NarraDetect: An annotated dataset for the task of narrative detection

Andrew Piper
Languages, Literatures, and Cultures
McGill University

Sunyam Bagga
School of Computer Science
McGill University

Abstract

Narrative detection is an important task across diverse research domains where storytelling serves as a key mechanism for explaining human beliefs and behavior. However, the task faces three significant challenges: (1) inter-narrative heterogeneity, or the variation in narrative communication across social contexts; (2) intra-narrative heterogeneity, or the dynamic variation of narrative features within a single text over time; and (3) the lack of theoretical consensus regarding the concept of narrative. This paper introduces the NarraDetect dataset, a comprehensive resource comprising over 13,000 passages from 18 distinct narrative and non-narrative genres. Through a manually annotated subset of 400 passages, we also introduce a novel theoretical framework for annotating for a scalar concept of "narrativity." Our findings indicate that while supervised models outperform large language models (LLMs) on this dataset, LLMs exhibit stronger generalization and alignment with the scalar concept of narrativity.

1 Introduction

Narrative detection is an essential task in NLP and the subfield of computational narrative understanding (Bamman et al., 2019; Zhu et al., 2023; Piper, 2023; Antoniak et al., 2023; Abdessamed et al., 2024). A growing body of research is developing across a variety of domains that focus on storytelling as a key mechanism for explaining human beliefs and behavior (Gottschall, 2012). Being able to detect where, when and to what degree the act of narration is taking place among textual outputs will support research into the function of narration across a range of fields.

We see three core challenges facing the task of narrative detection. First is the high degree of variety surrounding the social contexts of storytelling. This is called "situatedness" by Herman (2009) and is one of the four essential elements of narrative in his scheme. Stories can appear in the news media, on social media, in both fiction and non-fiction books, online fan writing sites, scattered throughout large cultural heritage archives, and multi-modally as well (graphic novels, comic books, and children's books) to name a few. While certain components of narrative behavior will likely change across contexts, we also expect some core behavior should remain consistent. We call this the problem of "inter-narrative" heterogeneity.

The second main challenge is what we call "intra-narrative" heterogeneity, i.e. the degree to which narrative communication can differ over narrative time. Narrative practices do not consist of a single, fixed set of behaviors that occur always and everywhere in a story, but rather a dynamic combination of features that may wax and wane.

One of the principal theoretical shifts to occur in the field of narratology over the past few decades has been this shift from understanding narrative as a matter of kind to one of degree (Herman, 2009; Giora and Shen, 1994; Pianzola, 2018). "Narrativity" according to this theoretical framework is a quality that can best be understood not as a global binary class (a document either is or is not a narrative), but as a local, multi-dimensional scalar property (Ochs et al., 2009). A narrative document, such as a novel, may exhibit greater or lesser degrees of narrativity at different moments in the text, just as ostensibly non-narrative documents, such as scientific reports, may also exhibit degrees of narrativity and in different ways.

This stylistic heterogeneity introduces the third challenge facing the task of narrative detection, which is the theoretical heterogeneity underlying the task. The concept of "narrative" consists of a complex set of dimensions and different sources have proposed different frameworks for its study. Not surprisingly, narrative continues to be understood and operationalized in different ways. A key goal for the field moving forward will be the devel-

Proceedings of the The 7th Workshop on Narrative Understanding, pages 1–7

opment of more standardized narrative models.

In this paper, we introduce the *NarraDetect* dataset, which aims to make the following contributions:

1. Address the social diversity of narrative communication by compiling a large collection of over 13,000 passages from 18 different narrative and non-narrative genres. This dataset captures a wide variety of narrative communication from significantly different social contexts.
2. Address intra-narrative diversity by introducing a novel theoretical framework for the annotation of a scalar concept of "narrativity." This framework is then used for the manual annotation of a subset of ca. 400 passages from the large corpus.
3. Validate our data on the task of narrative detection using both supervised and unsupervised models. We show that supervised models outperform LLMs on our data but generalize less well on other data. LLMs also illustrate solid understanding of our scalar concept of narrativity, suggesting good calibration with our theoretical framework.

