Finding Common Patterns in Domestic Violence Stories Posted on Reddit

Mohammad Shokri[1], Emily Klapper[2], Jason Shan[2], Sarah Ita Levitan[2]

[1]The Graduate Center, CUNY
[2]Hunter College, CUNY

Abstract

Domestic violence survivors often share their experiences in online spaces, offering valuable insights into common abuse patterns. This study analyzes a dataset of personal narratives about domestic violence from Reddit, focusing on event extraction and topic modeling to uncover recurring themes. We evaluate GPT-4 and LLaMA-3.1 for extracting key sentences, finding that GPT-4 exhibits higher precision, while LLaMA-3.1 achieves better recall. Using LLM-based topic assignment, we identify dominant themes such as psychological aggression, financial abuse, and physical assault which align with previously published psychology findings. A co-occurrence and PMI analysis further reveals the interdependencies among different abuse types, emphasizing the multifaceted nature of domestic violence. Our findings provide a structured approach to analyzing survivor narratives, with implications for social support systems and policy interventions.

1 Introduction

Narratives are central to human communication, proven to foster empathy, shared beliefs, and persuasiveness. With the growth of internet use globally, individuals increasingly share personal stories online, seeking empathy and emotional support from the online community. Domestic violence stories are a striking example of this trend. The abundance of domestic violence stories on the internet provides a unique opportunity for computational analysis to identify commonalities and variations in how these experiences are narrated. By examining these stories at scale, we can uncover recurring patterns in them, such as how survivors describe the progression of abuse, the typology of abuse, the role of legal interventions, or the types of support they seek.

Identifying common patterns in domestic violence narratives opens the door to various applications, ranging from privacy protection to early intervention strategies. For instance, detecting outlier patterns could help develop systems that prevent individuals from sharing stories that might inadvertently reveal their identities. Additionally, recognizing progressions in abuse-related narratives could contribute to predictive models that identify when relationships are at risk of escalating into more severe abuse. Beyond these, computational insights from these stories could be applied to support systems, legal frameworks, and advocacy efforts, ultimately improving both understanding and response strategies for domestic violence cases.

To enable these potential applications, we first need to distinguish common patterns from unique details within domestic violence stories. This paper focuses on learning the recurring structures in these narratives by identifying the key events that define interactions between the victim and the perpetrator. Events are central to narrative structure, and understanding which events frequently co-occur allows us to detect broader storytelling patterns. We hypothesize that domestic violence stories share a high degree of similarity, particularly in the progression of events that characterize abusive relationships.

Leveraging recent advancements in natural language processing (NLP), we explore the ability of large language models (LLMs) to extract and analyze these key events. In this paper, we propose a fully LLM-based method for processing stories and attributing topics to the events, with the goal of clustering and finding similar patterns. Specifically, we use LLaMA-3.1 (Dubey et al., 2024) and GPT-4 (Achiam et al., 2023) to extract those sentences from a narrative that capture interactions between the victim and the perpetrator. We use these LLMs to assign topics to the extracted sentences, which facilitates learning topic progressions in the stories. We analyze topic co-occurrence and topic n-grams from the stories to find similar patterns between our

Proceedings of the The 7th Workshop on Narrative Understanding, pages 15–25
May 4, 2025 ©2025 Association for Computational Linguistics

set of stories. We collected a large set of domestic violence stories from Reddit, consisting of more than 11,100 posts which we filtered for story-like posts, using a pre-trained classifier (Antoniak et al., 2023). Our dataset is available upon request.

2 Related Work

Narrative is commonly defined as a sequence of events that unfolds over time (Labov and Waletzky, 1997; Eisenberg and Finlayson, 2021). Events are the fundamental building blocks of narratives, providing structure and coherence by linking actions, participants, and consequences (Zhang et al., 2021). Earlier studies in the literature took a verb-based perspective on events, primarily focusing on extracting predicate-argument triples to represent narrative progression (Mousavi et al., 2023; Chaturvedi et al., 2017; Chambers and Jurafsky, 2008). More recent works have employed supervised learning, transfer learning, and sequence-to-sequence models for developing models that can extract events from a piece of text (Lu et al., 2021; Li et al., 2021; Sims et al., 2019; Uddin et al., 2024; Huang et al., 2017). Li et al. (2022) presents an extensive survey of deep learning-based methods for event extraction. Identifying recurring event structures allows researchers to analyze narrative evolution, uncover causal dependencies, and detect common thematic patterns across large story datasets.

While event extraction focuses on explicit actions, states, and participants, topic modeling provides a higher-level view of recurring themes within narratives, and it enables researchers to model narrative schema and arcs across large datasets (Min and Park, 2016; Schmidt, 2015; Boyd et al., 2020; Mathewson et al., 2020; Antoniak et al., 2023). As an example, Antoniak et al. (2019) used topic modeling to find clear patterns of events that occur in birth stories and used the learned topic transition probabilities to find outlier stories. Wagner et al. (2022) proposed a Point wise Mutual Information (PMI) based method to capture topic segmentation for Holocaust testimonies.

