Vrittanta-EN: A Benchmark Dataset for Event Trigger Detection and Classification Advancing Event Understanding in English Narrative Discourse

Chaitanya Kirti, Ashish Anand, Prithwijit Guha
Indian Institute of Technology Guwahati
Assam, India
{ckirti, anand.ashish, pguha}@iitg.ac.in

Abstract

Event trigger detection and classification involve identifying meaningful occurrences and categorizing them into predefined event types within narrative text. Despite extensive research on English event extraction in factual domains like news and biomedical text, narrative prose, such as, short stories has received comparatively little attention. To bridge this gap, *Vrittanta-EN* introduces a manually annotated English corpus comprising 11,272 event instances extracted from diverse short stories. The dataset captures a wide range of communicative, cognitive, and physical actions typical of narrative discourse. A comprehensive evaluation is conducted across a wide range of models, including classical machine learning baselines (SVM, Naive Bayes), neural sequential models (LSTM, BiLSTM, BiLSTM-CRF), encoder-only transformers (BERT, RoBERTa, ALBERT, DistilBERT, DeBERTa, ELECTRA), and encoder-decoder models (T5, BART), along with large language models (GPT-4.1, DeepSeek-V3.2-Exp, Claude Sonnet 4) under both zero-shot and five-shot settings. Experimental results show that ELECTRA achieved the highest overall performance for event trigger detection with an F1-score of 90.61%, while RoBERTa demonstrated superior performance for event classification with a macro F1 of 74.71%. These findings highlight the robustness of contextual transformer-based architectures for modeling narrative event structures in English short stories.

Keywords: Event Trigger Detection, Event Classification, Short Stories, English, Dataset, Evaluation

1. Introduction

Event understanding plays a pivotal role in enabling machines to interpret textual narratives by transforming unstructured stories into structured representations of actions, states, and changes (Chambers and Jurafsky, 2008; Ranade et al., 2022). Within this broader goal, event trigger detection identifies lexical units that signal an event, while event classification assigns them to predefined semantic types. These two subtasks form the foundation for downstream applications such as temporal reasoning, question answering, and narrative comprehension. An example illustrating an event trigger and its corresponding class is shown in Figure 1.

Sentence: The blue whale **left** the party accompanied by a large shark.

Event Trigger: *left*
Event Class: `MOVEMENT`

Figure 1: Example illustrating an event trigger and its corresponding class. The word "left" signals a `MOVEMENT` event.

Despite the remarkable success of event extraction in factual domains such as newswire (Walker et al., 2005; Knuth et al., 2015; Mitamura and Liu) and biomedicine (Kim et al., 2012; Uzuner et al., 2006), its application to narrative prose, particularly short stories, remains limited. Narrative texts differ fundamentally from news or clinical reports: events unfold through implicit causality, emotions, and subjective perception rather than explicit temporal or factual markers. These properties make narrative event extraction inherently challenging, especially when dealing with figurative expressions, nested event references, and indirect discourse.

Recent research (Sims et al., 2019) has highlighted the need for literary-domain corpora to explore such complexities. However, existing datasets are either limited in size or confined to specific genres, restricting their use for evaluating modern architectures ranging from neural sequence models to large language models (LLMs). Furthermore, English, despite being well-resourced overall, lacks a benchmark specifically curated for event trigger detection and classification in short stories.

To address this gap, *Vrittanta-EN* introduces a manually annotated corpus of 11,272 event instances drawn from English short stories across multiple thematic sources. Each event is labeled with its corresponding trigger span and categorized into one of seven predefined event classes, reflecting both physical and cognitive actions typical of narrative discourse. The corpus captures a diverse range of communicative, cognitive, and life-event scenarios that mirror the richness of storytelling.

A comprehensive evaluation is conducted across multiple paradigms—ranging from classical ma-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7755–7765
11–16 May 2026.

chine learning models and neural sequence architectures to contextual transformer encoders and generative encoder-decoders. In addition, a comparative analysis with LLMs under zero-shot and five-shot settings provides insights into their emergent event understanding capabilities.

The key contributions of this paper are summarized as follows:

- *Vrittanta-EN*, a manually annotated English dataset for event trigger detection and classification in the narrative domain, is presented.
- Comprehensive annotation guidelines are developed to capture events in English short stories.
- Baselines are established using classical, neural, and transformer-based architectures, and the performance of multiple pre-trained language models and large language models is evaluated.
- The strengths, limitations, and error patterns observed from the best-performing model are analyzed, providing insights to advance future research in narrative event understanding.

The remainder of this paper is structured as follows. Section 2 reviews prior work on event detection and classification. Section 3 details the dataset construction and key statistics. Section 4 describes the experimental setup and evaluation protocol. Section 5 reports and discusses the results, followed by Section 6, which analyzes common error patterns. Finally, Section 7 concludes the paper and outlines directions for future work.

