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

People share stories online for a myriad of purposes, whether as a means of self-disclosure, processing difficult personal experiences, providing needed information or entertainment, or persuading others to share their beliefs. Better understanding of online storytelling can illuminate the dynamics of social movements, sensemaking practices, persuasion strategies, and more. However, unlike other media such as books and visual content where the narrative nature of the content is often overtly signaled at the document level, studying storytelling in online communities is challenging due to the mixture of storytelling and non-storytelling behavior, which can be interspersed within documents and across diverse topics and settings. We introduce a codebook and create the Storytelling in Online Communities Corpus, an expert-annotated dataset of 502 English-language posts and comments with labeled story and event spans. Using our corpus, we train and evaluate an online story detection model, which we use to investigate the role of storytelling in different social contexts. We identify distinctive features of online storytelling, the prevalence of storytelling among different communities, and the conversational patterns of storytelling.

This is a non-peer-reviewed preprint.

1 Introduction

Storytelling is an important mechanism for human connection and information sharing. Given its ubiquity across time and cultures, it is not surprising that storytelling has come to play an important role in online social communication. While it has been less studied when compared to more traditional book or visual based narratives, studying online storytelling has led to important insights about collective sensemaking practices (Mamykina et al., 2015; Bietti et al., 2019; Antoniak et al., 2019; Young and Miller, 2019), social movements (Gallagher et al., 2019), and political framing (Stammbach et al., 2022), among other topics.

Example Texts Containing Stories
I remembered something that happened to me a few months ago. I left work after a long day at around 9pm and, still wearing a my work clothes, drove to the nearest grocery store to grab something easy for dinner. I was walking up and down different aisles looking for something good when I heard the recognizable sound of a saxophone somewhere in the store...
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.
Hello! Last week I upgraded to a viper after grinding for a long time. Wanting to go back to HQ I found that the jumps were too big, so I thought that I could upgrade my jump drive. Took a while but having finally saved up what I needed, I get to Star One, upgrade my engine. I head back and try to stear towards my destination but it's impossible! Okay, go back to the port and sell all my items. Still can't jump anywhere. I don't know what to do, can any of you help me?

Table 1: Paraphrased examples from the Storytelling in Online Communities Corpus demonstrating our annotation process, showing event and story spans. Examples illustrate the diversity of verb tenses, topics, and story lengths as well as the abundance of first-person singular narration in this data.

One of the biggest challenges facing the study of online storytelling, however, is the lack of explicit markers that indicate when stories begin and end. As we show in Table 1, stories can arguably be as short as a sentence or as long as an entire document. Unlike traditional media where titles and other paratextual cues (e.g. titles, genre) often signal to audiences whether an artifact is explicitly narrative-driven, storytelling in online communities can be fluid and cooccurring with a range of other kinds of discourse (such as seeking information, providing advice; Yang et al., 2019). While

this is also true for long-form narratives, where novels for example can engage in different types of non-narrative discourse such as dialogue, description and meta-level statements, the distinctiveness of online storytelling warrants its own approach.

In this paper, we provide **a codebook for the manual annotation of storytelling spans in online social media data** (§2). Building from prior work in narrative detection (Piper et al., 2021a; Piper and Bagga, 2022), we explicitly tie our story definition to the related task of *event detection* (Sims et al., 2019). Table 1 shows how event spans usually co-occur with story spans. Unlike prior work that has focused on passage- and document-level annotations related to specific domains (Eisenberg and Finlayson, 2017; Ganti et al., 2022; Piper and Bagga, 2022), we focus on identifying sentence-level boundaries between storytelling and non-storytelling behavior across multiple online communities.

Second, we create and make publicly available[1] **the Storytelling in Online Communities Corpus (SOC Corpus), an annotated dataset of diverse Reddit posts and comments** that includes expert-annotated story and event spans (§3). This dataset ranges across 300 popular subreddits and many different topics, resulting in 273 story-spans and 1,739 event-spans over 502 Reddit posts and comments in English. Using this data, we analyze which features occur significantly more in story versus non-story texts (§4). We find strong confirmation of the "deictic theory" of narrative that emphasizes storytelling's rootedness in linguistic expressions of agency, concretization, eventfulness, and temporal distantiation (Piper and Bagga, 2022), and we also find patterns specific to online storytelling, such as the predominance of first-person narratives. We show that this high-quality dataset is useful for automatic story detection in new data (§5), with significantly better performance attained by a fine-tuned model than zero-shot or even few-shot prompting.