Recent advancements in Transformer-based language models (Vaswani, 2017) have enhanced computational narrative understanding. Piper and Bagga (2024) examined ways in which LLMs could contribute to understanding core narrative features. Wagner et al. (2024) used GPT-4 thanks to its long context window (128k tokens) to extract trajectory mappings from a set of Holocaust testimonies. Heddaya et al. (2024) fine-tuned LLaMA (Dubey et al., 2024) and used GPT-4 in few-shot and zero-shot settings for detecting causal micro-narratives within a sentence.

Despite their abundance and importance, domestic violence narratives have not been studied extensively in the NLP community. Schrading et al. (2015) developed classifiers using n-grams and semantic roles as features for detecting posts on reddit discussing domestic abuse. Karlekar and Bansal (2018) used CNN-RNN architectures to classify between narratives containing different forms of sexual harassment shared online through a forum called SafeCity. Calderwood et al. (2017) studies physiological responses of readers reacting to abuse survivors studies. Shokri et al. (2024) focused on extracting common events from a small set of domestic violence stories and developed a classifier to classify between domestic violence stories and non-domestic violence stories based on a vector distance metric. In this paper, we introduce a large set of personal domestic violence stories from Reddit, and use LLMs to extract the events from stories and identify their topics.

3 Dataset

To collect personal stories about domestic violence, we turned to Reddit, specifically the subreddit *r/domesticviolence*, where users share their experiences and receive support from others. This community provides information and emotional support for victims, with members offering insights based on their personal experiences rather than professional opinions. We scraped this publicly available subreddit and archived 11,176 posts spanning from 2005 to 2021 to construct our dataset. To ensure anonymity, we only keep the posts' text.

An initial exploration of the dataset revealed that not all posts contain personal experiences. Some posts are general discussions or rants about domestic violence and its effects, without explicitly describing eventful personal narratives. To filter out non-narrative posts, we use *StorySeeker* (Antoniak et al., 2023), a pretrained RoBERTa model (Liu, 2019) designed for binary classification of stories vs. non-stories. Applying this model to our dataset, we identified 9,872 posts as stories (see Table 1).

To understand the structure of the collected stories, we analyzed the distribution of sentence

Category	Count
Non-story posts	1,304
Posts classified as stories	9,872
Total posts collected	11,176

Table 1: Summary of collected Reddit posts and distribution of story vs. non-story labels based on StorySeeker classification output.

counts per post. As shown in Figure 1, the majority of stories are relatively short, with a steep drop-off in frequency as sentence count increases. The median story length is around 16 sentences, with 25% of stories having fewer than 9 sentences and 75% having fewer than 28 sentences. While most stories contain only a few sentences, there are outliers with significantly higher sentence counts, reflecting variations in detail and narrative style. The distribution suggests that while many users share brief experiences, others provide in-depth narratives describing complex events. After extracting events from the stories (see Section 4), we only keep stories with at least 5 sentences to ensure working with story-like posts.

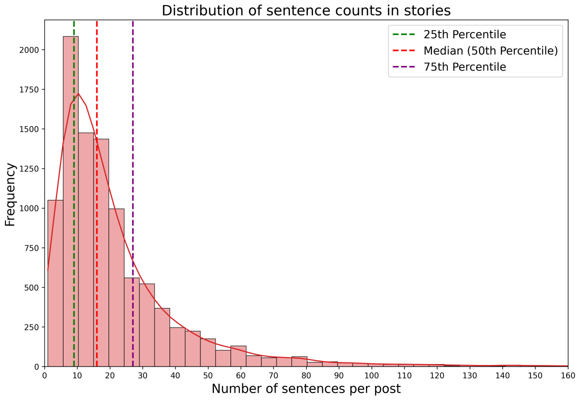

Figure 1: Distribution of sentence counts per post in the dataset. The majority of posts are short, with a few containing significantly more sentences. The x-axis is limited to 160 sentences to improve readability. The maximum number of sentences per post in our dataset is 477.

4 Extracting Events

After collecting the stories from Reddit, we aimed to extract events from them to enable an analysis of themes in the stories. Events are fundamental building blocks of a story, yet they are not unanimously and clearly defined in the literature. Most prevailing conceptions of events are based on changes in state (Vauth et al., 2021; Sims et al., 2019; Aguilar et al., 2014; Sims et al., 2019). Vendler (1957) categorized the relationship between verbs and time into four types: *activities*, *achievements*, *accomplishments*, and *states*. Sims et al. (2019) classifies activities, achievements, accomplishments, and changes of state as "events". Building on this, Antoniak et al. (2023) developed a more flexible event span annotation framework that includes not only real events but also hypothetical and recurring actions. We adopt the definition from Antoniak et al. (2023) and modify it to incorporate verbal interactions, as verbal abuse is a prevalent form of domestic abuse and we observed that it frequently appears in our dataset. The definition of event is provided in the Appendix section A.1.

People share their personal stories with varying levels of detail; some provide extensive background on their own or their partner's lives, while others narrate in detail the sequence of events leading up to instances of domestic violence. We focus on events involving both the victim and the perpetrator because we are most interested in uncovering patterns that characterize abusive relationships.