2. Related Work

Event extraction has been extensively studied within structured and factual domains such as newswire and biomedical text. Foundational corpora like ACE (Walker et al., 2005), ERE (Aguilar et al., 2014), and TAC-KBP (Mitamura et al., 2016) provided standardized ontologies and annotation frameworks that have guided research for nearly two decades. Early methods relied on statistical classifiers and hand-crafted linguistic features (Ahn, 2006; Ji and Grishman, 2008), while the advent of neural approaches introduced architectures such as CNNs (Nguyen and Grishman, 2015), RNNs (Chen et al., 2015), and attention-based models (Liu et al., 2018), which improved contextual representation learning. However, these systems were designed primarily for explicit, factual events in newswire data and often fail to generalize to narrative prose, where events are described through figurative or implicit expressions.

Event understanding in narrative text requires models to capture implicit causality, character intentions, and temporally interdependent actions. Literary narratives differ from factual reports in that events often emerge through dialogue, perception, or metaphor, rather than explicit factual cues. Early work such as LitBank (Bamman et al., 2020) and the EventStoryLine Corpus (Caselli et al., 2017) advanced the study of event semantics in literary text by annotating events and their relations across narrative contexts. More recent research has explored discourse-aware and commonsense-augmented event modeling (Pustejovsky and Stubbs, 2019; Li et al., 2021) to handle contextual and causal dependencies in stories. Despite these efforts, most available corpora remain limited in size, genre diversity, and annotation granularity, leaving a need for a comprehensive benchmark that captures the variety of events occurring in short stories.

Transformer-based encoders such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021) have significantly advanced event extraction by capturing deep bidirectional context and improving generalization across domains. Domain-specific variants such as BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019) demonstrated that task-specific pretraining can further enhance performance. In addition, encoder-decoder architectures like T5 (Raffel et al., 2020b) and BART (Lewis et al., 2020b) have enabled end-to-end event extraction and generation, unifying detection and classification under a single generative framework. These models have shown strong performance on open-domain event benchmarks and have become the foundation for evaluating modern event understanding systems.

The recent emergence of large language models such as GPT-4, Claude Sonnet, and DeepSeek has shifted event extraction toward instruction-following and in-context learning paradigms (Chen et al., 2024). These models exhibit strong zero-shot and few-shot reasoning abilities, capable of detecting and labeling event triggers without task-specific fine-tuning. While promising, systematic evaluations of such models on narrative texts remain scarce.

Unlike prior datasets focused on news or scientific literature, *Vrittanta-EN* emphasizes narrative complexity, implicit event expression, and stylistic diversity across English short stories. It bridges the methodological gap between structured event extraction and literary narrative understanding, offering a unified platform for evaluating classical models, neural architectures, and LLMs under consistent experimental conditions. This benchmark serves as an essential step toward advancing story-aware event understanding in English narrative discourse.

3. Dataset Construction

The following subsections describe the dataset overview, the process of collecting short stories, the detailed annotation procedure, and the annotated dataset statistics.

3.1. Dataset Overview

The *Vrittanta-EN* dataset is developed to support research in event trigger detection and classification within the narrative domain. It focuses on English short stories, which exhibit rich event semantics through dialogue, description, and moral progression. While prior English datasets have concentrated on factual reporting or newswire text, *Vrittanta-EN* captures the complexity of narrative event structures involving both realis and cognitive processes. The dataset contains a total of 11,272 annotated event instances derived from 200 short stories. Each story was segmented into sentences and tokenized before annotation, ensuring linguistic consistency and contextual clarity. The dataset was designed to facilitate fine-grained event understanding in literary discourse rather than domain-specific fact extraction.

3.2. Collection of Short Stories

The short stories included in *Vrittanta-EN* were curated from eight publicly available sources representing India's rich literary and moral storytelling tradition, including *Panchatantra*, *Champak*, *Tenali Raman*, *Akbar–Birbal*, *Betal Pachisi*, *Hitopadesh*, *Jataka Tales*, and *Singhasan Battisi*. These collections were selected to capture diverse narrative styles ranging from fables and dialogues to moral parables, thereby ensuring broad coverage of communication, cognitive, and physical event types. This diversity contributes to a balanced and culturally grounded corpus of narrative discourse.

3.3. Annotation Guidelines

The annotation of *Vrittanta-EN* follows a linguistically motivated framework designed to identify event expressions within English short stories. The guidelines were constructed to ensure consistency in event boundary marking and accurate categorization across narrative contexts. Annotators were instructed to label tokens that explicitly denote an event occurrence while preserving contextual coherence.

