Vrittanta-AS: Dataset Development and Benchmarking for Event Trigger Detection and Classification in Assamese

Chaitanya Kirti, Dhrubajyoti Pathak, Ashish Anand, Prithwijit Guha
Indian Institute of Technology Guwahati
Assam, India
{ckirti, drbj153, anand.ashish, pguha}@iitg.ac.in

Abstract

Event trigger detection and classification aim to identify and categorize events within unstructured text. While prior research has primarily focused on news or biomedical corpora, the literary domain, especially short stories, remains largely underexplored. This gap is particularly pronounced for low-resource languages such as Assamese, where limited annotated data and complex narrative structures hinder progress. To address this challenge, we introduce *Vrittanta-AS*, a manually curated Assamese event trigger detection and classification dataset comprising 13,171 annotated events extracted from short stories. The dataset is designed to advance research in information extraction and narrative understanding for low-resource Indian languages. We conduct a comprehensive evaluation using classical machine learning methods, neural sequential architectures, pre-trained transformer models, and large language models (LLMs) on the proposed dataset. Experimental results demonstrate that IndicBERT v2 achieves the highest performance for both event trigger detection (85.86% micro-F1) and classification (65.21% macro-F1). *Vrittanta-AS* serves as an important step toward developing benchmark resources for event trigger detection and classification in Assamese literary text.

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

1. Introduction

Event trigger detection and classification are two crucial and challenging tasks within the domain of information extraction. Event trigger detection focuses on identifying words or phrases in unstructured text that signal the occurrence of specific events, allowing the extraction of structured information from natural language. The extracted information can then be used for various downstream applications, such as information retrieval (Sankepally, 2019), question answering (Souza Costa et al., 2020), text summarization (Zhang et al., 2021), knowledge graph construction (Guan et al., 2023), temporal analysis (Ahmad et al., 2020), and relation extraction (Wen and Ji, 2021). An event is typically defined as a particular occurrence at a specific time and place involving one or more participants and often characterized by a change of state (Peinelt et al., 2020). The event detection and classification tasks are not widely covered in low-resource language such as Assamese. Assamese language (Glottocode: assa1263) is an Indo-Aryan language. Assamese has 15 million native speakers (Census, 2020 (accessed April, 2020) and is the official language of Assam, a state in Northeast India.

Figure 1 illustrates an example sentence for event trigger detection and classification in Assamese. The sentence contains one event, COMMUNICATION, triggered by "কৈছিল" /koisil/ 'told'. Previous research on event extraction tasks mainly focused on datasets from domains such as news and biomedical texts, where event structures tend to be explicit and syntactically straightforward. However, event detection and classification in literary text, especially children's short stories, remain relatively unexplored. This gap is particularly significant for Assamese, where limited annotated corpora and the narrative complexity of literary texts hinder model generalization. Detecting events in children's short stories is particularly challenging due to the unique narrative structures and the intended young audience. These tales often include dialogues between animals or even inanimate objects. The depiction of events in these narratives is deeply intertwined with the story's genre and style.

কুকুৰে খেলিবলৈ ভাল পাইছিল , সেয়েহে তেওঁ লগে লগে হয় বুলি **কৈছিল** । COMMUNICATION
/kukure khelibolɔj bʰal paɪsil, sejɛhe tɛw lɔge lɔge hɔj buli kɔɪsil/
The dog liked to play, so he immediately said yes.

Figure 1: Example illustrating the event trigger and its class. Bold word represents event trigger, and the corresponding box denotes its class.

To address this gap, we introduce *Vrittanta-AS*, a manually curated Assamese event trigger detection and classification dataset extracted from short stories. The dataset spans a wide range

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7581–7591
11–16 May 2026.

of literary themes, authors, and stylistic variations, capturing diverse linguistic and narrative expressions of events. It is designed to foster research in information extraction, narrative understanding, and language modeling for low-resource Indian languages. To maintain annotation consistency and reliability, comprehensive annotation guidelines were designed to capture narrative context, event semantics, and the morphological complexity of Assamese.

We further benchmark the dataset using a comprehensive suite of models, ranging from classical machine learning algorithms to neural sequential architectures, pre-trained transformer models, and large language models (LLMs).

The key contributions of this paper are summarized as follows:

- We present *Vrittanta-AS*, a manually annotated Assamese dataset for event trigger detection and classification in the literary domain.
- Comprehensive annotation guidelines are designed to capture events in Assamese short stories.
- Baselines are established using classical, neural, and transformer-based approaches, and the performance of multiple pre-trained language models and LLMs is evaluated.
- The strengths, limitations, and error patterns observed by the best-performing model are analyzed. Insights are provided to guide future research in low-resource event extraction.

Our contributions provide valuable linguistic and computational resources to the research community in the form of a curated dataset, annotation guidelines, trained models, and baseline evaluations. As Assamese is a morphologically rich and inflectional language, the findings of this study can also assist resource development and event extraction methodologies for other low-resource Indian languages sharing similar linguistic characteristics.