We do not assume all aspects of these stories are alike, given the numerous ways relationships start and people's diverse life backgrounds. Therefore, to identify the commonalities we believe exist in domestic violence narratives, we first extract sentences that describe events or actions that directly involve both the victim and the perpetrator. We prompt LLaMA-3.1 8B (`meta-llama/Llama-3.1-8B-Instruct`) and GPT-4 (`GPT-4-Turbo`) with the definition of events and a description of the task. We provide three examples in the prompt to clarify the task and serve as few-shot examples. The prompt we used for this task is available in the Appendix section (A.1). We set the temperature = 0.0 while prompting both models.

4.1 Annotation

In order to evaluate the LLM-based event extraction, we asked two members of our research team to read the stories and extract the sentences which *describe an event or action that happened in the story which involved the victim and the perpetrator*. We randomly selected 50 stories from our dataset and asked the annotators to find eventful sentences. The total number of sentences in the stories were 1587. In cases where the annotators' labels disagreed, we conducted a consolidation session, during which both annotators discussed their reasoning to resolve conflicts. Final labels were assigned based on mutual agreement, ensuring a consistent and high-quality labeled dataset. There were 431 sentences extracted as eventful sentences.

The inter-annotator agreement calculated as Cohen's kappa (Cohen, 1960) score was 0.67 which indicates substantial agreement. Although a high level of inter-annotator agreement was observed, certain disagreements arose during the classification of events. Variations in narrative styles across the stories contributed to ambiguity in identifying specific events. In numerous instances, the narrator's commentary implied an event without explicit mention, leading to interpretive differences. Additionally, disagreements emerged when analyzing sentences involving individuals beyond the victim and perpetrator (such as bystanders, law enforcement, etc.), as well as in cases where stories featured multiple victims or perpetrators. These complexities highlight the nuanced nature of event classification within this dataset.

4.2 Evaluation of Event Extraction

The results of our sentence extraction task are shown in Table 2. Our results highlight key differences between LLaMA-3.1 and GPT-4 in terms of precision, recall, and F1-score, both for eventful and non-eventful sentences.

For eventful sentences (positive class), GPT-4 achieves a slightly higher F1-score (0.5374) compared to LLaMA-3.1 (0.5355), despite having much lower recall (0.4084 vs. 0.6729). This indicates that GPT-4 is more selective, extracting fewer irrelevant sentences (higher precision: 0.7857 vs. 0.4448), but LLaMA-3.1 captures a broader range of eventful sentences due to its higher recall, though at the cost of more false positives.

For sentences not containing description of events (negative class), both models perform strongly, with GPT-4 achieving an F1-score of 0.8797 and LLaMA-3.1 scoring 0.7594. Notably, GPT-4 excels in recall (0.9585), identifying nearly all non-eventful sentences, while LLaMA-3.1 shows a better balance between precision (0.8492) and recall (0.6869).

Looking at the overall macro averages, GPT-4 outperforms LLaMA-3.1 with a higher F1-score (0.7086 vs. 0.6475), achieving better balance across both eventful and non-eventful classes. These results suggest that LLaMA-3.1 is better suited when comprehensive coverage (high recall) is essential, while GPT-4 is preferable when precision is critical, minimizing false positives and extracting more reliable eventful sentences.

	GPT-4			LLaMA-3.1		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Event Class (Positive)	0.7857	0.4084	0.5374	0.4448	0.6729	0.5355
Non-Event Class (Negative)	0.8129	0.9585	0.8797	0.8492	0.6869	0.7594
Macro Average	0.7993	0.6834	0.7086	0.6470	0.6799	0.6475

Table 2: Comparison of GPT-4 and LLaMA-3.1 Performance on Event Sentence Extraction

Figure 2 presents the distribution of the number of sentences extracted by GPT-4 and LLaMA-3.1 for our dataset. Consistent with the performance metrics discussed earlier in this section, we observe a key difference in the extraction tendencies of the two models. GPT-4 produces a more concentrated distribution, with a median of 3 extracted sentences per story and a mean of 3.4, suggesting that the model is more selective in identifying eventful sentences. This aligns with its higher precision (0.79), as it extracts fewer sentences overall, reducing false positives but potentially missing relevant details.

On the other hand, LLaMA-3.1 demonstrates a much broader distribution, with a median of 7 extracted sentences per story and a mean of 10.9. This reinforces the previously observed higher recall (0.68) of LLaMA-3.1, indicating that it tends to classifies a larger number sentences as relevant, even at the cost of lower precision. The figure suggests that using the same prompt, LLaMA-3.1 often extracts significantly more sentences per story, capturing a wider range of contextual information, albeit with more noise.

We filter our dataset to retain only stories with at least five sentences extracted by GPT-4 to ensure that there are sufficient descriptions of events between a victim and perpetrator so we can identify patterns of such events in a meaningful way. This resulted in 1576 stories. The remaining analysis in this paper considers only this set of stories.

5 Generating Topics for Sentences

After extracting sentences containing events, we generated topics for those sentences in order to uncover patterns in topics across stories.