Using our story detection model, we then investigate three research questions (§6) about storytelling across online communities: (1) Where (in what kinds of online communities) do people tell stories? (2) How distinctive is storytelling language by community? (3) In what conversational contexts to do people tell stories?

[1] https://github.com/maria-antoniak/stories-online-communities

We highlight that **the prevalence of storytelling in contentful texts differs strongly among different online communities** (where contentful texts have coherent sentences leading to a summary statement), ranging from low storytelling communities (in categories like religion, news and politics) to high (in categories like healthcare, mental health, and relationships and in explicit story-telling subreddits). We also look at conversational patterns of storytelling (Georgakopoulou, 2007), specifically the relationship of storytelling to post-comment flow. We find that storytelling is far more likely to appear in posts rather than comments, suggesting that online storytelling appears to be more of an initiatory form of communication than a responsive one, though there are strong community-based differences here as well.

Overall, our work constructs annotative and analytical frameworks that highlight the promise of in-domain expert-annotated datasets to study storytelling not for a single context but across online communities with diverse topics and members.

2 Designing a Story Codebook

Storytelling is a broad concept that has been explored by fields as diverse as economics (Shiller, 2020), literary theory (Bal and Van Boheemen, 2009), and NLP (Eisenberg and Finlayson, 2017; Piper et al., 2021b; Ranade et al., 2022). Arriving at agreement on a single story definitions is a challenging task. We briefly summarize prior work on constructing story definitions before introducing our codebook, which was designed for annotating statements from online communities. Our definition needs to work across diverse sets of online communities while recognizing features that make storytelling in these contexts potentially unique from storytelling in literary or visual contexts.

2.1 Background

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: a) teller, b) audi-

ence, c) mode of telling, d) social context, e) agent, f) action, g) object, f) location, g) time-frame, and h) rationale.

NLP research on narratives and storytelling has deployed a wide range of definitions. Many different definitional dimensions have been proposed, the most common being requirements for the presence of sequences of events arranged temporally, causally related events leading to resolutions, and entities or characters, while a smaller number require 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; Roos and Reccius, 2021; Piper et al., 2021a; Piper and Bagga, 2022).

2.2 Codebook Definitions

Our research team worked iteratively to code data, compare labels, and design a codebook containing story and event definitions that fit our context (diverse online communities) and research goal (measuring storytelling). Drawing from this prior work, we focus our story definition on agent-centered events with a recognizable narrator.

Our task is two-fold: first, the annotation of event spans, and second, the annotation of story spans. Our strong emphasis on events while annotating story spans is novel (to our knowledge, no prior work has annotated both of these spans at once) 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 by reconsidering this task within the context of the story annotation task, leading us to make context-based updates to previous event definitions.

Story Definition *A story describes a sequence of events involving one or more people as told by an individual.* 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 (see examples in Table 1). A story can thus appear as a portion of a larger passage and is not limited to only those posts that are entirely focused on storytelling. Unlike some prior work, we do not include world-building or setting in our definition, as we found that such descriptions are rarely included in the Reddit data.

Event Definition Our events definition draws heavily from the event definition in Sims et al. (2019), summarized as *an event is a singular occurrence at a particular place and time*. We modify this definition 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 negative verbs to be labeled as events. These changes reflect both the different norms on Reddit (e.g., using present or future conditional tenses to tell stories), as opposed to literature, and the story-detection intention underlying our event labeling (e.g., the importance of some negative events in a story sequence). We provide examples of our annotations in Table 1.

3 Creating a Multi-Community Corpus

We develop a new story dataset that covers a large number of online communities: the Storytelling in Online Communities Corpus (SOC Corpus). To detect storytelling across diverse online communities, we require training and evaluation data that span a variety of topics and settings and that contain narrative and non-narrative communication mixed together by topic, context, and even within a single document. Prior work in story annotation has either focused on specific discourse domains such as healthcare (Ganti et al., 2022), book publishing (Piper and Bagga, 2022), or blogs (Eisenberg and Finlayson, 2017) and has largely focused on passage- or document-level annotations. Additionally, much past annotation work is not publicly available. We expand on these works by widening our view to many communities and topics and by using a span-based approach.

