Where Do People Tell Stories Online? Story Detection Across Online Communities

Maria Antoniak[♣] **Joel Mire**[♢] **Maarten Sap**[♢♣] **Elliott Ash**[♠] **Andrew Piper**[♡]

[♣]Allen Institute for AI [♢]Carnegie Mellon University [♠]ETH Zürich [♡]McGill University

Abstract

Story detection in online communities is a challenging task as stories are scattered across communities and interwoven with non-storytelling spans within a single text. We address this challenge by building and releasing the 🔭 StorySeeker toolkit, including a richly annotated dataset of 502 Reddit posts and comments, a detailed codebook adapted to the social media context, and models to predict storytelling at the document and span levels. Our dataset is sampled from hundreds of popular English-language Reddit communities ranging across 33 topic categories, and it contains fine-grained expert annotations, including binary story labels, story spans, and event spans. We evaluate a range of detection methods using our data, and we identify the distinctive textual features of online storytelling, focusing on storytelling spans, which we introduce as a new task. We illuminate distributional characteristics of storytelling on a large community-centric social media platform, and we also conduct a case study on *r/ChangeMyView*, where storytelling is used as one of many persuasive strategies, illustrating that our data and models can be used for both inter- and intra-community research. Finally, we discuss implications of our tools and analyses for narratology and the study of online communities.

1 Introduction

Automatic detection of stories online is an important but complex task, as online stories can be topically varied and embedded within a longer document. For example, in Table 1, stories can make up only a small portion of an advice-seeking post, which at first glance may not look like it contains a story. Successful storytelling detection could enable large-scale analyses of storytelling patterns across online communities, deepening our understanding of the social functions of stories.

Several challenges have hindered the large scale analysis of storytelling online. First, defining storytelling, i.e., what is a story and what is not a story, is a difficult task that the field of narratology has been concerned with for decades (Bal and Van Boheemen, 2009). Second, unlike traditional media where titles, genres, and other paratextual cues can signal its presence, storytelling in online communities can be fluid and co-occur with many other kinds of discourse (such as seeking information, providing advice; Yang et al., 2019a). Third, existing story datasets are not appropriate for training and evaluating a cross-community story detector; some datasets are not publicly available (Ganti et al., 2022, 2023), others contain sparse positive labels (Gordon and Swanson, 2009), some are not available in English (dos Santos et al., 2017), some include topical confounders (Prize, 2019), and some are largely related to book-based narratives (Piper and Bagga, 2022). Finally, no storytelling detection work has tackled detecting stories embedded within documents.

The mods removed my post last week, very frustrating. Anyway, my major is in Information Science and I'm entering my senior year. [I began school in CS, but then I switched to the iSchool because I discovered that the topics were more interesting for me.] I know I shouldn't worry about this, but I feel like my IS degree could hurt my chances of getting into a CS graduate program. I thought you all might have input about my options.

Table 1: A motivating example that shows event and [story] spans and illustrates the difficulty of determining story boundaries and event sequences.

To bridge this gap, we formalize the task of story detection by publicly releasing[1] 🔭 StorySeeker: the first dataset, codebook, and models for detecting stories and story boundaries across diverse online communities. In doing so, (1) we release a public, detailed codebook that can be used across a variety of data types, (2) we explicitly tie our story definition to the related task of *event detec-*

[1] https://github.com/maria-antoniak/storyseeker

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7104–7130

Text	Commentary
Im trying to lose 4kg over 2 months or 0.5kg a week. The caloric intake calculator says i should aim for about 2560 calories per day. It also breaks it down into three categories carbs (gm) protein (gm) and fats (gm), which it says is 333 147 91 respectively. Am i right in assuming that (gm) is grams because ive only ever seen it (g). And if so is this about accurate? If it is it makes it easier for me because all the foods here say on the label how many grams of these 3 macros are in the food. I understand this might be a bit basic but [ive searched around the internet and here but havent found an answer though ive probs missed it].	Our codebook allows for very short stories composed of as few as two meaningfully connected events. By annotating these minimal spans of narrativity within larger texts, downstream modelling and application can be more attuned to the ways that stories are intermixed with other discourses. Additionally, we can more precisely analyze which textual features are associated with story spans (vs. a less precise approach that considers textual features across the entire text).
I'm a grad student for context. [It starts out very general, you will spend a year or two really just catching up on literature since, you know, your advisor has been along for the ride for about 15 years so comparatively you're an academic infant. So you read a paper which answers questions, but also creates more questions for you. But that's ok and is par for the course, so you look up more papers which answer those questions, and this is a cycle for your entire career. At some point you will think, "why didn't they do this?" but this time... this wondrous time, you find there is no further information on the topic. But it's a good thing you spent the past two 2 years catching up on it, and now hopefully you can try and answer the question yourself. Now you read others papers to get answers, which leads to more questions. Suddenly it's your own set of questions and answers that drive the field.] I am more typical I think and would say after about 12-18 months I started asking questions that didn't have really well worked out answers, and that's when the real work begins! I think this diagram is a perfect explanation of how it goes.	Although prior work on literary event detection focused solely on realis events (Sims et al., 2019), which are asserted to have actually happened, our codebook recognizes future-tense and hypothetical stories in certain cases, as in this case, where an author's experience informs a story-like sequence of events told in a hypothetical grammatical mode. By accounting for these cases in our codebook, we support a more capacious notion of storytelling, which is not limited to a specific grammatical mode.
Hey all, guess I'll be coming here more often from now on. I finally ordered an Xbox. I also bought an HDMI adapter on Amazon. Anything I should know? I heard about the undervolting, so I might try that.	This example demonstrates that the presence of multiple events does not always imply that there is a story. Stories must contain a *sequence* of events that are meaningfully related.

Table 2: Examples of texts that are difficult to classify as either containing a story or not. Our codebook guides our annotations for these edge cases, dealing with issues such as exceptionally short stories, hypothetical stories, and texts with events that are are not clearly interrelated.

tion (Sims et al., 2019; Vauth et al., 2021) by annotating event spans, (3) our codebook handles token-level boundaries between storytelling and non-storytelling spans, and (4) our data includes large and diverse sets of online communities rather than focusing on a single online setting, furthering our goal of providing a generalizable story annotation framework.

We create the StorySeeker dataset through iterative rounds of open coding, codebook development, and expert annotation, culminating in discussions that assign a consensus label to every text in our dataset, drawing on prior work in narratology (Herman, 2009; Piper et al., 2021b). The final dataset includes expert-annotated Reddit posts and comments with story labels, story boundaries, and event labels. This dataset ranges across hundreds of popular subreddits drawn from 33 broad topic categories, resulting in 235 storytelling documents and 1,739 event-spans over 502 English-language Reddit posts and comments.

StorySeeker opens up new research questions for computational story analysis. While prior work has focused on storytelling in specific communities and topics (e.g., healthcare communities; Ganti et al., 2022, 2023), many larger research questions about online storytelling require the ability to detect stories across large and diverse sets of communities. Prior work has not been able to answer important questions such as where storytelling happens online, which community features lead to more storytelling, or how storytelling is used rhetorically across social contexts — and until now, resources have not existed to support such research.

Our contributions include the following.

- We release the StorySeeker dataset, codebook, and models to detect storytelling documents and text spans.
- Using StorySeeker, we quantify storytelling rates across many online communities for the first time, finding that the prevalence of storytelling varies widely, ranging from low

storytelling rates in religion-focused communities to high storytelling rates in healthcare-focused communities.

- We identify text features that are more frequently present in storytelling spans.
- We map communities by their storytelling rates and the distinctiveness of their stories, providing insights for researchers interested in studying particular community categories.
- We illustrate the effectiveness of our toolkit not only for inter-community analyses but also demonstrate how storytelling can be examined as a rhetorical strategy in a specific community through a case study of *r/ChangeMyView*.

2 Related Work

Narratology Storytelling is a broad concept that has been explored by fields as diverse as economics (Shiller, 2020), literary theory (Bal and Van Boheemen, 2009), sociology (Berger and Quinney, 2004), and NLP (Eisenberg and Finlayson, 2017; Piper et al., 2021b; Ranade et al., 2022). Arriving at agreement on a single story definition is a challenging task.

In the field of narrative theory ("narratology"), storytelling has been defined by its emphasis on sequences of events, aspects of change and/or conflict, and embodiment or "feltness" (Herman, 2009; Bruner, 1991; 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." Piper et al. (2021b) propose a minimum schema for capturing storytelling that emphasizes the presence of ten basic elements, including an agent, action, and location in time and space.

Storytelling datasets Operationalizing storytelling schemas for NLP is an active area of research (Roos and Reccius, 2021; Piper et al., 2021a; Piper and Bagga, 2022). NLP research on narratives and storytelling has deployed a wide range of definitional dimensions, the most common being sequences of events arranged temporally, causally related events leading to resolutions, and the presence of entities or characters, while a smaller number include a rhetorical purpose for the text and world building (Ceran et al., 2012; Yao and Huang, 2018; Eisenberg and Finlayson, 2017; Castricato et al., 2021; Alzahrani et al., 2016). See A.6 for an enumeration of story definition features used in prior work.

Prior work in story annotation has mostly focused on specific discourse domains such as healthcare (Ganti et al., 2022), argumentation (Falk and Lapesa, 2022, 2023), book publishing (Piper and Bagga, 2022), bereavement (Doyle et al., 2024), or blogs (dos Santos et al., 2017; Gordon and Swanson, 2009). To our knowledge, all prior work on story detection has focused on passage- or document-level annotations, except for a set of preliminary annotation guidelines for embedded narratives without an associated dataset by Eisenberg and Finlayson (2021). Additionally, much past annotation work is not publicly available (Ceran et al., 2012; Ganti et al., 2023), due to data agreements, sensitive content, and other constraints.

We expand on these works by widening our view to many communities and topics, by using a span-based approach, and by publicly releasing our annotations. Our annotation guidelines presented in §3 most closely resemble the guidelines provided by Eisenberg and Finlayson (2017) and Eisenberg and Finlayson (2021), which rely on events and characters, and draw most inspiration from the narrative annotation guidelines provided by Piper et al. (2021a) and the event annotation guidelines provided by Sims et al. (2019).

Story detection Most prior work on automatic story detection has focused on using feature-based classification approaches, relying on features like n-grams, POS tags, and coreference chain length (Ceran et al., 2012; Gordon and Swanson, 2009; Yao and Huang, 2018; Eisenberg and Finlayson, 2017; Piper et al., 2021a). See A.6 for an enumeration of story features used for prediction in prior work. These studies either had an explicit goal of using interpretable methods or were conducted prior to the arrival of large language models (LLMs). Two newer works have attempted narrative detection using LLMs (Ganti et al., 2022, 2023). Both studies hand-annotate a small set of texts from online healthcare communities with binary story labels, and both find that fine-tuned BERT-based models perform better than classical models at the document-level story prediction task.

