Benchmarking Bangla Causality: A Dataset of Implicit and Explicit Causal Sentences and Cause-Effect Relations

Anonymous ACL submission

Abstract

Causal reasoning is central to language understanding, yet remains under-resourced in Bangla. In this paper, we introduce the first large-scale dataset for causal inference in Bangla, consisting of over 11663 sentences annotated for causal sentence types (explicit, implicit, non-causal) and token-level spans for causes, effects, and connectives. The dataset [1] captures both simple and complex causal structures across diverse domains such as news, education, and health. We further benchmark a suite of state-of-the-art instruction-tuned large language models, including LLaMA 3.3 70B, Gemma 2 9B, Qwen 32B, and DeepSeek, under zero-shot and three-shot prompting conditions. Our analysis reveals that while LLMs demonstrate moderate success in explicit causality detection, their performance drops significantly on implicit and span-level extraction tasks. This work establishes a foundational resource for Bangla causal understanding and highlights key challenges in adapting multilingual LLMs for structured reasoning in low-resource languages.

1 Introduction

Understanding causality in natural language is fundamental to cognitive reasoning and machine comprehension. Causal relationswhere one event (the *cause*) leads to another (the *effect*)(Chan et al., 2002). Detection of causal relation from text has many analytical and predictive applications(Sheikh et al., 2023; Liu et al., 2023; Jiang et al., 2024; Tan et al., 2023; Hershowitz et al., 2024; Zhao et al., 2016; Yu et al., 2019). Few of these are crucial in supporting applications such as explainable question answering, discourse understanding, decision support systems, and narrative generation, detecting cause-effect relations in medical documents, learning about after effects of natural disasters, learning causes for safety related incidents etc. However, to build a meaningful application that can detect an event from texts and predict its possible effects, there is a need to curate large volume of cause-effect event pairs (Sorgente et al., 2013; Blanco et al., 2008; Do et al., 2011; Girju, 2003; Hobbs, 2005; Asghar, 2016; Low et al., 2001; Ittoo and Bouma, 2011).

Causality is often expressed in natural language through explicit markers such as "because," "therefore," "as a result," or implicitly inferred through world knowledge and discourse structure (See Table 3). Identifying causal sentences and extracting their internal cause-effect structure is a two-step problem: a) **Causal Sentence Classification:** Determine whether a given sentence encodes a causal relationship and b) **Cause-Effect Extraction:** Identify the specific spans in the sentence that represent the cause and the effect.

While significant progress has been made in developing computational methods for causal relation extraction in English and other high-resource languages, the same cannot be said for **Bangla**, the seventh most spoken language in the world, with over 230 million speakers. Despite its widespread use, Bangla remains underrepresented in higher-level NLP tasks due to the lack of fundamental linguistic and annotated resources including annotated corpora with sentence-level and span-level causal labels.

This paper presents a comprehensive effort to develop linguistic resources for identifying and extracting causal relations from Bangla text. We describe the construction of a gold-

[1] https://github.com/anonymous-2344/aacl-ijcnlp

Domain	# Sent.	# Causal Sent.	# Implicit Causal Sent.
Politics	4907	3796	142
Editorials	1980	1489	114
Sports	1581	802	136
International	746	602	35
Entertainment	648	345	76
Finance	528	483	13
Science&tech	461	343	27
Story	812	470	281
Total	11663	8330	824

Table 1: Data distribution across domains

standard annotated corpus, the design of annotation guidelines sensitive to Bangla syntax and semantics, and a taxonomy of explicit and implicit causal connectives. Our work aims to establish a foundation for research and applications in Bangla causality understanding and discourse analysis. We did intensive experimentations with parts of the dataset using some of the openly available LLMs, which will be discussed in the following sections.

2 Related Works

Early work on causality detection from text includes (Khoo et al., 1998; Do et al., 2011; Girju, 2003; Hobbs, 2005; Grishman, 1988; Garcia, 1997). Machine learning approaches were introduced in (Bui et al., 2010; Khoo, 1995; Khoo et al., 1998, 2001), with growing emphasis on domain-independence and scalability (Girju et al., 2002; Low et al., 2001; Chan et al., 2002; Bui et al., 2010; Girju et al., 2009). Implicit causality extraction was explored by Ittoo et al. (Ittoo and Bouma, 2011), while Radinsky et al. used statistical inference and clustering to predict events (Radinsky et al., 2012). Deep learning methods emerged in (Xu et al., 2015; Zhao et al., 2016; Dasgupta et al., 2018; Yu et al., 2019; Li et al., 2021), including SCITE, which uses Self-attentive BiLSTM-CRF with Transferred Embeddings. Guo et al. applied unsupervised learning to link pressure injuries with risk factors (Guo et al., 2020) and build causal graphs (Veitch et al., 2019). Surveys on causal relation extraction are available in (Asghar, 2016; Yang et al., 2021).

