thanked individuals as heroes, while RoBERTa only occasionally did this, and Llama never did. Resolutions were also frequently mislabeled as heroes by both DeepSeek models and LlamaFew. LlamaFew further misclassified "the Dakar Declaration" as a hero, although we instructed models against labeling non-resolution documents.

Bias Several models displayed patterns suggesting biases, such as RoBERTa tending to over-label official entities (e.g. "council," "the Secretary-General," "my government") as heroes. More concerning is that RoBERTa often assigned the role of victim to women without justification. LlamaZero demonstrated a similar problem: In the phrase "training must target not only girls and women, but also boys and men"[5], it assigned both "hero" and "victim" to "girls and women," while labeling "boys and men" as only "hero." In the following example, all models annotated women only as victims, rather than as both victims and heroes:

(5) "[R]esolution 1325 (2000) represented an enormous step forward in the protection of **women** and highlighted the importance of their role [...]"

[5] UNSC_2004_SPV.5066Resumption1_spch026

In (6), DeepSeekZero failed to annotate the entity at all, while DeepSeekFew even labeled “the armed forces” as "hero" despite their role in sexual violence:

(6) "[...] the adoption of a plan of action to combat sexual violence among **the armed forces** [...]"

The reasoning by DeepSeekFew offers an explanation for the misclassification:

> *"[...] the military justice system are Heroes as they’re part of the government’s efforts."*

It simply conflated the perpetrators as a part of the government with the heroic efforts of the government, overlooking an even explicit accusation. This example illustrates why relying solely on LLMs for tasks like character role identification can be dangerous.

While some models, like RoBERTa, generally perform well in terms of entity recognition, challenges remain in handling complex cases involving articles, pronouns, multi-word entities, and role shifts. The few-shot prompting approaches showed some improvement, but inconsistencies and biases still persist. These findings highlight the need for careful consideration of bias when applying LLMs to role classification tasks, particularly in contexts involving gender and institutional entities. Further refinements and research to determine where these biases originate from is crucial for improving their performance.

5 Conclusion

Inspired by the central question of how women are depicted in a discourse centered on them, this paper explored how entities are framed in UNSC debates. Using topic modeling, clustering, and manual annotations of 54 speeches, we found that women are the main characters and primary victims. However, they also appear as heroes, providing valuable knowledge and playing a significant role in peacekeeping. Frequent co-occurrence with "girls" and "children" reinforces their portrayal as innocent civilians, while they are never framed as villains.

Men appear far less often and are predominantly grouped with women and children, either as co-victims or co-heroes, highlighting the importance of cooperation between genders. Meanwhile, the UNSC often portrays itself as a hero by highlighting its contributions, though occasionally they acknowledge their own shortcomings. Apart from this, while some non-state actors are explicitly called out, perpetrators largely remain unnamed, and no Member State is explicitly accused. This reflects the diplomatic tone of UN discourse, where positive, optimistic visions are encouraged.

To assess the feasibility of automating role classification, we evaluated RoBERTa, and DeepSeek and Llama in zero- and few-shot settings. While RoBERTa performed reliably and sometimes even outperformed human annotations, the chat models often ignored key rules from the annotation guidelines and misclassified even obvious labels.

Despite limitations, our findings demonstrate the potential of framing analysis in institutional discourse and lay the groundwork for scalable, automated narrative role detection in international diplomatic settings, while emphasizing that human validation remains essential.

Future work could build on these insights by exploring country- and time-specific framing differences, addressing class imbalance through data augmentation, and experimenting with alternative model architectures. Additionally, further research is needed to identify and mitigate sources of bias in large language models to improve their effectiveness and fairness in this context.

Ethics Statement

This study uses publicly available political speeches from the United Nations Security Council and does not involve sensitive personal data. Three of the authors served as annotators; while manual annotation of character roles is inherently subjective, we attempted to mitigate bias through collaborative guideline development and consensus on labels. Given the highly subjective nature of the task, complete objectivity cannot be guaranteed.

We acknowledge that both the discourse analyzed and the computational models used may contain biases reflecting existing power structures and social inequalities. Since automated narrative role detection carries risks of oversimplification and misclassification, we emphasize that such methods should never replace, but rather complement, human interpretation.

All computational models used are publicly available, ensuring reproducibility and transparency.

Limitations

One of the primary limitations of this study is the scope of our annotations. Our dataset size was limited by time constraints, reducing model robustness. While the annotation process itself was not overly time-consuming, significant effort was required to define the annotation guidelines and curate the speeches collaboratively, making this stage particularly resource-intensive. A more extensive dataset would have provided additional training data and enabled a more robust evaluation of our approaches.