3 Designing a Story Codebook

Our guidelines must work across diverse online communities while recognizing features that make storytelling in these contexts potentially unique

from storytelling in literary or visual contexts.

Our interdisciplinary team worked iteratively to code data, compare labels, and design a codebook containing a story description that fit our context (diverse online communities) and research goal (measuring storytelling). Drawing from prior work, we focused our story identification guidelines on agent-centered events, leading to a two-step annotation process: first, the annotation of event spans, and second, the annotation of story spans. We provide our full codebook instructions in A.7.

Our strong emphasis on events while annotating story spans is novel and adds consistency to the annotation of a diverse set of texts. It also allows us to build on prior work on event span annotation (Sims et al., 2019) by reconsidering this task within the context of storytelling, leading us to a more flexible description of events than prior work.

Importantly, our codebook should be considered not as presenting complete or final definitions of either stories or events but rather providing actionable advice that captures features useful for the consistent annotation of these phenomena.

Story annotation guidelines During our story annotation process, we look for texts containing *a sequence of events involving one or more people.* Notably, as long as a text meets these requirements, we annotate it as a story, regardless of whether it is as short as one sentence or as long as the entire post. Storytelling is thus not limited to only posts that are entirely focused on storytelling. Unlike some prior work, we do not include world-building or setting as features in our guidelines, as such descriptions are rare in our data, and we also do not include the presence of a narrator, as all Reddit posts and comments by default have their authors as narrators. When selecting story spans, we do not include the post title, introductory text about the subreddit, or explanations and discussions external to the story, but we do include non-event text that sets the stage, summarizes the story, or ends with a lesson learned.

Event annotation guidelines Our events guidelines draw heavily from the guidelines in Sims et al. (2019), summarized as *an event is a singular occurrence at a particular place and time*. We modify this guideline in three ways: (1) We do not require verbs to be in the past tense, (2) we sometimes allow hypothetical verbs to be labeled as events, and (3) we sometimes allow verbs with negation to be labeled as events. These changes reflect both the different norms in online communities (e.g., the frequency of using present tense to tell stories), and the story-detection intention underlying our event labeling (e.g., the importance of some negative events in a story sequence). Some changes, like the increased attention to stative verbs, reflect choices also made by Vauth et al. (2021), a follow-up work on literary event detection.

4 Creating a Multi-Community Corpus

We developed the StorySeeker dataset, a collection of Reddit posts that introduces the story span annotation task and covers a large number of online communities. To detect storytelling across diverse settings, we required training and evaluation data that contained both storytelling and non-storytelling communication mixed together by topic, context, and even within a single document.

Data source We sample texts from Webis-TLDR-17, a corpus of 3.8 million Reddit posts and comments (Völske et al., 2017). This dataset was designed for summarization research, and so unlike a random sampling of Reddit data, each text in this dataset contains contentful texts (texts with coherent sentences leading to a summary statement, rather than images, links, or other kinds of texts).

Sampling across the 500 most frequent subreddits in the dataset, we follow an open-coding approach to categorize the subreddits into 33 categories (e.g. *professional advice*). We use these categories to remove sensitive (e.g., *r/confessions*), toxic (e.g., *r/pettyrevenge*), explicit (e.g., *r/sex*), and non-English (e.g., *r/mexico*) subreddits from the human annotation tasks, and we use these categories to structure our analyses (§6). We show the full set of categories and their member subreddits in A.2. We sample a balanced set of five texts from each of the filtered subreddits, downsampling *gaming*[2] and requiring that each text contains at least 100 tokens and no more than 500 tokens.

Annotation process Our annotation process included highlighting both story and event spans, as we found that first identifying events was crucial in making the story span decision. After several rounds of codebook construction, two of the authors independently annotated the target data using Prodigy; each annotator annotated the full set of texts.[3] Annotation was challenging, as texts can

[2]The *gaming* category has twice the number of subreddits of the next most frequent category, *hobbies* (100 versus 49).

[3]https://prodi.gy/

Story Annotation: N=502, Cohen's k=0.66 (binary), 0.72 (span)				
Type	**Story Label**	**# Docs**	**% Docs**	**# Tokens / Span**
post	story	137	64%	119 (mean)
	non-story	78	36%	–
comment	story	98	34%	92 (mean)
	non-story	189	66%	–

Table 3: Overview of the annotated data.

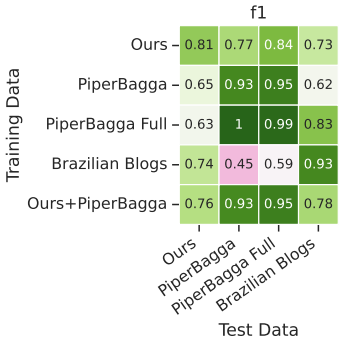

Figure 1: Comparison of document classification performance across datasets using RoBERTa model fine-tuned on our *consensus* labels or the narrative labels in the associated datasets, which were split into 70-30 finetuning and test sets. Cells show F1 scores for the story label.

contain many events, and we included frequent rounds of group deliberation for difficult examples as well as a final round of discussion to arrive at a single consensus label for each document.

Final corpus Our final StorySeeker dataset includes 502 texts with event- and story-spans (see Table 3), sampled randomly from the larger dataset and removing eight toxic texts that escaped our automatic filters. Our inter-annotator agreement for both spans is in the traditional "substantial" range (Cohen's k = 0.65 for event spans, 0.72 for story spans). 47% of the texts included story spans according to the final consensus labels.

5 Developing a Story Detection Model

We release two fine-tuned classifiers as part of StorySeeker to detect storytelling documents and spans across diverse online communities.

Prediction methods To predict the presence of storytelling, we evaluate a variety of prediction models and methods to cover reasonable baselines that vary in cost, GPU usage, and need for labeled data. These methods include a SVMs baseline using TF-IDF features, a fine-tuned RoBERTa[4] model (Liu et al., 2019), and zero-shot and few-shot GPT-4 prompting (OpenAI, 2023).

For the document classification task, we predict the binary presence of storytelling in the document, while for the span detection task, we predict whether each token is part of a story. We divide our expert-annotated data into a training/prompting set of 301 texts, a validation set of 100 texts, and a test set of 101 texts. We rely on the consensus labels, i.e., if the annotators agree after discussion that a text contains a story span, then we use this as either (i) a positive instance of storytelling or (ii) an indicator of whether to include the union of the annotator's spans for token prediction.

We experimented with different prompts using our validation set, including variations of the below task that included examples, chain-of-thought questions, and guidelines.

```
Task: A story describes a
sequence of events involving one
or more people. Does the
following text contain a story?
Answer yes or no, and then explain
your reasoning.
Text: <TEXT>
Answer:
```

For more details and the full set of prompts, see A.8. For few-shot tests with OpenAI models, we interleave two positive and two negative examples in the prompt, using model versions GPT-4-0314 and GPT-3.5-turbo-0613.

Evaluation results We find the best overall story-detection performance from the finetuned RoBERTa model (Table 4), but the GPT-4 prompts are sometimes comparable. GPT-4 performed better than GPT-3.5 and chain-of-thought prompting (Camburu et al., 2018) did not yield consistent improvements. Given our small expert set, these results are averaged over $k = 5$ cross-validation folds, and we show both the means and standard deviations. We examine model errors in A.1.

Cross-dataset prediction performance Overall, StorySeeker performs better on our Reddit data than identical classifiers fine-tuned using

[4]We used the Hugging Face library with the `roberta-base` model for both document and sequence prediction, using `RobertaForSequenceClassification` for document prediction and `RobertaForTokenClassification` for sequence prediction from the `transformers` library.

Model	P	R	F1	P	R	F1
Document Classification	*Story*			***Not** Story*		
SVM with TF-IDF	0.82 ± 0.06	0.69 ± 0.11	0.74 ± 0.08	0.77 ± 0.06	0.87 ± 0.03	0.81 ± 0.03
Fine-tuned RoBERTa	0.87 ± 0.02	0.85 ± 0.08	0.86 ± 0.04	0.88 ± 0.05	0.89 ± 0.03	0.88 ± 0.02
GPT-4 Zero-Shot	0.83 ± 0.02	0.76 ± 0.06	0.79 ± 0.03	0.80 ± 0.04	0.87 ± 0.02	0.83 ± 0.02
GPT-4 Few-Shot	0.84 ± 0.05	0.70 ± 0.06	0.76 ± 0.03	0.77 ± 0.04	0.88 ± 0.04	0.82 ± 0.03
GPT-4 C-o-T	0.56 ± 0.05	0.96 ± 0.03	0.71 ± 0.04	0.91 ± 0.06	0.35 ± 0.02	0.50 ± 0.03
GPT-3.5-Turbo Zero-Shot	0.93 ± 0.02	0.37 ± 0.08	0.53 ± 0.08	0.64 ± 0.04	0.97 ± 0.01	0.77 ± 0.03
GPT-3.5-Turbo Few-Shot	0.79 ± 0.06	0.53 ± 0.10	0.63 ± 0.09	0.68 ± 0.07	0.88 ± 0.03	0.77 ± 0.05
GPT-3.5-Turbo C-o-T	0.56 ± 0.07	0.92 ± 0.03	0.69 ± 0.05	0.85 ± 0.06	0.36 ± 0.07	0.50 ± 0.05
Token Classification	*Story* ($N = 7,756$)			***Not** Story* ($N = 20,477$)		
Fine-tuned RoBERTa	0.77 ± 0.05	0.79 ± 0.08	0.78 ± 0.04	0.90 ± 0.03	0.88 ± 0.03	0.89 ± 0.01
GPT-4 Few-Shot	0.52 ± 0.06	0.86 ± 0.04	0.64 ± 0.05	0.88 ± 0.04	0.57 ± 0.04	0.69 ± 0.04

Table 4: Cross-validation results ($k = 5$) for document and token classification performance across methods, broken apart by the binary labels where the presence of a story is determined using the *consensus* label. For each score, we show the mean and standard deviation across the k folds. We used GPT model versions gpt-4-0314 and gpt-3.5-turbo-0613; see A.8 for GPT prompts.

other story datasets, demonstrating the importance of creating a customized codebook, dataset, and model. Figure 1 shows the results across the story-annotated datasets to which we could reasonably compare. We find that our training data generalizes to these non-Reddit datasets reasonably well, with the lowest performance for the set of Brazilian blog posts (Ceran et al., 2012), a dataset that we automatically translated from Portuguese to English using Google Translate for this experiment. The PiperBagga datasets (Piper and Bagga, 2022) include both a full dataset, with a large number of automatic genre-based labels, and a smaller dataset containing 394 hand-annotated texts mostly from literary sources. Storytelling in the PiperBaggas datasets is strongly correlated with text genre, resulting in models that are overfit when finetuned and tested on the same datasets.