We make all of our data and annotations available in a long-term repository following the best practices of open science (Collaboration, 2015).[1]

2 Prior Work

The creation of narrative datasets within the field can be divided into two principal areas: the first is the development of domain specific collections of stories or story dimensions for the purposes of narrative understanding. These include news stories (Chambers and Jurafsky, 2008), cultural heritage material (Underwood et al., 2020; Bagga and Piper, 2022; Hamilton and Piper, 2023), novels (Brahman et al., 2021; Iyyer et al., 2016), birth stories (Antoniak et al., 2019), and artificial stories (Mostafazadeh et al., 2016), to name but a few.

Datasets for the task of narrative detection are far more scarce and rely on both positive and negative examples. Antoniak et al. (2023) have created one of the few publicly available narrative detection datasets. The *StorySeeker* corpus consists of an annotated dataset of narratives at the sentence level on a set of 502 Reddit posts and comments drawn from over 100 different subreddits from the Webis-TLDR-17 dataset (Völske et al., 2017). They use a binary model of annotation applied to sentence spans and following Sims et al. (2019) define a narrative as "a sequence of events involving one or more people."

Doyle et al. (2024) have created a collection of 750 manually annotated Reddit posts for the presence of narrative from the *r/SuicideBereavement* subreddit. Following (Smith, 2001), they annotate posts based on the following categories: the presence of a plot, characters, the author as a character, and a clear beginning, middle, and end.

Ganti et al. (2022) annotated a collection of 849 Facebook posts related to the topic of breast cancer for the presence of narratives. In a follow-up study, Ganti et al. (2023) annotated a collection of 3,000 tweets drawn from the *ANTiVax* (Hayawi et al., 2022) and *CMU-MisCov19* (Memon and Carley, 2020) datasets respectively. They annotate tweets for the presence of "narrative style," which they define as: "the presentation of a sequence of events experienced by a character or characters" following (Dahlstrom, 2021).

Narrative detection datasets to date can thus be characterized by the following qualities: narrative has only been operationalized as a binary category; annotation has largely been undertaken with respect to a specific domain (social media); and different theoretical constructs have been used to inform annotation, with events and event sequences being the most predominant category.

3 The NarraDetect Dataset

3.1 Large Corpus: Binary genre-labeled collection

We introduce two corpora to support the task of narrative detection. The first is a large collection of 13,543 text passages drawn from 18 different genres as described in Table A2 in the Appendix. Genres are labeled according to a binary scheme of narrative or non-narrative. Narratives consist of both fictional and non-fictional stories from different social contexts (social media, contemporary publishing, cultural heritage material, and online experimental writing like flash fiction). Non-narrative passages are drawn from a range of informational documents such as Supreme Court decisions, academic articles and abstracts, book reviews, and legal contracts. All passages are randomly sampled from respective documents and consist of five sentences in length.

While this collection has the advantage of size

[1] https://doi.org/10.5683/SP3/HEEEKN

and diversity compared to other manually annotated datasets, it still utilizes a binary conception of narrative. Additionally, because we are sampling passages rather than full documents (to align with our interest in "narrativity") it is possible that passages in the narrative genres may exhibit low-levels of narrativity and vice versa. For this reason, we recommend using the scalar corpus in the next section as the test set. We observe that 6% of passages in the manually annotated scalar corpus are misaligned with their categorical labels, giving users some sense of the possible mislabel rate in the large corpus. Despite these limitations, the large corpus provides researchers with a diverse cross-section of storytelling behavior for the purposes of model training and narrative understanding.

3.2 Scalar Corpus: Human Annotated Collection

As mentioned above, narrative theorists have emphasized the concept of "narrativity" to capture the idea of narrative as one of degree rather than kind. Such a scalar concept is one way of capturing the intra-narrative stylistic diversity that attends narrative communication, though others may be proposed. We develop our annotation framework from one of the foundational handbooks in narrative theory (Herman, 2009). While we do not directly annotate passages over narrative time, our passage-level annotations can be used for estimating changes in narrativity over narrative time.

