Mrs. Dalloway Said She Would Segment the Chapters Herself

Peiqi Sui[1,2], **Lin Wang**[2], **Sil Hamilton**[3], **Thorsten Ries**[1], **Kelvin Wong**[2], and **Stephen T. Wong**[2]
[1] University of Texas at Austin
[2] System Medicine and Bioengineering (SMAB), Houston Methodist Cancer Center
[3] McGill University
peiqisui@utexas.edu

Abstract

This paper proposes a sentiment-centric pipeline to perform unsupervised plot extraction on non-linear novels like Virginia Woolf's *Mrs. Dalloway*, a novel widely considered to be "plotless." Combining transformer-based sentiment analysis models with statistical testing, we model sentiment's rate-of-change and correspondingly segment the novel into emotionally self-contained units qualitatively evaluated to be meaningful surrogate pseudo-chapters. We validate our findings by evaluating our pipeline as a fully unsupervised text segmentation model, achieving a F-1 score of 0.643 (regional) and 0.214 (exact) in chapter break prediction on a validation set of linear novels with existing chapter structures. In addition, we observe notable differences between the distributions of predicted chapter lengths in linear and non-linear fictional narratives, with the latter exhibiting significantly greater variability. Our results hold significance for narrative researchers appraising methods for extracting plots from non-linear novels.

1 Introduction

What is the shape of a story? Narratologists have long been fascinated with reducing narratives to a compelling linear visual rhetoric: the narrative arc, a line chart that smoothly demonstrates the (emotional) rise and fall of the story (Freytag, 1895; Campbell, 1949; Propp, 1968). Recent scholarship has introduced emotive expressions and affect as a vital analytical tool for the construction of such narrative arcs (Kleres, 2011; Keen, 2011; Winkler et al., 2023). The digital humanities community has shown great interest in operationalizing this problem as a sentiment analysis task across various literary corpora (Jockers 2015; Underwood, 2015; Elkins, 2022). The success of this approach has recently been extended beyond the literary domain to encompass a wider range of inquiries driven by social science (Boyd et al., 2020; Chun 2021).

Meanwhile, existing methods for sentiment-based narrative arc extraction tend to underperform on what literary scholars call non-linear narratives (Richardson, 2000). We posit that literary works often assume varying degrees of clarity and straightforwardness when conveying a story, an explicative quality known as narrativity — computationally, it has been defined as a scalar measuring the success of a work in conveying a linear sequence of events as narrative discourse (Piper et al., 2021). While some novels may convey their story-worlds with relative transparency via chronological accounts of their fictional agents' actions, others may withhold it from the audience for artistic purposes (Pianzola, 2018). This non-linearity has been considered a hard problem for narratology, by both computational (Elkins and Chun, 2019; Bhyravajjula et. al, 2022) and traditional (Ryan, 2005) approaches. Virginia Woolf's 1925 novel *Mrs. Dalloway*, in particular, has been identified as an especially recalcitrant text to model with existing methods (Elkins, 2022), possibly due to its renowned stream-of-consciousness style.

This study takes on the challenging task of performing unsupervised plot extraction on *Mrs. Dalloway*, a novel widely held and celebrated by literary scholars to be essentially "plotless." We hypothesize that it is possible to excavate latent plot structures from nonlinear fiction if we use sentiment data to statistically model the notion of non-linearity itself. To avoid the pitfall of imposing

Proceedings of the The 5th Workshop on Narrative Understanding, pages 92–105
July 14, 2023 ©2023 Association for Computational Linguistics

linear narrative arcs onto non-linear narratives via smoothing-based de-noising techniques, we propose a sentiment-centric pipeline which instead aims to embrace the "noise" inherent to a non-linear and highly fragmented novel like *Mrs. Dalloway*. The goal of this pipeline is to capture the full expression of non-linearity in sentiment data. Leveraging the softmax probability distributions of pre-trained language models like BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020), we perform a paired t-test that models sentiment's rate-of-change to identify breakpoints and correspondingly segment the novel into emotionally coherent parts. Our approach finds 19 such surrogate "chapters" in *Mrs. Dalloway*, which we then qualitatively evaluate to assess their literary and narratological coherence. To further verify the validity of our results, we quantitatively evaluate our pipeline on a linear fiction dataset to determine its ability to restore existing chapter structures, while also contributing a generative approach to the task of text segmentation in the literary domain.

Main contributions: 1) a segment-based approach to plot extraction, designed to address the challenge of modeling non-linear fiction 2) a sentiment-centric pipeline for fully unsupervised chapter segmentation 3) an attempt to consider literary theoretical claims as falsifiable hypotheses that could inform model design, in the hope for the greater inclusion of literary scholarship in the collaboration pipeline for NLP research.

2 Background and Related Work

2.1 Definitions

Our study will operate under the following narratological definitions:

- **Plot:** We define plot not as a fixed structure but a gradual process of structuration, a dynamic development that actualizes and amends itself as the narrative unfolds and constantly reshapes the experience of reading (Brooks, 1984; Phelan, 1989). The concept of structuration highlights the need to examine not just the general distribution of sentiment scores, but also as the relative rate of change between measured points of sentiment.
- **Linearity:** Colloquially, linear narratives often refer to storylines that are aligned with chronological order. In narratology, linearity is the side of plot that relates to causality (Forster, 1927), and linear narratives are the framing of fictional "action[s] as a chronological, cause-and-effect chain of events occurring within a given duration and a spatial field" (Bordwell, 1985:49). The metaphor of the "chain" necessitates something that comes before together with a subsequent, a sequence of events that becomes coherent for the audience through the clear cognition of time and a correspondence between the two. Linearity, therefore, is the plot made visible via having its causal sequences ordered in the plain sight of chronological time.
- **Sentiment:** We borrow our definition of sentiment from leading narrative theorist Patrick Hogan, who views emotion as the hallmark of non-linearity: "our emotion systems respond to perceptual fragments […] these cluster into incidents that provoke emotional spikes in emotional experiences that are, like time, not smoothly continuous but jagged," (2011:66). In making this claim, Hogan draws a distinction between objective, universal clock time and our non-uniform experience of temporality. Just as objective time orders causal chains into a linear plot that makes sense to the audience, subjective narrative time is organized by emotional fluctuations into coherent units, which we hope to segment with sentiment analysis. In the context of our study, Hogan's argument implies that our sentiment analysis pipeline would be expected to extract a set of "jagged" distributions from non-linear novels, instead of a smooth line, to represent the non-linear narrative arc.