We do not include comparisons to the ICWSM 2009 Spinn3r Dataset (Kevin Burton and Soboroff, 2009; Gordon and Swanson, 2009) because of its very low ratio of stories to non-stories, nor do we include comparisons to the Hewlett Essays (Prize, 2019) because of their strong topical confounders (storytelling is almost perfectly correlated with essay prompt). Other datasets were not accessible for comparison. None of the other datasets include span annotations, and so we cannot draw comparisons about story spans.

Measure	Effect Size (d)	Direction	p-value
expert-annotated events	2.122***	story	$p < 0.001$
past tense	1.742***	story	$p < 0.001$
realis events	1.655***	story	$p < 0.001$
1st-person singular pronouns	1.051***	story	$p < 0.001$
3rd-person singular pronouns	0.459***	story	$p < 0.001$
entity mentions	0.346***	story	$p < 0.001$
concreteness	0.329***	story	0.001
non-past tense	1.296***	non-story	$p < 0.001$
is comment (vs. post)	0.612***	non-story	$p < 0.001$
2nd-person pronouns	0.551***	non-story	$p < 0.001$
sentence length	-	-	0.345
1st-person plural pronouns	–	–	0.345

Table 5: Results of t-tests comparing features between texts labeled as containing stories vs. not containing stories in the StorySeeker dataset. The story group is composed solely by the story spans, as opposed to the entire text labeled as containing a story. We control for multiple comparisons using the Holm method (***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$).

6 Analysis

6.1 What are the signs of storytelling spans?

Based on prior literature and our iterations of annotation and codebook refinement, we identify a set of features that we expect may be associated with storytelling in social media. These include entity and pronoun rates (Eisenberg and Finlayson, 2017; Piper and Bagga, 2022), events (Hühn, 2009; Gius and Vauth, 2022; Sap et al., 2022; Sims et al., 2019), verb tense, concreteness (Piper and Bagga, 2022; Brysbaert et al., 2013), whether a text is a post or comment, and average sentence length. See A.9 for

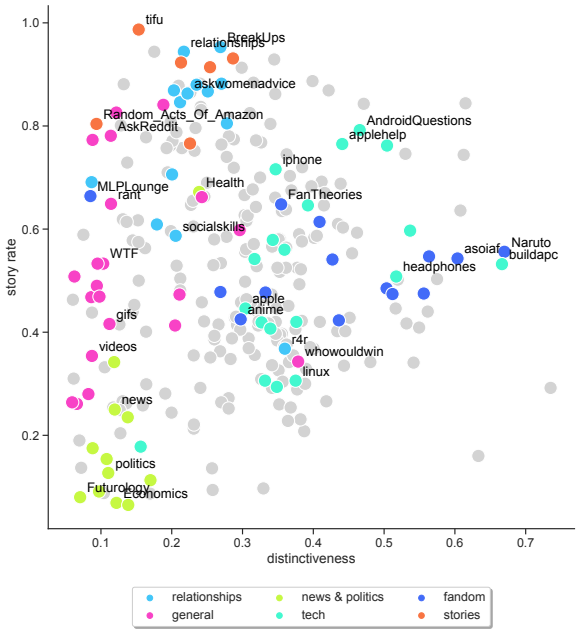

Figure 2: Some subreddits contain stories on specific topics, while others draw a wide range of storytelling topics. We show the 500 most frequent subreddits, colored by category. The y-axis shows the rate of storytelling in the subreddit (predicted by our classifier), and the x-axis shows the distinctiveness of the subreddit's vocabulary (calculated only for texts containing stories). See Figure 7 in the Appendix to see overall category rankings and variation.

context on prior work related to these features, the precise definitions we adopted for this study, and a feature comparison test with another narrative detection dataset composed of mostly literary texts.

Using our consensus labels to compare story spans with non-storytelling texts, we test whether each feature is more prominent in one group than the other. Specifically, we run t-tests on the features, applying the Holm method (Holm, 1979) to account for multiple comparisons (see Table 5). The tests indicate that, in decreasing order of effect size, the frequency of our expert-annotated events, past-tense verbs, *realis* events (Sims et al., 2019), first-person singular pronouns, third-person singular pronouns, entity mentions, and concrete terms are significantly more frequent in stories. Conversely, present- and future-tense verb tenses, the text type being a comment (instead of a post), and second-person pronouns are significantly less frequent in stories.

6.2 Where do people tell stories?

We use the fine-tuned RoBERTa model, which we release as part of StorySeeker, to identify storytelling across a larger set of texts from the Webis-TLDR-17 dataset, sampling a balanced set of 1k texts at random from each subreddit that contains at least that number of texts, resulting in a set or 291 subreddits for prediction. For this experiment, we assign each text a binary prediction for the presence of storytelling.

Using our predicted story labels, we find meaningful differences in the rate at which individuals tell stories across different communities. We find a high of 0.98 stories per all posts and comments (*r/tifu*, i.e., *Today I F*-ed Up*) and a low of 0.11 (*r/Futurology*). Overall, 52% of all the texts were predicted to contain stories, with a macro average across categories of 58%, suggesting that storytelling is indeed a frequent communicative behavior across different kinds of communities.

Figure 7 shows the subreddit categories ranked by their storytelling rates. The *stories*, *addiction*, *animals*, and *healthcare* categories are ranked highest, while *countries*, *news & politics*, *software dev*, and *religion* are ranked lowest; these categories include subreddits whose mean storytelling rate is relatively low. Some categories, e.g., *professional advice*, have wide variation in the storytelling rates of their subreddits. In general, storytelling is more prevalent within communities focused on personal issues related to health and relationships.

When interpreting these results, it is important to keep in mind that our comparisons are only across texts in the Webis-TLDR-17 dataset, i.e. *coherent* texts with summaries. Since many subreddits predominately include photos or other content, our storytelling rates should be interpreted as relative rankings rather than as absolute rates of storytelling. We provide additional results in Figure 5 in A.5 that are not restricted to texts with summaries, validating that general patterns hold when sampling over the full space of subreddit posts.

6.3 How distinctive are stories by community?

Drawing on work on dataset cartography (Swayamdipta et al., 2020), we calculate the *distinctiveness* of the stories shared in different communities to help researchers make decisions about story detection methods for specific subsets of communities. Distinctiveness measures how similar story vocabulary is in comparison to a background vocabulary distribution and has been used in prior work to map Reddit and scholarly communities (Zhang et al., 2017; Lucy et al., 2023). Following Zhang et al. (2017), we first

calculate the specificity S of each word used in a community,

$$S_c(w) = log\frac{P_c(w)}{P_C(w)} \quad (1)$$

where the score compares the probability of each word (w) in a single subreddit (c) versus its probability across all of the subreddits (C). To measure differences in storytelling behavior, we average the specificity scores across the vocabulary, arriving at a single *distinctiveness* score for each subreddit. Importantly, we calculate distinctiveness only for the texts predicted as containing stories because we are interested in the language used in stories, not in the subreddit overall.

In Figure 2, we map communities across two axes: their *story rate*, predicted by our story detector, and the *distinctiveness* of their vocabulary. Categories of subreddits form interpretable clusters. We show 291 subreddits (those matching our filtering criteria, see §5), and Table 9 in A.5 shows the "corners" of this plot, i.e., the subreddits with the most and least storytelling and distinctiveness. For example, subreddits in the *stories* category, such as *r/Glitch_in_the_Matrix*, tend to have both high rates of storytelling and low distinctiveness — these subreddits elicit stories that do not use consistently distinctive language. Some categories contain a wide range, e.g., the *technology* category contains subreddits varying across half the distinctiveness range, from *r/apple* (not at all distinctive) to *r/buildapc* (very distinctive stories). In Figure 6 in A.5, we show the full ranking of subreddit categories, which can aid researchers in determining whether to build a custom detector for their target domain or rely on a mixture of finetuning data to align with the diverse storytelling in some categories.

7 Case Study: *r/ChangeMyView*

We demonstrate StorySeeker's usefulness not only when measuring storytelling across communities but also when measuring storytelling across topics within a single community, by sharing a focused case study of how one subreddit uses storytelling as a rhetorical strategy.

r/ChangeMyView is a forum dedicated to good faith debate, in which posters share opinions, commenters compete to persuade the posters to change their view, and posters assign awards for any successful persuasion. Commenters use various rhetorical strategies, including storytelling, to persuade

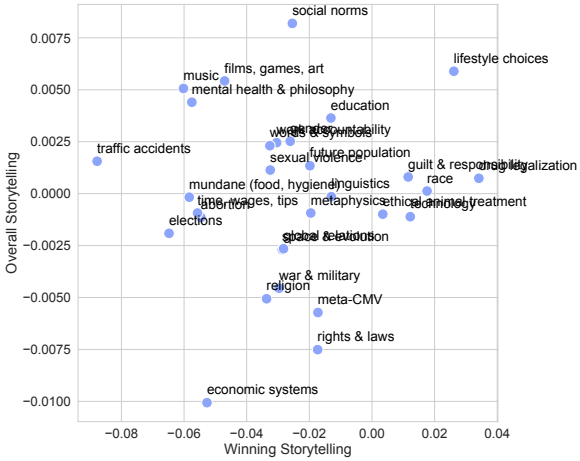

Figure 3: The debate topics plotted by the overall rate of storytelling arguments they receive (y-axis) vs. the rate of storytelling in winning comments (x-axis). Some post topics, like *music*, elicit many storytelling comments but very few of those comments are persuasive, while other topics like *race* elicit fewer stories but responses containing stories more often tend to be persuasive. See Figure 4 in A.5 for detailed views of both axes with standard deviation bounds.

the poster, and this community has frequently been the subject of research about persuasion (Falk and Lapesa, 2023). Using the Winning Arguments Corpus (Tan et al., 2016), we explore how storytelling is used to persuade.

Figure 3 shows the post topics that elicit comments with more or less storytelling and whose storytelling comments are more or less likely to receive a winning "delta" point, indicating that the poster was convinced by the comment's argument. Topics are trained on the posts via a latent Dirichlet allocation (LDA) (Blei et al., 2003) model ($k = 30$), and after examination of the highest probability words and documents, we assign each topic a descriptive name. The y-axis shows the difference between the mean post topic probabilities for storytelling and non-storytelling comments, while the x-axis shows the difference between the proportion of winning comments that include storytelling and the proportion of non-winning comments that include storytelling. More details about our methods and the topic model are given in A.3.

We find that social and personal topics like *lifestyle choices* and *music* receive more storytelling comments, while abstract topics like *economic systems* receive fewer storytelling comments. However, topics that receive more storytelling comments are not necessarily the same topics where storytelling comments are more likely to win delta

points. Examining the outliers in Figure 3, we find that for topics like *drug legalization* and *words & symbols*, storytelling comments are more likely to be persuasive even though commenters are less likely to use storytelling in their arguments. Conversely, topics like *social norms* receive many storytelling comments but those comments are less likely to win. Our results add fine-grained distinctions to prior work that has examined "personal and anecdotal" arguments on *r/ChangeMyView* (Poulsen and DeDeo, 2023).