Moreover, seeing as the task was subjective, the annotators inevitably brought their own biases into the annotation process. This likely influenced both the evaluation and the model's performance, as the training data was shaped by our individual interpretations. These subjective factors further highlight the challenges of automating entity role classification and the need for continuous refinement of guidelines and methodologies. The subjective nature of detecting character frames is also reflected in the relatively low inter-annotator agreement. However, after extensive discussions, all three annotators reached consensus on the final labels, integrating different perspectives and thus improving data reliability.

Another limitation pertains to the computational and accessibility constraints associated with the models used. Many cutting-edge LLMs are not freely available and require subscription-based access, limiting their usability for broader research applications. Working locally restricted hyperparameter tuning; while our small dataset eased training, it also limits generalizability.

References

Çağlayan Başer. 2024. Women's role in violence and un women, peace, and security agenda. *Alternatif Politika*, 16(1):1–30.

Kelly Bergstrand and James M. Jasper. 2018. Villains, victims, and heroes in character theory and affect control theory. *Social Psychology Quarterly*, 81(3):228–247.

R. Charli Carpenter. 2005. "women, children and other vulnerable groups": Gender, strategic frames and the protection of civilians as a transnational issue. *International Studies Quarterly*, 49(2):295–334.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Stanley Cohen. 2011. *Folk devils and moral panics*. Routledge.

Victoria Martín de la Rosa and Luis Miguel Lázaro. 2019. How women are imagined through conceptual metaphors in united nations security council resolutions on women, peace and security. *Journal of Gender Studies*, 28(4):373–386.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Lea Frermann, Jiatong Li, Shima Khanehzar, and Gosia Mikolajczak. 2023. Conflicts, villains, resolutions: Towards models of narrative media framing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8712–8732, Toronto, Canada. Association for Computational Linguistics.

Kai Gehring and Matteo Grigoletto. 2023. Analyzing climate change policy narratives with the character-role narrative framework. Technical report, CESifo Working Paper.

Sheri Lynn Gibbings. 2011. No angry women at the united nations: Political dreams and the cultural politics of united nations security council resolution 1325. *International Feminist Journal of Politics*, 13(4):522–538.

Diego Gomez-Zara, Miriam Boon, and Larry Birnbaum. 2018. Who is the hero, the villain, and the victim? detection of roles in news articles using natural language techniques. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, pages 311–315.

Francesca Grasso, Stefano Locci, and Manfred Stede. 2025. Applying the character-role narrative framework with llms to investigate environmental narratives in scientific editorials and tweets. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 49–67.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

James M Jasper, Michael Young, and Elke Zuern. 2018. Character work in social movements. *Theory and Society*, 47:113–131.

Michael D Jones and Mark K McBeth. 2010. A narrative policy framework: Clear enough to be wrong? *Policy studies journal*, 38(2):329–353.

Stephen Karpman. 1968. Fairy tales and script drama analysis. *Transactional analysis bulletin*, 7(26):39–43.

Orrin E Klapp. 1954. Heroes, villains and fools, as agents of social control. *American sociological review*, 19(1):56–62.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.

Ludovic Kun, Jayesh Bankoti, and David Kiskovski. 2022. Logically at the constraint 2022: Multimodal role labelling. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 24–34, Dublin, Ireland. Association for Computational Linguistics.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

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. *Preprint*, arXiv:1907.11692.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2329–2334.

Mirco Schoenfeld, Steffen Eckhard, Ronny Patz, Hilde van Meegdenburg, and Antonio Pires. 2019. The UN Security Council Debates.

Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Findings of the CONSTRAINT 2022 shared task on detecting the hero, the villain, and the victim in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics.

Pranaydeep Singh, Aaron Maladry, and Els Lefever. 2022. Combining language models and linguistic information to label entities in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 35–42.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Jacqui True and Antje Wiener. 2019. Everyone wants (a) peace: the dynamics of rhetoric and practice on 'women, peace and security'. *International Affairs*, 95(3):553–574.

UNDP. 2023. Parliament as partners: Supporting the women, peace, and security agenda - a global handbook. Accessed: 2025-03-24.

Erika Allen Wolters, Michael D. Jones, and Kathryn L. Duvall. 2021. A narrative policy framework solution to understanding climate change framing research. *Narratives and the Policy Process : Applications of the Narrative Policy Framework*.

Yue Wu, Xuan Tang, Tom M. Mitchell, and Yuanzhi Li. 2024. Smartplay: A benchmark for llms as intelligent agents. *Preprint*, arXiv:2310.01557.