5.1 TopicGPT

We use TopicGPT (Pham et al., 2023) to generate topics for the sentences extracted from stories. TopicGPT is a prompt-based framework that uses LLMs to uncover latent topics in a text collection

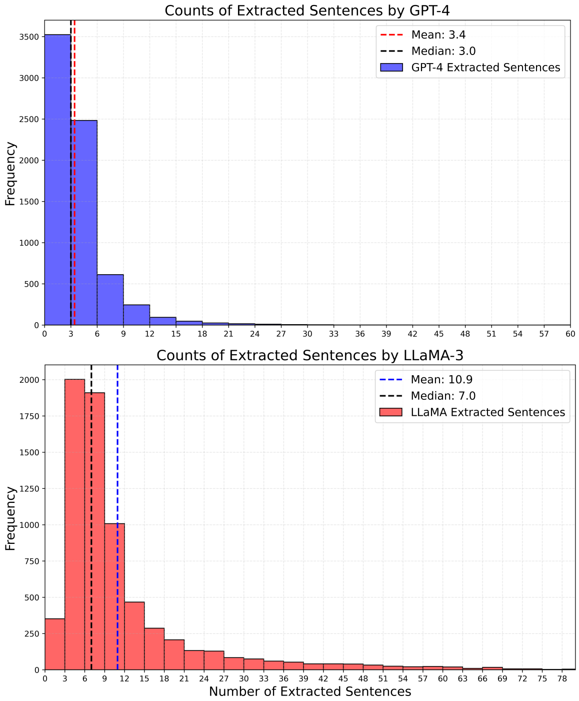

Figure 2: Distribution of the number of extracted sentences per story by GPT-4 and LLaMA-3 in our dataset.

(Pham et al., 2023). Given a corpus and some manually-curated example topics, TopicGPT identifies additional topics in each corpus document. For each document, the model is instructed to either assign a document to an existing topic or generate a new topic that better describes the document and add it to the list of topics. The framework then refines the list by merging repeated topics and removing infrequent topics. Once the set of topics are established, given the generated topics, an LLM assigns the most relevant topic to each document.

Previous studies have utilized dependency parsing to capture the main verb of the sentence to represent as the main event (Chaturvedi et al., 2017). However, with this approach, some contextual and useful information is lost in complex sentences which contain more than one verb. The advantage of using TopicGPT is that it assigns topics to sentences which are closely aligned with human categorizations and this approach sustains more context (Pham et al., 2023). Additionally, it allows us to inject our prior knowledge about topics that are extremely likely to be seen in the documents. To craft the initial set of topics which will improve TopicGPT's performance, we look at scientific works on domestic violence.

5.2 Initial Set of Topics

Intimate partner violence and its typologies have been studied extensively (Ali et al., 2016; Chapman and Gillespie, 2019; Krebs et al., 2011). The world health organization defines IPV as "behavior within an intimate relationship that causes physical, sexual or psychological harm, including acts of physical aggression, sexual coercion, psychological abuse and controlling behaviors" (Organization et al., 2010). One of the most commonly used measures of IPV is the revised conflict tactics scale (CTS2) (Straus et al., 1996). These scales were created to objectively measure the prevalence and frequency of tactics used by partners to resolve conflicts in dating, cohabiting, or marital relationships (Chapman and Gillespie, 2019). The CTS2 includes scales to measure four conflict tactics: *physical assault*, *psychological aggression*, *negotiation*, and *sexual coercion*. Each scale is divided into two subscales—*minor* and *severe*—with negotiation further including emotional and cognitive components. These eight high-level topics form our initial set of topics which we pass to the model as part of our topic generation process.

5.3 Generating Topics

To generate topics for the sentences which were extracted in the previous section, we used a slightly modified version of TopicGPT. The prompt we used is available in the Appendix section (A.2).

First, we passed the extracted sentences to the LLM individually. Next, instead of running the framework in two separate phases (generation and assignment), we provided the model with a predefined set of initial topics and instructions to assign one of the provided topic(s) or generate a topic for the sentence if there is no topic to which the model belongs. At each iteration, a newly generated topic is retained only if it is not too similar to an existing topic. To measure topic similarity, we use Sentence-BERT (Reimers, 2019) to capture topic embeddings. Figure 3 summarizes the the number of unique topics found after processing all 1576 stories with different similarity thresholds. As seen in Figure 3, using similarity thresholds in the set {0.5, 0.6, 0.7} will lead to a stable number of unique topics after processing around 300-600 stories for both models, whereas setting the similarity threshold to higher values generates unbounded number of topics as the number of stories grows. We set the similarity threshold to 0.7 to limit the