3.3.1. Event Expression

The annotation framework focuses on identifying words or phrases that denote realis events, emphasizing grammatical and semantic cues that indicate an occurrence, action, or state. Finite and non-finite verbs (e.g., **painted**, **wanted**), participles such as **retired** or **playing**, and nominalized forms like **marriage** or **attack** are typically annotated as event triggers. Adjectival forms expressing emotional or resultant states (e.g., **terrified**) are also included when functioning as event indicators. Bold words represent annotated event triggers, while underlined words mark event-like expressions that are excluded, as they either describe non-occurring actions or sub-events embedded within a broader event.

3.3.2. Event Categories

Each identified event trigger is assigned to one of seven predefined event types, encompassing a broad range of narrative actions:

- **`COMMUNICATION (COM)`:** Acts of speech or dialogue where participants exchange information.
- **`GENERAL-ACTIVITY (GEN)`:** Routine or habitual human actions observed in everyday life.
- **`MOVEMENT (MOV)`:** Events denoting motion or change of location, such as walking, flying, or traveling.
- **`COGNITIVE-MENTAL-STATE (CMS)`:** Mental or emotional states, including thinking, remembering, perceiving, or feeling.
- **`LIFE-EVENT (LE)`:** Significant life occurrences such as birth, illness, injury, death, or marriage.
- **`CONFLICT (CON)`:** Disagreements, confrontations, or fights, either verbal or physical.
- **`OTHERS (OTH)`:** A residual category for events that do not clearly belong to the above classes.

3.4. Annotation Process

Annotation was performed using the BRAT web-based annotation tool (Stenetorp et al., 2012). Each story was manually annotated for event triggers and their corresponding event types following a structured guideline. The guidelines were designed to handle narrative-specific challenges such as indirect speech, implicit events, and cognitive or emotional expressions embedded in dialogue.

Two trained annotators, both with expertise in English linguistics and computational annotation, participated in the process. Inter-annotator agreement (IAA) was computed on a randomly selected subset of 20 stories, yielding an agreement score of 93.8%, which reflects strong consistency in event

Sl.No	Statistics	Count
1.	Total tokens in the dataset	174,559
2.	Total unique tokens in the dataset	10,484
3.	Total sentences in the dataset	10,259
4.	Average tokens per story	873
5.	Average sentences per story	51
6.	Average tokens per sentence	17
7.	Total events in the dataset	11,272
8.	Average events per story	56

Table 1: Statistics of the annotated *Vrittanta-EN* dataset.

Sl.No	Event Type	Count
1.	COMMUNICATION	3,456
2.	GENERAL-ACTIVITY	2,951
3.	MOVEMENT	2,192
4.	COGNITIVE-MENTAL-STATE	1,528
5.	LIFE-EVENT	535
6.	OTHERS	490
7.	CONFLICT	120

Table 2: Distribution of different types of events in *Vrittanta-EN*.

identification and labeling. Disagreements were resolved through iterative discussion and consensus adjudication. Table 1 summarizes the core statistics of the *Vrittanta-EN* dataset.

3.5. Annotated Dataset Statistics

Table 1 presents the quantitative characteristics of the *Vrittanta-EN* dataset. The corpus comprises 174,559 tokens and 10,259 sentences distributed across 200 short stories, with an average of 56 annotated events per story. The relatively high token count and sentence density highlight the narrative diversity of the English stories, which include dialogues, moral reflections, and dynamic physical interactions. On average, each story contains 51 sentences, making the dataset sufficiently rich for modeling long-range contextual dependencies in narrative event structures.

The class-wise distribution of event types is reported in Table 2. COMMUNICATION and GENERAL-ACTIVITY are the most prevalent categories, illustrating the dominance of dialogic and routine human actions in narrative storytelling. MOVEMENT and COGNITIVE-MENTAL-STATE events occur frequently, reflecting the physical dynamics and introspective nature typical of English fables and moral tales. In contrast, LIFE-EVENT, CONFLICT, and OTHERS appear less often, introducing a natural class imbalance that mirrors the narrative focus of the source texts rather than a sampling artifact. Such imbalance poses challenges for classification models, particularly in low-frequency categories such as CONFLICT, and remains a key issue for future research.

Trigger Word	Count	Event Rate
said	1,298	96%
asked	415	98%
went	409	95%
came	291	82%
told	190	92%
saw	179	92%
replied	170	98%
thought	148	81%
took	142	78%
hearing	113	92%

Table 3: Top 10 event trigger words with their frequency and event rate in *Vrittanta-EN*. The event rate represents the percentage of instances where these words are labeled as events compared to their total occurrences in the corpus.

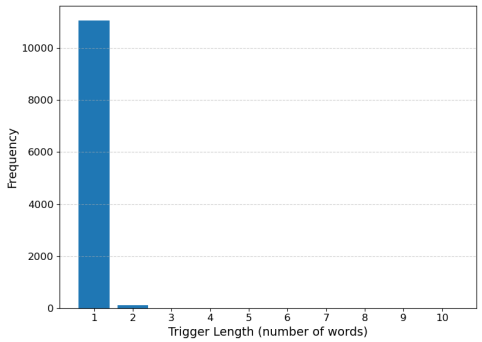

Figure 2: Trigger length (number of words) vs. frequency plot for *Vrittanta-EN*.