Hypothesis: Operationalizing Hogan's (2011) theory of affective narratology, which heavily emphasizes an underlying connection between plot, non-linearity, and sentiment, we propose a conception of plot as a continuous process of structuration with two components: the easily

observable[1], time-dependent arm as a causal chain of events ordered in objective time, and the latent, time-independent arm as the fragmented, non-linear, yet internally coherent, narrative arc concealed in emotion. These two arms are not always present in all narratives. Rather, they are two ways for a plot to be expressed, and if they happen to coexist like in linear narratives, their structure tends to synchronize because they essentially describe different aspects of the same plot. This narratological unity they share enables the use of the observable arm as gold-standard ground truth to validate inferences made from the latent arm. Through a combination of qualitative and quantitative evaluations of our pipeline's output in Section 4, we aim to holistically validate our use of sentiment as a plausible approach to plot extraction.

2.2 Narrative Arc Construction with Sentiment Analysis

Prior research in this area has heavily relied on smoothing techniques to identify linear and human-readable patterns in the noisy sentiment data of long-form texts. Jockers' (2015) *Syuzhet* utilizes fast Fourier transform and discrete cosine transformation to extract sentiment arcs from its lexicon-based sentiment models. Gao et al. (2016) build on Jockers' work by employing a more complex model for smoothing with an adaptive filter. More recently, Chun (2021) proposes an ensemble approach that combines the outcomes of multiple sentiment models to mitigate model and dataset bias, while still requiring smoothing with simple moving average. To the best our knowledge, we are the first study to extract narrative arcs from sentiment data without any involvement of smoothing. For our intent and purposes, smoothing is problematic because it seeks to reduce non-linear narratives to a clean yet oversimplified line.

2.3 Text Segmentation in Fiction

Since our pipeline outputs a segmented narrative arc, it also contributes to the broader problem of text segmentation in the literary domain. Recent studies have fine-tuned pre-trained language models to perform chapter segmentation, and their methods tend to use classification-based, reducing the problem to the binary classification of each potential breakpoint candidate as a predicted chapter boundary or not. Pethe et al. (2020) fine-tune BERT's next sentence prediction model as a binary classifier for chapter break prediction, and use the inference's confidence score to rank all breakpoint candidates in each novel to select the top P as predicted chapter breaks, P being the number of ground truth chapters. Their approach outperforms all non-transformer baselines. Virameteekul (2022) further improves Pethe et al.'s performance by utilizing a XLNet and a CNN model instead of BERT.

Although our quantitative evaluations in Section 4.2 perform the same task as these existing approaches, we cannot use them as baselines due to significant differences in methods and experimental setting. This includes: 1) our pipeline is fully unsupervised, without any knowledge of P during inference 2) our sentiment models are not fine-tuned any chapter break training data, and 3) the paired t-test in our study could only infer segments on the multiples of the initial sequence length *alpha*, making it arithmetically impossible to locate most exact chapter breaks. Nonetheless, our study could be considered as a generative approach to text segmentation, an alternative to existing classification-based methods.

3 Methods

This section describes the experimental designs of our pipeline. It takes a text file of *Mrs. Dalloway* and outputs an unsmoothed narrative arc segmented into surrogate "chapters," as shown in Figure 1 below.

3.1 Literary Domain Fine-Tuning for Sentiment Analysis

For sentiment analysis, we fine-tune a pre-trained ELECTRA model on a Victorian fiction sentiment dataset (Kim, 2022), the only open-source fiction dataset we find with sentence-level sentiment labels. ELECTRA is selected over BERT because it reports better performance on benchmark sentiment analysis datasets (Clark et al, 2020). We also implement a popular BERT-based sentiment model obtained from HuggingFace fine-tuned on product reviews (nlptown, 2019) as a general-domain reference. Using an additional model allows us to troubleshoot the question of

[1] "Observable" here means both being visible on the page, i.e., chapter boundaries, and being causally cognizable, i.e., the chain-of-thought summaries of "what happened" ordered chronologically.