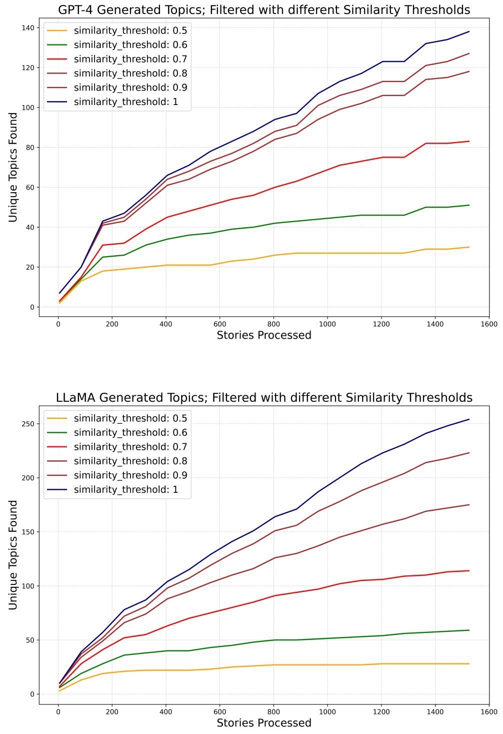

Figure 3: Number of unique topics found using different similarity thresholds.

number of generated topics but also to allow for more nuance in the generated topics. This resulted in 83 different topics.

6 Analysis

After identifying eventful sentences and generating topics for them, we then aimed to identify patterns of topics across stories.

6.1 Topic Co-occurrence

To find patterns within the stories, we investigate the topics that co-occur most frequently together within a story. Figure 5 presents a Pointwise Positive Mutual Information (PPMI) heatmap, capturing the relationships between the top 10 most frequent topics in the stories.

A notable pattern is the strong connection between "emotional manipulation" and "financial neglect", suggesting that financial and emotional control often co-occur within survivor narratives. Similarly, "economic abuse" frequently appears alongside "minor psychological aggression". The association between "substance use" and "legal protections" suggests that intoxication often precipitates conflicts or incidents that result in legal interventions, such as protective orders or law enforcement involvement.

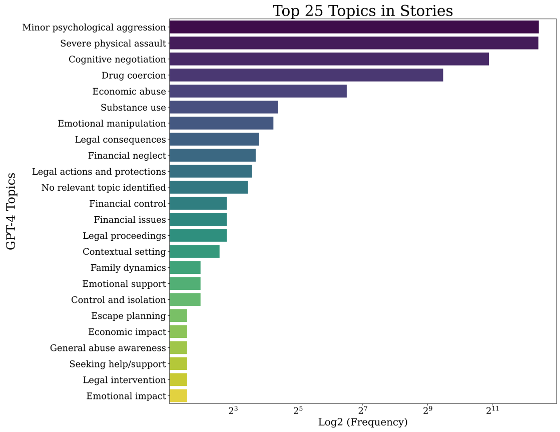

Figure 4: The top 25 most frequent topics generated by GPT-4. The x-axis represents the log_2-transformed frequency of each topic.

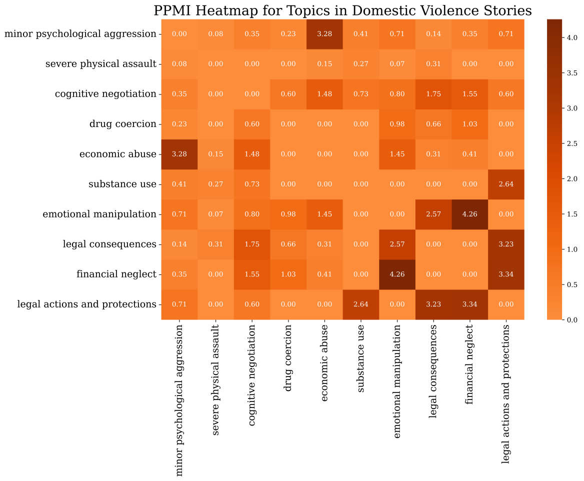

Figure 5: PPMI heatmap showing the relationships between the top 10 most frequent topics assigned to extracted sentences. Darker shades indicate stronger-than-expected associations between topics.

Interestingly, some topics have low or zero co-occurrence with others, such as "severe physical assault", which does not show strong connections with many of the top topics. This suggests that descriptions of physical violence may often appear in isolation, rather than alongside financial or psychological abuse in the same sentence-level context.

Overall, this heatmap highlights the interconnected nature of abuse forms, showing how certain patterns of violence, manipulation, and financial control frequently emerge together in survivor accounts. The strong positive PMI values for certain topic pairs reinforce the idea that domestic abuse is often multidimensional, rather than consisting of isolated forms of harm.

6.2 Topic N-grams and Sequential Patterns

To identify meaningful topic patterns beyond simple frequency biases, we employed a Monte Carlo-based significance analysis (Robert et al., 1999). In our data so far, we have reduced each story into its eventful constituent sentences and each sentence into its dominant topic(s), constructing a set of topic sequences. In this section, we construct topic sequences with lengths of two, three, four, and five, and we refer to them as "topic n-grams". Since certain topics occur more frequently overall (see Figure 4), raw frequency counts of topic n-grams are insufficient for detecting meaningful patterns. To account for this, we generated a null distribution by randomly shuffling topics across sentences and stories while preserving the original dataset's structure. To preserve the dataset's structure, we maintain the number of stories, the number of sentences per story, and the occurrences of each topic within a sentence. By running multiple Monte Carlo simulations under these constraints, we computed the expected frequency of each topic n-gram under random shuffles within each sentence. The most distinctive topic n-grams were identified as those whose observed frequency in the real dataset was significantly greater than their expected frequency under the null distribution, as determined by statistical significance testing. Statistical significance was determined using Z-scores and one-tailed p-values from a normal approximation, ensuring that the extracted patterns reflect genuine structural relationships in the data rather than simple topic frequency effects.