2. Background and Related Work

Event trigger detection and classification have long been central to information extraction research, aiming to identify event triggers and assign them to predefined event types. Early approaches relied on manually crafted linguistic and lexical patterns (Hogenboom et al., 2011; Cao et al., 2018), which demanded extensive expert knowledge and lacked scalability across domains. Traditional machine learning methods later emerged, where handcrafted lexical, syntactic, and dependency features were used to train classifiers such as Support Vector Machines and Conditional Random Fields (Li et al., 2013; Hong et al., 2011).

With the advent of deep learning, neural architectures such as Convolutional Neural Networks (CNNs) (Chen et al., 2015; Zeng et al., 2014) and Recurrent Neural Networks (RNNs) (Wang, 2018) significantly advanced event detection. CNNs effectively captured local contextual cues, while RNNs and their variants, especially Bidirectional Long Short-Term Memory (BiLSTM) networks, modeled longer dependencies and sequential context. Hybrid models combining CNNs and BiLSTMs further improved representation learning by leveraging both local and global semantic information (Feng et al., 2018). Subsequent frameworks such as JMEE (Liu et al., 2018) and JRNN (Nguyen et al., 2016) employed attention and graph-based architectures to jointly extract event triggers and arguments, achieving notable improvements over earlier models.

The introduction of the Transformer architecture (Vaswani et al., 2017) marked a paradigm shift in NLP. Transformers utilize multi-head self-attention to capture complex dependencies without recurrence, enabling efficient contextual representation learning. They have since become foundational for event detection (Qin et al., 2021), especially when trained on large-scale corpora.

Pre-trained Language Models (PLMs) such as BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu et al., 2019) have revolutionized NLP by leveraging large-scale unsupervised pre-training followed by supervised fine-tuning. PLMs dynamically capture contextual meaning, eliminating the limitations of static embeddings like Word2Vec and FastText. Their success has inspired multilingual variants such as mBERT, XLM-R, MuRIL, and IndicBERT, which extend coverage to Indian languages. IndicBERT v2, in particular, offers robust representations for Indic scripts and low-resource languages by incorporating data from diverse linguistic families. More recently, large language models such as GPT (Radford et al., 2018) have extended this paradigm through generative pre-training and zero-shot generalization.

Despite these advances, most prior work on event detection remains concentrated on high-resource languages and domains such as newswire and biomedical texts (Yang et al., 2019; Meghatria et al., 2020). The literary domain, characterized by implicit events, figurative expressions, and complex narrative dependencies, remains largely neglected. Moreover, Assamese has received very limited attention in the context of event detection and classification. This scarcity

of data and models constrains research on structured event understanding in Assamese literature.

3. Dataset Construction

This section outlines the methodology for collecting and annotating the dataset for event detection and classification. It describes each stage of the process, from guideline development to annotation, and includes illustrative examples for transparency and reproducibility. The section also presents key dataset statistics, highlighting its structural and distributional characteristics.

3.1. Dataset Overview

The proposed dataset *Vrittanta-AS* is manually annotated for event detection and classification that includes 13,171 events extracted from 200 Assamese short stories.

Motivated by the previous study (Sims et al., 2019), which benchmarked models in 100 texts, this work also introduces a sufficiently large dataset to allow meaningful benchmarking of our datasets. The short stories have been collected from diverse sources representing various narrative structures and stylistic forms, ensuring broad coverage of linguistic and contextual variations across the Assamese literary domain.

3.2. Collection of Short Stories

The *Vrittanta-AS* corpus was compiled from eight classic Indian storytelling sources—Panchtantra, Champak, Tenali Raman, Akbar–Birbal, Betal Pachisi, Hitopadesh, Jataka Tales, and Singhasan Battisi. These texts collectively represent India's moral, philosophical, and didactic narrative tradition. The stories span diverse contexts, combining folklore, wisdom, and humor, which introduces linguistic and contextual variability in event representation. This diversity makes event detection challenging due to implicit actions and culturally grounded expressions.

3.3. Annotation Guidelines

Event triggers in Assamese short stories are primarily expressed through grammatical categories such as verbs and their combinations. In some cases, nouns, adjectives, or auxiliaries attached to verbs also function as event-inducing elements. These variations are carefully annotated to capture both single-word and multi-word triggers that signify real-world occurrences. We define seven event classes in the dataset, each representing a distinct semantic type of occurrence. The classes are as follows:

COMMUNICATION (COM): It represents acts of speech or dialogue where one or more participants exchange information.
GENERAL-ACTIVITY (GEN): It encompasses routine or habitual human actions observed in everyday life.
MOVEMENT (MOV): It includes events denoting motion or change of location through walking, flying, or traveling.
COGNITIVE-MENTAL-STATE (CMS): It covers mental and emotional states such as thinking, remembering, perceiving, or feeling.
LIFE-EVENT (LE): It represents significant life occurrences such as birth, illness, injury, death, or marriage.
CONFLICT (CON): It captures disagreements, confrontations, or fights, either verbal or physical.
OTHERS (OTH): It serves as a residual category for events that do not clearly belong to the above classes.