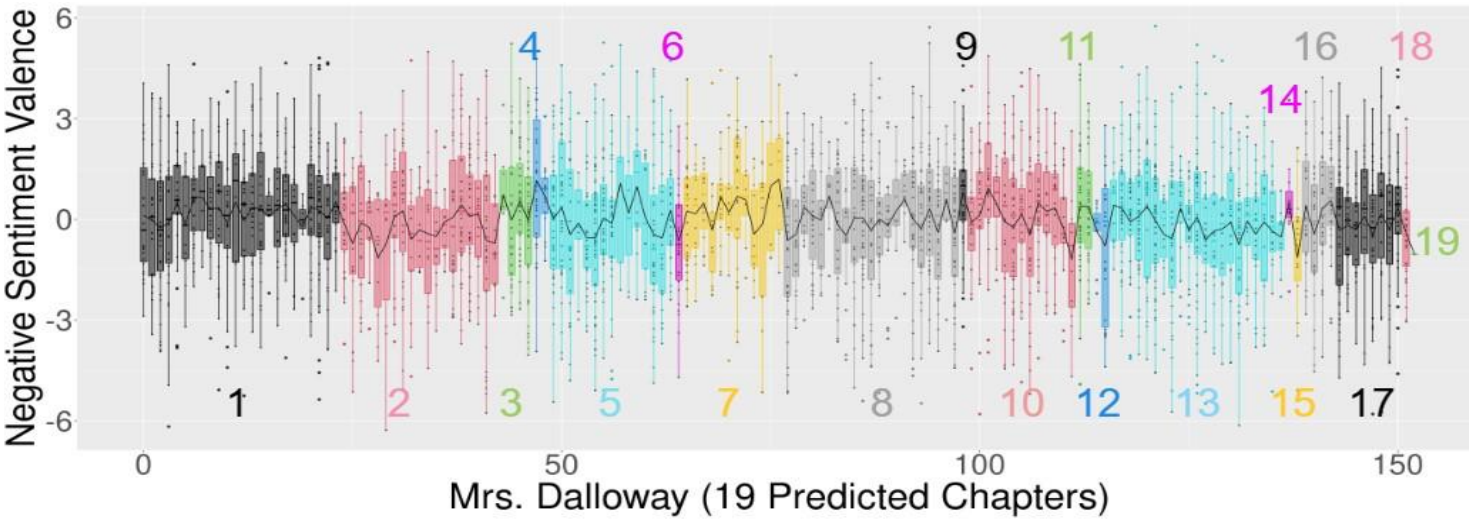
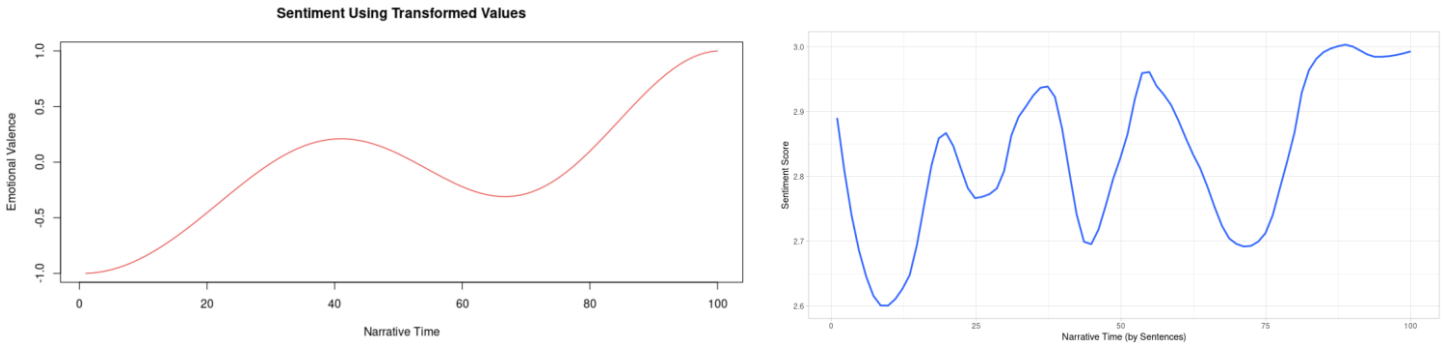


Figure 1. Segmented narrative arc the BERT model extracts from *Mrs. Dalloway*, with 19 predicted chapters (use of the same color does not indicate the continuation of the same chapter)



(a) (b)

Figure 2. *Mrs. Dalloway*'s sentiment visualized with existing methods: (a) fast Fourier transform from Jockers' *Syuzhet*, (b) Loess smoothing (span=0.3)

cross-domain consistency, as the performance of many NLP systems has been demonstrated to "drop precipitously" when applied to the literary domain (Bamman et al., 2019:2141). This is supported by our model evaluations, as the ELECTRA model's testing metrics on its hold-out test set (accuracy=0.49, ovr AUC=0.734) significantly outperforms that of the BERT model's inference on the same test set (accuracy=0.21, ovr AUC=0.613).

Inference: We run each sentence of *Mrs. Dalloway* (n=3626, no preprocessing) through the sentiment classifiers and generate a 5-dimensional probability distribution to represent the sentence's sentiment (details in Section 4.2). Despite the gap in testing performance, we pass the output of both models into our subsequent pipeline to further experiment with cross-domain consistency, since recent studies have demonstrated that there is no one-size-fit-all sentiment model for constructing narrative arcs (Elkins, 2022).

3.2 Breakpoint Analysis with Statistical Testing

When implemented on non-linear narratives like *Mrs. Dalloway*, past studies' use of smoothing (Section 2.2) often produces oversimplified and unconvincing results that lack explainability (Figure 2a & 2b). To address this shortcoming, we perform segment detection to extract a fully explainable narrative arc: any statistically significant movement on the sentiment arc could be traced back to the corresponding sentence where a breakpoint is identified, making it possible to explain model decision with close reading, the gold standard for evidentiary claims in literary studies (Felski, 2008).

Dimensionality Transformation: The default output of both the BERT and the ELECTRA model is a sentiment score on a 5-point Likert scale, a monolithic representation that often fails to capture the nuanced sentiment of literary texts. The neutral sentiment label, in particular, is

Input Sentence	ELECTRA Softmax	PCA (PC1, PC2)
"And then, thought Clarissa Dalloway, what a morning-fresh as if issued to children on a beach."	0.107, 0.453, 0.413, 0.002, 0.025	1.066, -1.942
"What a lark!"	0.118, 0.284, 0.0034, **0.595**, 0.178	-0.54, -0.457
"For it was the middle of June."	0.247, 0.131, 0.145, **0.367**, 0.111	3.899, 3.17

Table 1. Dimensionality transformation of sentiment data

responsible for many trivial zero values in narrative arc plotting whose presence does not necessarily correlate with actual emotional neutrality (Elkins, 2022). To avoid this pitfall and mitigate the issue of data oversimplification, we configure the sentiment models to directly output the 5-dimensional probability vector, instead of taking the argmax of the softmax probabilities of each sentence. This approach allows for a more holistic representation of ambiguous sentiment by transforming a discrete sentiment scale into a continuous one, as shown in Table 1. We also experiment with the unnormalized logits tensors prior to the softmax operation and observe similar outcomes in subsequent procedures.