3 Data Collection Methodology

The construction of our Bangla causality corpus followed a structured multi-phase methodology involving corpus design, sentence selection, annotation protocol development, and quality assurance. The primary objective was to create a representative dataset for training and evaluating models capable of identifying causal sentences and extracting cause-effect relations in Bangla.

We compiled text data from diverse public sources to ensure linguistic and topical variety. News articles from major Bangla outlets like *Prothom Alo*[2] contributed event-driven and policy-oriented content. Formal language was captured from Bangla Wikipedia [3] and educational texts. To include implicit causal narratives, we sourced childrens storybooks and folk tales [4]. Contemporary informal usage was represented through selected social media posts. After preprocessing and deduplication, the final corpus comprised approximately 11,663 clean sentences. Details of the dataset is reported in Table 1.

Preprocessing: We perform a number of preprocessing over the collected dataset. The first stage of preprocessing involves identifying which sentences are probably candidates for cause-effect identification out of a body of text. This involves looking for the presence of at least one causal connective in the sentence under consideration. Following the work of (Xuelan and Kennedy, 1992) and (Blanco et al., 2008) we create an initial list of 27 Bangla causal connectives (see Table 5). We further expand the list by adding common phrases that contain one or more of these words. For example, the seed word **কারণ** is extended to include phrases like **কারণে, সে কারণে, যার কারণে, কারণ হলো, যেসব কারণে** etc. This gives us an extended connective list of 310 words/phrases. Table 6 shows a few examples of seed words and new terms added to the list. After preprocessing, we finally obtained a dataset of 18K sentences for annotation in terms of their cause, effect and causal connectives.

To construct a balanced dataset, we conducted an initial round of binary classification where three annotators labeled 2,000 randomly sampled sentences as causal or non-causal. Based on the estimated causal incidence rate of approximately 30-35%, we per-

[2]https://www.prothomalo.com/archive
[3]https://bn.wikipedia.org/wiki/
[4]https://rabindra-rachanabali.nltr.org/node/6584

formed stratified sampling to create a dataset of 11,663 sentences, ensuring sufficient representation of causal phenomena. Each sentence was assigned a unique identifier and tagged with its source type and domain metadata.

The Annotation Process: The above sentences are presented to six expert annotators. The experts were asked to complete the following two tasks. a) Identify whether a given sentence contains a causal event (either cause/effect) and b) Annotate each word in a sentence in terms of the four labels cause (C), effect(E), causal connectives(CC) and None. An illustration of the annotated dataset is depicted in Table 4.

The annotation framework was organized into two layers. At the sentence level, annotators assigned a binary label indicating the presence or absence of a causal relationship. For sentences marked as causal, annotators further identified specific textual spans corresponding to the cause and effect. These were annotated using the standard BIO format, assigning tags such as `B-Cause`, `I-Cause`, `B-Effect`, and `I-Effect` to relevant tokens, while the remaining tokens were labeled as `O`.

Annotation was performed using DocAnno tool[5] supporting multi-annotator workflows, token-level highlighting, and review logs. Annotators were trained through detailed orientation sessions and practiced on a pilot set before production annotation. The connectives were categorized by their function (cause-introducing or result-introducing), grammatical role (e.g., conjunctions, adverbials), and frequency in the corpus. Both explicit forms such as **কারণ** and **ফলে**, and implicit indicators including idiomatic expressions and clause-level cues, were included.

In some of the candidate sentences, it is observed that a single sentence contains multiple cause-effect pairs, some of which are even chained together. In order to handle multiple instances of causality present in the same sentence, sentences are split into sub-sentences. For example, **উন্নয়নশীল দেশে মোট রোগের প্রায় পাঁচ ভাগের চার ভাগই জলবাহিত রোগের কারণে ঘটে, এবং শিশুমৃত্যুর প্রধান কারণ হলো ডায়রিয়া।**(*In developing countries, four-fifths of all diseases are caused by waterborne diseases, with diarrhea being the leading cause of childhood death*) (Hendrickx et al., 2009). This sentence has two distinct causes and their corresponding effects : **Cause 1:** **জলবাহিত রোগ** → Effect 1: **মোট রোগের পাঁচ ভাগের চার ভাগ ঘটে** four-fifths of all the illnesses are caused by water-borne diseases; **Cause 2:** **ডায়রিয়া**→ Effect 2: **শিশুমৃত্যুর প্রধান কারণ** (diarrhea being the leading cause of childhood death). We have also observed a number of cases where a single sentence contains a chain of causal events where a cause event e1 results the effect of another event e2 which in turn causes event e3. In such cases e2 will be marked as both effect for e1 and cause for e3. for example **টানা বৃষ্টির কারণে নদীর জল বেড়ে যায়, ফলে গ্রামে জলাবদ্ধতা দেখা দেয়, যার ফলে বহু মানুষ ঘরছাড়া হয়।**. We extract multiple relationships from the sentence, and then treat each relationship as a separate sentence.