3.4. Annotation Process

The annotation of events in selected Assamese texts were carried out following the compiled annotation guidelines. The BRAT annotation tool (Stenetorp et al., 2012) is used for annotation. Two co-authors of this paper, both native Assamese speakers with linguistic expertise, performed the annotation. They underwent a calibration phase to establish a shared understanding of event semantics and boundary identification before annotating the full corpus. This process ensured annotation uniformity and reduced subjectivity.

To assess inter-annotator agreement (IAA), 20 randomly selected stories containing 753 events were independently annotated by both annotators. The resulting IAA score was 92.5%, indicating strong consistency in identifying event triggers and their corresponding classes. Each annotated event trigger was assigned to one of seven predefined event categories. The combination of linguistic expertise, detailed guidelines, and double-annotation validation contributed to the overall reliability of the dataset.

3.5. Annotated Dataset Statistics

Table 1 summarizes the quantitative characteristics of *Vrittanta-AS*. The dataset contains 139,214 tokens and 13,861 sentences across 200 stories, with an average of 66 annotated events per story. The relatively high number of sentences and unique tokens highlights the linguistic richness of Assamese short stories.

The class-wise distribution of event types is shown in Table 2. The most frequent categories are COMMUNICATION and

Sl.No	Statistics	Count
1.	Total tokens in the dataset	139,214
2.	Total unique tokens in the dataset	14,410
3.	Total sentences in the dataset	13,861
4.	Average tokens per story	786
5.	Average sentences per story	69
6.	Average tokens per sentence	11
7.	Total events in the dataset	13,171
8.	Average events per story	66

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

Sl.No	Event Type	Count
1.	COMMUNICATION	3,052
2.	GENERAL-ACTIVITY	1,811
3.	MOVEMENT	1,146
4.	COGNITIVE-MENTAL-STATE	1,824
5.	LIFE-EVENT	178
6.	OTHERS	2,804
7.	CONFLICT	67

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

GENERAL-ACTIVITY, reflecting the dialogic and action-oriented nature of children's narratives. COGNITIVE-MENTAL-STATE events also occur frequently, representing introspection and moral reflection, while CONFLICT and LIFE-EVENT appear less often, consistent with the instructive tone of these stories. This inherent scarcity leads to a class imbalance, which is a consequence of the domain's narrative characteristics rather than a sampling flaw. Consequently, models show higher error dispersion on these low-frequency classes, which is a key challenge for future research.

Table 3 lists the top 10 most frequent event triggers and their event rates. The high rates for triggers such as "কৈছিল" /koisil/ (said), "ক'লে" /kole/ (said), and "সুধিলে" /xud^{h}ile/ (asked) confirm the prominence of communication-based events in Assamese literature, especially within dialogue-heavy narratives.

The frequency distribution of trigger spans based on the number of words are computed to analyze the linguistic characteristics of event triggers. As shown in Figure 2, most triggers are composed of one or two words, with a sharp decline beyond three words. The dataset contains 4,320 one-word triggers and 5,267 two-word triggers, indicating that Assamese events are typically expressed through concise verbal expressions or short verb phrases. Multi-word triggers longer than four tokens are rare, reflecting the syntactic compactness of event expressions in Assamese narrative discourse.

Trigger Word	Count	Event Rate
কৈছিল	906	96%
ক'লে	762	95%
সুধিলে	364	97%
গৈছিল	185	92%
কয়	154	81%
অনুভৱ কৰিছিল	133	91%
দেখিছিল	116	90%
ভাবিছিল	101	91%
সুধিছিল	94	97%
আহিছিল	91	87%

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

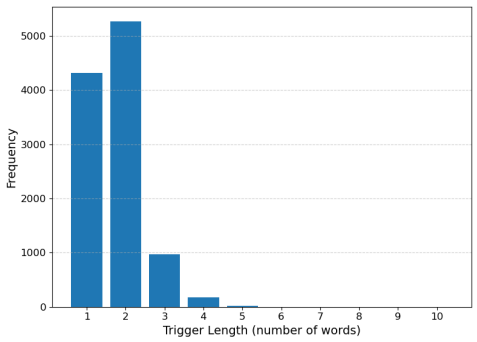

Figure 2: Trigger length (number of words) vs. frequency plot for *Vrittanta-AS*.

4. Dataset Evaluation

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

4.1. Task Description

Event detection and classification aim to identify event trigger expressions in text and assign them to one of several predefined event types. In the *Vrittanta-AS* dataset, each token in a sentence is annotated with an event boundary and type label. We formulate this as a sequence labeling problem using the standard BIO tagging scheme. Given an input sentence $\mathbf{S} = [w_1, w_2, \ldots, w_n]$ consisting of n tokens, the objective is to predict a corresponding label sequence $\mathbf{L} = [l_1, l_2, \ldots, l_n]$, where each label $l_i \in O, B\text{-}X, I\text{-}X$ and X denotes the event type. Here, $B\text{-}X$ marks the beginning of an event trigger of type X, $I\text{-}X$ marks tokens inside the trigger span, and O marks tokens outside any event boundary.