These patterns likely are driven not only by the topic's subject matter but also the particular framing used in this subreddit; topics like *ethical animal treatment* could be imagined to elicit to many stories, but do not in this community.

8 Discussion

Implications for narratology The features that are most strongly associated with storytelling support prior work's emphasis on storytelling's basis in agent-centered, event-driven forms of communication that are grounded in concrete settings. These features are believed to support social coordination by fostering joint attention around virtual events (known as the "deictic theory" of storytelling; Piper and Bagga, 2022).

Nevertheless, considering narratology's historical focus on long-form literary forms like the novel, Elinor Ochs's famous assertion that the "mundane conversational narratives of personal experience constitute the prototype of narrative activity rather than the flawed byproduct of more artful and planned narrative discourse" highlights the opportunities for narrative theorists to attend to "small stories" (Georgakopoulou, 2007), such as social media stories profiled here. We hope that the 🔭 StorySeeker codebook, dataset, and models can serve as a cross-disciplinary bridge for narrative theorists, computational social scientists, and NLP researchers interested in large-scale analysis of narrativity across social media.

Implications for storytelling in online communities Compared to prior work that detects storytelling in online healthcare communities (Ganti et al., 2022, 2023), our model predicts an overall *lower* rate of storytelling, likely owing to the diversity of topics in our dataset. Our classification performance is lower than the performance reported in those works, despite using similar fine-tuning techniques, signalling the challenge of identifying stories across diverse topics and communities. While storytelling may be a sign of trust as a form of self-disclosure (Ma et al., 2019), we do not find a significant relationship between the overall storytelling rate in a specific community and the same community's toxicity (measured via the Perspective API[5]), size (number of members), or user activity (posts per member). Finally, we provide some additional experiments about overall post and comment patterns in A.4.

9 Conclusion

We formalize the task of story detection and story span detection by releasing 🔭 StorySeeker, a framework that contains a dataset, codebook, and fine-tuned models for online story detection. Our analysis showcases how these tools can be used both across many diverse Reddit communities and to probe a specific community. Using our annotated story spans, we provide the first results indicating how story spans differ from non-storytelling spans cooccurring in the same texts, and we map where people tell stories. We hope our tools and analysis spark further research into online storytelling.

10 Limitations

The Webis-TLDR-17 dataset provided coherent texts across a range of topics, communities, rhetorical goals, and time periods — important qualities for our study — but it also comes with limitations. It only includes texts that include a "TL;DR" summary statement, and it only includes data through 2016. Running our story detection system over a larger dataset would allow us to (a) study chronological patterns in storytelling and (b) study more fine-grained conversational dynamics, given the full post and comment threads contextualizing each target text. However, our initial experiments shown in Figure 5 in the Appendix, which were sampled from the full set of posts in those subreddits, indicates that our rankings are reliable beyond the Webis-TLDR-17 dataset.

In addition, our dataset and analysis are restricted to English-language texts on Reddit. An analysis of cultural patterns in storytelling would be important follow-up work to our study and would require an expansion across different languages. Likewise, analyses of and comparisons across other online forums and social media platforms could help designers in understanding user behavior.

[5] https://perspectiveapi.com/

Our annotated dataset is small, with only 500 annotated texts. However, we emphasize the length of these texts (each text can contain up to 500 tokens) as well as the arduous nature of our annotation task, which involved multiple levels of annotation (both event and story spans), an extensive codebook with many edge cases, the need to use subjective interpretation for the annotation task, and multiple discussions to arrive at high quality consensus labels. Despite these challenges, we achieved inter-annotator reliability scores in what is traditionally understood as the "substantial agreement" for both events and stories, but this required significant time and effort.

Our annotation procedure can include multiple contiguous spans of story spans for a given text. When a story is interrupted by non-story text, we highlight the story spans and do not highlight the interrupting non-story span. We observed many cases like this in the dataset, emphasizing the importance of our span annotations (rather than labeling the entire text with a single binary label). However, we do not capture whether non-contiguous spans are part of the same story. Future work could augment our annotations with such story span coreference information; this would be a valuable addition to our dataset.

11 Ethical Considerations

Online forums like Reddit often contain toxic, explicit, and sensitive text. For example, texts can include calls for violence, ethnic slurs, sexually explicit discussion, and private health information. Depending on their exposure, these texts can harm both their readers (annotators, researchers) and/or their authors, if they did not intend their texts to be shared out of their original context.

While we share Reddit IDs and their corresponding annotations produced in this study, we do not share replications of user IDs, post or comment texts, or other user information. The post and comment IDs can be used to "rehydrate" the document annotations, and we release spans upon request. The 🔭 StorySeeker models can be used without downloading any data.

All of the texts are used in our automatic analysis, but we attempt to remove the most potentially harmful texts from our annotators. After manually categorizing the subreddits (see §4), we filter a set of categories from the annotation task: *banned*, *children*, *confessions*, *mental health*, *NSFW*, *queer culture*, *relationships*, *drugs*, *gender*, *sexuality*, and *toxic*. Our goal is to (1) protect the annotators from toxic content and (2) protect the Reddit users from having their sensitive information used in an annotation task. We also hand-select specific subreddits to filter from across the other categories (a full list is given in our public code repository). We additionally remove 8 texts by hand from our annotated dataset after filtering; these texts were toxic, violent, and/or explicit but were posted in subreddits that we had not filtered. We do not omit this data from our analysis but only from our annotated dataset.

All texts quoted in this article are paraphrased amalgamations of texts in our dataset; this avoids revealing information publicly that was shared in the context of a specific community, and it preserves the ability of Reddit users to edit or delete their texts.

Our study was considered exempt by the IRB at the Allen Institute for AI.

References

Sultan Alzahrani, Betul Ceran, Saud Alashri, Scott W Ruston, Steven R Corman, and Hasan Davulcu. 2016. Story forms detection in text through concept-based co-clustering. In *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom)*, pages 258–265. IEEE.

Mieke Bal and Christine Van Boheemen. 2009. *Narratology: Introduction to the theory of narrative*. University of Toronto Press.

Ronald J Berger and Richard Quinney. 2004. *Storytelling sociology: Narrative as social inquiry*. Lynne Rienner Publishers.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Jerome Bruner. 1991. The narrative construction of reality. *Critical Inquiry*, 18(1):1–21.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Louis Castricato, Spencer Frazier, Jonathan Balloch, and Mark Riedl. 2021. Fabula entropy indexing: Objective measures of story coherence. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 84–94, Virtual. Association for Computational Linguistics.

Betul Ceran, Ravi Karad, Ajay Mandvekar, Steven R Corman, and Hasan Davulcu. 2012. A semantic triplet based story classifier. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 573–580. IEEE.

Henrique DP dos Santos, Vinicius Woloszyn, and Renata Vieira. 2017. Portuguese personal story analysis and detection in blogs. In *Proceedings of the International Conference on Web Intelligence*, pages 709–715.

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. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):354–366.

Joshua Eisenberg and Mark Finlayson. 2017. A simpler and more generalizable story detector using verb and character features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2708–2715, Copenhagen, Denmark. Association for Computational Linguistics.

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

Lisa Capps Elinor Ochs. *Living Narrative: Creating Lives in Everyday Storytelling*. Harvard University Press.

Neele Falk and Gabriella Lapesa. 2022. Reports of personal experiences and stories in argumentation: datasets and analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5530–5553, Dublin, Ireland. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2023. StoryARG: a corpus of narratives and personal experiences in argumentative texts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2350–2372, Toronto, Canada. Association for Computational Linguistics.

Monika Fludernik. 2002. *Towards a "natural" narratology*. Routledge.

Jolene Galegher, Lee Sproull, and Sara Kiesler. 1998. Legitimacy, authority, and community in electronic support groups. *Written communication*, 15(4):493–530.

Achyut Ganti, Eslam 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*, Singapore, Singapore. Association for Computational Linguistics.

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.

Alexandra Georgakopoulou. 2007. *Small stories, interaction and identities*, volume 8. John Benjamins Publishing.

Evelyn Gius and Michael Vauth. 2022. Towards an event based plot model. a computational narratology approach. *Journal of Computational Literary Studies*, 1(1).

Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Third international conference on weblogs and social media, data challenge workshop, San Jose, CA*, volume 46, pages 16–23.

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

Ismail Harrando, Pasquale Lisena, and Raphael Troncy. 2021. Apples to apples: A systematic evaluation of topic models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 483–493, Held Online. INCOMA Ltd.

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

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70.

Alexander Miserlis Hoyle, Pranav Goel, Rupak Sarkar, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peter Hühn. 2009. Event and eventfulness. *Handbook of Narratology*, 19:80.

Adam N Joinson, Ulf-Dietrich Reips, Tom Buchanan, and Carina B Paine Schofield. 2010. Privacy, trust, and self-disclosure online. *Human–Computer Interaction*, 25(1):1–24.

Akshay Java Kevin Burton and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. In *Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*, San Jose, CA. AAAI. <a href=.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Li Lucy, Jesse Dodge, David Bamman, and Katherine Keith. 2023. Words as gatekeepers: Measuring discipline-specific terms and meanings in scholarly publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6929–6947, Toronto, Canada. Association for Computational Linguistics.

Xiao Ma, Justin Cheng, Shankar Iyer, and Mor Naaman. 2019. When do people trust their social groups? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Xiao Ma, Jeffery T Hancock, Kenneth Lim Mingjie, and Mor Naaman. 2017. Self-disclosure and perceived trustworthiness of airbnb host profiles. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 2397–2409.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

OpenAI. 2023. Gpt-4 technical report.

Andrew Piper and Sunyam Bagga. 2022. Toward a data-driven theory of narrativity. *New Literary History*, 54(1):879–901.

Andrew Piper, Sunyam Bagga, Laura Monteiro, Andrew Yang, Marie Labrosse, and Yu Lu Liu. 2021a. Detecting narrativity across long time scales. *Proceedings of the Computational Humanities Workshop*, 1613:0073.

Andrew Piper, Richard Jean So, and David Bamman. 2021b. 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.

Victor Møller Poulsen and Simon DeDeo. 2023. Large language models in the labyrinth: Possibility spaces and moral constraints.

Automated Student Assessment Prize. 2019. The Hewlett Foundation: Automated essay scoring.

Priyanka Ranade, Sanorita Dey, Anupam Joshi, and Tim Finin. 2022. Computational understanding of narratives: A survey. *IEEE Access*, 10:101575–101594.

Michael Roos and Matthias Reccius. 2021. Narratives in economics. *Journal of Economic Surveys*.