Appendix A UNSC Quote Filenames

Quote #	Filename
(1)	UNSC_2003_SPV.4852Resumption1_spch035.txt
(2)	UNSC_2019_SPV.8649_spch030.txt
(3)	UNSC_2015_SPV.7533_spch015.txt
(4)	UNSC_2013_SPV.6984_spch057.txt
(5)	UNSC_2011_SPV.6642_spch055.txt
(6)	UNSC_2016_SPV.7704_spch087.txt

Table 3: Corresponding filenames for numbered in-text examples.

Appendix B Interannotator Agreement

	Annotator 1	**Annotator 2**	**Annotator 3**	**Curator**
Annotator 1	–	0.30	0.44	0.47
Annotator 2		–	0.38	0.48
Annotator 3			–	0.70
Curator				–

Table 4: Krippendorff's Alpha (unitizing/character offsets) for agreement on role annotation (Hero, Victim, Villain) computed pairwise. Three of the authors served as annotators and worked collaboratively to define a gold standard, which is reflected in the Curator column.

Appendix C Heroes, Victims, and Perpetrators in WPS: Annotation Guidelines

i. Document Level Rules Exclude all president speeches

ii. Character Level Rules

a. We include entire noun phrases (NP): This includes any numerical or descriptive modifiers.
Example: "60 million Africans"
b. We annotate restrictive relative clauses fully.
Example: "parties that perpetuate acts of violence against women and children"
c. We include possessive modifiers within the NP:
Example: "my delegation"
d. When multiple characters are listed together as part of the same NP and share a common role or action, we consider the entire sequence as one entity, so we not split or annotate each item separately.
Example: "Allow me to begin by thanking **the Secretary-General, Mr. Kevin Hyland, Mr. Yury Fedotov and Ms. Ilwad Elman** for their briefings."
e. We do not include punctuation at the end of annotations.
Example: "**men and boys**."
f. We do not annotate predicate nominatives or any other descriptions of characters.
Example: "**Women** are not only victims." Here, 'victims' should not be annotated.
Example: "**women** as agents of peace"
g. We do not annotate positions or groups when they are only mentioned in the abstract, without references to the achievements of the particular entity.
Example: "the establishment of the post of the Special Representative of the Secretary General on sexual violence in situations of armed conflict"

Role-Definitions

In German, the constellation of hero, victim, and perpetrator is called the drama triangle, first introduced by Karpman (1968). It is a psychological model used in transactional analysis to describe interactions between individuals or groups as adopting one of three stereotypical roles: victim, perpetrator, or hero, which simplifies the linking of behaviors with stereotypical roles learned at an early stage (Karpman, 1968; Gomez-Zara et al., 2018). The link to stereotypes, the framing of the characters with said roles, influences the readers to have a certain bias and to develop a non-reflective opinion through the stereotypical knowledge, which is influenced by the author alone (Gomez-Zara et al., 2018). Hero, victim, and perpetrator are also of great importance in affect control theory and character theory (Bergstrand and Jasper, 2018). The following are definitions we will work with for the WPS-project:

Victim

Victims tend to be portrayed as weak, good, innocent people who are in need of protection (Jasper et al., 2018). Due to these characteristics, they often motivate and encourage action towards a specific cause and can help make aware of injustices which are worth com bating. Jasper et al. (2018) state that victim's sufferings are often elaborated in detail to arouse more moral emotions and indignation. "Popular" victims, as they get the most sympathetic reactions in the modern world due to their cultural innocence, are children (Jasper et al., 2018).

Task-specific additions (based on the WPS corpus):

- Someone who is excluded from decision-making processes/someone who is not given the recognition/power that they deserve.
- Someone who suffers acts of sexual violence/physical harm/displacement, etc.
- Someone who is not given the same equal rights as other parties

Perpetrator/Villain

Perpetrators are people whose moral reputation turns or has turned negative (Jasper et al., 2018). They are considered to be people who spread anxiety and fear, cause people to lose their daily routines, and make them sacrifice their lives, for example, within wars (Bergstrand and Jasper, 2018). Perpetrators often share the same characteristics as heroes, such as being strong, brave, and intelligent. However, their description tends to be more like that of beast-like predators: powerful, threatening, and delinquent (Cohen, 2011).

Task specific additions (based on the WPS corpus):

- Someone who is responsible for causing anxieties, damage, and crimes.
- Someone who stands in the way of equal rights and justice for victims.

Hero

Heroes are people who, by helping victims (and defeating the villains (Jasper et al., 2018; Klapp, 1954)), can become heroes. They are defined as people who increase agreement within groups and boost commitment to a cause. They tend to be well-intentioned people, who recognize injustice, try to resolve and fight it, as well as protect others. However, this does not mean that heroes are completely independent. Jasper et al. (2018) state that even a hero might be in need of help from an even more experienced hero. Furthermore, they are often put in the context of success (Klapp, 1954).