4.2. Evaluation Models

We evaluate the dataset using a diverse set of models, ranging from traditional machine learning algorithms to transformer-based and large language models.

Classical models Statistical classifiers such as Support Vector Machine (SVM) and Naïve Bayes (NB) were employed as initial baselines. These models rely on handcrafted lexical and contextual features to identify event trigger tokens and are effective for small-scale, feature-rich datasets.

Neural sequence models To capture sequential dependencies, we experimented with Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and BiLSTM with a Conditional Random Field layer (BiLSTM+CRF). The BiLSTM+CRF architecture achieved the best token-level sequence labeling performance among neural models, establishing a strong baseline encoder for subsequent transformer-based experiments.

Pre-trained language models To leverage contextual representations, we evaluated several encoder-only transformers pre-trained using the Masked Language Modeling (MLM) objective: multilingual BERT (mBERT) (Kenton and Toutanova, 2019), MuRIL (Khanuja et al., 2021), XLM-RoBERTa (XLM-R) (Conneau et al., 2020), IndicBERT (Kakwani et al., 2020), IndicBERT v2 (Doddapaneni et al., 2023), IndicBART (Dabre et al., 2022), and AxomiyaBERTa (Nath et al., 2023), a monolingual Assamese BERT model trained on the IndicCorpV2 Assamese subset. mBERT is trained on Wikipedia data from 104 languages, enabling strong cross-lingual generalization. XLM-RoBERTa is trained on 2.5 TB of CommonCrawl data from 100 languages and provides improved multilingual performance, particularly for low-resource settings. MuRIL (Khanuja et al., 2021) is optimized for Indian languages using both monolingual and parallel corpora in 17 languages. IndicBERT and IndicBERT v2, based on the ALBERT architecture (Lan et al., 2019), are trained on the large-scale IndicCorp (Madanbhavi et al., 2024) and IndicCorpV2 corpora and have demonstrated state-of-the-art performance on the IndicXTREME benchmark. IndicBART is a multilingual sequence-to-sequence model designed for Indian language generation and understanding tasks.

Large language models To evaluate zero-shot and few-shot generalization without fine-tuning, we tested DeepSeek-V3.2-Exp [1] and GPT-5-mini [2] using prompt-based setups. For each model, two configurations were tested: zero-shot (0-shot) inference, where the model directly predicts event types from plain text, and few-shot (5-shot) inference, where five in-context examples from *Vrittanta-AS* were provided within the prompt. A 5-shot per class setup (35 examples in total across 7 event types for event classification) was adopted to ensure balanced exposure of all classes during in-context learning (Schick and Schütze, 2021). The complete set of prompts can be accessed here. These evaluations measure how well frontier LLMs can adapt to Assamese event detection without supervised optimization.

[1] https://api-docs.deepseek.com/news/news250929

[2] https://platform.openai.com/docs/models/gpt-5-mini

Model	P	R	F1
SVM	68.67	84.51	75.77
NB	63.22	93.29	75.37
LSTM	68.74	64.28	66.44
BiLSTM	80.11	79.95	80.03
BiLSTM+CRF	81.25	78.56	79.88
mBERT	80.76	88.56	84.48
MuRIL	79.33	90.65	84.61
XLM-RoBERTa	80.05	87.64	83.67
IndicBERT	81.58	77.00	79.22
IndicBERT v2	86.92	84.82	**85.86**
IndicBART	74.25	89.60	81.20
AxomiyaBERTa	76.84	87.25	81.71
DeepSeek-V3.2-Exp (0-shot)	25.54	38.65	30.75
DeepSeek-V3.2-Exp (5-shot)	35.51	30.31	32.70
GPT-5-mini (0-shot)	24.78	28.96	26.72
GPT-5-mini (5-shot)	29.37	32.18	30.73

Table 4: Performance comparison of different models on *Vrittanta-AS* for event trigger detection task. The best F1-score is highlighted in bold.

4.3. Experimental Setup

All classical and neural models were implemented in Python using Scikit-learn and PyTorch. Transformer-based architectures were fine-tuned using the HuggingFace Transformers library with the AdamW optimizer. Each model was trained for 30 epochs with a batch size of 32 and a learning rate of 2e-5, employing early stopping based on validation F1-score. The dataset was divided into training, validation, and test splits in a 60:10:30 ratio. For event trigger detection, performance was measured using micro-averaged precision, recall, and F1-score, while for event classification, both micro and macro averages were reported to address class imbalance across event categories.

5. Results and Discussion

This section presents the results and analysis of model performance across event detection and event classification tasks on the *Vrittanta-AS* dataset.