Maarten Sap, Anna Jafarpour, Yejin Choi, Noah A. Smith, James W. Pennebaker, and Eric Horvitz. 2022. Quantifying the narrative flow of imagined versus autobiographical stories. *Proceedings of the National Academy of Sciences*, 119(45):e2211715119.

Alexandra Schofield, Måns Magnusson, and David Mimno. 2017a. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 432–436, Valencia, Spain. Association for Computational Linguistics.

Alexandra Schofield and David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300.

Alexandra Schofield, Laure Thompson, and David Mimno. 2017b. Quantifying the effects of text duplication on semantic models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2737–2747, Copenhagen, Denmark. Association for Computational Linguistics.

Robert J Shiller. 2020. *Narrative economics: How stories go viral and drive major economic events*. Princeton University 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.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of WWW*.

Michael Tomasello. 2010. *Origins of human communication*. MIT press.

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

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, Copenhagen, Denmark. Association for Computational Linguistics.

Diyi Yang, Robert E. Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019a. Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA. Association for Computing Machinery.

Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019b. The channel matters: Self-disclosure, reciprocity and social support in online cancer support groups. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Wenlin Yao and Ruihong Huang. 2018. Temporal event knowledge acquisition via identifying narratives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 537–547, Melbourne, Australia. Association for Computational Linguistics.

Justine Zhang, William Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):377–386.

A Appendix

A.1 Error Analysis

For the fine-tuned RoBERTa model, we observe the following categories of errors, using open coding to categorize false positives and false negatives.

Stories misclassified as non-stories sometimes contain cognitive verbs, such as "plan," "decide," or "notice," which we annotated as events, using context to decide whether the verb met our criteria for specificity and sequentiality. This category also includes stories made up of hypothetical verbs (we annotated these only when they were strongly storylike, but their occurrence is rare) and very short stories (one sentence or less).

Non-stories misclassified as stories often contain general or repeating events or describe a state without sequence. These texts often include pronouns, entities, and concrete language like place descriptions, making the texts appear more story-like. These mistakes reflect some of the many edge cases that our codebook was designed to avoid but whose sparsity and ambiguity make it difficult for automatic methods to capture.

A.2 Additional Information About Dataset

Category	Example Subreddits
gaming	*r/leagueoflegends*, *r/gaming*, *r/DotA2*
hobbies	*r/poker*, *r/photography*, *r/MakeupAddiction*
tech	*r/technology*, *r/linux*, *r/AndroidQuestions*
fandom	*r/asoiaf*, *r/doctorwho*, *r/StarWars*
general	*r/AskReddit*, *r/pics*, *r/Showerthoughts*
informative	*r/explainlikeimfive*, *r/math*, *r/RealEstate*
news & politics	*r/PoliticalDiscussion*, *r/changemyview*, *r/Economics*
professional advice	*r/legaladvice*, *r/AskDocs*, *r/graphic_design*
professional sports	*r/nfl*, *r/hockey*, *r/LiverpoolFC*
relationships	*r/relationships*, *dating_advice*, *r/Parenting*
fitness	*r/running*, *r/bodybuilding*, *r/keto*
religion	*r/atheism*, *r/Christianity*, *r/DebateReligion*
stories	*r/tifu*, *r/TalesFromRetail*, *r/Dreams*
cities	*r/toronto*, *r/Seattle*, *r/LosAngeles*
countries	*r/canada*, *r/japan*
drugs	*r/Drugs*, *r/LSD*, *r/opiates*
mental health	*r/depression*, *r/ADHD*, *r/socialanxiety*
queer culture	*r/lgbt*, *r/ainbow*, *r/bisexual*
gender	*r/TwoXChromosomes*, *r/AskMen*, *r/AskWomen*
self help	*r/introvert*, *r/GetMotivated*, *r/INTP*
software dev	*r/programming*, *r/cscareerquestions*, *r/gamedev*
confessions	*r/offmychest*, *r/DoesAnybodyElse*, *r/confession*
finance	*r/personalfinance*, *r/Bitcoin*, *r/investing*
academic	*r/EngineeringStudents*, *r/college*, *r/GradSchool*
addiction	*stopdrinking*, *stopsmoking*, *cripplingalcoholism*
animals	*r/Pets*, *r/Dogtraining*, *r/cats*
healthcare	*r/BabyBumps*, *r/diabetes*, *r/SkincareAddiction*
other	banned subreddits, subreddits about children, NSFW and toxic subreddits, miscellaneous

Table 6: The subreddit categories and examples of member subreddits. These categories were developed via an open-coding approach followed by consolidation. Categories are shown in order of descending frequency by number of member subreddits.

A.3 Stories, Toics, and Winning Arguments in *r/ChangeMyView*

We trained a latent Dirichlet allocation (LDA) topic model (Blei et al., 2003) on 3,046 posts from *r/ChangeMyview* distributed as part of the Winning Arguments Corpus (Tan et al., 2016). While newer topic modeling methods like BERTopic have become popular (Grootendorst, 2022), LDA's performance

on human coherence evaluation tests is still very strong if not stronger (Harrando et al., 2021; Hoyle et al., 2022). We removed punctuation, normalized numbers, lower-cased the text, removed duplicate documents (Schofield et al., 2017b), and did not stem or remove stop words (Schofield and Mimno, 2016; Schofield et al., 2017a). After training we examined the highest probability words and documents for each topic to qualitatively assign a label to each topic. We show the final set of 30 topics in Table 7.

We calculate two metrics to rank the post topics, shown by the bar plots in Figure 4. In each case we are ranking post topics by a measurement on comments responding to post topics. The first metric measures *overall storytelling*; we calculate the mean topic probability for all posts that storytelling comments respond to, and we subtract the the mean topic probability for all posts that non-storytelling comments respond to. The second metric measures *winning storytelling*; for each topic, we find all the posts topic probability over 0.1, and we find the proportion of storytelling comments for that post that win a delta point and subtract the proportion of non-storytelling comments that win a delta point.

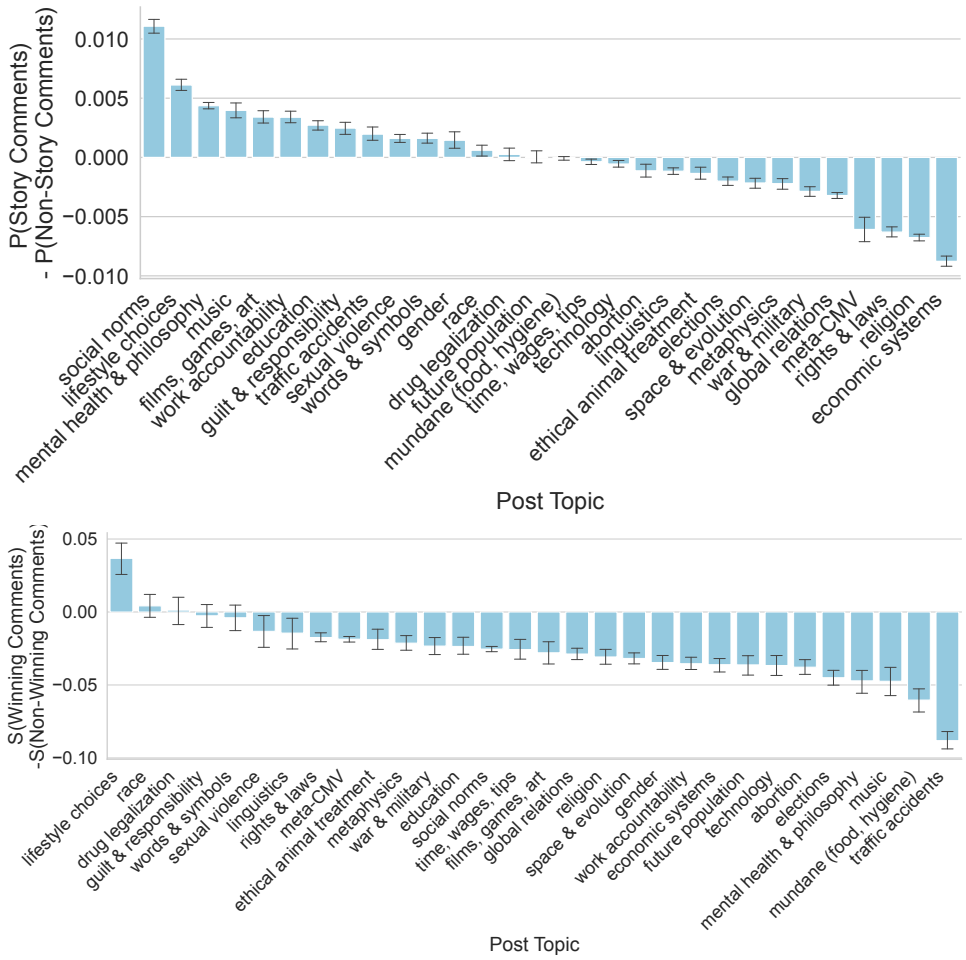

Figure 4: The first plot shows the difference in mean topic probabilities between story and non-story comments for each post topic; higher bars indicate more overall storytelling for that topic. The second plot shows the difference in proportions of winning (delta-awarded) comments containing stories and losing comments containing stories; higher bars indicate more winning comments containing storytelling.