Here we describe the process of building the SOC Corpus, which we make publicly available for further study.

3.1 Data Source

We use Webis-TLDR-17, a dataset 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).

We take the 500 most frequent subreddits in the dataset and follow an open-coding approach to categorize the subreddits into 33 categories (e.g. *professional advice*, *news & politics*, *sports*). Some categories contain subreddits on specific topics while others contain subreddits on a wide range of topics (e.g., *hobbies*). We use these categories

Story Annotation (N=502, Annotators=2, Cohen's k=0.66)			
Type	**Story Label**	**# Texts**	**% Texts**
post	story	150	70%
	non-story	65	30%
comment	story	123	43%
	non-story	164	57%

Event Annotation (N=502, Annotators=2, Cohen's k=0.65)		
Story Label	**Event Label**	**Mean # Events / Text**
story	event	2.10
non-story	non-event	0.13

Table 2: Overview of the annotated data.

to remove very 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 analysis of storytelling across communities (§6). We sample a balanced set of five texts from each of the filtered subreddits, downsampling the largest category, *gaming*,[2] and requiring that each text contain at least 100 tokens and no more than 500 tokens.

3.2 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. Annotation is arduous, as texts can be long and contain many events, and we included frequent rounds of group deliberation for difficult examples. We used Prodigy[3] as the interface for our span-based annotations.

3.3 Final Corpus

Our final dataset includes 502 texts with event- and story-spans, and our inter-annotator agreement for both spans is in the traditional "substantial" range (Cohen's k = 0.65 for events, 0.66 for events). 54% of the texts were tagged as stories by at least one annotator, and the texts contain of mean of one event per text, with a maximum rate of eight events per 100 words. Table 2 provides more details. The value of our data is that with the focus on spans we can better understand story boundaries, while the focus on diverse social media posts allows us to classify more ambiguous storytelling practices such as hypothetical and future tense stories. We make our full codebook and dataset publicly available.

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

[3]https://prodi.gy/

Measure	**Effect Size (d)**	**Direction**	**p-value**
union of expert-annotated events	1.857***	story	$p < 0.001$
realis events	1.466***	story	$p < 0.001$
past tense	1.406***	story	$p < 0.001$
first-person singular pronouns	1.03***	story	$p < 0.001$
concreteness	0.347**	story	0.002
third-person singular pronouns	0.324**	story	0.002
non-past tense	0.887***	non-story	$p < 0.001$
second-person pronouns	0.625***	non-story	$p < 0.001$
is comment (vs. post)	0.573***	non-story	$p < 0.001$
first-person plural pronouns	–	–	0.778
entity mentions	–	–	0.049
sequentiality	–	–	0.778
post length	–	–	0.778
sentence length	–	–	0.115

Table 3: Table with t-tests showing effect sizes of features that correlate with storytelling or not storytelling. Summary of differences between stories and non-stories, according to proposed measures. We indicate significance after controlling for multiple comparisons using the Holm method (***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$).

4 What are the tell-tale signs of online social storytelling?

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. To test our hypotheses, we first sort manually annotated posts into two groups ('story' or 'not story') based on the union of the expert annotators' labels. Between these two groups, we then test whether each feature is more prominent in one group than the other.

4.1 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. 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).[4] These metrics allow us to compare our new 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. As in prior work, we expect past tense verbs to be a sign of storytelling, but as we observed in our updates to the events definition, this dataset also includes present-tense and other kinds of verbs in our annotated story spans.

Flow Sap et al. (2022) introduce *sequentiality* to measure narrative flow. It measures the difference between the log-likelihood of a text conditioned on (1) a topic and preceding context sentences versus (2) the topic only. Sequentiality is scored sentence-by-sentence and the overall sequentiality for a post is the average sequentiality score of the post's sentences. In our setup, we query GPT-4 [5] for short 1-2 sentence summaries of the posts to serve as topics. Additionally, for (1), we use the full context history (i.e. all preceding sentences) available to each sentence when calculating the sequentiality metric.