Model	Metric	COM	GEN	MOV	CMS	LE	OTH	CON	Macro F1
SVM	**P**	53.02	38.46	45.45	65.21	0.00	37.55	0.00	
	R	21.10	4.62	19.32	26.38	0.00	28.34	0.00	
	F1	30.19	8.26	27.11	37.56	0.00	32.30	0.00	19.35
Naive Bayes	**P**	47.91	0.00	0.00	78.37	0.00	45.45	0.00	
	R	9.77	0.00	0.00	7.28	0.00	7.82	0.00	
	F1	16.23	0.00	0.00	13.33	0.00	13.35	0.00	6.13
LSTM	**P**	83.70	34.04	57.70	63.21	43.13	39.87	50.00	
	R	80.78	31.01	47.20	60.40	21.35	31.48	15.38	
	F1	82.20	32.45	51.92	61.78	28.57	35.18	23.52	45.09
BiLSTM	**P**	80.70	45.74	60.76	70.68	47.22	54.43	55.00	
	R	89.00	43.39	60.40	69.39	49.51	50.51	28.20	
	F1	84.65	44.53	60.58	70.03	48.34	52.40	37.28	56.83
BiLSTM-CRF	**P**	81.70	46.60	61.90	71.70	48.10	55.30	56.20	
	R	90.20	44.30	61.50	70.50	50.50	51.40	28.80	
	F1	85.73	45.43	61.70	71.10	49.27	53.28	38.08	57.80
mBERT	**P**	85.35	41.78	60.56	70.82	55.88	57.76	55.00	
	R	92.55	52.50	69.18	82.27	58.16	52.21	30.55	
	F1	**88.80**	46.53	64.58	76.12	57.00	54.85	39.28	61.02
MuRIL	**P**	85.87	41.61	60.92	73.20	52.59	59.74	0.00	
	R	94.34	58.69	79.20	87.03	68.93	63.39	0.00	
	F1	89.91	48.70	**68.86**	**79.52**	59.66	61.51	0.00	58.31
XLM-RoBERTa	**P**	85.17	50.14	62.34	72.42	58.53	59.34	57.14	
	R	92.94	49.64	72.78	81.38	69.90	61.66	30.76	
	F1	88.88	49.89	67.16	76.64	**63.71**	60.48	40.00	63.82
IndicBERT	**P**	86.58	23.34	57.76	69.15	56.60	45.70	50.00	
	R	82.03	40.06	54.52	59.88	30.61	47.91	5.55	
	F1	84.25	29.49	56.09	64.18	39.73	46.78	10.00	47.22
IndicBERT v2	**P**	86.20	54.28	59.84	71.27	50.00	58.81	44.68	
	R	93.11	52.01	76.00	84.41	66.99	69.00	53.84	
	F1	89.52	**53.12**	66.96	77.29	57.26	**63.50**	**48.83**	**65.21**
IndicBART	**P**	81.75	41.87	57.71	63.85	57.79	58.52	51.72	
	R	92.77	53.56	68.53	87.78	64.28	56.51	41.66	
	F1	86.91	47.00	62.66	73.92	60.86	57.50	46.15	62.14
AxomiyaBERTa	**P**	85.99	37.54	53.70	72.74	56.38	59.32	56.52	
	R	91.45	58.11	75.00	77.36	54.08	48.72	36.11	
	F1	88.64	45.62	62.58	74.98	55.20	53.50	44.06	60.65
DeepSeek-V3.2-Exp (0-shot)	**P**	6.54	7.38	12.78	15.56	7.40	0.00	8.16	
	R	18.70	7.23	23.60	13.87	1.94	0.00	10.25	
	F1	9.70	7.30	16.58	14.67	3.07	0.00	9.09	8.63
DeepSeek-V3.2-Exp (5-shot)	**P**	9.50	10.00	18.20	21.20	9.10	3.30	9.80	
	R	26.00	9.80	33.00	18.00	2.70	3.30	12.60	
	F1	13.91	9.90	23.46	19.47	4.16	3.30	11.03	12.18
GPT-5 mini (0-shot)	**P**	7.30	8.00	14.00	16.70	8.00	2.50	8.90	
	R	20.50	7.90	25.50	15.00	2.20	2.50	11.30	
	F1	10.75	7.95	18.08	15.80	3.45	2.50	9.96	9.78
GPT-5 mini (5-shot)	**P**	8.61	9.44	16.52	19.71	9.44	2.95	10.50	
	R	24.19	9.32	30.09	17.70	2.60	2.95	13.33	
	F1	12.70	9.38	21.33	18.65	4.07	2.95	11.75	11.55

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

5.1. Event Detection

Table 4 presents the results of event detection across classical, neural, and transformer-based models, including zero-shot and few-shot large language models. Among the classical approaches, Naïve Bayes achieved the highest recall (93.29%) but at the cost of precision, while SVM demonstrated a more balanced trade-off with an F1-score of 75.77%. These models, however, rely heavily on lexical co-occurrence features and lack the ability to generalize across varying narrative contexts in Assamese text.

Neural sequence models offered a significant improvement by capturing sequential and contextual dependencies. The BiLSTM model achieved an F1-score of 80.03%, outperforming the simpler LSTM architecture (66.44%). The BiLSTM-CRF model performed comparably (79.88%), showing that explicit transition modeling through the CRF layer offers marginal gains in boundary consistency. This indicates that Assamese event triggers, often multiword and morphologically rich, benefit from bidirectional contextual encoding.