Topic	Label	Highest Probability Tokens	Example Post Text
0	traffic accidents	car, driving, cars, NUM, drive, traffic, stop, risk, dog, seat	Take a look at these statistics from Wikipedia: > According to the U.S. Natio...
1	drug legalization	death, drugs, crime, society, health, alcohol, drug, illegal, consent, life	The official death toll in the Mexican drug war states that sixty thousand peopl...
2	economic systems	money, would, pay, people, government, work, tax, income, business, job	## Section I: Why is Basic Income Increasingly Popular? "Basic income" is a poli...
3	war & military	war, us, country, states, government, military, gun, united, guns, countries	Let me be as clear as possible: I'm talking only about the Army, the main branch...
4	time, wages, tips	NUM, time, year, sports, number, hour, every, team, three, football	Here is an outline of my reasoning with MS paint diagrams: http://i.imgur.com/nY...
5	ethical animal treatment	animals, food, meat, eat, eating, animal, humans, vegetarian, healthy, diet	Given that livestock consume more water, feed, create more greenhouse gasses, an...
6	mundane (food, hygiene)	power, water, would, use, workers, gt, demand, paper, could, employers	Benn Wyatt made several good points about the virtues of calzones. For those tha...
7	linguistics	culture, use, like, term, language, way, change, often, people, words	Some cases in point: - The US/English pronunciation of the name Rothschild as "R...
8	education	school, college, education, students, job, schools, work, student, high, learn	Note: Although I do hold this opinion, it feels prejudiced to me so I am genuine...
9	guilt & responsibility	people, person, someone, bad, think, kill, even, self, good, lives	Batman's strict no-killing policy has lead to the deaths of hundreds of people. ...
10	mental health & philosophy	like, reddit, think, read, much, hate, seems, pretty, etc, mean	When discussing people in altered states, including those brought about through ...
11	rights & laws	right, rights, law, laws, free, gt, freedom, case, legal, argument	I believe the modern Libertarian (as defined by people like Ron Paul) is hypocri...
12	future population	NUM, us, years, world, new, time, gt, better, population, means	*"Will robots inherit the earth? Yes, but they will be our children." - Marvin M...
13	work accountability	people, would, work, go, get, life, time, could, much, day	I understand that child rearing is important, but how can I have equal respect f...
14	films, games, art	game, games, art, play, video, movie, movies, show, character, characters	I believe 3D is a cancer upon the film industry for the following reasons: * 3D ...
15	lifestyle choices	even, many, go, end, ever, less, age, years, real, least	I have been drinking various craft beers for more than 20 years now. In the past...
16	technology	use, ads, phone, buy, used, price, computer, store, internet, apple	You probably get this a lot, but I was thinking about it in the shower today. Ye...
17	abortion	child, children, parents, kids, abortion, life, mother, birth, pro, family	Both the man and woman are equally responsible for an unplanned pregnancy. My re...
18	music	music, great, even, time, like, sound, much, value, many, english	Why do sound technicians ruin so many indoor gigs by turning the whole volume up...
19	sexual violence	rape, victim, information, issues, victims, google, internet, see, reddit, sexual	The argument from miracles is the argument that there have been miracles which a...
20	social norms	like, people, get, want, really, know, see, re, think, would	This is something that's always bothered me, all the way back to when I was a ki...
21	race	white, black, people, race, racism, racist, news, group, media, social	I had a heated ideological debate last night with 2 sociologists and a feminist ...
22	global relations	society, many, state, world, believe, social, major, countries, science, philosophy	It appears to me that when people talk about the rise of China as a global force...
23	metaphysics	one, believe, think, based, evidence, matter, also, understanding, true, know	First of all, the "you" you identify with is probably the summation of biologica...
24	elections	vote, people, political, would, system, voting, party, politics, democracy, politicians	People who have no interest in politics, and are not interested in learning abou...
25	gender	women, men, sex, gender, male, gay, woman, man, female, sexual	Gender, as defined by Wikipedia, is "the range of characteristics pertaining to,...
26	words & symbols	word, use, definition, gt, using, used, logic, said, first, considered	EDIT: Please refrain from making anymore comments about controlled demolitions, ...
27	meta-CMV	view, edit, people, think, would, change, changed, believe, point, post	Lately (especially since the invasion from /r/adviceanimals), there have been a ...
28	space & evolution	would, one, could, life, likely, make, police, environment, might, human	First off, let me clarify, I am not defending the actions of Ferguson Police Off...
29	religion	god, religion, human, religious, believe, moral, good, would, beliefs, belief	The typical Christian resolution to the problem of evil is to state that it is h...

Table 7: The 30 topics derived from a topic model trained the posts in *r/ChangeMyView*. We show the 10 words with highest probability as well as the prefix of an example post text for each topic.

A.4 In what conversational contexts do people tell stories?

Prior theoretical work has emphasized the importance of turn-taking and interaction among actors as a key determinant of narrative behavior (Georgakopoulou, 2007). According to this paradigm, stories depend on the social interactions that elicit and modulate their telling (Herman, 2009). Similarly, prior work on trust in online communities (Galegher et al., 1998) suggests that self-disclosures, such as personal stories, occur more frequently in communities where trust is higher (Yang et al., 2019b; Ma et al., 2017; Joinson et al., 2010), an important metric of community health.

As an initial foray, we find that *posts* are more likely to contain stories than *comments* across most communities, with a mean ratio of storytelling rates in posts versus comments of 2.28. Some subreddits that ranked very low for overall storytelling nevertheless have relatively high rates of storytelling in posts; e.g., *r/askscience* has an overall low storytelling rate (0.03) but ranks highest for storytelling in posts

versus comments (0.56 in posts, 0.09 in comments). Question-asking subreddits can be found at both ends of the ranking (e.g., *r/NoStupidQuestions* has a high ratio of storytelling in posts versus comments while *r/AskReddit* has a low ratio), and via a Pearson correlation test, we do not find a significant correlation between rates of question-asking (i.e., the rate of question mark characters) and storytelling in posts ($p > 0.05$). More results are in Table 8.

Subreddit	**Ratio**	$P(s\|p)$	$P(s\|c)$
*Subreddits with **highest** post:comment storytelling ratio*			
r/askscience	6.35	0.56	0.09
r/philosophy	3.83	0.36	0.10
r/legaladvice	3.58	0.97	0.27
r/NoStupidQuestions	3.47	0.68	0.19
r/summonerschool	3.40	0.62	0.18
r/LeagueofLegendsMeta	3.38	0.41	0.12
r/Bitcoin	3.08	0.53	0.17
r/applehelp	3.07	0.77	0.25
r/techsupport	3.02	0.82	0.27
r/poker	2.88	0.77	0.27
*Subreddits with **lowest** post:comment storytelling ratio*			
r/nfl	1.21	0.50	0.41
r/LifeProTips	1.21	0.62	0.52
r/weddingplanning	1.20	0.92	0.77
r/TalesFromRetail	1.19	0.97	0.82
r/DoesAnybodyElse	1.16	0.77	0.66
r/travel	1.16	0.76	0.66
r/harrypotter	1.16	0.81	0.70
r/AskReddit	1.10	0.88	0.80
r/SquaredCircle	1.10	0.70	0.64
r/Random_Acts_Of_Amazon	1.01	0.90	0.89

Table 8: Ranking of the subreddits by storytelling in posts versus comments. Also shown are the probabilities of storytelling s given either a post p or comment c.

Finally, over a targeted set of eight subreddits expected to have high rates of self-disclosure (*relationships*, *healthcare*) and eight subreddits expected to have low rates of self-disclosure (*technical questions*, *machine learning*), we confirm that the high self-disclosure subreddits also contain more storytelling, according to predictions generated by our fine-tuned RoBERTa model. Results are shown in Figure 5.

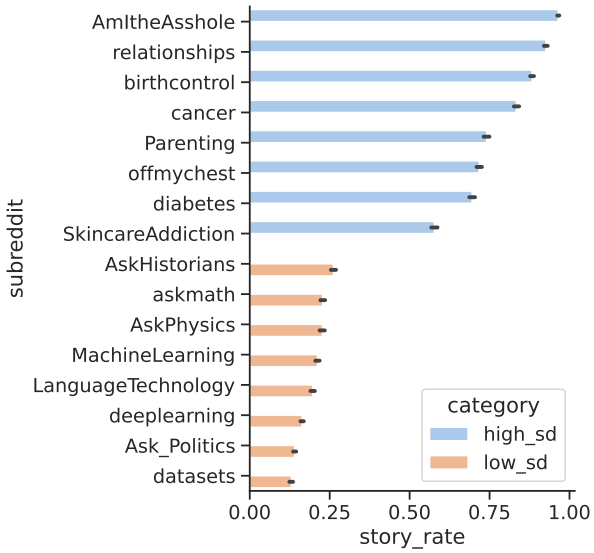

Figure 5: Subreddits with expected *high* or *low* rates of self-disclosure and their predicted storytelling rates, using the fine-tuned RoBERTa model and our *consensus* annotations. As expected, subreddits with higher rates of self-disclosure also have higher rates of storytelling. In this plot, we are showing results for the a set of 500 posts randomly sampled from each subreddit rather than limiting our analysis to *coherent* posts from the Webis-TLDR dataset, providing validation that our other ranksings are consistent.

A.5 Additional Results

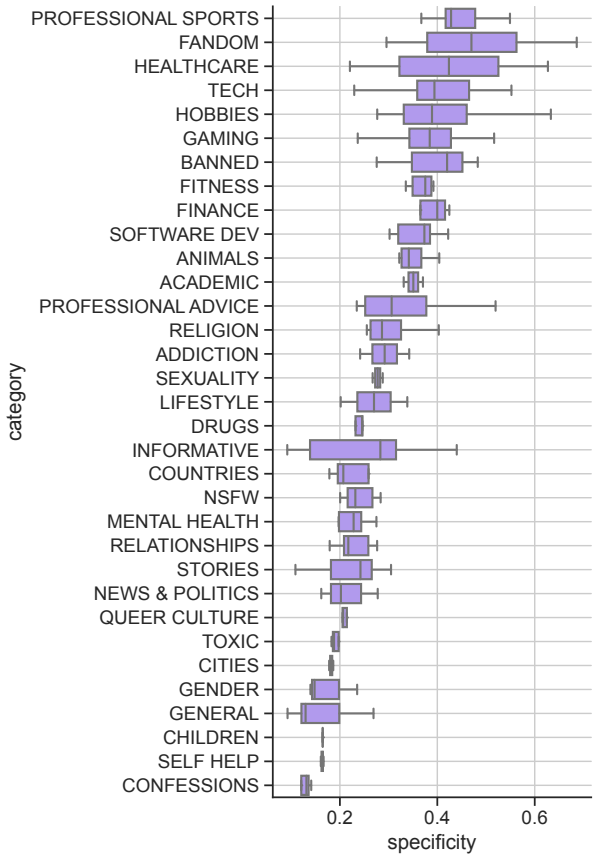

Figure 6: The subreddit categories ranked by their distinctiveness scores.

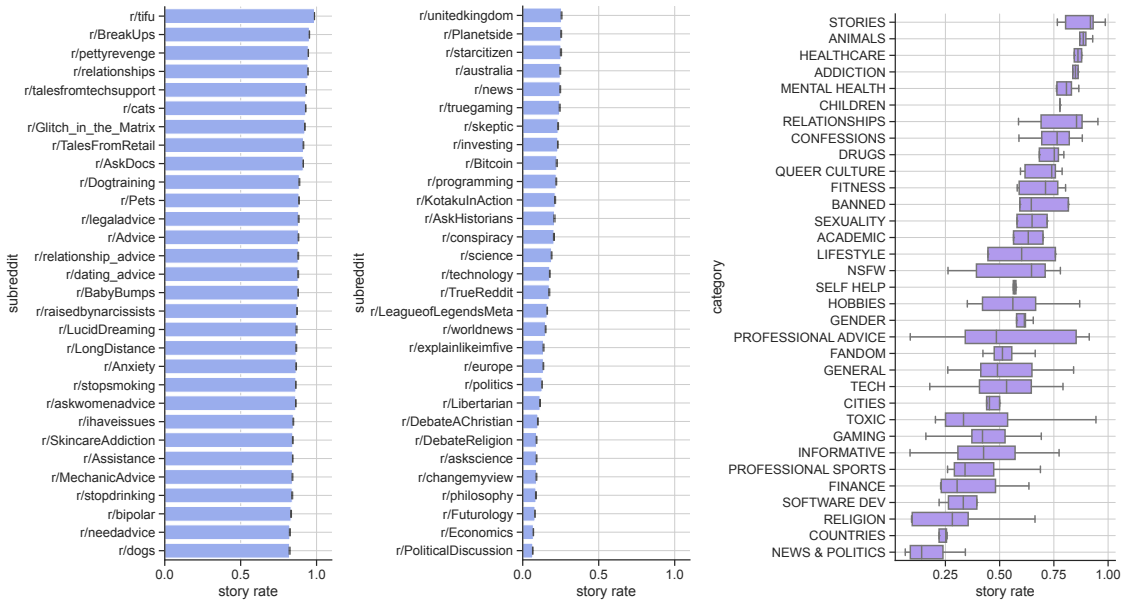

Figure 7: Subreddits (left) and categories of subreddits (right) ranked by their rate of texts (posts and comments) containing stories, as predicted by StorySeeker. Results represent 20 bootstrapped samples of the texts for each subreddit.