Table 3 lists the ten most frequent event trigger words along with their event rates. High-frequency triggers such as "*said*," "*asked*," and "*went*" confirm that communication and motion-based expressions dominate English storytelling, similar to classical fable and moral genres. This lexical distribution reinforces the narrative orientation of the dataset toward dialogue-driven and action-centered plots.

An analysis of trigger span lengths, illustrated in Figure 2, shows that most event triggers consist of a single word, with a smaller number of two-word phrases and very few longer expressions. The dominance of single-word triggers indicates that English events in narrative text are generally encoded through concise lexical units. They are most often verbs or verbal predicates, reflecting the syntactic economy of event expression in English discourse.

4. Dataset Evaluation

This section presents the task description, evaluation models, and the experimental setup used to benchmark the *Vrittanta-EN* dataset.

Model	P	R	F1
SVM	69.18	86.68	76.95
Naïve Bayes	90.16	57.67	70.34
LSTM	82.38	82.20	82.29
BiLSTM	80.62	85.28	82.89
BiLSTM–CRF	85.74	84.68	85.21
BERT	86.95	93.00	89.88
RoBERTa	86.85	93.04	89.84
ALBERT	89.09	90.37	89.73
DistilBERT	84.47	94.91	89.39
DeBERTa	83.88	82.20	83.03
T5	87.28	92.24	89.69
BART	83.74	94.62	88.85
ELECTRA	88.65	92.65	**90.61**
GPT-4.1 (0-shot)	43.30	89.90	58.50
GPT-4.1 (5-shot)	49.84	85.67	63.05
DeepSeek-V3.2-Exp (0-shot)	43.80	93.70	59.70
DeepSeek-V3.2-Exp (5-shot)	45.38	89.37	60.19
Claude Sonnet 4 (0-shot)	45.50	94.30	61.40
Claude Sonnet 4 (5-shot)	49.16	91.58	63.96

Table 4: Performance of different models on event trigger detection in *Vrittanta-EN*.

4.1. Task Description

Event trigger detection and classification involve identifying the word or phrase in a sentence that signals the occurrence of an event and assigning it to an appropriate event type. The task is modeled as a sequence labeling problem, where each token in a sentence is assigned a label following the BIO tagging format. Specifically, each token w_i in an input sentence $\mathbf{S} = [w_1, w_2, \ldots, w_n]$ is labeled as $l_i \in \{O, B\text{-}X, I\text{-}X\}$, where $B\text{-}X$ denotes the beginning of an event trigger of type X, $I\text{-}X$ marks continuation tokens, and O indicates non-event tokens.

4.2. Evaluation Models

To benchmark *Vrittanta-EN*, a diverse range of models was evaluated, covering classical, neural, and transformer-based paradigms.

Classical Baselines: Support Vector Machine (SVM) and Naïve Bayes (NB) were implemented using TF–IDF representations to establish non-contextual baselines for event detection and classification.

Neural Sequence Models: Standard Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and BiLSTM–Conditional Random Field (BiLSTM–CRF) models were trained to capture temporal dependencies and label transitions effectively.

Encoder-Only Transformers: Multiple pre-trained contextual encoders were evaluated, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), DistilBERT (Sanh et al., 2019), DeBERTa (He et al., 2020), and ELECTRA (Clark et al., 2020). These models, trained on large-scale corpora, were fine-tuned for both event detection and classification.

Encoder–Decoder Transformers: T5 (Raffel et al., 2020a) and BART (Lewis et al., 2020a) were employed to explore text-to-text and denoising generation frameworks, respectively, for event-related sequence tagging and type prediction.

Large Language Models (LLMs): Generative LLMs including GPT-4.1 [1], DeepSeek-V3.2-Exp[2], and Claude Sonnet 4[3] were evaluated under both zero-shot and five-shot prompting settings. Few-shot prompts were constructed by providing five example sentences per class, followed by an unseen test sentence for prediction.

4.3. Experimental Setup

All classical and neural models were implemented in Python using Scikit-learn and PyTorch frameworks. For transformer-based architectures, the HuggingFace Transformers library was used with default configurations and AdamW optimization. All transformer models, whether encoder-only or encoder–decoder, were used in their base variants. Models were trained for 30 epochs with a batch size of 32 and a learning rate of 2e-5, employing early stopping based on validation F1. The dataset was split into training, validation, and test sets with a 60:10:30 ratio. For event trigger detection, model performance was evaluated using micro-averaged precision (P), recall (R), and F1-score across all event types. For classification, micro as well as macro-averaged precision, recall, and F1 were computed to account for class imbalance.