N-gram	Total N-grams	Statistically Significant N-grams
3-grams	540	213
4-grams	934	484
5-grams	1511	737

Table 3: Number of statistically significant n-grams in the dataset based on Monte Carlo simulations (α = 0.05, one-tailed test with Z > 1.645).

The results presented in Table 3 indicate that a substantial proportion of topic n-grams exhibit statistically significant deviations from the null distribution, suggesting the presence of structured topic sequences in the dataset. The relatively high proportion of significant topic n-grams across all levels reinforces the idea that topic transitions are not random, but rather follow discernible patterns, reflecting underlying thematic structures in the stories.

Table 4 and Table 5 present the top tri-grams and four-grams respectively. The tables highlight the key narrative structures that emerge across stories, reinforcing the presence of natural topic progressions that differ from random assignment of topics. Many of these statistically significant n-grams encapsulate intuitive thematic patterns that summarize recurring story structures at an abstract level. As an example, in the Table 4, the sequence *"seeking help/support → emotional support → preparation for emergencies"* represent coherent progressions of events that naturally align with real-world experiences.

Tri-gram	Z-score
Seeking help/support - Emotional support - Preparation for emergencies	23.42
Legal and custodial actions - Legal consequences - Severe physical assault	10.40
Emotional support - Preparation for emergencies - Minor psychological aggression	10.10
Minor psychological aggression - Legal and custodial actions - Legal consequences	8.37
Cognitive negotiation - Legal actions and protections - Economic impact	7.50

Table 4: Top statistically significant trigrams based on Monte Carlo simulations ($\alpha = 0.05$, one-tailed test with $Z > 1.645$).

Four-gram	Z-score
Drug coercion - Economic abuse - Drug coercion - Severe physical assault	7.85
Financial neglect - Minor psychological aggression - Severe physical assault - Minor psychological aggression	4.64
Severe physical assault - Drug coercion - Cognitive negotiation - Cognitive negotiation	4.17
Severe physical assault - Emotional manipulation - Cognitive negotiation - Minor psychological aggression	4.08
Minor psychological aggression - Severe physical assault - Emotional manipulation - Cognitive negotiation	4.06

Table 5: Top statistically significant four-grams based on Monte Carlo simulations ($\alpha = 0.05$, one-tailed test with $Z > 1.645$).

Overall, these results indicate that topic sequences in the dataset are not merely driven by individual topic frequencies, but rather follow predictable, structured progressions that characterize different forms of conflict, abuse, and crisis response.

7 Conclusion

In this paper, we analyzed a large dataset of domestic violence stories posted on Reddit. We investigate LLMs' ability to extract events which involve main characters of the story. Our findings suggest that despite LLMs showing remarkable performance across various NLP tasks, they still fall short of human-level performance for extracting events that meet specific conditions. We used a modern LLM-based topic modeling approach, TopicGPT, and find it suits our task well, as is able to assign coherent and interpretable topics to sentences in the story. Our proposed method, an LLM based pipeline for extracting sentences and assigning topics to them, reduces each story into a structured topic sequence, facilitating narrative analysis. Using Monte Carlo simulations, we examined the topic sequences generated by our method, and found them to contain meaningful structures which are significantly different than any random assignment of the assigned topics. The results validate that our pipeline extracts structural patterns that are highly interpretable. In future work, we will analyze the stories with a generative approach and develop techniques for identifying narratives that deviate from predominant topic progression patterns.

Limitations

Despite the valuable insights gained from our analysis of domestic violence narratives, our approach has several limitations. First, the limited number of human-annotated examples for event extraction constrains the quality of model supervision, potentially affecting the accuracy of both tasks. Expanding the annotation set could lead to better understanding of LLMs' performance for event extraction. Second, our approach is susceptible to error propagation, as inaccuracies in event extraction directly impact the quality of topic assignments. For instance, if the LLM fails to identify a key event, the resulting topic sequence may misrepresent the narrative's structure, leading to misleading conclusions about topic progression patterns. Lastly, while we modeled topic transitions using a sequence-based approach, other methods of sequential analysis, such as Hidden Markov Models (HMMs), Recurrent Neural Networks (RNNs) could provide alternative perspectives on narrative structures. Exploring these methods in future work could enhance our understanding of how domestic violence narratives evolve over time.

Ethical Considerations

We use publicly available Reddit posts while adhering to the platform's terms of service, but we recognize the sensitive nature of the content. To protect individuals' anonymity, we do not disclose usernames, personal identifiers, or specific excerpts that could lead to the identification of survivors.