Transformer-based pre-trained language models (PLMs) achieved the highest overall performance. IndicBERT v2 obtained the best F1-score of 85.86%, surpassing MuRIL (84.61%) and XLM-RoBERTa (83.67%). The superior performance

of mBERT can be attributed to its broader multilingual coverage and stable cross-lingual embeddings, which align well with Assamese syntax and morphology despite its absence in the original training corpus. mBERT (84.48%) and IndicBART (81.2%) also performed competitively, benefiting from domain-specific pretraining on Indic languages, while AxomiyaBERTa, a monolingual Assamese model, achieved an F1-score of 81.71%, highlighting the effectiveness of in-language pretraining even on a smaller corpus.

The zero-shot and few-shot results from DeepSeek-V3.2-Exp and GPT-5-mini were significantly lower, with DeepSeek-V3.2-Exp (5-shot) achieving 32.7% and GPT-5-mini (5-shot) 30.73% F1. Despite their general reasoning capabilities, both models struggled to identify event triggers consistently in morphologically complex and low-resource Assamese text without explicit supervision. In summary, the results indicate that while PLMs offer substantial improvements, further fine-tuning on narrative-domain Assamese text is essential to achieve robust event boundary detection.

5.2. Event Classification

The event classification task, which requires assigning detected triggers to one of seven semantic event categories, exhibited a similar performance hierarchy. As shown in Table 5, classical models such as SVM and Naïve Bayes performed poorly, with macro F1-scores of 19.35% and 6.13%, respectively, due to their inability to capture complex contextual and semantic relations between event triggers and surrounding words.

Among neural models, BiLSTM and BiLSTM-CRF demonstrated notable improvements, with macro F1-scores of 56.83% and 57.80%, respectively. The bidirectional structure allowed these models to integrate contextual information from both preceding and succeeding words—an essential property for disambiguating event classes such as COMMUNICATION and COGNITIVE-MENTAL-STATE. The CRF layer offered consistent label transitions but did not yield large gains, suggesting that the inherent variability of literary narratives may limit the effect of sequence-level constraints.

Pre-trained transformer models achieved the best overall performance. IndicBERT v2 achieved the highest macro F1-score of 65.21%, followed closely by XLM-RoBERTa (63.82%) and IndicBART (62.14%). These results demonstrate the advantage of Indic-specific pretraining, as IndicBERT v2 captures subword-level morphological and syntactic nuances unique to Assamese. mBERT (61.02%) and AxomiyaBERTa (60.65%) also performed strongly, with mBERT

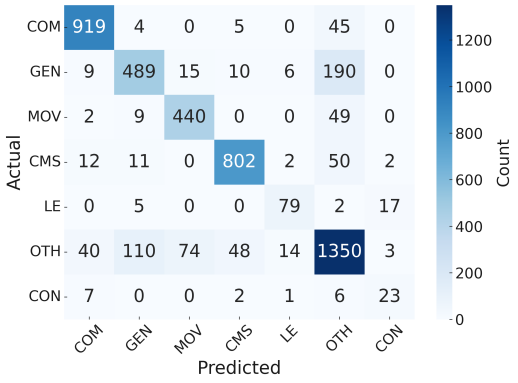

Figure 3: Confusion matrix of the best performing model (IndicBERT v2) on the event classification task.

excelling in COMMUNICATION (F1 = 88.80%) class, while MuRIL performed well on MOVEMENT and COGNITIVE-MENTAL-STATE categories. IndicBERT v2 demonstrated improved performance for the previously misclassified OTHERS and CONFLICT classes.

Large language models under zero-shot and few-shot prompting conditions underperformed significantly, with macro F1-scores between 8.63% and 12.18%. DeepSeek-V3.2-Exp (5-shot) slightly outperformed its zero-shot variant, indicating that in-context demonstrations provide minor benefits but remain insufficient without token-level fine-tuning. GPT-5-mini (5-shot) improved the results slightly but lacked consistent event-type mapping, often misclassifying subtle psychological and discourse events, probably due to limited exposure to Assamese linguistic patterns.

Overall, these findings confirm that contextual pretraining on Indic languages is crucial for event understanding in Assamese. While LLMs demonstrate emerging potential for low-resource inference, domain-specific fine-tuning and integration with structured annotation signals remain necessary to achieve competitive performance on event classification tasks.

6. Error Analysis

IndicBERT v2 achieved the highest F1-scores for both event detection and classification. The corresponding confusion matrix is shown in Figure 3. The model demonstrates strong class-wise accuracy, with high true positives for COMMUNICATION (919), OTHERS (1350), and COGNITIVE-MENTAL-STATE (802). Moderate confusion is observed for GENERAL-ACTIVITY (489) and MOVEMENT (440), which are often misclassified as OTHERS. Low-frequency classes such as LIFE-EVENT (79) and CONFLICT (23) exhibit higher error dispersion, reflecting data imbal-

ance.