Concreteness Concreteness has been found to be a strong indicator of narrativity in book-based forms (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 post that simultaneously appear in the lexicon. The term weights are derived from the lexicon and represent the degree to which a given term is concrete.

Text Type We consider whether the text is a post or comment.

Length We consider the number of tokens in the text and the mean number of tokens in the sentences.

4.2 Results

Table 3 shows the results of our experiments, comparing various textual features and the story labels of the expert annotators. We run t-tests on the features, applying the Holm method (Holm, 1979) to account for multiple comparisons.

The tests indicate that, in decreasing order of effect size, the frequency of events (with event detection by either the experts or the *realis* event tagger), past-tense verbs, first-person singular pronouns, concrete terms, and third-person singular pronouns are significantly more frequent in stories. Conversely, present- and future-tense verb tenses, second-person pronouns, and text type being a comment (instead of a post) are significantly less frequent in stories.

5 Story Detection

The ability to detect story spans is a prerequisite for more in-depth understanding of storytelling within online communities. As mentioned above, this is a challenging task given a) the informal nature of online communication and b) the intermittent nature of online storytelling (i.e. it may or may not encompass an entire post or comment).

Most prior work on automatic story detection has focused on using feature-based classification approaches, relying on features like n-grams, POS tags, coreference chain length, LIWC categories, and verb classes (dos Santos et al., 2017; Ceran et al., 2012; Gordon and Swanson, 2009; Yao and Huang, 2018; Eisenberg and Finlayson, 2017; Piper et al., 2021a; Piper and Bagga, 2022). 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 large language models (LLMs). Ganti et al. (2022) annotate a set of 849 Facebook posts about breast cancer with binary story labels and then fine-tune three

[4] 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.

[5] version: gpt-4-0314

BERT-based models on the posts and labels, finding that these classifiers have comparable performance to one another and that they all perform better than classical models like SVMs or logistic regression. Similarly, Ganti et al. (2023) annotate 3,000 health misinformation tweets and compare BERT-based classifiers to prompting methods. We follow these approaches by also comparing a classical model to a BERT-based model and prompt-based methods using generative models from OpenAI.

Model Training We evaluate a series of models for our story detection task, including a SVMs baseline using TF-IDF features, a fine-tuned RoBERTa model (Liu et al., 2019), and zero-shot and few-shot GPT-4 prompting (OpenAI, 2023). 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. As in the previous section, we use the *union* of the two annotators' labels to determine positive story labels, i.e., if either annotator labels a text as containing a story span, then we use this as a positive training and evaluation instance of storytelling.[6]

Evaluation Results We show evaluation results in Table 4. Given our small expert set, these results are averaged over bootstrapped samples of the test set, and we show both the mean evaluation scores and their standard deviations. The TF-IDF baseline achieves an F1 score of 0.73, lower than the fine-tuned RoBERTa classifier's F1 score of 0.86. Prompting methods largely yield results comparable to the TF-IDF baseline, with GPT-4 performing better than GPT-3.5 and chain-of-thought prompting not yielding consistent improvements. Compared to prior work by Ganti et al. (2023), we find that our performance scores are consistently lower across the different models and methods, likely owing to the diversity of topics in our dataset.

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 the 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 storylike. 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.

[6]This decision arises from (a) error analysis of classifiers trained using the intersection versus the union of our labels and (b) our codebook, which supports subjective judgments.

	P	R	F1
Majority	0.29 ± 0.02	0.50 ± 0.00	0.37 ± 0.02
SVM with TF-IDF	0.74 ± 0.05	0.73 ± 0.05	0.73 ± 0.05
Fine-tuned RoBERTa	$\mathbf{0.86} \pm 0.03$	$\mathbf{0.86} \pm 0.03$	$\mathbf{0.86} \pm 0.03$
GPT-4 Zero-Shot	0.77 ± 0.04	0.76 ± 0.04	0.75 ± 0.04
GPT-4 Few-Shot	0.75 ± 0.04	0.72 ± 0.04	0.71 ± 0.05
GPT-4 C-o-T	0.78 ± 0.04	0.69 ± 0.03	0.67 ± 0.03
GPT-3.5-Turbo Zero-Shot	0.77 ± 0.03	0.65 ± 0.02	0.59 ± 0.04
GPT-3.5-Turbo Few-Shot	0.69 ± 0.06	0.67 ± 0.05	0.65 ± 0.06
GPT-3.5-Turbo C-o-T	0.72 ± 0.04	0.68 ± 0.04	0.66 ± 0.04