5. Results and Discussion

This section reports and analyzes the performance of various models across both event trigger detection and event classification tasks on the *Vrittanta-EN* dataset.

5.1. Event Trigger Detection

Table 4 presents the performance of all evaluated models on the event trigger detection task, measured in terms of precision (P), recall (R), and F1-score. Classical models such as SVM and Naïve Bayes demonstrated moderate performance, achieving F1-scores of 76.95% and 70.34%, respectively. Their limitations arise primarily from their reliance on surface-level lexical cues without contextual understanding.

[1] https://platform.openai.com/docs/models/gpt-4.1
[2] https://api-docs.deepseek.com/news/news250929
[3] https://www.anthropic.com/news/claude-4

Model	Metric	COM	GEN	MOV	CMS	LE	OTH	CON	Macro F1
SVM	**P**	82.49	39.82	51.31	53.53	23.19	18.83	41.86	
	R	90.59	62.73	79.08	80.03	55.90	39.28	64.28	
	F1	86.35	48.72	62.24	64.16	32.78	25.46	50.70	52.92
Naive Bayes	**P**	96.41	79.61	91.00	83.87	0.00	0.00	0.00	
	R	73.04	15.84	30.22	10.37	0.00	0.00	0.00	
	F1	83.11	26.42	45.38	18.47	0.00	0.00	0.00	24.77
LSTM	**P**	87.09	49.11	64.35	47.08	40.25	18.61	2.19	
	R	87.46	66.53	71.70	70.85	39.75	25.00	21.42	
	F1	87.27	56.51	67.83	56.57	40.00	21.34	3.98	47.64
BiLSTM	**P**	85.29	55.50	67.25	60.64	35.35	16.16	4.34	
	R	87.87	61.34	73.63	71.05	43.47	19.28	14.28	
	F1	86.56	58.27	70.30	65.44	38.99	17.58	6.66	49.11
BiLSTM-CRF	**P**	90.73	67.82	77.18	79.72	68.53	44.59	77.77	
	R	90.07	61.72	74.34	69.86	37.88	23.57	25.00	
	F1	90.40	64.63	75.73	74.46	48.80	30.84	37.83	60.38
BERT	**P**	92.44	70.20	83.92	78.55	66.25	36.43	93.75	
	R	94.56	83.90	87.17	82.63	67.08	33.57	53.57	
	F1	93.49	76.44	85.51	80.54	**66.66**	34.94	68.18	72.25
RoBERTa	**P**	91.65	74.71	82.92	76.34	60.00	43.55	82.60	
	R	94.14	82.76	89.63	87.62	65.21	50.71	67.85	
	F1	92.88	**78.53**	**86.14**	81.59	62.50	**46.86**	**74.50**	**74.71**
ALBERT	**P**	91.83	72.50	82.51	78.35	55.29	36.58	72.22	
	R	95.19	78.20	85.41	83.83	58.38	42.85	46.42	
	F1	93.48	75.24	83.93	81.00	56.79	39.47	56.52	69.49
DistilBERT	**P**	91.03	73.58	81.42	77.57	55.37	40.15	77.77	
	R	94.46	79.08	86.29	84.23	63.97	37.85	50.00	
	F1	92.71	76.23	83.78	80.76	59.36	38.97	60.86	70.38
DeBERTa	**P**	84.60	45.20	67.73	49.91	22.09	0.00	0.00	
	R	86.72	55.51	70.47	55.68	11.80	0.00	0.00	
	F1	85.65	49.82	45.16	52.64	15.38	0.00	0.00	35.52
T5	**P**	99.42	68.16	80.61	78.92	56.60	40.20	0.00	
	R	99.06	78.96	86.99	82.23	55.90	27.85	0.00	
	F1	**99.24**	73.16	83.68	80.54	56.25	32.91	0.00	60.83
BART	**P**	90.32	70.70	82.52	75.04	58.33	45.16	100.00	
	R	94.67	79.84	89.63	88.22	60.86	40.00	50.00	
	F1	92.44	75.00	85.93	81.10	59.57	42.42	66.66	71.87
Electra	**P**	91.41	70.91	83.25	79.40	63.57	40.60	73.07	
	R	94.56	82.50	88.22	85.42	59.62	38.57	67.85	
	F1	92.96	76.27	85.66	**82.30**	61.53	39.56	70.37	72.66
GPT-4.1 (0-shot)	**P**	10.87	13.90	24.35	23.04	13.90	4.55	15.66	
	R	35.40	11.69	44.00	21.40	4.07	9.06	18.60	
	F1	16.58	12.72	31.38	22.19	6.32	6.06	17.00	16.04
GPT-4.1 (5-shot)	**P**	14.68	18.11	33.02	31.66	18.43	6.17	21.29	
	R	47.43	15.46	61.56	29.30	5.40	12.22	25.65	
	F1	22.32	16.69	43.03	30.44	8.40	8.21	23.28	21.77
DeepSeek-V3.2-Exp (0-shot)	**P**	11.45	12.10	20.92	25.60	12.18	3.96	13.80	
	R	32.47	12.32	40.23	23.21	3.39	7.87	17.33	
	F1	16.79	12.21	27.51	24.36	5.32	5.29	15.39	15.27
DeepSeek-V3.2-Exp (5-shot)	**P**	15.69	15.85	29.08	34.30	22.00	10.80	18.35	
	R	43.19	16.51	53.10	31.57	10.50	12.20	22.70	
	F1	23.01	16.17	37.58	32.88	14.19	11.47	20.30	22.23
Claude Sonnet 4 (0-shot)	**P**	9.96	10.82	18.53	22.57	10.96	3.44	12.10	
	R	27.82	10.71	34.60	20.36	2.99	6.78	15.32	
	F1	14.49	10.77	24.05	21.43	4.66	4.57	13.52	13.36
Claude Sonnet 4 (5-shot)	**P**	14.15	15.21	25.97	32.26	15.06	4.78	16.43	
	R	37.95	14.90	49.16	28.19	4.13	9.51	20.63	
	F1	20.62	15.05	33.98	30.09	6.49	6.36	18.29	18.70