To preprocess and denoise the 5-dimensional sentiment scores for segment detection, we utilize *principal component analysis* (PCA) to identify the two most significant emotional dimensions of any given distribution that are linearly independent from each other. PCA is an orthogonal linear transformation technique that renders the greatest variance of all projections of the data onto the first coordinate, then the second greatest variance on the second coordinate, and so on. This simplifies the dataset while maximizing data preservation, as PCA finds that the first two principal components could explain more than 99% of the variability in both the BERT and the ELECTRA model's sentiment predictions on *Mrs. Dalloway*[2]. We only keep these two dimensions, given that they contain the most substantial information regarding the data and contribute the highest variance. Since each pair of PC1 and PC2 can be traced back to their originating sentence, PCA allows for a more explainable interpretative framework than previous studies' use of smoothing and numerical filtering techniques.

Segment Detection: We design a statistical model that recurrently performs paired samples t-tests to trace sentiment's rate-of-change across a novel and predict potential locations for breakpoints. The t-test draws from sentence-level sentiment scores and groups them by paragraph. For one group of sentiment scores from passage *P*1 with an average of *M1* and its next group *P2* with an average of *M2*, the null hypothesis $M1 = M2$ is assumed to be true unless the p-value is less than a critical value of 0.1, in which case the alternative hypothesis $M1 \neq M2$ would be established. Since a paired test would require the two groups *P1* and *P2* to have the same length, the model would also take one hyperparameter *alpha*, the number of paragraphs in each passage. To give the model a higher degree of freedom, we set *alpha*=5 for inference on *Mrs. Dalloway*, since the end of the fifth paragraph of the novel is qualitatively determined to be the earliest semantic focal point to break off the initial "chapter." From here on, the model treats each following *alpha* paragraphs as the *P2* and evaluates it against *P1*, concatenating them into a longer *P1* if the difference is not statistically significant, and marking a new chapter if it is while simultaneously making the paragraph in question the new *P1*.

Aside from our definition of plot as structuration, our decision to focus on the rate-of-change is also motivated by narratologist Tzvetan Todorov's equilibrium-disruption model (Todorov, 1971). Todorov posits that literary narratives typically 1) start out in a state of stability, 2) disrupted by an often unexpected event, and 3) iterate through multiple attempts to restore the initial equilibrium as new disruptions arise. This interplay between equilibrium and disequilibrium is often accompanied by rapid shifts in emotions, or fluctuations in sentiment scores that our pipeline captures as a non-linear analog of these movements. Our approach lends especially well to *Mrs. Dalloway*, as literary scholars have noted that such "interrupt[ions]" and "reinstate[ments]" occur recurrently in Woolf's fiction (Richardson, 2000: 686). These interruptions occur in a cyclic manner emblematic of the novel's non-linearity.

[2] This is also the case for all other novels we use for quantitative evaluations in Section 4.

4 Results and Discussion

To holistically evaluate the validity and limitations of our pipeline's output on *Mrs. Dalloway* shown in Figure 1, we follow the approach of Wang and Iyyer (2019) to present the outcomes for literary close reading alongside quantitative metrics.

4.1 Qualitative Evaluations

We perform a domain expert review of the predicted chapter divisions. Contrary to our expectations, the general domain BERT model returns more explainable results on *Mrs. Dalloway* when being compared against the ELECTRA model, extracting the boundaries of 19 reasonable segments that could be thought of as surrogate "chapters" (Table 6). This suggests that domain-specific fine-tuning with the Victorian fiction dataset could not be transferred to *Mrs. Dalloway*, a modernist and non-linear novel. By extension, different literary time periods could be considered as different domains, which is supported by the conclusion of an existing body of research in digital humanities (Underwood, 2013).

Our pipeline succeeds in capturing recurrent disruptions in fictional narratives. We observe that six of the sentences marking the beginning or ending of our predicted segments involve Dr. Holmes or Sir William Bradshaw, both of whom are particularly disruptive characters heavily involved in the suicide of Septimus, one of the novel's protagonists. Peter's conversation with Sally in "chapters" 17 and 18, for instance, represents how a thematically coherent whole could be interrupted by the appearance of the Bradshaws at Clarissa's party. Similarly, Bradshaw being sent for at the beginning of "chapter" 12 interrupts Septimus' last happy moment with Rezia paragraphs before his suicide. Moreover, we note that the first appearance of Holmes that ends "chapter" 7 opens the scene that formally initiates Septimus' radical downward spiral. While Woolf scholars have hypothesized the Bradshaws as the vital link between the Clarissa and Septimus storylines (Joyes, 2008), our findings take this a step further by dissecting the novel through its affective substratum to show his structural significance on an empirical level. We provide the detailed predicted chapters list in the appendix.

Datasets	Linear plot (ground truth segmentations)	Non-linear plot (predicted segmentations)
Linear fiction (9 novels, Section 4.2)	Gold-standard	Pipeline validation
Non-linear fiction (5 novels, Section 4.3)	Does not exist	Pipeline inference

Table 2. Schematic of the relations between linear vs non-linear datasets and the linear vs non-linear distinction in plot defined in Section 2.1.