A qualitative analysis of the model's predictions reveals several consistent error patterns. These errors primarily stem from Assamese morphological complexity, contextual ambiguity, and semantic overlap between event categories. Representative examples from the prediction outputs are provided below to illustrate these challenges. The following portion categorizes these errors for clearer understanding.

Boundary Detection Errors The model often failed to correctly detect the full span of multi-word event triggers, especially in compound verb phrases common in Assamese. For example, in the sentence , "তেওঁ মল্লযোদ্ধাজনক বাহিৰ লৈ যোৱাৰ বিষয়ে সম্পূৰ্ণ নিশ্চিত আছিল ।" /tɛo mɔllɔyodʰyajɔnɔk bahir loi joar bixoye sampurnɔ nicchitɔ asil/ (*He was absolutely sure about taking the wrestler out)*, the model detected only the beginning of the trigger and missed the auxiliary components "লৈ যোৱাৰ" /loi joar/ (to take out), resulting in a partial boundary mismatch.

Event Type Confusion IndicBERT v2 sometimes confused semantically close event types, particularly between GENERAL-ACTIVITY, OTHERS, and COGNITIVE-MENTAL-STATE. For instance, in the sentence "তেওঁ মল্লযোদ্ধাজনক পৰাজিত কৰিছিল ।" /tɛo mɔllɔyodʰyajɔnɔk pɔrajitɔ korisil/ (*He defeated the wrestler*), the event trigger "পৰাজিত কৰিছিল" /pɔrajitɔ korisil/ (*defeated*) belongs to GENERAL-ACTIVITY, but it is predicted as OTHERS. In an another example sentence, "ৰজাই লজ্জিত অনুভৱ কৰিছিল ।" /rɔjai lɔjjitɔ ɔnubʰɔb kɔrisil/ (*The king felt ashamed*), the event trigger "অনুভৱ কৰিছিল" /anubʰɔb kɔrisil/ (*felt*) belongs to COGNITIVE-MENTAL-STATE, but it is predicted as OTHERS.

The model tends to underperform when emotional or mental cues are contextually subtle or dispersed across the sentence. This broad scope of GENERAL-ACTIVITY and the inclusive nature of OTHERS often lead to substantial confusion among annotators, resulting in a higher incidence of misclassification between these two classes.

Several false classifications occurred when generic verbs with wide semantic coverage were used in ambiguous contexts. For example, in the sentence, "তেওঁ মল্লযুদ্ধৰ বিষয়ে চিন্তা কৰিছিল ।" /tɛo mɔllɔyodʰyar bixoye sinta korisil/ (*He was thinking about wrestling*), the event trigger "চিন্তা কৰিছিল" /sinta korisil/ (*was thinking*), belongs to COGNITIVE-MENTAL-STATE, but it is misclassified to OTHERS. The model misinterpreted the mental activity verb "চিন্তা কৰিছিল" /sinta korisil/ (*was thinking*), as a generic activity due to context truncation or insufficient semantic cues.

Contextual Ambiguity in Communication Events Communication events were sometimes missed when embedded in quoted or indirect speech structures. For evidence, in the sentence "হালকুৱে কৈছিল , 'তেতিয়া মই তেওঁৰ উৎপীড়ন সহ্য কৰিব লাগিব ।" /halkuye koisil, 'tɛtia mɔi tɛor utpidɔn xoyya kɔribɔ lagibɔ/ (*Halku said , 'Then I will have to endure his persecution*), the quotation boundary likely disrupts contextual dependency tracking, leading to missed detections of verbs sucha as "কৈছিল" /koisil/ (*said*).

Morphological Variants and Inflected Verbs IndicBERT v2 showed reduced sensitivity to inflected or dialectal verb forms, which are common in colloquial Assamese. For instance, in the sentence, "মই আপোনাক আলুৰ চিপচ আনিবলৈ কৈছিলো ।" (*I told you to bring potato chips*) /moi aponak alur sips aniblɔı kɔısilu /, the variant "কৈছিলো" /kɔısilu/ (*told*) was missed, suggesting that the tokenizer or pretraining data lacked sufficient morphological coverage.

The analysis shows that while IndicBERT v2 exhibits strong generalization and lexical understanding, it falters in deeper semantic interpretation and context-dependent reasoning. Most errors stem from subtle contextual shifts, long-range dependencies, and overlapping event semantics that challenge even robust multilingual PLMs. The model performs well on explicit event cues but struggles with morphologically complex or implicit triggers. These results underscore the need to integrate richer morphological features, contextual span modeling, and discourse-level reasoning to improve event extraction in low-resource languages like Assamese.