Table 4: Comparison of classification performance across methods. We show results for models trained on the expert annotations, where the presence of a story is determined using the *union* of the annotators' positive labels. For few-shot tests with OpenAI models, we prompt with two positive and two negative examples in the prompt. We use model versions gpt-4-0314 and gpt-3.5-turbo-0613.

Prediction We use the fine-tuned RoBERTa model 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 305 subreddits for annotation. We assign each post and comment a binary prediction about the presence of storytelling in the text.

Interpretation When interpreting our predicted storytelling rates, 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, short texts, or structured texts, our storytelling rates should not be interpreted as overall storytelling rates in the subreddit but rather the rate of storytelling among longer, coherent posts and comments in that subreddit. The rates should be interpreted in a comparative way (e.g., rank-

ing of subreddit categories) rather than as absolute rates of storytelling in the subreddit.

6 Analysis

In this section we discuss our findings regarding the prevalence, distinctiveness, and social dynamics of storytelling using our best model.

6.1 Where do people tell stories online?

Overall, 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, 40% of all the texts were predicted to contain stories, with a macro average across subreddits of 60% and a macro average across categories of 68%, suggesting that storytelling behavior is indeed a frequent communicative behavior across different kinds of communities.

Using our predicted story labels, Figure 1 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, relationships, sexuality, among other topics.

6.2 How distinctive are stories by community?

We map communities across two axes: their *story rate*, predicted by our story detector, and the *distinctiveness* of their vocabulary.

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

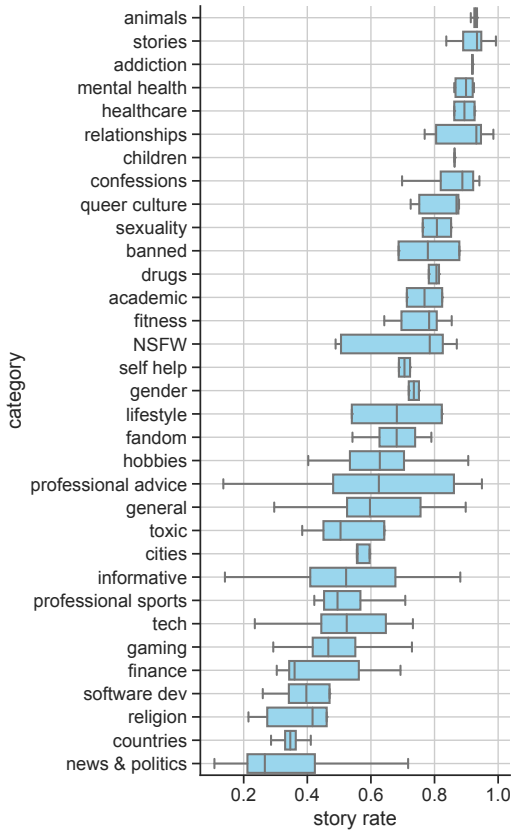

Figure 1: Categories of subreddits ranked by their rate of texts (posts and comments) containing stories, as predicted by our classifier. Results represent 20 bootstrapped samples of the texts for each subreddit.

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.