4.2 Quantitative Evaluations

Data: For quantitative evaluation, we assemble two fictional datasets from Project Gutenberg (Gutenberg, n.d.): 1) 9 linear novels to vertically validate our pipeline's ability to accurately segment fictional narratives, 2) 4 additional non-linear novels to horizontally validate our findings in *Mrs. Dalloway*. For the purposes of quantitative testing, we define linear fiction as novels already divided into chapters by their authors. Conversely, non-linear fiction refers to novels published without existing chapter structures[3], usually out of aesthetic choices (Pianzola, 2018), accompanied by a greater degree of narrative fragmentation that makes them harder to model. Table 2 demonstrates the relation between these surface-level operational definitions and their conceptual counterparts defined in Section 2.1: by definition, non-linear novels do not have linear plot, while linear novels contain both, one observable (chapters) and one latent (sentiment). The non-linear narrative arc that our pipeline extracts is not mutually exclusive with linear narrative features like chapters — linear novels, too, are often embedded in latent emotional spaces, carrying a hidden sentiment arc that co-exists alongside the linear organization of plot through chapters.

The linear fiction dataset only contains Victorian novels, to maintain domain consistency with the ELECTRA model's fine-tuning set. We use Chapterize (Reeve, 2016) to extract from each novel's Gutenberg text file a list of paragraph indices that represent the locations of chapter breakpoints. All 9 lists are then manually curated to

[3] Some non-linear novels are segmented into sections or parts that are not labeled as chapters by the author. They are usually done out of editorial convenience, and do not have a chapter's commitment to thematic coherence and fictional causality. Therefore, we do not consider them as ground truth segmentations.

Novel	F1 (exact location)	*Alpha* (optimal initial chapter length)	F1 (rounded *alpha*)	Predicted chapters (actual chapters)
Adam Bede	0.197	14	0.691	54 (55)
Great Expectations	0.229	7	0.667	59 (59)
Little Dorrit	0.25	6	0.557	153 (70)
North and South	0.172	17	0.738	51 (52)
Lady Audley's Secret	0.217	10	0.633	79 (41)
Oliver Twist	0.29	4	0.641	135 (53)
The Woman in White	0.179	14	0.658	68 (51)
Vanity Fair	0.164	15	0.712	67 (67)
Pride and Prejudice	0.232	7	0.494	57 (61)
All	*0.214*	-	*0.643*	-

Table 4. The ELECTRA model's segmentation performance with tuned *alpha*

ensure that the annotations of chapter boundaries are correct, a step necessary due to the known header alignment issues in Project Gutenberg documents (Pethe et al., 2020). The curated output will be considered as the gold-standard ground truth labels for chapter segmentation.

Chapter Segmentation: The predicted chapter segmentation results from 4.1 could not be directly evaluated with quantitative metrics, due to the absence of author-assigned ground truth chapter segmentation labels in *Mrs. Dalloway*. To overcome this limitation, we opt for indirectly evaluating our results, by testing the ability of our pipeline to restore the existing chapter boundaries of linear novels. In doing so, we hope to validate our approach of extracting emotive plot itself, that it is indeed a form of plot, and generally of its linear counterpart.

We remove all chapter headers and related signals from the texts (the only input text preprocessing step in our study) and apply our pipeline to the linear fiction dataset. To match the inference with the format of the dataset's ground truth chapter segmentations for evaluation, we adjust the pipeline to output from each novel a list of paragraph indices where each predicted chapter begins.

We follow Pethe et al.'s use of F1[4] to report the performance of exact break prediction, with one key caveat: due to the nature of the paired samples t-test, the only potential breakpoint candidates would be the multiples of the hyperparameter *alpha*, which makes it arithmetically impossible for our pipeline to predict the exact location of most

Algorithm	F1 (exact location)	F1 (general area)
Random	0.037	0.101
Dummy	0.028	0.134
BERT	0.095	0.335
ELECTRA (ours)	**0.202**	**0.513**

Table 3. Segmentation performance when *alpha=5*

pipeline, we also compute a general area F1, where the locations of ground truth chapters are rounded up to the closest multiple of *alpha*.

Table 3 compares the performance of our pipeline when utilizing the ELECTRA and BERT sentiment models, with *alpha* set to 5 to remain consistent with the findings of Section 4.1. The literary domain ELECTRA model significantly outperforms the general domain BERT model. To further substantiate this result, we incorporate the following baselines into our evaluation:

- Random: *P* breakpoints are randomly selected from each novel, where *P* denotes the number of ground truth chapters.
- Dummy regressor: *P* breakpoints are randomly selected from all available multiples of *alpha* in each novel. This baseline is designed as an ablation of the use of sentiment analysis.

Since both baselines are randomly generated, we report their average F1 over 10 iterations. Even with the hint of *P* provided as supervision, the baselines' performance remains insignificant. This validates the complexity of the task and the effectiveness of our pipeline.

Table 4 reports the performance of the ELECTRA model on each novel when F1 is not

[4] Since our pipeline is not supervised with the correct number of chapters, it may not predict the same number of segments as the ground truth. This constraint does not meet the input data requirements for other commonly used metrics in text segmentation. Evaluative approaches that we are unable to appropriately utilize include sliding window-based methods, inter-annotator agreement measures, and geometric distances.

Dataset	Variance (ELECTRA)	Variance (BERT)	CV (ELECTRA)	CV (BERT)
Linear fiction	269.91	1025.49	92.11%	106.84%
Non-linear fiction	**744.83**	**1509.62**	109.16%	113.53%

(a) Sentiment models

Dataset	Variance (Random)	Variance (Dummy)	CV (Random)	CV (Dummy)
Linear fiction	3420.97	3875.62	89.06%	93.9%
Non-linear fiction	4157.5	5355.96	89.42%	94.3%

(b) Baselines (averaged iterations = 10)

Table 5. Predicted chapter lengths distribution

fixed. To explore *alpha* as a tunable hyperparameter, we experiment with *alpha* values ranging from 1 to 30, and use the exact location F1 to select the optimal value to compute the general area F1. The improvement provided by the optimal *alpha* is not significant, as the exact location F1 of most *alpha* values tend to be similar. A smaller *alpha* results in a larger number of predicted breaks, covering more ground truth breaks (true positives), while also predicting more breaks where one does not exist (false positives). Conversely, a larger *alpha* means fewer predicted breaks, fewer correct predictions, but also fewer mistakes. Nonetheless, the optimal *alpha* has a significant impact on the accuracy of predicting the number of chapters. The inference of 5 of the 9 optimal *alpha* falls within plus/minus 1 of the ground truth chapter count *P*, while all *alpha* values from 1 to 30 average a Manhattan difference of 53 from *P*.