	Less Storytelling	More Storytelling
Generic	*r/politics* *r/explainlikeimfive* *r/PoliticalDiscussion* *r/Futurology*	*r/tifu* *r/pettyrevenge* *r/Glitch_in_the_Matrix* *r/Advice*
Distinctive	*r/asoiaf* *r/summonerschool* *r/Naruto* *r/fantasyfootball*	*r/SkincareAddiction* *r/LucidDreaming* *r/techsupport* *r/MechanicAdvice*

Table 9: Subreddits with the most or least storytelling and most or least distinctive vocabulary.

A.6 Prior Definitions of Stories

The following story definitions are drawn from prior work in NLP.

"*A narrative is a discourse presenting a coherent sequence of events which are causally related and purposely related, concern specific characters and times, and overall displays a level of organization beyond the commonsense coherence of the events themselves, such as that provided by a climax or other plot structure.*" – Eisenberg and Finlayson (2021)

"*A narrative is a discourse presenting a coherent sequence of events which are causally related and purposely related, concern specific characters and times, and overall displays a level of organization beyond the commonsense coherence of the events themselves. In sum, a story is a series of events effected by animate actors... at least two key elements to stories, namely, the plot (fabula) and the characters (dramatis personae) who move the plot forward (Abbott, 2008).*" – Eisenberg and Finlayson (2017)

"*It is generally agreed in narratology (Forster, 1962; Mani, 2012; Pentland, 1999; Bal, 2009) that a narrative presents a sequence of events arranged in their time order (the plot) and involving specific characters (the characters).*" – Yao and Huang (2018)

"*A narrative is the recounting of a sequence of events that have a continuant subject and constitute a whole (Prince, 2003).*" – Castricato et al. (2021)

"*A sequence of related events, leading to a resolution or projected resolution.*" – Ceran et al. (2012)

"*Situatedness: narrativity depends on the social context in which it occurs. Event sequencing: narrativity depends on temporally ordered events. World making: narrativity depends on the fact of disequilibrium*

such that we can observe a change in the world." – Piper et al. (2021a)

"Operationalizing Smith (2001)'s characteristics, our codebook had four inclusion criteria: the presence of a plot, characters, the author as a character, and a clear beginning, middle,and end. Additionally, we included three exclusion criteria. Posts were marked as non-narrative that were: purely informational, entirely providing resources, or entirely composed of a question posed to the subreddit community." – Doyle et al. (2024)

Features Used in Definition	**Prior Work**
sequences of events arranged temporally	Piper et al. (2021a) Yao and Huang (2018) Castricato et al. (2021) Doyle et al. (2024)
causally related events leading to resolutions	Eisenberg and Finlayson (2017) Ceran et al. (2012) Alzahrani et al. (2016)
entities or characters	Eisenberg and Finlayson (2017) Piper et al. (2021a) Yao and Huang (2018) Alzahrani et al. (2016) Doyle et al. (2024)
rhetorical purpose	Eisenberg and Finlayson (2017) Roos and Reccius (2021) Castricato et al. (2021)
world building or setting	Piper et al. (2021a)

Table 10: Features used in story **definitions** from prior work.

Features Used for Prediction	**Prior Work**
n-gram	dos Santos et al. (2017) Piper et al. (2021a) Gordon and Swanson (2009) Ceran et al. (2012)
part of speech	Yao and Huang (2018) Piper et al. (2021a) Ceran et al. (2012)
coreference chain length	Eisenberg and Finlayson (2017) Yao and Huang (2018)
LIWC	dos Santos et al. (2017) Yao and Huang (2018)
readability	dos Santos et al. (2017)
verb classes	Eisenberg and Finlayson (2017)
syntactic production rules	Yao and Huang (2018)
verb sequence perplexity	Yao and Huang (2018)
dependency tags	Piper et al. (2021a)
tense	Piper et al. (2021a)
mood	Piper et al. (2021a)
voice	Piper et al. (2021a)
amount of dialog	Piper et al. (2021a)
named entities	Ceran et al. (2012)
stative verbs	Ceran et al. (2012)
semantic triples	Ceran et al. (2012)

Table 11: Examples of features used to **predict** storytelling in prior work.

A.7 Full Codebook

This codebook drew from Sims et al. (2019) and Piper et al. (2021a) with significant modifactions for our online setting, story detection task, and span-based annotation.

Does this text contain a story? *Use the guidelines below to support your decisions, but ultimately, follow your best judgment as there are many edge cases.*

A story describes a sequence of events involving one or more people.

- Stories can be fictional or real, exciting or mundane.

- Focus only on the current text. Don't worry about whether there might be a story before or after this text. References to stories aren't stories.
- Stories describe the experiences of one or more specific people.
  - "People" can include animals, aliens, etc.
  - "People" can include groups as long as these are specific groups of people that exist at a specific time and place.
  - "People" includes the first person narrator.
- Stories must include multiple, specific events.
  - These events should be sequential: one event happens, then another event happens. It's ok if the events are narrated out of order, but there should still be a clear sequence.
  - These events should be connected: they might be about the same people, they might be causally connected, they might describe an overall change or transformation in the state of the world, they might describe a single experience.
  - Jumbles of events that are unordered and/or unconnected (like lists of examples) are not stories.
- What are events?
  - Events are "a singular occurrence at a particular place and time."
  - General, repeating, isolated, or hypothetical situations, states, and actions are usually not events, unless they appear together in a strongly story-like sequence.
  - 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.
    - For example: "I tried to leave the room, but the door wouldn't open."
    - For example: "He asked her to make cookies for his birthday today, but she didn't make him cookies because she doesn't have an oven at home." ("didn't make" is an event but "doesn't have" is not an event)
  - Events are usually verbs but can also be nouns and adjectives.
  - When are states events? See Sims et al. (2019).
- When highlighting any spans:
  - Do not highlight spans in the post title.
- When highlighting the story spans:
  - Include all the text you think is part of the story. This should include not just events but also text that sets the stage, summarizes the story, ends with a lesson learned, etc.
  - Text that usually shouldn't be included in the story span:
    - introductory text about the subreddit, why they're posting, etc.
    - questions about the story
    - explanations, discussion, hypotheses external to the story
  - Ask yourself: Is this text necessary if I were writing a summary of the story?
- When highlighting the event spans:
  - Highlight only one word per event. For example, highlight only "am" in the phrase "am walking slowly".
  - As a rule, never highlight infinitives

– The "ing" form of a verb should usually never be highlighted if it is acting as a noun (e.g. "Demanding forgiveness is unfair"). If it is acting as an adjective (e.g. "His demanding look") it may or may not be highlighted, depending on context. If it is acting as a present participle in an event's verb phrase (e.g. "They are demanding a response"), then highlight "are".

A.8 GPT Prompts

A.8.1 Few-Shot Document Classification

We modify the below prompt based on the setting. For zero-shot, we exclude the "Examples" section. We test two few-shot settings: $k = 2$ and $k = 4$ (where k is the number of examples). In each case, we sample from the training dataset until we have a total of k examples, evenly split between positive (i.e. story) and negative (i.e. not story) instances from the training dataset. Because we use k-fold cross validation, we sample examples from the fold-specific training dataset. We then interleave the negative and positive samples in the "Examples" section.

We also try excluding the Guidelines section. We find that including the Guidelines section improves performance in the zero-shot setting. In contrast, the best performing few-shot setting for each model used $k = 4$ examples and excluded the Guidelines section.

```
Guidelines:

Use the guidelines below to support your decisions, but ultimately,
follow your best judgment as there are many edge cases.

A story describes a sequence of events involving one or more people.

Stories can be fictional or real, exciting or mundane. Stories describe the
experiences of one or more specific people. "People" can include animals,
aliens, etc. "People" can include groups as long as these are specific groups
of people that exist at a specific time and place. "People" includes the
first person narrator.

Stories must include multiple, specific events. These events should be
sequential: one event happens, then another event happens. It's ok if
the events are narrated out of order, but there should still be a clear
sequence. These events should be connected: they might be about the same
people, they might be causally connected, they might describe an overall
change or transformation in the state of the world, they might describe a
single experience. Jumbles of events that are unordered and/or unconnected
(like lists of examples) are not stories.

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. Events are usually verbs but
can also be nouns and adjectives.

Examples:

Text: <TEXT>
Answer: <YES/NO>

Task:
```

A story describes a sequence of events involving one or more people. Does the following text contain a story? Answer yes or no, and then explain your reasoning.

Text:
Answer:

A.8.2 Chain-of-Thought Document Classification

We use the following prompt template to present a simplified breakdown of the classification task into two subtasks: identifying qualifying characters and identifying qualifying events.

Your task is to decide whether a text contains a story. You should follow the guidelines below and think step by step.

Use the guidelines below to support your decisions, but ultimately, follow your best judgment as there are many edge cases.

A story describes a sequence of events involving one or more people.

Stories can be fictional or real, exciting or mundane. Stories describe the experiences of one or more specific people. "People" can include animals, aliens, etc. "People" can include groups as long as these are specific groups of people that exist at a specific time and place. "People" includes the first person narrator.

Stories must include multiple, specific events. These events should be sequential: one event happens, then another event happens. It's ok if the events are narrated out of order, but there should still be a clear sequence. These events should be connected: they might be about the same people, they might be causally connected, they might describe an overall change or transformation in the state of the world, they might describe a single experience. Jumbles of events that are unordered and/or unconnected (like lists of examples) are not stories.

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. Events are usually verbs but can also be nouns and adjectives.

Using the definitions for 'people' and 'events' given in the guidelines above, answer the following questions

Question 1: Does the text contain 'people'? If so, list them.

Question 2: Does the text contain a sequence of causally connected 'events'. If so, list them.