Task specific additions (based on the WPS corpus):

- Someone who has the potential to save others, given their abilities, knowledge, positionality/perspective, but who is not able to do so because of structural discrimination.
- Someone who makes a call to action
- Someone who is recognizing the unjust treatment and violation of victims or is calling it out.

Further Thoughts

It can be seen that the victim, hero, and perpetrator definitions align very much with stereotypes we are taught and confronted with since our childhood. However, roles are not fixed, as heroes can transform into victims or perpetrators (Jasper et al., 2018). The definitions are tried to be kept as simple and generally adaptive as possible, because depending on the source one works with, representations can change. Jasper et al. (2018), for example, state that in news articles, victims are kept nameless in contrast to heroes and perpetrators. In the WPS dataset, however, we observed that it is the perpetrators who most commonly remain unnamed.

Entities to tag

a. People
b. Organisations
c. Countries
d. Groups
e. UN Resolutions. Only annotate United Nations Resolutions if they are explicitly personified.
Example: In “Resolution 1325 calls for action,” the resolution should be tagged as a hero. In “working towards the implementation of resolution 1325,” the resolution should not be marked.

Entities NOT to tag

a. Abstract concepts: Abstract ideas or symbolic references.
Examples: “International cooperation,” “sexual violence,” “**women’s** participation.” In the latter example, women are the entities that are taking heroic action, while their participation refers to a concept.
b. Entities hoping for/welcoming/thanking/commending something good. These are passive actions.
c. Laws or Treaties.

Annotation of character role labels

We only annotate generic role terms (like “victim”) when no other specification is included.

Example: In “Victims of these atrocious crimes have been waiting for justice”, we would tag “victims of these atrocious crimes” as Victim. However, if specific entities are mentioned, like in “Victims of these atrocious crimes, namely women, have been waiting for justice” we only annotate “women” as Victim.

Annotation of multiple roles

In certain cases, characters may be portrayed with multiple roles simultaneously. When this happens, we annotate the entity with a combined role. Furthermore, the same entity can take on different roles throughout the speech. For instance, while “women” might be classified as "victim" in one sentence, the same entity can also be classified as "hero" in a different part of the speech.

Example: “The equal right to decision-making and participation, along with **women’s** empowerment, is crucial to ensure a functioning society and peace and justice in the aftermath of conflicts.”

Explanation: In this case, “women” are portrayed both as victims (since they need external help to be empowered) as well as heroes (since they contribute to a peaceful society). When an entity fits into multiple roles based on the context, we use combined tags.

iii. Annotation examples and explanations

Legend: Hero Perpetrator Victim

1. "We commend the work that has been done by the United Nations Children’s Fund in reintegration projects that has led to the release of girls from the armed forces in various countries."

In this example, "we" is not annotated as a hero because they are just thanking another organization.

However, the organization in question, the "United Nations Children's Fund," is marked as a hero because they are responsible for these projects that help girls. The "girls," on the other hand, are marked as victims because, although it could be argued that they are the benefactors of the situation (since they have been released), we believe that the fact that they were abducted in the first place carries more weight, also because they are in danger of being abducted again at any time. Most intuitively, "the armed forces" are the perpetrators, since they were responsible for the girls' abduction.

2. "Members of the Council note that although women have begun to play an important role in conflict resolution, peacekeeping and peace- building, they are still under-represented in decision-making in regard to conflict. If women are to play an equal part in security and maintaining peace, they must be empowered politically and economically, and represented adequately at all levels of decision-making, both at the pre-conflict stage and during hostilities, as well as at the point of peacekeeping, peace-building, reconciliation and reconstruction."

This example demonstrates that the same entity (in this case, "women") can take on multiple roles within one sentence. In the first instance ("women"), they are marked as heroes due to their important role in conflict resolution. However, in the second part of the sentence, where they are referred to as "they," the focus shifts to their underrepresentation in decision-making, marking them as victims. The conjunction "although" introduces this contrast, highlighting both aspects of the situation. In the second sentence, the focus remains on women as victims ("if women are to play an equal part in security"), implying that they are not currently able to do so. The need for empowerment, both politically and economically, is presented because they are not given equal opportunities as men. Therefore, they are portrayed as victims in this context.

3. "Those who commit crimes against women, including the peacekeeping personnel, should be brought to book. Let us heed the women's cry for an equal opportunity to voice their ideas in official peace negotiations. And let us act now."

In the last two sentences of this example, "us" is marked as a hero since the speaker is engaging in a call to action. The speaker portrays "us" as a hero by attempting to encourage others to acknowledge and act upon the unjust inequality faced by women.