Using the optimal *alpha*, the ELECTRA model achieves a general area F1-score of 0.643, indicating its ability to predict the location of most chapter boundaries within the margin of a few paragraphs, which is more than adequate given the room for ambiguity in literary works. Our quantitative findings validate the hypothesis put forward in Section 1 that the emotional patterns underlying fictional narratives often correspond with the linear arm of plot, evident in the number of breakpoints that sentiment analysis shares with the existing chapter segmentations of linear novels. This correspondence, in turn, supports the validity of using the sentiment-centric pipeline for inference on non-linear novels like *Mrs. Dalloway*, where the visualizations like Figure 1 serve as the surrogate of linear plot by making a novel's latent emotional space observable.

4.3 Towards Quantifying Non-Linearity in Fiction

Table 5a compares the distribution of predicted chapter lengths (counted by the paragraph) in the linear and non-linear fiction datasets, with *alpha* set again to 5 to maintain consistency with previous experiments. Notably, the lengths of counterparts. However, their coefficient of variation (CV), a metric measured against the mean, does not exhibit a significant difference. This suggests that non-linear novels have more variable chapter lengths compared to linear ones in terms of absolute variability, while the relative variability of the two groups is similar.

We validate this pattern with the baselines from Section 4.2. As Table 5b shows, the random and dummy baselines also produce similar CVs and different variances between linear and non-linear fiction, though the difference is less substantial than that of the sentiment models. This indicates that the difference in variance pertains to the two fictional corpora instead of methods for extraction. Furthermore, the fact that all 4 models produce similar CVs might undermine its effectiveness as a metric in this experiment.

One potential explanation for the discrepancy between variance and CV is that our pipeline identifies more outlier chapters in non-linear novels. The 1% longest chapters the ELECTRA model extracts from the non-linear set contain 7.8% of all paragraphs in the corpus, compared to 5.8% for the chapters in the 99th percentile in length obtained from the linear set. The length of "chapters" in non-linear novels is not constrained by the need to fit a linear plot, therefore containing more outliers that lead to greater variability and fragmentation.

This result further validates our findings in Section 4.1. The BERT model that outputs Figure 1 reports Var=1647.84 and μ=28.15 from the lengths its predicted chapters on *Mrs. Dalloway*, which are consistent with the averages of the non-linear fiction dataset (Var=1509.62, μ=34.97). This offers some support for the generalizability of the outcomes of Section 4.1 to other non-linear novels, if similar qualitative analysis is to be performed on them by domain experts.

5 Conclusion and Future Work

With a pipeline capable of excavating non-linear plot from both non-linear and linear novels, this study takes the first steps to 1) investigate the

hypothesis proposed in Section 1, and 2) explore the positive impact literary theory could have on model design for narrative understanding. We demonstrate that it is possible to extract a narrative arc with coherent segments from non-linear narratives like *Mrs. Dalloway*, and the explainability of our approach affords actionable outcomes for literary studies—explainable results promote empirical theory testing. We validate our findings with both qualitative and quantitative evaluations, achieving a F1 0.643 (general area) and a 0.214 (exact) after hyperparameter tuning. In doing so, we also uncover some evidence for a potential correspondence between the linear (chronological, causal) and nonlinear (emotional) arms of plot in the linear fiction dataset. We further discover that the chapters we extract from non-linear fiction tend to vary more in length, which we understand as a corpus-level difference.

The qualitative analysis in Section 4.1 shows that the general domain BERT model produces more explainable results than the ELECTRA model fine-tuned on Victorian novels, while ELECTRA quantitatively outperforms BERT in Section 4.2. This is not so much a contradiction as a guidance for future research: perhaps the "literary domain" is not a monolith, but an umbrella term for a collection of domains that are significantly different from each other. Is the domain barrier between linear and non-linear fiction? If so, then ELECTRA could be considered as an "in-domain" model for experiments in 4.2 because the object of inference is linear fiction, while not for 4.1 since it concerns non-linearity. It is possible that if ELECTRA is fine-tuned on a non-linear fiction dataset with sentiment labels, it could further improve upon the findings of 4.1.

Aside from these questions, other potential directions for our future work include 1) designing more robust methods for quantifying non-linearity in fiction, which could be leveraged for a wide range of inquiries in digital humanities, 2) combining our sentiment-based pipeline with existing semantic-based approach to improve the performance of chapter segmentation, and 3) expanding this research to narratives beyond the literary domain, which is also a key interest of contemporary narratology. We also hope to open up the discussion of non-linearity in fictional narratives to event-centric and character-centric approaches to better understand the interplay between causal and emotional dimensions of plot.

6 Limitations

Due to various constraints, our experiments are only able to cover five non-linear novels and nine linear novels, listed in Table 7. This pales in comparison to the thousands of novels typically expected of large-scale studies in digital humanities, whose scale allows them to make generalizable claims regarding narratives or literary history (Piper et al., 2021). We hope to make up for this gap in our future work. One key challenge to scaling our dataset would be data availability. The use of non-linearity in fiction is predominantly a 20th century phenomenon, which suggests that many non-linear novels will not be in the public domain for some time to come.