We utilize the following three categories:

Agency. Narrative is first and foremost language addressing individual experience (Fludernik, 2002). As Herman (2009) writes, "Narrative roots itself in the lived, felt experience of human or human-like agents interacting in an ongoing way with their surrounding environment" (21). Narrativity thus depends on the prominence of a few distinct agents actively experiencing events in the passage.

Event Sequencing. Narrative is about time and process (Ricoeur, 2012). As Herman (2009) writes, "Narrative is a basic human strategy for coming to terms with time, process, and change." One of the principal ways this can occur is through the sequencing of events. Narrativity thus depends on the clarity with which sequences of events are presented.

World Building. Narratives are not just about individuals and events, but as Herman (2009) argues they are also about *lived experience*. Narrativity thus depends on the extent to which an experiencable world is constructed, one that can be clearly seen and felt by the reader.

We trained three undergraduate literature students to code passages using a detailed codebook. After multiple training rounds, they rated each passage on a 5-point Likert scale. In the final round, they annotated 394 passages representing approximately 20 documents per genre.

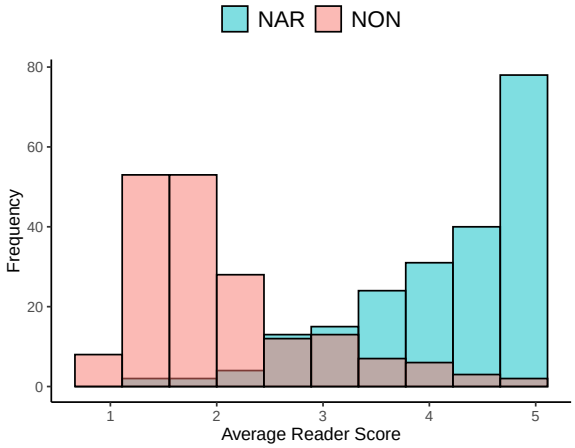

Figure 1: Histogram of average reader narrativity scores across all three categories by positive and negative labels for the scalar corpus.

Figure 1 shows a bimodal distribution of reader scores, clustering below 2 and above 4. Inter-rater agreement, measured using the average deviation index (O'Neill, 2017), yielded a median of 0.37 and a mean of 0.41 (+/- 0.31), indicating strong consistency within half a Likert point. We found no association between narrativity score and agreement levels. Table A3 provides examples of passages rated for high, medium, and low narrativity, while Figure A4 shows the full distribution of reader scores across our three narrative dimensions.

4 Evaluating the NarraDetect Corpus for Narrative Detection

We evaluate the utility of the NarraDetect dataset using both supervised and unsupervised methods. For supervised models, we experiment with two feature representations: (1) a semantically neutral feature space derived from part-of-speech (POS) tags excluding punctuation and (2) contextual embeddings obtained from the BERT large cased model. An SVM with a Gaussian kernel serves as the classifier in both cases.

In order to disentangle narrativity-related fea-

tures from genre-specific signals, we employ an adversarial learning approach. A shared feature extractor, implemented as a feedforward neural network, generates input representations optimized for narrativity classification. The primary narrativity classifier predicts whether a passage is narrative or non-narrative, while an auxiliary genre predictor identifies the passage's genre. A gradient reversal layer between the extractor and genre predictor suppresses genre-specific signals, with a combined loss function balancing narrativity and genre prediction using a trade-off parameter λ. This approach enables the model to learn features capturing narrativity independently of genre.

The adversarial learning process achieves an F1 score of 0.87 / 0.97 for narrativity classification using POS / BERT features, while keeping genre prediction accuracy low at 0.18 / 0.19 on our manually annotated test set. These results demonstrate the model's ability to extract narrativity-relevant features with minimal genre interference.