4. "My Government attaches great importance to the protection and security of women and girls, both in situations of armed conflict and in peace."

In this case, "my government" is marked as a hero because, while the action is indirect, they recognize the importance of the issue and stand up for it. "Women and girls" are marked as victims since they require protection and security, as they cannot protect themselves (see victim definition above). Although they benefit from the protection measures (and thus could be seen as benefactors), we chose to prioritize their need for protection, which highlights their vulnerable position and role as victims in society.

5. "Being direct victims of violence and discrimination, women have gained a great understanding of the need to address peace comprehensively."

Here, "women" are simultaneously considered heroes and victims. The victim annotation is indicated by the context of them being the "direct victim," while the hero annotation relies on the positionality of women–their potential to provide valuable knowledge.

Appendix D RoBERTa Model

Appendix D.1 BIO-Annotation Example

Token	"[...]	my	country	has	[...]	defending	the	victims	of	terrorism	."
Labels	O	HERO	HERO	O	O	O	VICTIM	VICTIM	VICTIM	VICTIM	O
BIO Labels	O	B-HERO	I-HERO	O	O	O	B-VICTIM	I-VICTIM	I-VICTIM	I-VICTIM	O

Table 5: Example of the BIO annotation scheme applied to a sample sentence from `UNSC_2015_SPV.7585_spch010.txt`. The whole sentence is: *Council members know that my country has been a standard-bearing country for defending the victims of terrorism.* Entity spans are labeled as "B-" (begin), "I-" (inside), or "O" (outside)

Appendix D.2 Hyperparameter Optimization and Model Selection

During fine-tuning we limited the amount of values to ensure computational feasibility within the available resources.

	Learning Rate	**Epochs**	**Optimizer**
m1	5e-5	3	AdamW
m2	5e-5	30	AdamW
m3	1e-5	3	AdamW
m4	1e-5	30	AdamW
m5	5e-5	30	SGD
m6	3e-5	30	SGD

Table 6: Hyperparameters for m1-m6

The performance of the fine-tuned models on the development dataset is shown in Table 7, based on their F1 values for models m1-m4. Models m5 and m6 are excluded from the table as their results were identical to those of m4.

	F1				**S**
	m1	m2	m3	m4	
HERO	0.59	0.62	0.63	**0.64**	122
VICTIM	0.67	0.72	**0.73**	**0.73**	77
VILLAIN	**0.54**	0.52	0.5	0.52	10
HERO_VICTIM	0.45	**0.46**	0.36	0.4	26
HERO_VILLAIN	0	0	0	0	0
VICTIM_VILLAIN	0	0	0	0	3
micro avg	0.6	0.62	0.62	**0.64**	238
macro avg	0.45	**0.46**	0.44	0.38	238
weighted avg	0.59	0.62	0.62	**0.63**	238

Table 7: F1 scores of the models m1, m2, m3, and m4. S (support) represents the number of instances per class in the dataset.

The results show that models m3 and m4 achieve the highest F1 scores for the two most frequently occurring classes, heroes (m3: 0.63; m4: 0.64) and victims (m3 & m4: 0.73). m4 yields the highest weighted average F1 score (0.63). These findings served as the primary rationale for selecting the hyperparameters of m4 for the final fine-tuning process. The final model was fine-tuned using both the train and the dev set.