Figure 2 shows the 306 subreddits in our dataset mapped by their story rate (as predicted by our story detector) and their distinctiveness. Table 5 shows the "corners" of this plot, i.e., the subreddits with the most and least storytelling and distinctiveness. Categories of subreddits form interpretable clusters. 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 — while subreddits in the *fandom* category such as *r/asoiaf* (dedicated to the series of novels, *A Song of Ice and Fire*) tend to have less storytelling, but when storytelling happens, the stories use relatively distinc-

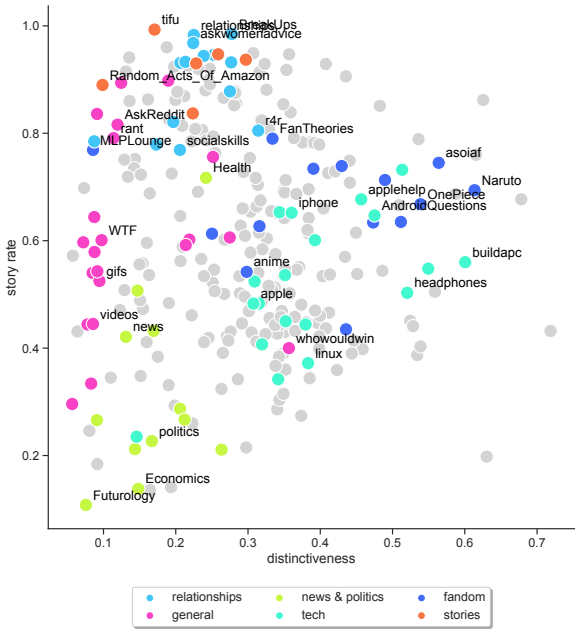

Figure 2: Some subreddits contain stories on specific topics, while others draw a wide range of storytelling topics. We show the 500 most frequent subreddits, colored by category. The y-axis shows the rate of storytelling in the subreddit (predicted by our classifier), and the x-axis shows the distinctiveness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

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

Table 5: Subreddits with the most or least storytelling and most or least distinctive vocabulary.

tive language. Some categories, like *professional advice*, range across the plot, with member subreddits like *askscience* (containing little storytelling and low distinctiveness) and *MechanicAdvice* (containing frequent storytelling and highly distinctive language).

6.3 In what conversational contexts do people tell stories?

One of the affordances of working with social media data is we can observe the conversational dynamics of storytelling. 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 are not isolated events but depend on the social interactions that elicit and modulate their telling (see also (Herman, 2009) for an emphasis on "situatedness" for story meaning).

In this section, we condition on the relationship between *posts* and *comments* as an initial way of understanding the conversational patterns of online storytelling. 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. This suggests that storytelling is more often a way of initiating new threads rather than a means of responding to existing posts.

Nevertheless, we also find that the likelihood of comments containing stories differs strongly across communities. Table 6 shows the subreddits highest and lowest ranked by their ratio of storytelling in posts versus comments. We observe that some subreddits 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 (measured via the rate of question mark characters) and storytelling in posts ($p > 0.05$).

7 Discussion

In this paper, our aim has been to contribute to the larger area of computational narrative understanding and modeling (Ranade et al., 2022; Clark et al., 2022; Piper, 2023; Eisenberg and Finlayson, 2017; Ganti et al., 2023) through an examination of storytelling behavior across online communities. In addition to providing new annotation frameworks for facilitating further work in this area, we demonstrate significant model accuracy for the task of online storytelling detection (F1 of 0.86). As with prior work, continued evidence suggests that storytelling is a distinctive form of human communication and is detectable even among the fluid and informal forms of communication that dominate social media.

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 6: 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.

Defining Online Storytelling The features that are most strongly associated with storytelling largely confirm prior work's emphasis on storytelling being grounded in agent-centered, event-driven forms of communication that ground those events in concrete settings (known as the "deictic theory" of storytelling (Piper and Bagga, 2022)). However, unlike prior work, we find significantly more usage of first-person singular pronouns as well as significantly more storytelling in posts in contrast to comments. Future work will want to continue to assess the accuracy of models with more nuanced and higher level narrative features.

We also find high levels of storytelling among our sampled documents, with 40% of all texts predicted to exhibit storytelling behavior and a macro average of 60% across all sampled subreddits. This suggests that prior work's indication of the importance of storytelling to online social communication remains consistent. Storytelling is not the only form of online communication but is indeed a common form. For some communities it is an essential form of communication and information sharing. This confirms our belief that storytelling should be seen as a key component for understanding online discourse more generally in future work.