In terms of experiment design, an important limitation of the quantitative evaluations in Section 4.2 is its assumption that a novel's chapter divides provided by its author could be thought of as a form of "gold standard" labels for model validation. This claim of authorial control and "authority" over the text has been thoroughly problematized in literary studies since the emergence of poststructuralism (Barthes, 1967; Foucault, 1969), while analogous suspicions have been raised in natural language generation against the assumed reliability of human evaluators (Clark et al., 2021). Unfortunately, the author's input is the only operationalizable criteria for ground truth available to us within the scope of this study.

Acknowledgement

We sincerely thank Fangyuan Xu, the workshop organizers, and the anonymous reviewers for their generous time, attention, and helpful feedback on this paper. Peiqi Sui, Lin Wang, Kelvin Wong, and Stephen T. Wong were supported by T. T. & W. F. Chao Foundation and the John S Dunn Research Foundation.

References

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Volume 1: Long Papers*), pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Roland Barthes. 1967. The death of the author. *S/Z*. Hill and Wang.

David Bordwell. 2013. *Narration in the fiction film*. Routledge.

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).

Peter Brooks. 1984. *Reading for the plot: Design and intention in narrative*. Harvard University Press.

Sriharsh Bhyravajjula, Ujwal Narayan, and Manish Shrivastava. 2022. Marcus: An event-centric nlp pipeline that generates character arcs from narratives. In *Proceedings of Text2Story: Fourth Workshop on Narrative Extraction from Texts held in conjunction with the 43rd European Conference on Information Retrieval (ECIR 2021)*, pages 67-74.

Joseph Campbell. 1949. *The hero with a thousand faces*. Princeton University Press.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2019. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv: 2003.10555*.

Jon Chun. 2021. SentimentArcs: A novel method for self-supervised sentiment analysis of time series shows SOTA transformers can struggle finding narrative arcs. *arXiv preprint arXiv: 2110.09454*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Katherine Elkins. 2022. *The shapes of stories: Sentiment analysis for narrative*. Cambridge University Press.

Katherine Elkins and Jon Chun. 2019. Can sentiment analysis reveal structure in a "plotless" novel?. *arXiv preprint arXiv:1910.01441*.

Michel Foucault. 1969. What is an Author. In *Language, Counter-memory, Practice: Selected Essays and Interviews by Michel Foucault*. Cornell University Press.

E. M. Forster. 1927. *Aspects of the novel*. London: E. Arnold.

Gustav Freytag. 1895. *Technique of the drama: An exposition of dramatic composition and art*. S. Griggs.

Jianbo Gao, Matthew L. Jockers, John Laudun, and Timothy Tangherlini. 2016. A multiscale theory for the dynamical evolution of sentiment in novels. In *2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*, pages 1-4, Durham, North Carolina.

Project Gutenberg. n.d. www.gutenberg.org.

Patrick Colm Hogan. 2011. *Affective narratology: The emotional structure of stories*. University of Nebraska Press.

Matthew L. Jockers. 2015. Revealing sentiment and plot arcs with the syuzhet package. www.matthewjockers.net/2015/02/02/syuzhet.

Kaley Joyes. 2008. Failed witnessing in Virginia Woolf's Mrs. Dalloway. *Woolf Studies Annual*, 14: pp. 69–89.

Suzanne Keen. 2011. Introduction: Narrative and the emotions. *Poetics Today*: 32 (1): 1–53.

Hoyeol Kim. 2021. VictorianLit. https://github.com/elibooklover/VictorianLit.

Jochen Kleres. 2011. Emotions and narrative analysis: A methodological approach. *Journal for the theory of social behavior*, 41:182-202.

nlptown. 2020. Bert-base-multilingual-uncased-sentiment. https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment.

Charuta Pethe, Allen Kim, and Steve Skiena. 2020. Chapter captor: Text segmentation in novels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8373–8383, Online. Association for Computational Linguistics.

James Phelan. 1989. *Reading people, reading plots: Character, progression, and the interpretation of narrative*. University of Chicago Press.

Federico Pianzola. 2018. Looking at narrative as a complex system: The proteus principle. *Narrating complexity*, 101–122.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and

Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrew Piper, Sunyam Bagga, Laura Monteiro, Andrew Yang, Marie Labrosse, and Yu Lu Liu. 2021. Detecting narrativity across long time scales. In *Proceedings of the 2021 Computational Humanities Research Conference*, pages 319-332.

Vladimir Iakovlevich Propp. 1968. *Morphology of the folktale*. University of Texas Press

Jonathan Reeve. 2016. Chapterize. https://github.com/JonathanReeve/chapterize.

Brian Richardson. 2000. Linearity and its discontents: Rethinking narrative form and ideological valence. College English, 62(6): 685–95.

Felski, Rita. 2008. *Uses of literature*. Wiley-Blackwell.

Marie-Laurie Ryan. 2005. On the theoretical foundations of transmedial narratology. *Narratology beyond literary criticism*, 6: 1–24.

Tzvetan Todorov. 1971. The 2 principles of narrative. *Diacritics*, 1(1): 37-44.

Ted Underwood. 2015. Free research question about plot. tedunderwood.com/2015/04/01/free-research-question-about-plot.

Ted Underwood. 2013. *Why literary periods mattered: Historical contrast and the prestige of English studies*. Stanford University Press.

Paveen Virameteekul. 2022. Paragraph-level attention based deep model for chapter segmentation.” *PeerJ Computer Science*, 8:e1003.

Shufan Wang and Mohit Iyyer. 2019. Casting light on invisible cities: Computationally engaging with literary criticism. *arXiv preprint arXiv: 1904.08386*.