Appendix E Figures

Appendix E.1 Annotation Insights

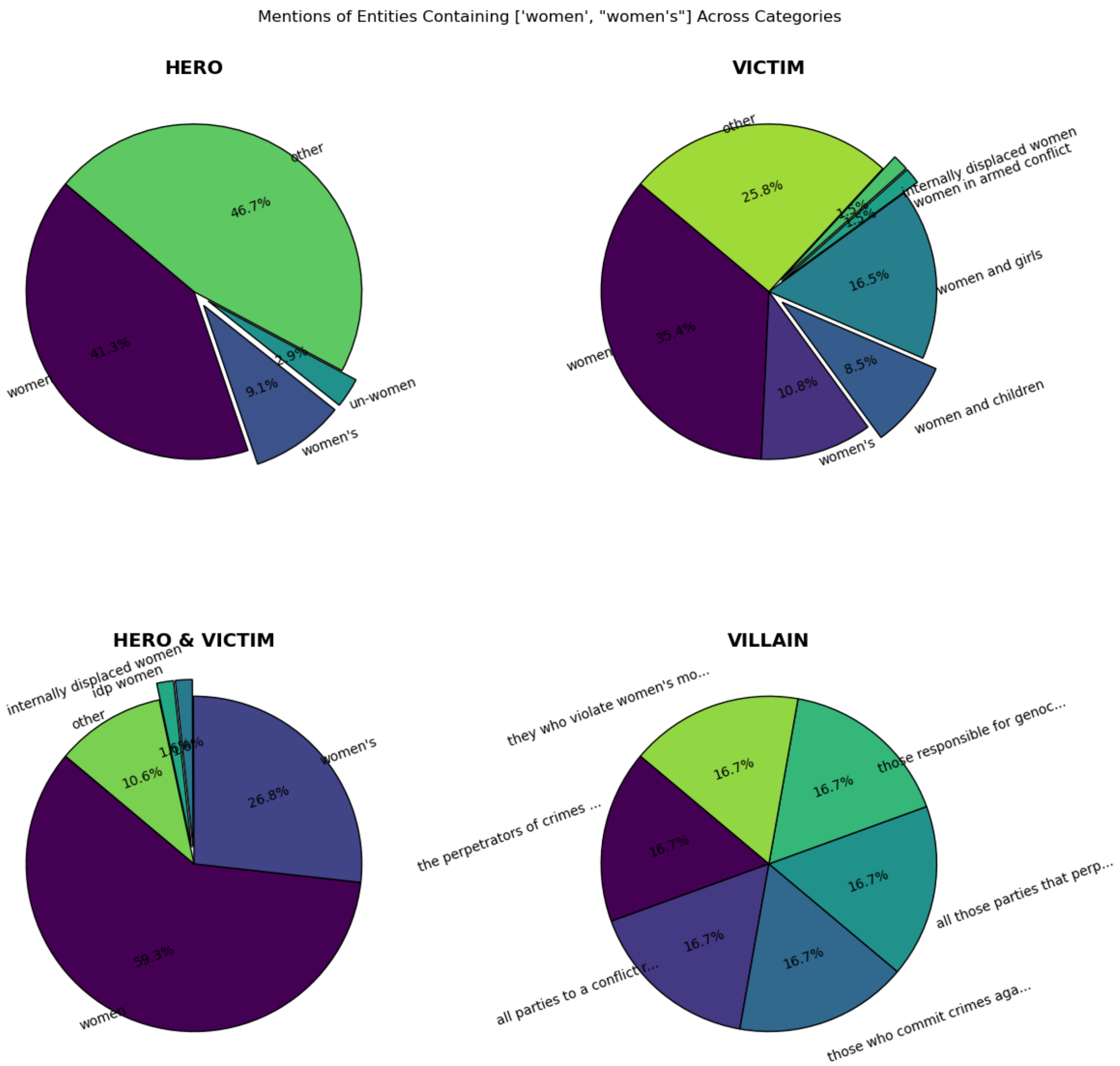


Figure 4: Distribution of Characters for Women Entities

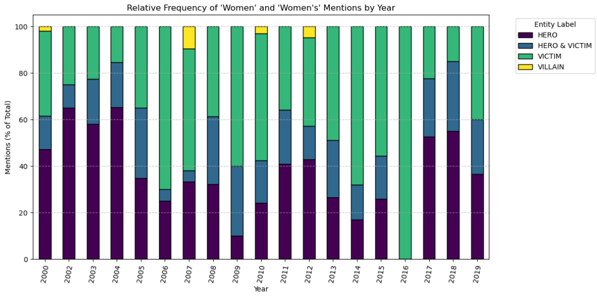


Figure 5: Character Labels for "women" and "women’s" Across the Years

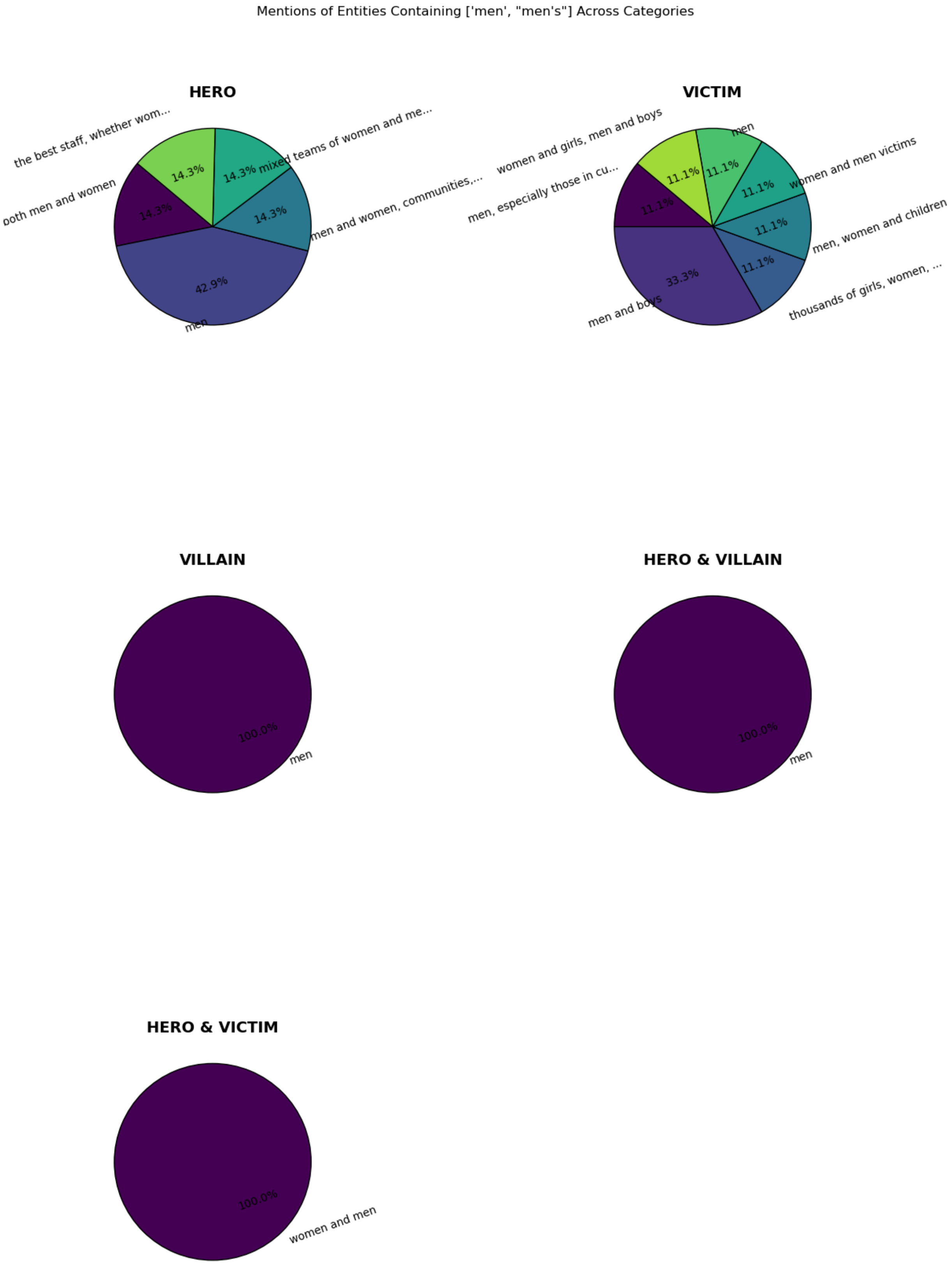


Figure 6: Distribution of Characters for Men Entities

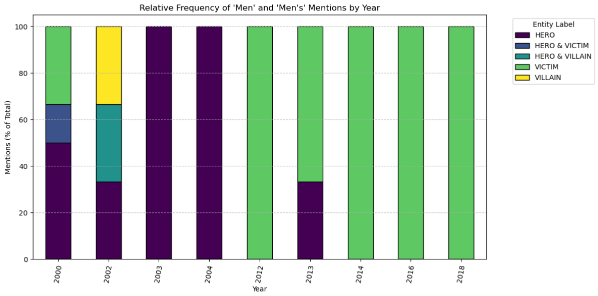


Figure 7: Character Labels for "men" and "men’s" Across the Years

Appendix E.2 Meta-Analysis