Finally, based on the guidelines and your answers to Question 1 and Question 2, decide whether the text contains a story. Respond 'Yes' or 'No'.

Two examples are provided below:

Example 1: 'Yesterday, I went to the store to buy a jug of milk so that I could make pancakes. When I arrived, the cashier told me that the store lost power the previous night, so all the milk spoiled. So much for pancakes!'

Question 1 Answers: [‘I’, ‘cashier’].
Question 2 Answers: [‘went’, ‘arrived’, ‘told’, ‘lost’, ‘spoiled’].
Story Decision: Yes

Example 2:
‘Not a good FDIC tip. Coverage levels start at $250k and can be increased by certain multipliers–so that’s not really a concern unless you’ve got several million in the bank.’

Question 1 Answers: [].
Question 2 Answers: [].
Story Decision: No

Your turn.
Text: <TEXT>

A.8.3 Few-Shot Story Boundary Detection

Formulating a story boundary detection task for GPT-4 is difficult, due the proclivity for responses to include unrequested mutations to the original source string, which can lead to token misalignment. We mitigate this issue through careful prompt engineering, including manually-constructed few-shot examples that demonstrate desired behavior on different formats observed in our data, and setting the temperature to 0 in OpenAI chat completion requests. The results support the token-level story span detection results (see Table 4). We include the prompt below. Finally, we note that in consideration of the poor performance of this method relative to the Fine-tuned RoBERTa model, in general we do not recommend this approach for for token-level discourse boundary detection tasks.

A story describes a sequence of events involving one or more people. Stories can be fictional or real, exciting or mundane. Stories describe the experiences of one or more specific people. “People” can include animals, aliens, etc. “People” can include groups as long as these are specific groups of people that exist at a specific time and place. “People” includes the first person narrator. Stories must include multiple, specific events. These events should be sequential: one event happens, then another event happens. It’s ok if the events are narrated out of order, but there should still be a clear sequence. These events should be connected: they might be about the same people, they might be causally connected, they might describe an overall change or transformation in the state of the world, they might describe a single experience. Jumbles of events that are unordered and/or unconnected (like lists of examples) are not stories. 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. Events are usually verbs but can also be nouns and adjectives.

Stories may or may not span the entire text. Below are some guidelines for determining what belongs in a story span. 1. Story spans include all of the text you think is part of the story. This should include not just events but also text that sets the stage, summarizes the story, ends with a lesson learned, etc. 2. Parts of texts that usually do not belong to a story span include introductory text about the subreddit, why they’re posting, etc; questions about the story; explanations, discussion, hypotheses external to

the story 3. Text that would be necessary to summarize the story usually belongs in the story span.

The text below may or may not contain one or more stories. If it does contain one or more stories, those stories may or may not span the entire text. Your task is to annotate any story spans in the text. To mark the start of a story span, insert «S». To mark the end of a story, insert «E». If you insert an «S», you must eventually insert an «E». The first marker you insert must always be «S» and you may never insert an «E» unless the last marker you inserted was an «S».

Aside from optionally inserting «S» and «E» markers, you must return the entire original text exactly as it was presented to you. Even if the input text starts with a title in brackets or HTML tags (e.g. ['Title' or 'Title:') or has line breaks, HTML tags, grammatical errors, or random sequences of characters, you must not make unauthorized changes and output the entire original text, including any title. This is the most important rule you must follow no matter what. Think carefully before your respond to make sure you are following my instructions.

Example 1: [Title: Forgot my lunch at home]

Hi everyone, this is my first post here, so please be nice. Anyway, yesterday I forgot my lunch at home. I am a picky eater, so I decided to buy a snack from a vending machine rather than eat the cafeteria food at the office. One of my coworkers made a comment, which I thought was rude. What do you all think?

Example 1 Answer: [Title: Forgot my lunch at home]

Hi everyone, this is my first post here, so please be nice. «S»Anyway, yesterday I forgot my lunch at home. I am a picky eater, so I decided to buy a snack from a vending machine rather than eat the cafeteria food at the office. One of my coworkers made a comment, which I thought was rude.«E» What do you all think?

Example 2: Comment:Does anyone have major political disagreements with a close family member? Thanksgiving is coming up and I'm curious how people in this situation are planning to handle awkward conversations. What do you all think?

Example 2 Answer: Comment:Does anyone have major political disagreements with a close family member? Thanksgiving is coming up and I'm curious how people in this situation are planning to handle awkward conversations.

Example 3: Title: I [24F] am about to fail my Bio class... again. Help. (

Post: Idk what to do AGGHHHH!!!!?$$@#. Help please. I started studying religiously two weeks ago, and I already covered 4 chapters, so maybe I'm on the right track? I talked to the professor, but he didn't offer much help. One TA sent me a list of additional resources, but it's not realistic for me to read all of those in time for the exam.

Example 3 Answer: Title: I [24F] am about to fail my Bio class... again. Help. (

Post: Idk what to do AGGHHHH!!!!?$$@#. Help please. «S»I started studying religiously two weeks ago, and I already covered 4 chapters, so maybe I'm on the right track? I talked to the professor, but he didn't offer much help. One TA sent me a list of additional resources, but it's not realistic for me to read all of those in time for the exam.«E»

Example 4: I went to the mall yesterday and walked into Sephora for the first time in years. I knew it was expensive, but it's out of control. I ended up trying samples for a few minutes, and then caved and bought a lip liner. There went $20. Oh well

Example 4 Answer: «S»I went to the mall yesterday and walked into Sephora for the first time in years. I knew it was expensive, but it's out of control. I ended up trying samples for a few minutes, and then caved and bought a lip liner. There went $20. Oh well«E»

Example 5: Comment:They are terrible to their customers! I truly don't understand. I went their last month and they didn't pay any attention to me until I hunted down the hostess to ask for a table. Then the food took an hour to arrive. And it wasn't even good fwiw.

Example 5 Answer: Comment:«S»They are terrible to their customers! I truly don't understand. I went their last month and they didn't pay any attention to me until I hunted down the hostess to ask for a table. Then the food took an hour to arrive. And it wasn't even good fwiw.«E»

Text: <TEXT>

A.9 Additional Context on Story Feature Analysis

A.9.1 Story Features

Entities Entity and pronoun rates have been used in prior work to both define and detect storytelling (Eisenberg and Finlayson, 2017; Piper and Bagga, 2022). To capture entities, we compute the proportion of several pronoun groups in the texts, including first-person singular, first-person plural, second person, and third-person singular.[6] Additionally, we consider the entity mention rate, defined as the proportion of third-person singular pronouns plus the number of times the spaCy EntityRecognizer detects a *PERSON* entity in the text.

Events Prior work has emphasized eventfulness as a key predictor of storytelling behavior (Hühn, 2009; Gius and Vauth, 2022). We consider event rates in stories based on two event detection methods. First, we use the union of the event labels from the two expert annotators. Second, following Sap et al. (2022), we use a BERT *realis* event tagger trained on a dataset of realis events in literary texts (Sims et al., 2019).[7] These metrics allow us to compare our event definition to past event definitions and how these interact with our story labels.

Verb Tense Previous research has suggested that temporal distance is a key function in establishing the state of joint attentionality (Tomasello, 2010) among narrator and audience members (Piper and Bagga, 2022). To evaluate the importance of verb tense, we sort verbs into two groups: past-tense and not past-tense. Our heuristic for detecting verb tense is based on a partition of the six Penn Treebank (Marcus et al., 1993) verb subtypes employed by the spaCy part-of-speech tagger. We assign *VBD* and *VBN* to the past tense group and all other verb tags to complementary group.

Concreteness Concreteness has been found to be a strong indicator of narrativity in books (Piper and Bagga, 2022). Our concreteness rating for texts is based on the lexicon from Brysbaert et al. (2013). We take the weighted proportion of terms in the text that appear in the lexicon.

Text type is the text's status as post or comment, and **length** is the number of tokens in the text and the mean number of tokens in the sentences.

[6] first-person singular: 'i', 'me', 'my', 'myself', 'mine'; first-person plural: 'we', 'us', 'our', 'ourselves', 'ours'; second-person: 'you', 'your', 'yours', 'yourself', 'yourselves'; third-person singular: 'he', 'she', 'his', 'her', 'him', 'hers', 'himself', 'herself'

[7] We adapted the BERT tagger in https://github.com/maartensap/ACL2019-literary-events. After training, the model achieved F-1 scores of 0.776 and 0.717 on the validation and test sets, respectively.

Measure	Effect Size (d)	Direction	p-value	Effect Size (d)	Direction	p-value
	StorySeeker Dataset			*Piper-Bagga Dataset*		
expert-annotated events	1.899***	story	$p < 0.001$	n/a	n/a	n/a
realis events	1.429***	story	$p < 0.001$	1.687***	story	$p < 0.001$
past tense	1.408***	story	$p < 0.001$	1.207***	story	$p < 0.001$
1st-person singular pronouns	1.009***	story	$p < 0.001$	0.84***	story	$p < 0.001$
concreteness	0.439***	story	$p < 0.001$	1.526***	story	$p < 0.001$
3rd-person singular pronouns	0.397***	story	$p < 0.001$	1.281***	story	$p < 0.001$
entity mentions	0.285**	story	0.006	1.085***	story	$p < 0.001$
non-past tense	0.947***	non-story	$p < 0.001$	–	–	0.639
is comment (vs. post)	0.612***	non-story	$p < 0.001$	n/a	n/a	n/a
2nd-person pronouns	0.444***	non-story	$p < 0.001$	0.285*	story	0.017
sentence length	0.259*	non-story	0.012	0.828***	non-story	$p < 0.001$
1st-person plural pronouns	–	–	0.106	–	–	0.154
text length	–	–	0.106	0.825***	non-story	$p < 0.001$

Table 12: Results of t-tests comparing features between texts labeled as containing stories vs. not containing stories in the StorySeeker dataset and the PiperBagga dataset, which contains mostly literary texts (Piper and Bagga, 2022). Because the PiperBagga dataset does not include story span annotations, we consider the entire text labeled as containing a story for the story group for the StorySeeker tests, for comparison purposes. We control for multiple comparisons using the Holm method (***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$).

A.9.2 Story Feature Comparison Across Datasets

In §6.1, we conduct t-tests comparing features between texts labeled as containing stories vs. not containing stories in the StorySeeker dataset where the story group is composed solely by the story spans, as opposed to the entire text labeled as containing a story. Here, we share complementary test where full texts are used (rather than just story spans) for the story group. We perform the same tests on another narrative detection dataset, which which contains mostly literary texts (Piper and Bagga, 2022).

Our findings confirm many of those found using a hand-annotated story dataset ($N = 394$) from Piper and Bagga (2022), which included mostly literary and non-social media texts. However, some differences emerged; e.g., the rate of non-past-tense is significantly lower in our stories, but there is no significant relationship in the PiperBagga dataset.