7. Conclusion

This paper introduced *Vrittanta-AS*, a manually curated Assamese dataset for event trigger detection and classification in the literary domain. The corpus comprises 13,171 annotated events from 200 short stories, along with detailed annotation guidelines. We established comprehensive benchmarks spanning classical models (SVM, NB), neural sequence architectures (LSTM, BiLSTM, BiLSTM-CRF), pretrained language models (mBERT, MuRIL, XLM-R, IndicBERT, IndicBERT v2, IndicBART, AxomiyaBERTa), and LLMs under zero-shot/few-shot prompting. Results show that transformer-based PLMs are the most effective: IndicBERT v2 attains the highest F1 for event trigger detection (85.86% micro-F1) as well as for event classification (65.21% macro-F1). Error analyses highlight challenges arising from morphologically complex triggers, discourse/contextual ambiguity, and

semantic overlap between GENERAL-ACTIVITY, OTHERS, and COGNITIVE-MENTAL-STATE. Zero-shot/few-shot LLMs underperform without token-level supervision in this low-resource, narrative setting. Future directions include enhancing multi-word trigger detection through morphology-aware tokenization, incorporating broader discourse context via long-context encoders, and applying semi-supervised learning to address class imbalance.

Ethical Considerations

To properly benchmark the dataset, certain ethical principles must be adhered to. We have taken several actions to ensure the annotation of ethical data. Usually, data sources that contain gender, racial, or community bias are susceptible to producing dangerous results. To mitigate this issue, we have chosen stories from sources that are followed by a wide range of audiences and are devoid of the preceding biases. The stories are openly available so that data source transparency is also achieved. Also, copyright and licensing issues will not affect the creation of the dataset. Annotator biases were mitigated, and the event annotations were double-checked with the assistance of a language specialist.

Limitations

Although *Vrittanta-AS* provides a valuable resource for Assamese event understanding, it is limited by its domain scope and size. The dataset focuses exclusively on short stories, which may not generalize to other genres such as news or social media. Certain event categories remain underrepresented, leading to class imbalance. Moreover, token-level annotations may not fully capture implicit or cross-sentence events, and the current PLMs struggle with long-range context and morphological variation inherent to Assamese. Despite exploring several prompt engineering strategies, more effective formulations may exist that could further boost LLM performance on event trigger detection and classification.

Data and Code Availability

The *Vrittanta-AS* 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

Haseeb Ahmad, Mudassar Ahmad, Waqar Ahmad, Nadeem Faisal, et al. 2020. Extraction of temporal events' frequency from online news channels. In *2020 30th International Conference on Computer Theory and Applications (ICCTA)*, pages 109–116. IEEE.

Kai Cao, Xiang Li, Weicheng Ma, and Ralph Grishman. 2018. Including new patterns to improve event extraction systems. In *The Thirty-First International Flairs Conference*.

Census. 2020 (accessed April, 2020). *ABSTRACT OF SPEAKERS' STRENGTH OF LANGUAGES AND MOTHER TONGUES - 2011*.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M Khapra, and Pratyush Kumar. 2022. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Xiaocheng Feng, Bing Qin, and Ting Liu. 2018. A language-independent neural network for

event detection. *Science China Information Sciences*, 61:1–12.

Saiping Guan, Xueqi Cheng, Long Bai, Fujun Zhang, Zixuan Li, Yutao Zeng, Xiaolong Jin, and Jiafeng Guo. 2023. What is event knowledge graph: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7569–7589.

Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska De Jong. 2011. An overview of event extraction from text. *DeRiVE@ ISWC*, pages 48–57.

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 1127–1136.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

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*.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82.

Xiao Liu, Zhunchen Luo, and He-Yan Huang. 2018. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256.

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*.

Lalitha Madanbhavi, Padmashree Desai, Neha Dhirendra Sirur, Ananya Deshpande, Risheek V Hiremath, and Chetan M Patil. 2024. An efficient multilingual text classification using indiccorp dataset. In *2024 5th IEEE Global Conference for Advancement in Technology (GCAT)*, pages 1–6.

Riadh Meghatria, Chiraz Latiri, and Fahima Nader. 2020. Event nugget detection using pre-trained language models. *Procedia Computer Science*, 176:320–329.

Abhijnan Nath, Sheikh Abdul Mannan, and Nikhil Krishnaswamy. 2023. Axomiyaberta: A phonologically-aware transformer model for assamese. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 300–309.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tbert: Topic models and bert joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055.

Yanxia Qin, Jingjing Ding, Yiping Sun, and Xiangwu Ding. 2021. A transformer-based model for low-resource event detection. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part IV 28*, pages 452–463. Springer.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Rashmi Sankepally. 2019. Event information retrieval from text. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1447–1447.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

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.

Tarcísio Souza Costa, Simon Gottschalk, and Elena Demidova. 2020. Event-qa: A dataset for event-centric question answering over knowledge graphs. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 3157–3164.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Baoxin Wang. 2018. Disconnected recurrent neural networks for text categorization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2311–2320.

Haoyang Wen and Heng Ji. 2021. Utilizing relative event time to enhance event-event temporal relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10431–10437.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5284–5294.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 2335–2344.

Junsheng Zhang, Kun Li, Changqing Yao, and Yunchuan Sun. 2021. Event-based summarization method for scientific literature. *Personal and Ubiquitous Computing*, 25:959–968.