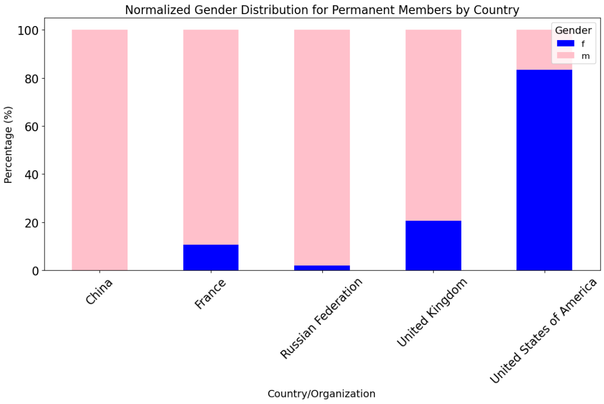
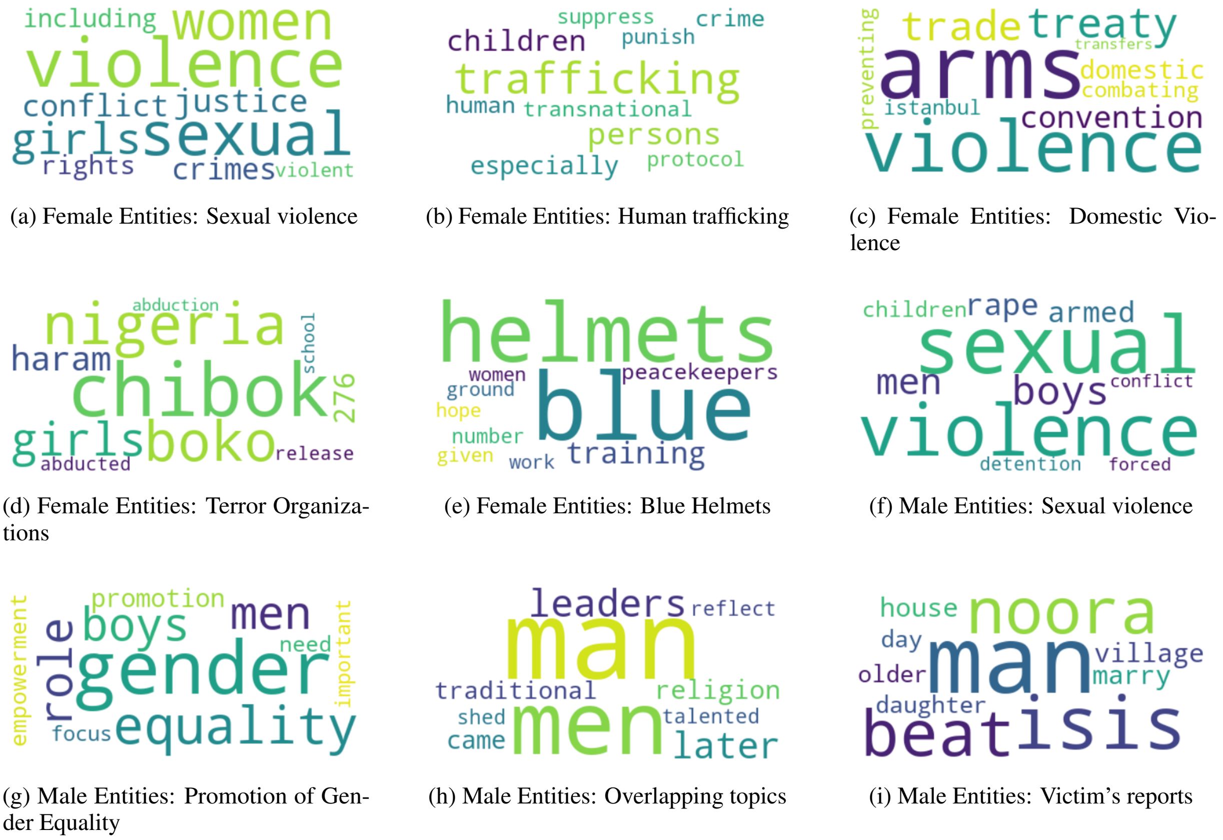


Figure 8: Gender Distribution of Speakers Among Permanent Members of the UNSC

Appendix E.3 BERTopic



(a) Female Entities: Sexual violence

(b) Female Entities: Human trafficking

(c) Female Entities: Domestic Violence

(d) Female Entities: Terror Organizations

(e) Female Entities: Blue Helmets

(f) Male Entities: Sexual violence

(g) Male Entities: Promotion of Gender Equality

(h) Male Entities: Overlapping topics

(i) Male Entities: Victim's reports

Figure 9: BERTopic visualizations for male and female entities.

Appendix E.4 Clustering

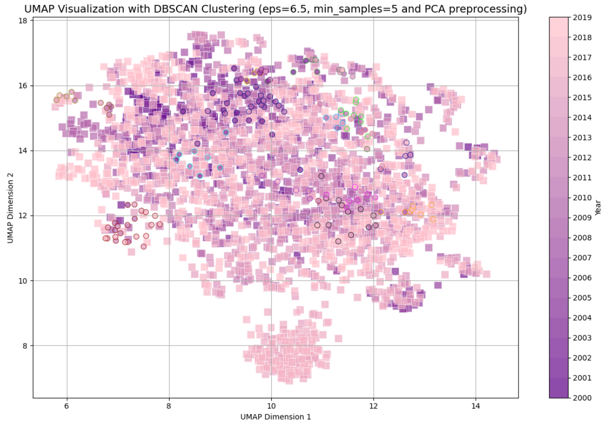


Figure 10: Clusters as identified using DBSCAN. Each point represents a speech. Points that are square-shaped with a white outline represent outliers, while circle-shaped points represent speeches that are part of different clusters, distinguished by the color of the outline. The fill color of each point represents the year that the speech was held.