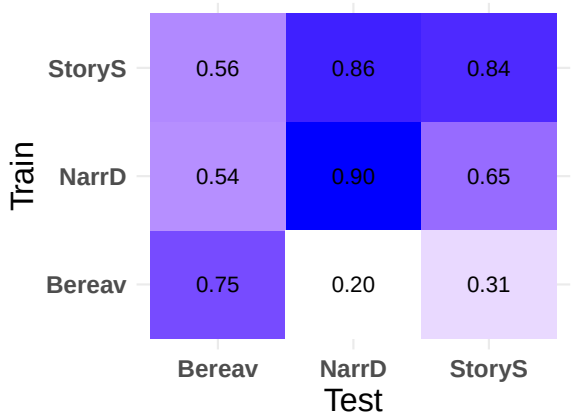

Figure 2: Heatmap of F1 scores using different train and test set combinations using the BERT feature space.

Next we test our data alongside two other datasets discussed in prior work: StorySeeker (Antoniak et al., 2023) and the r/Bereavement data (Doyle et al., 2024). Once again using our SVM classifier and two feature representations, we rotate through all train / test splits and measure F1 scores for each scenario. As shown in Figure 2, there is high within-group accuracy, coupled with considerable decline on out-of-domain data. The exception is the StorySeeker data which generalizes well to our data though the reverse is not the case. The r/Bereavement data shows the lowest generalizability of all sets.

For our unsupervised training, we employ GPT-4 (`gpt-4-0613`) as our frontier model and Llama3.1:8B as our open-weight model. We use a zero shot prompting framework: "Is this passage from a story? Answer only with a number: 1 if yes, 0 if no." For the scalar task, we use similar prompts to what our human annotators received (e.g. "How strongly do you agree with this statement: This passage is organized around sequences of events that occur over time"). Table 1 shows the performance of our two models on the different datasets in both the binary task (F1) and correlation with the scalar task (ρ) as illustrated in Figure A3.

	GPT-4		Llama3.1	
Dataset	F1	ρ	F1	ρ
NarraDetect	0.87	0.81	0.89	0.78
StorySeeker	0.84	-	0.74	-
Bereavement	0.58	-	0.59	-

Table 1: F1 scores for our two candidate LLMs for binary classification and Spearman's ρ for our scalar model comparing LLMs to human annotations.

5 Conclusion

To advance the goal of narrative detection, we introduce the *NarraDetect* dataset, which formalizes "narrativity" theoretically and includes two sub-corpora. The large corpus captures diverse narrative practices across contexts, while the smaller, manually annotated dataset provides a novel scalar framework to address intra-narrative heterogeneity, grounded in foundational narrative theory (Herman, 2009).

Our models achieve high predictive accuracy, though supervised models show performance drops on out-of-domain data, warranting further investigation. Unsupervised LLMs, however, demonstrate robustness across narrative datasets and align well with human annotations, reinforcing the validity of our framework.

We hope *NarraDetect* enriches existing resources and aids in benchmarking LLMs for narrative understanding.

Limitations

Despite our data being drawn from numerous genres and social situations, the cultural contexts of storytelling are vast. Future work will want to continue to expand the number of situations, genres, and languages to facilitate the benchmarking of narrative detection at broader scales and in more

domains. As noted in the paper, researchers need to use caution in supervised learning scenarios both to control for genre effects and also on the appropriateness of out of domain data for the task.

One further limitation of this project is the limited amount of comparative data. We were only able to surface two other data sets for comparison, one of which appears to be not well aligned with the task of narrative detection given its low performance across models. The field will benefit from the creation of further manually annotated narrative datasets.

Finally, our work on unsupervised approaches was limited to two LLMs. Future work will want to do a cross-model assessment on all available models to assess the trade-offs between size and performance on this task. We also look forward to future iterations that are able to perform multilingual narrative detection.

Acknowledgements

We wish to thank the Social Sciences and Humanities Research Council of Canada (435-2022-089) for funding to support this research.

References

Yosra Abdessamed, Shadi Rezapour, and Steven Wilson. 2024. Identifying narrative content in podcast transcripts. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2631–2643, St. Julian's, Malta. Association for Computational Linguistics.

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*.

Sunyam Bagga and Andrew Piper. 2022. Hathi 1m: Introducing a million page historical prose dataset in english from the hathi trust. *Journal of Open Humanities Data*, 8.