Table 5: Performance comparison of different models on *Vrittanta-EN* for event classification task. The best F1-score is highlighted in bold.

Neural architectures achieved significant improvements, with BiLSTM and BiLSTM–CRF attaining F1-scores of 82.89% and 85.21%. The inclusion of the CRF layer in BiLSTM–CRF improved label consistency, particularly for multi-token triggers.

Among transformer-based encoders, all models performed exceptionally well, with BERT (89.88%), RoBERTa (89.84%), and ALBERT (89.73%) exhibiting comparable F1-scores. ELECTRA achieved the

highest overall detection performance with an F1-score of 90.61%, outperforming even larger models like BERT and RoBERTa. This gain can be attributed to its replaced-token detection pre-training objective, which better aligns with token-level tasks such as event trigger identification.

Encoder–decoder models (T5 and BART) also performed competitively, with F1-scores of 89.69% and 88.85%, respectively, indicating that generative text-to-text formulations can effectively capture trigger semantics in narrative English.

Large language models (LLMs) such as GPT-4.1, DeepSeek-V3.2-Exp, and Claude Sonnet 4 achieved relatively lower results under zero-shot conditions, reflecting the challenge of adapting instruction-tuned generative models to fine-grained token classification. However, in the five-shot setting, all LLMs showed marked improvement (average +4–5 F1 points), confirming that even limited in-context learning can enhance structured event recognition. Overall, the results suggest that contextual pre-training and language model adaptability are key for achieving robust performance in event detection tasks.

5.2. Event Classification

The results of the event classification task are summarized in Table 5. Classical approaches such as SVM and Naïve Bayes yielded F1-scores of 52.92% and 24.77%, respectively, indicating that shallow lexical models are insufficient for capturing nuanced event semantics in narrative prose.

Sequential models, including BiLSTM–CRF, achieved improved performance (49.11%), benefiting from contextual sequence modeling and transition dependencies. However, transformer-based architectures demonstrated the most significant advancement, with BERT, ALBERT, and BART achieving macro-F1 scores around 70%. RoBERTa outperformed all other models with a macro-F1 of 74.71%, demonstrating its superior ability to encode contextual dependencies and narrative coherence in English storytelling. ELECTRA and T5 followed closely, indicating the effectiveness of hybrid token and sequence modeling for event understanding.

Among LLMs, GPT-4.1, DeepSeek-V3.2-Exp, and Claude Sonnet 4 again performed modestly in zero-shot conditions, with average macro-F1 scores around 13–16%. The five-shot configuration substantially improved LLM outputs (average +4–7 F1), though they remained below the performance of fine-tuned transformer models. This highlights the continued gap between prompt-based inference and supervised fine-tuning in structured event classification.

The experimental findings highlight clear distinctions in model capabilities. Transformer-based encoders, particularly ELECTRA and RoBERTa, demonstrated strong generalization in both detection and classification tasks, indicating their robustness in capturing the rich semantic and syntactic structure of English narrative text. Sequential models like BiLSTM-CRF, while effective, struggled with longer and multi-event sentences. Classical models exhibited limited contextual sensitivity, and LLMs faced challenges in token alignment and label consistency. These results collectively underscore the need for joint modeling frameworks and discourse-aware representations to further improve event comprehension in narrative domains.

6. Error Analysis

This section presents the error analysis for the best-performing models in the two subtasks: ELECTRA for event trigger detection and RoBERTa for event classification.