Our findings highlight common patterns in domestic violence narratives based on event and topic analysis. However, we stress that these patterns should not be used to invalidate or discredit stories that deviate from them, as every survivor's experience is unique. A story that does not follow the typical narrative structure identified in our study is not inherently inaccurate or less credible. Our analysis aims to provide insights into common themes, not to impose a rigid framework for assessing narrative authenticity.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. 2014. A comparison of the events and relations across ace, ere, tac-kbp, and framenet annotation standards. In *Proceedings of the second workshop on EVENTS: Definition, detection, coreference, and representation*, pages 45–53.

Parveen Azam Ali, Katie Dhingra, and Julie McGarry. 2016. A literature review of intimate partner violence and its classifications. *Aggression and violent behavior*, 31:16–25.

Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27.

Maria Antoniak, Joel Mire, Maarten Sap, Elliott Ash, and Andrew Piper. 2023. Where do people tell stories online? story detection across online communities. *arXiv preprint arXiv:2311.09675*.

Ryan L Boyd, Kate G Blackburn, and James W Pennebaker. 2020. The narrative arc: Revealing core narrative structures through text analysis. *Science advances*, 6(32):eaba2196.

Alexander Calderwood, Elizabeth A Pruett, Raymond Ptucha, Christopher Homan, and Cecilia Ovesdotter Alm. 2017. Understanding the semantics of narratives of interpersonal violence through reader annotations and physiological reactions. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 1–9.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Harriet Chapman and Steven M Gillespie. 2019. The revised conflict tactics scales (cts2): A review of the properties, reliability, and validity of the cts2 as a measure of partner abuse in community and clinical samples. *Aggression and violent behavior*, 44:27–35.

Snigdha Chaturvedi, Mohit Iyyer, and Hal Daume III. 2017. Unsupervised learning of evolving relationships between literary characters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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

Joshua D Eisenberg and Mark Finlayson. 2021. Narrative boundaries annotation guide. *Journal of Cultural Analytics*, 6(4).

Mourad Heddaya, Qingcheng Zeng, Chenhao Tan, Rob Voigt, and Alexander Zentefis. 2024. Causal micronarratives. *arXiv preprint arXiv:2410.05252*.

Lifu Huang, Heng Ji, Kyunghyun Cho, and Clare R Voss. 2017. Zero-shot transfer learning for event extraction. *arXiv preprint arXiv:1707.01066*.

Sweta Karlekar and Mohit Bansal. 2018. Safecity: Understanding diverse forms of sexual harassment personal stories. *arXiv preprint arXiv:1809.04739*.

Christopher Krebs, Matthew J Breiding, Angela Browne, and Tara Warner. 2011. The association between different types of intimate partner violence experienced by women. *Journal of Family Violence*, 26:487–500.

William Labov and Joshua Waletzky. 1997. Narrative analysis: Oral versions of personal experience.

Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, et al. 2022. A survey on deep learning event extraction: Approaches and applications. *IEEE Transactions on Neural Networks and Learning Systems*.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. *arXiv preprint arXiv:2104.05919*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. *arXiv preprint arXiv:2106.09232*.

Kory Wallace Mathewson, Pablo Samuel Castro, Colin Cherry, George Foster, and Marc G Bellemare. 2020. Shaping the narrative arc: Information-theoretic collaborative dialogue. In *Proceedings of the 11th International Conference on Computational Creativity*, pages 9–16.

Semi Min and Juyong Park. 2016. Mapping out narrative structures and dynamics using networks and textual information. *arXiv preprint arXiv:1604.03029*.

Seyed Mahed Mousavi, Shohei Tanaka, Gabriel Roccabruna, Koichiro Yoshino, Satoshi Nakamura, and Giuseppe Riccardi. 2023. Whats new? identifying the unfolding of new events in narratives. *arXiv preprint arXiv:2302.07748*.

World Health Organization et al. 2010. *Preventing intimate partner and sexual violence against women: Taking action and generating evidence*. World Health Organization.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2023. Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*.

Andrew Piper and Sunyam Bagga. 2024. Using large language models for understanding narrative discourse. In *Proceedings of the The 6th Workshop on Narrative Understanding*, pages 37–46.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Christian P Robert, George Casella, and George Casella. 1999. *Monte Carlo statistical methods*, volume 2. Springer.

Benjamin M Schmidt. 2015. Plot arceology: A vector-space model of narrative structure. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 1667–1672. IEEE.

Nicolas Schrading, Cecilia Ovesdotter Alm, Raymond Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on reddit. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2577–2583.

Mohammad Shokri, Allison Bishop, and Sarah Ita Levitan. 2024. Is it safe to tell your story? towards achieving privacy for sensitive narratives. In *The 6th Workshop on Narrative Understanding*, page 47.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3623–3634.

Murray A Straus, Sherry L Hamby, SUE Boney-McCoy, and David B Sugarman. 1996. The revised conflict tactics scales (cts2) development and preliminary psychometric data. *Journal of family issues*, 17(3):283–316.

Md Nayem Uddin, Enfa Rose George, Eduardo Blanco, and Steven Corman. 2024. Asking and answering questions to extract event-argument structures. *arXiv preprint arXiv:2404.16413*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Michael Vauth, Hans Ole Hatzel, Evelyn Gius, and Chris Biemann. 2021. Automated event annotation in literary texts. In *CHR*, pages 333–345.