David Bamman, Snigdha Chaturvedi, Elizabeth Clark, Madalina Fiterau, and Mohit Iyyer. 2019. Proceedings of the first workshop on narrative understanding. In *Proceedings of the First Workshop on Narrative Understanding*.

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. "let your characters tell their story": A dataset for character-centric narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1734–1752.

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

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Michael F Dahlstrom. 2021. The narrative truth about scientific misinformation. *Proceedings of the National Academy of Sciences*, 118(15):e1914085117.

Dylan Thomas Doyle, Jay K Ghosh, Reece Suchocki, Brian C Keegan, Stephen Voida, and Jed R Brubaker. 2024. Stories that heal: Characterizing and supporting narrative for suicide bereavement. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 354–366.

Monika Fludernik. 2002. *Towards a 'Natural' Narratology*. Routledge.

Achyutarama Ganti, Eslam Ali Hassan Hussein, Steven Wilson, Zexin Ma, and Xinyan Zhao. 2023. Narrative style and the spread of health misinformation on twitter. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4266–4282.

Achyutarama Ganti, Steven Wilson, Zexin Ma, Xinyan Zhao, and Rong Ma. 2022. Narrative detection and feature analysis in online health communities. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 57–65, Seattle, United States. Association for Computational Linguistics.

Rachel Giora and Yeshayahu Shen. 1994. Degrees of narrativity and strategies of semantic reduction. *Poetics*, 22(6):447–458.

Jonathan Gottschall. 2012. *The storytelling animal: How stories make us human*. Houghton Mifflin Harcourt.

Sil Hamilton and Andrew Piper. 2023. Multihathi: A complete collection of multilingual prose fiction in the hathitrust digital library. *Journal of Open Humanities Data*, 9.

Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public health*, 203:23–30.

David Herman. 2009. *Basic elements of narrative*. John Wiley & Sons.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings*

of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1534–1544.

Shahan Ali Memon and Kathleen M Carley. 2020. Characterizing covid-19 misinformation communities using a novel twitter dataset. *arXiv preprint arXiv:2008.00791*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Elinor Ochs, Lisa Capps, et al. 2009. *Living narrative: Creating lives in everyday storytelling*. Harvard University Press.

Thomas A O'Neill. 2017. An overview of interrater agreement on likert scales for researchers and practitioners. *Frontiers in psychology*, 8:264983.

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

Andrew Piper. 2023. Computational narrative understanding: A big picture analysis. In *Proceedings of the Big Picture Workshop*, pages 28–39.

Paul Ricoeur. 2012. *Time and Narrative, Volume 1*. University of Chicago press.

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, Florence, Italy. Association for Computational Linguistics.

Carlota S Smith. 2001. Discourse modes: aspectual entities and tense interpretation. *Cahiers de grammaire*, 26(1):183–206.

Ted Underwood, Patrick Kimutis, and Jessica Witte. 2020. Noveltm datasets for english-language fiction, 1700-2009. *Journal of Cultural Analytics*, 5(2).

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63.

Lixing Zhu, Runcong Zhao, Lin Gui, and Yulan He. 2023. Are NLP models good at tracing thoughts: An overview of narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10098–10121, Singapore. Association for Computational Linguistics.

Appendix

Narrative	# Docs
Artificial Stories (ROC)	976
AskReddit	971
Biographies	898
Fables	258
Fairy tales	740
Flash fiction	390
Histories	979
Memoirs	935
Novels (19C)	998
Novels (Contemporary)	776
Short Stories	451
Non-narrative	
Academic Articles (Phil)	519
Academic Articles (Lit)	468
Aphorisms	462
Book reviews	776
Contracts	1054
Scientific Abstracts	950
U.S. Supreme Court Decisions	942

Table 2: Table of narrative and non-narrative genres in the Large Corpus.