6.1. Event Trigger Detection Errors

Despite achieving the highest detection F1-score (90.61%), the ELECTRA model exhibited certain systematic errors. These errors can be grouped into the following types:

Boundary Detection Errors ELECTRA occasionally failed to capture the full span of multiword event triggers, especially in phrasal verb constructions. For example, in the sentence “He had finally made up his mind to leave the kingdom.”, the annotated trigger was *“made up”*, but the model predicted only *“made”*. This indicates a potential limitation in handling rare compositional events that span multiple tokens, where semantic interpretation depends on the complete phrase.

False Positives in Hypothetical Contexts ELECTRA tended to over-predict events in sentences containing modal verbs or unrealized conditions. For instance, in the sentence, “If he could travel back in time, he would change his decision.”. Here, the token *“travel”* was incorrectly identified as an event trigger despite being hypothetical. This pattern mirrors modality confusion reported in earlier works (Sims et al., 2019), suggesting that fine-grained cues such as “could” or “would” remain difficult to disambiguate during contextual encoding.

Implicit or Missing Triggers. Some events with implicit or abstract realization were missed entirely. For example, in the sentence, “The people rejoiced at the victory.”, the true trigger *“rejoiced”* was missed due to its non-agentive form and lack of syntactic markers linking it to a clear subject-action

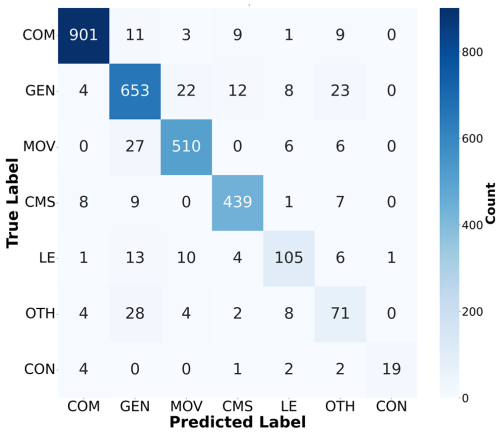

Figure 3: Confusion matrix of the best performing model (RoBERTa) on the event classification task.

relation. Such omissions are common in narrative text where emotional or state-change events are expressed implicitly.

Overall, ELECTRA demonstrates strong contextual precision but exhibits minor recall loss for semantically diffuse or compositional triggers. The model also struggles with disambiguating actual events from hypothetical events, a known challenge in narrative discourse analysis.

6.2. Event Classification Errors

RoBERTa achieved the best macro-F1 (74.71%) in event classification; however, qualitative analysis and the confusion matrix, as shown in Figure 3, reveal systematic misclassifications arising from *semantic overlap*, *contextual ambiguity*, and *class imbalance*.

Semantic Overlap between Classes The most frequent confusions occur between GENERAL-ACTIVITY, MOVEMENT, and OTHERS. For instance, in the sentence, "She ran to help the injured man", the correct label for *"ran"* is MOVEMENT, but the model predicted GENERAL-ACTIVITY. This confusion arises from overlapping semantics between voluntary actions and motion events, a challenge also noted in earlier cross-lingual studies (D'Odorico, 2013)

Cognitive–Communicative Confusions RoBERTa sometimes misclassified COGNITIVE-MENTAL-STATE as COMMUNICATION in several cases. For example, in the sentence, "He thought about telling her the truth", the trigger *"thought"* (true label COGNITIVE-MENTAL-STATE) was incorrectly tagged as COMMUNICATION. This confusion likely stems from the model's difficulty in differentiating internal cognition from externally expressed dialogue contexts.

Rare-Class Misclassifications Underrepresented categories such as LIFE-EVENT and CONFLICT exhibit high confusion with GENERAL-ACTIVITY and OTHERS. For instance, in the sentence, "The king's death changed the fate of the kingdom", the trigger *"death"* (true class LIFE-EVENT) was occasionally misclassified as OTHERS. The sparse distribution of such classes restricts the model's ability to learn discriminative representations for infrequent event types.

Overall, RoBERTa's contextual encoding performs robustly for frequent event categories like COMMUNICATION and GENERAL-ACTIVITY, but struggles with semantically diffuse and low-resource classes. Future improvements could include hierarchical class organization and semantic role augmentation to minimize confusion among conceptually adjacent event types.

Cross-Model Observations. A comparative review of ELECTRA and RoBERTa reveals complementary strengths in event understanding. ELECTRA excels in token-level precision, effectively capturing explicit and well-localized event triggers, while RoBERTa demonstrates superior capability in modeling higher-level event semantics and discourse context for accurate classification. However, both models exhibit sensitivity to narrative ambiguity, class imbalance, and compositional phrase structures. These observations suggest that integrating detection and classification through a joint or multi-task framework could further improve consistency across event boundaries and semantic types, paving the way for end-to-end event understanding in narrative texts.