Zeno Vendler. 1957. Verbs and times. *The philosophical review*, 66(2):143–160.

Eitan Wagner, Renana Keydar, and Omri Abend. 2024. Zero-shot trajectory mapping in holocaust testimonies. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes)@ LREC-COLING 2024*, pages 63–70.

Eitan Wagner, Renana Keydar, Amit Pinchevski, and Omri Abend. 2022. Topical segmentation of spoken narratives: A test case on holocaust survivor testimonies. *arXiv preprint arXiv:2210.13783*.

Xiyang Zhang, Muhao Chen, and Jonathan May. 2021. Salience-aware event chain modeling for narrative understanding. *arXiv preprint arXiv:2109.10475*.

A Prompts

We include the exact prompts used for LLaMA-3.1 and GPT-4 during event extraction and topic assignment to ensure the reproducibility of our experiments. These prompts guided the models to extract sentences involving specific characters and to assign topics to narrative segments. Below are the prompts we used.

A.1 Prompts Used for Event Extraction

Following is the prompt we used for the event extraction task with both our models:

```
[Event Definition]
Events are “a singular occurrence at
a particular place and time.” General,
repeating, isolated, or hypothetical
situations, states, and actions are
usually not events.
Most stories are told in the past tense.
Present and future tense can also be used,
but the bar is higher and the narrated
events need to be strongly story-like.
Most events are positively asserted as
occurring, but depending on the context,
negative verbs can also be events when
occurring at a specific time and place.
Verbal interactions could be events too.
Events are usually verbs but can also be
nouns and adjectives.


Read the story below and extract ALL the
sentences that describe an event which
only involves both the victim and the
perpetrator in the story.


[few-shot examples]
.
.
.

[Story]
{}


Please ONLY return the extracted
sentences.

[Your output]
Extracted sentences:
```

We provided three examples in the prompt for event extraction task. Due to space limitations, we didn’t write them in the above prompt. We chose three of the annotated stories as few-shot examples and provided as few-shot examples in the prompt.

A.2 Prompts Used for Topic Generation

Following is the prompt we used for topic generation for both models:

```
You will receive a sentence from a
domestic violence story posted on reddit
and a set of topics. Your task is to
identify topics within the sentence
which describe the sentence best. If
any relevant topics are missing from the
provided set, please add them. Otherwise,
output the existing topic as identified
in the sentence.


[Topics]
{}


[Instructions]
Step 1: Determine topics mentioned in
the sentence which describe the sentence
best. - The topics must reflect a SINGLE
topic instead of a combination of topics.
- The new topics must have a short general
label. - The topics must be broad enough
to accommodate future subtopics.


[Example]
Sentence: He strangled me and told me he
is going to kill me next time.
Topics:
1. Severe physical assault
2. Severe psychological agression


[Sentence]
{}


Please ONLY return the relevant or
modified topics.

[Your response]
Topics:
```

A.3 Example Output

Here we show an example from GPT-4’s output for both tasks related to the following story.

Content Warning: The following story contains language that may be offensive or disturbing to some readers.

```
Female, 19.dated my now ex-boyfriend (who is
20, turns 21 in a month) for a year and seven
months. After we broke up in October of 2010,
I was devastated. A lot of my friends didn’t
really understand why. They knew that he had
been emotionally/verbally abusive and of course
knew I was better off, but I never really came
to that conclusion until many, many months later.
He’d never outright call me fat or ugly, but he
definitely found indirect ways to tell me. He
even told me, after we’d been broken up for a
couple of months, that if I were to have sex with
anyone else but him, I’d be a slut. He’s the
only person I’ve ever slept with, and we were
in a committed relationship for awhile before we
started having sex.struggled with my self-esteem
a lot before I started dating him, but now it’s at
all all-time low. Everything he used to say burns
in the back of my mind. I go to the gym and work
out three days a week and do pilates twice a week
and try to eat healthy, but I feel like it will
never be enough. I’m 5’2” and 135 pounds. I’m a
size 4-6. But it’s become an unhealthy obsession
to win his approval even though I know I’m never
going to get it. I want to be smaller. Lighter.
Thinner.called me the other day and wanted to talk
with me, so we hung out for a bit. He again found
ways to call me a slut, and tell me that I’m still
not as thin as I could be. I don’t understand
why I need his approval so badly. Other guys have
told me that I have a great body and I’m pretty,
but it holds absolutely no weight.do I stop this
madness? I feel like I’m on a one-way path to
self-destruction and I don’t know how to stop,
only how to slow down or speed up. Please help.

GPT-4 extracted sentences:

1. He even told me, after we’d been broken up for
a couple of months, that if I were to have sex
with anyone else but him, I’d be a slut.
2. He called me the other day and wanted to talk
with me, so we hung out for a bit.
3. He again found ways to call me a slut, and
tell me that I’m still not as thin as I could be.

GPT-4 assigned topics:

1. minor psychological aggression
2. Cognitive negotiation
3. Minor psychological aggression
```