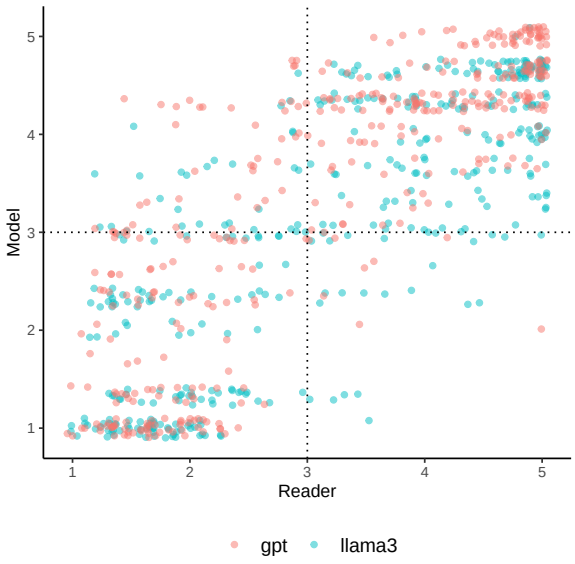

Figure 3: Correlation between average reader scores and GPT and Llama3.1:8B scores on the scalar task.

Score 5.0 / Deviation = 0.0
In the center of the town, the Mercedes stopped a second time, outside a charcuterie and an adjoining boulangerie. Again Keller sped past, but Gabriel managed to conceal himself in the lee of an ancient church. There he watched as the woman climbed out of the car and entered the shops alone, emerging a few minutes later with several plastic sacks filled with food.
Score = 3.0 / Deviation = 0.84
There were other dramatic glitches, too. Despite Cornell's love for the part, she was not suited to it. While Anouilh's Antigone epitomized the enfant terrible, Cornell was in her early fifties and brought to the role a calm, dignified strength, making it harder for the audience to feel that she was imperiled. Photographs of the production reveal her imposing, statuesque presence, precisely the opposite of "la petite maigre" called for by Anouilh.
Score = 1.2 / Deviation = 0.38
To understand a thing is to discover how it operates. The eternal forms of things are laws of natural action. Such are the law of gravitation, the laws of optics or of chemical combination. A static picture unless so interpreted must be at once valueless and meaningless. It follows that Thought and Discourse, in furnishing us with Knowledge, must themselves be active, and must in some way or other reproduce the activity of Nature.

Table 3: Examples of passages with high, medium, and low ratings for narrativity.

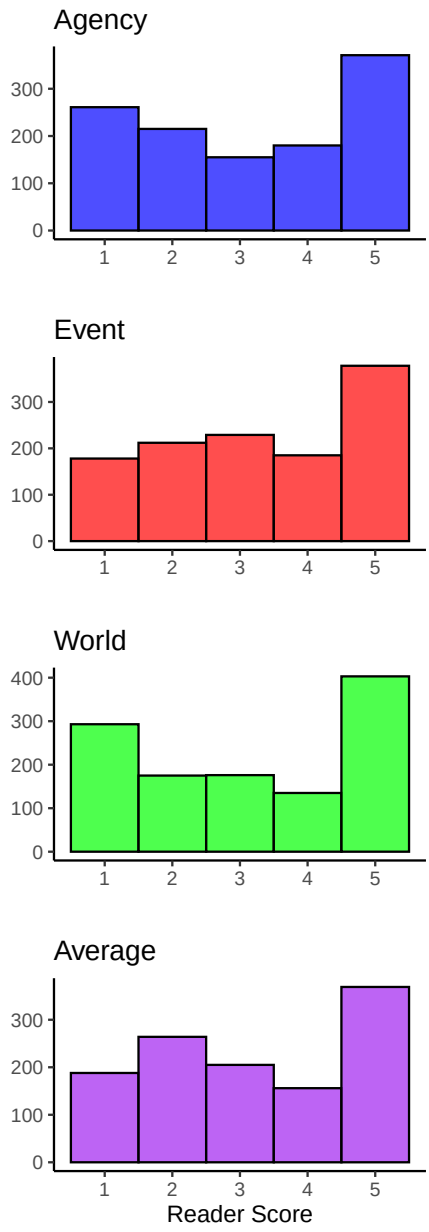

Figure 4: Distribution of reader scores across our three primary narrativity dimensions along with the average of all scores.