Julia R. Winkler, Markus Appel, Marie-Luise C.R. Schmidt, and Tobias Richter. 2023. The experience of emotional shifts in narrative persuasion. *Media psychology,* 26(2): 141-171

A Additional Tables

Chapter	Scene/Event	Starting Sentence(s)	Ending Sentence(s)	Chapter length (paragraphs)
1	Clarissa walking towards the flower shop; Septimus' back story; Clarissa returns home to prepare for the party	"Mrs. Dalloway said she would buy the flowers herself."	"'Star-gazing?' said Peter."	119
2	Clarissa lost in memories of Peter; Peter's surprise visit; Peter leaves, follows a young woman, falls asleep in Regent's Park, and dreams about woman figures	"It was like running one's face against a granite wall in the darkness! It was shocking; it was horrible!"	"So the elderly nurse knitted over the sleeping baby in Regent's Park. So Peter Walsh snored."	95
3	Peter links his dream to his memories of Clarissa	"He woke with extreme suddenness, saying to himself, 'The death of the soul.'"	"It was an extraordinary summer… Clarissa in bed with headaches."	19
4	Peter reminisces his parting with Clarissa	"The final scene, the terrible scene which he believed had mattered more than anything in the whole of his life…"	"...when the child ran full tilt into her, fell flat, and burst out crying."	9
5	Child runs towards Rezia in Regent's Park; Peter looks at the couple and thinks about Clarissa's marriage and his own; Septimus' romantic history	"That was comforting rather."	"Could she not read Shakespeare too? Was Shakespeare a difficult author? she asked."	75
6	Septimus' conditions and melancholia worsen	"One cannot bring children into a world like this."	"The verdict of human nature on such a wretch was death."	5
7	Dr. Holmes and Sir William Bradshaw's treatment of Septimus	"**Dr. Holmes** came again."	"But Rezia Warren Smith cried, walking down Harley Street, that she did not like that man."	64
8	Richard's lunch with Lady Bruton & Hugh, returns home, and a quick exchange with Clarissa; Clarissa laments on their distance in marriage, and thinks derogatively about Miss Kilman as Elizabeth leaves with her	"Shredding and slicing, dividing and subdividing, the clocks of Harley Street nibbled at the June day…"	"...upon the body of Miss Kilman standing still in the street for a moment to mutter, 'It is the flesh.'"	105
9	Miss Kilman resents Clarissa as well	"It was the flesh that she must control. Clarissa Dalloway had insulted her."	"...and she chose, in her abstraction, portentously, and the girl serving thought her mad."	8
10	Elizabeth starts to feel overwhelmed around Miss Kilman, and takes the omnibus home; Septimus and Lucrezia's moment of happiness in their apartment as a girl brings their evening papers	"Elizabeth rather wondered, as they did up the parcel, what Miss Kilman was thinking."	"He was very tired. He was very happy. He would sleep. He shut his eyes. But directly he saw nothing the sounds of the game became fainter and stranger and sounded like the cries of people…"	65

Table 6. Full list of predicted chapters (BERT) in *Mrs. Dalloway* (continues to next page). The corresponding narrative arc is displayed in Figure 1.

Chapter	Scene/Event	Starting Sentence(s)	Ending Sentence(s)	Chapter length (paragraphs)
11	Septimus fears the arrival of Holmes and Bradshaw	"He started up in terror."	"But this hat now. And then (it was getting late) **Sir William Bradshaw**."	9
12	Rezia shares a beautiful moment with Septimus before leaving; Holmes arrives the apartment; Septimus commits suicide	"She held her hands to her head, waiting for him to say…"	"'I'll give it you!' he cried, and flung himself vigorously, violently down on to Mrs. Filmer's area railings."	13
13	Guests arriving at the party	"'The coward!' cried **Dr. Holmes**, bursting the door open."	"She could not resist recalling what Charles Darwin had said about her little book on the orchids of Burma."	105
14	Clarissa talking to Lady Bruton about her lunch with Richard	"(Clarissa must speak to Lady Bruton.)"	"(Lady Bruton detested illness in the wives of politicians.)"	5
15	Peter's arrival at the party; Clarissa wants to talk but could not	"'And there's Peter Walsh!' said Lady Bruton"	"... she must go up to **Lady Bradshaw** and say . . ."	8
16	Clarissa hosting the party, then learns about Septimus' death and withdraws	"But **Lady Bradshaw** anticipated her."	"She must assemble. She must find Sally and Peter. And she came in from the little room."	29
17	Peter's conversation with Sally	"But where is Clarissa?' said Peter. He was sitting on the sofa with Sally."	"He made Sally laugh."	40
18	Peter and Sally looking at Elizabeth	"But **Sir William Bradshaw** stopped at the door to look at a picture."	"What is it that fills me with extraordinary excitement?"	5
19	"It is Clarissa, he said. For there she was."	"It is Clarissa, he said."	"For there she was."	2

Table 6 (continue). Full list of predicted chapters (BERT) in *Mrs. Dalloway* (continues from last page). The corresponding narrative arc is displayed in Figure 1.

Novel	Author	Type	Length (Paragraphs)
Mrs. Dalloway	Virginia Woolf	Non-linear	761
The Sound and the Fury	William Faulkner	Non-linear	3176
Swann's Way	Marcel Proust	Non-linear	1392
Good Morning, Midnight	Jean Rhys	Non-linear	1493
Ulysses	James Joyce	Non-linear	7444
Adam Bede	George Eliot	Linear	2563
Great Expectations	Charles Dickens	Linear	3898
Lady Audley's Secret	Elizabeth Braddon	Linear	3285
Little Dorrit	Charles Dickens	Linear	6610
North and South	Eliza Gaskell	Linear	3499
Oliver Twist	Charles Dickens	Linear	3900
Pride and Prejudice	Jane Austen	Linear	2081
The Woman in White	Wilkie Collins	Linear	4214
Vanity Fair	William Makepeace Thackeray	Linear	3432

Table 7. Full list of all novels used in this study