7. Conclusion

This paper presented *Vrittanta-EN*, a benchmark dataset for event trigger detection and classification in English short stories. The dataset contains 11,272 manually annotated event instances capturing diverse cognitive, communicative, and physical event expressions typical of narrative discourse. Extensive experiments were conducted using classical, neural, transformer-based, and large language models under zero-shot and few-shot settings. Among all models, ELECTRA achieved the best performance in event trigger detection, while RoBERTa performed best in event classification, establishing strong baselines for future research. Error analysis revealed challenges in handling multiword and context-dependent triggers, as well as underrepresented event categories.

The *Vrittanta-EN* corpus marks a step toward bridging structural and semantic event understanding in narrative English. Future work will focus on addressing class imbalance, exploring joint mod-

eling of detection and classification, incorporating discourse-level cues, and extending the framework to multilingual and cross-domain contexts.

Ethical Considerations

The *Vrittanta-EN* dataset is developed entirely from publicly available short stories that are in the open domain and free from copyright restrictions. All texts were sourced from verified repositories and manually screened to exclude offensive, discriminatory, or personally identifiable content. The dataset focuses exclusively on literary narratives and does not include any sensitive or private information. The dataset is intended strictly for research and educational use to advance event understanding in narrative text. The authors acknowledge the importance of responsible dataset release and encourage users to comply with fair-use and data ethics standards when employing *Vrittanta-EN* in downstream applications.

Limitations

Despite the comprehensive design and evaluation of *Vrittanta-EN*, several limitations remain. First, the dataset is confined to the domain of short stories and may not fully capture the diversity of event expressions present in other genres such as news, conversational dialogue, or social media text. As a result, the generalizability of models trained on this corpus to other narrative or non-narrative domains may be limited.

Second, while annotation consistency was ensured through detailed guidelines and inter-annotator agreement checks, subjective interpretations of abstract or implicit events, particularly those reflecting psychological or cognitive states, may introduce minor ambiguities.

Third, large language models (LLMs) were evaluated under zero-shot and five-shot settings without explicit fine-tuning, which might not reflect their full potential in domain-adapted scenarios. Finally, the current dataset focuses on event trigger detection and classification but does not include argument or temporal relation annotations. Extending the framework to incorporate these aspects would further enhance its applicability for downstream event understanding and narrative reasoning tasks.

Data and Code Availability

The *Vrittanta-EN* dataset, including annotation guidelines, preprocessing scripts, and baseline implementations, is available upon reasonable request to the corresponding author. All resources are accompanied by detailed documentation describing the data format, annotation schema, and usage instructions to support transparency and reproducibility.

References

Jacqueline Aguilar, Charley Beller, Paul McNamee, Stephanie Strassel, and Zhiyi Song. 2014. A comparison of the events and relations across ace, ere, and tac-kbp. In *Workshop on Events: Definition, Detection, Coreference, and Representation*.

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*.

David Bamman, Ted Underwood, et al. 2020. An annotated dataset of literary entities. In *Proceedings of LREC*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of EMNLP*.

Tommaso Caselli et al. 2017. Event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of ACL*.

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

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17772–17780.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of ACL*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Tommaso D'Odorico. 2013. *An ontological analysis of vague motion verbs, with an application to event recognition*. University of Leeds.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Pengcheng He et al. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL*.

Jin-Dong Kim, Ngan Nguyen, Yue Wang, Jun'ichi Tsujii, Toshihisa Takagi, and Akinori Yonezawa. 2012. The genia event and protein coreference tasks of the bionlp shared task 2011. *BMC bioinformatics*, 13(Suppl 11):S1.

Magnus Knuth, Jens Lehmann, Dimitris Kontokostas, Thomas Steiner, and Harald Sack. 2015. The dbpedia events dataset. In *ISWC (Posters & Demos)*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: A pre-trained biomedical language representation model for biomedical text mining. In *Bioinformatics*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Mike Lewis et al. 2020b. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*.

Sha Li et al. 2021. Future is not one-dimensional: Graph modeling of event structures for event prediction. In *Proceedings of ACL*.

Xiaodong Liu, Zhunchen Luo, and Heyan Huang. 2018. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of EMNLP*.

Yinhan Liu et al. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*.

Teruko Mitamura and Zhengzhong Liu. Overview of tac kbp 2015 event nugget track.

Teruko Mitamura et al. 2016. Overview of tac kbp 2016 event track. In *Text Analysis Conference (TAC)*.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of ACL*.

James Pustejovsky and Amber Stubbs. 2019. Towards a standard for annotating events in text. In *Proceedings of LREC*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Colin Raffel, Noam Shazeer, et al. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. In *JMLR*.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France. Association for Computational Linguistics.

Ozlem Uzuner, Peter Szolovits, and Isaac Kohane. 2006. i2b2 workshop on natural language processing challenges for clinical records. In *Proceedings of the Fall Symposium of the American Medical Informatics Association*. Citeseer.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2005. The ace 2005 (automatic content extraction) evaluation. In *Proceedings of the Linguistic Data Consortium (LDC)*.