Cross-Community Themes At the same time, we observe strong cross-community and thematic differences in the prevalence of storytelling online. Communities that focus more on personal issues such as health, relationships and sexuality tend to favor storytelling, while those that are oriented around politics, tech, and current events (news, gaming, finance, sports) tend to use less storytelling. Importantly, these rates are drawn from among contentful texts only, rather than all texts shared in the communities. They thus indicate the rate of storytelling amid coherent discussions.

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. This fits with our finding that healthcare communities have higher rates of storytelling, in contrast to other kinds of online communities. Our classification performance is also lower than the performance reported in both of those works, despite using similar fine-tuning techniques, signalling the challenge of (1) identifying stories across diverse topics and communities and (b) allowing story spans to be short and include hypothetical events and events with various verb tenses.

Conversational Patterns Our work undertakes an initial exploration of the conversational patterns of storytelling. Narrative theory has emphasized the interactive nature of storytelling through the concept of "small stories" (Georgakopoulou, 2007), where storytelling involves the interactive participation of multiple actors. Online storytelling similarly occurs in dynamic conversational contexts. Here we have modeled such patterns in two key ways: a) the relationship between posts and comments and b) the relationship between stories within a given community (i.e. "distinctiveness").

We find that storytelling tends to be more initiatory in nature, i.e. more likely to appear in posts rather than in comments, though with strong community differences. We also find that some communities are marked by very "distinctive" stories, meaning they tend to be highly specific to that forum rather than more general topics. Some communities are marked by high degrees of semantic coherence in their storytelling behavior while others are marked by an openness to diverse kinds of stories. There is much further work one could do to deepen our understanding of the social dynamics of online storytelling, which would help contribute

to an understanding of the social functions of storytelling behavior more generally.

Building a Story Detector Finally, our work demonstrates the feasibility of different methods, including fine-tuning with expert-annotated data, few-shot prompting, and zero-shot prompting, providing lessons for other researchers who would like to predict storytelling in new domains. If the highest precision and recall is desired, then our findings suggest that fine-tuning an LLM like RoBERTa yields highest performance. However, if annotation is costly and time-consuming (as we found), then a codebook used with zero-shot prompting of GPT-4 yielded performance comparable to a classifical TF-IDF-weighted classifier, confirming prior work by Ganti et al. (2023).

8 Conclusion

Our study of storytelling has begun to map the variation in storytelling across online communities. We designed a new story dataset specific to online discourse, built a story predictor, mapped where storytelling occurs, identified which features are signals of storytelling in this online context, and have begun to illustrate the conversational features of storytelling. While storytelling has been the subject of intense scrutiny for millennia, our work begins to make further headway into the computational analysis of storytelling across large online communities.

9 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.

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.

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, and the need to use subjective interpretation for the annotation task. 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.

10 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, but we do not share replications of user IDs, post or comment texts, or other user information.

All of the data is included in our automatic analysis, but we attempt to remove the most potentially harmful data from our annotators. After manually categorizing the subreddits (see §3.1), we filter a set of categories from the annotation task: *banned*, *children*, *confessions*, *mental health*, *NSFW*, *queer culture*, *relationships*, *drugs*, *gender*, *sexuality*, and *toxic*. 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.

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 also prevents readers from searching Reddit for those exact texts and identifying their authors.

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.

Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

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

Lucas M. Bietti, Ottilie Tilston, and Adrian Bangerter. 2019. Storytelling as adaptive collective sensemaking. *Topics in Cognitive Science*, 11(4):710–732.

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.

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.

Elizabeth Clark, Faeze Brahman, and Mohit Iyyer. 2022. Proceedings of the 4th workshop of narrative understanding (wnu2022). In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*.

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.

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.

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

Ryan J. Gallagher, Elizabeth Stowell, Andrea G. Parker, and Brooke Foucault Welles. 2019. Reclaiming stigmatized narratives: The networked disclosure landscape of metoo. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

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.

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.

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

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.

Lena Mamykina, Drashko Nakikj, and Noemie Elhadad. 2015. Collective sensemaking in online health forums. In *Proceedings of CHI*.

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. 2023. Computational narrative understanding: A big picture analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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.

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.

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.

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

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

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. 2019. 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.

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.

Alyson L. Young and Andrew D. Miller. 2019. "this girl is on fire": Sensemaking in an online health community for vulvodynia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

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.