Quality assurance: Based on the annotation scheme, each annotator received 2500 sentences. Out of these, 2000 sentences are unique and rest 500 are overlapping. Using these 500 common sentences, we measure the inter annotator agreement of the annotation using the Fleiss Kappa (Fleiss and Paik, 1981) measure. We have achieved the inter annotator agreement to be around 0.63. This implies that the expert annotated dataset is reliable to be used for further processing. Some of the discrepancies were resolved through adjudication by a linguistic expert, and automatic consistency checks were used to validate the BIO tag sequence integrity. The final dataset contained 11663 annotated sentences, of which 6863 were labeled as causal, with a total of 8,800 cause-effect span pairs. Approximately 21% of the causal instances lacked explicit connectives, making them valuable for evaluating models' ability to capture implicit causality.

4 Experiment and Baseline Models

The experiments are conducted in two stages: (i) sentence-level multi-class causal sentence identification, and (ii) phrase-level extraction of cause, effect, and connective spans.

Traditional baselines include fine-tuned transformer models such as **XLM-R + BiLSTM + CRF** (Zeng et al., 2024), **MuRIL + Linear**(Khanuja et al., 2021), a joint multi-

[5]https://doccano.github.io/doccano/

Table 2: Results depicting causal sentence classification and cause-effect relation extraction

	Causal Classification				Cause-Effect Extraction(F1)		
Model	**Accuracy**	**Precision**	**Recall**	**F1**	**Cause**	**Effect**	**Implicit**
LLaMA-3.3-70B (3-shot)	0.80	0.79	0.80	0.78	0.63	0.82	0.54
LLaMA-3.3-70B (0-shot)	0.78	0.77	0.78	0.76	0.54	0.77	0.69
LLaMA-3.1-8B (3-shot)	0.75	0.72	0.75	0.69	0.52	0.49	0.57
LLaMA-3.1-8B (0-shot)	0.79	0.76	0.79	0.77	0.49	0.59	0.62
Gemma-2-9B (3-shot)	0.62	0.75	0.62	0.62	0.57	0.79	0.51
Gemma-2-9B (0-shot)	0.48	0.75	0.48	0.53	0.50	0.76	0.46
DeepSeek-70B (3-shot)	0.56	0.68	0.56	0.61	0.39	0.65	0.54
DeepSeek-70B (0-shot)	0.59	0.74	0.59	0.65	0.39	0.61	0.54
Qwen3-32B (3-shot)	0.47	0.63	0.47	0.52	0.57	0.70	0.64
Qwen3-32B (0-shot)	0.58	0.71	0.58	0.60	0.53	0.76	0.66
XLM-R+BiLSTM	0.70	0.65	0.74	0.69	0.60	0.62	0.67
MuRIL	0.72	0.70	0.75	0.72	0.69	0.62	0.70
Multi-task	0.76	0.75	0.77	0.75	0.77	0.72	0.79

task network (Dasgupta et al., 2022), Self-attentive BiLSTM-CRF **(SCITE)** model, proposed in (Li et al., 2021). These models are trained on our annotated dataset. The sequence tagging head is responsible for extracting BIO-labeled spans, and the sentence classifier is trained to predict among `Causal`, `Non-Causal`, and `Implicit-Causal` labels. These baselines serve as reference points for evaluating span extraction and classification performance under strong supervision.

We further assess the zero-shot and three-shot prompting performance of state-of-the-art LLMs including: LLaMA 3.3 70B (Versatile) and LLaMA 3.1 8B (Instant), Gemma 2 9B (Instruction-tuned), Qwen 3 32B (Instant), DeepSeek R1-Distill (LLaMA-70B).

For both **0-shot** and **3-shot** scenarios, prompts are constructed using Bangla examples. Each model receives a causality instruction followed by Bangla text, and is tasked with (a) classifying the sentence, and (b) extracting cause-effect-connective spans using plain-text output.

Results and Discussion: Experiment 1-Causal Sentence Classification: As shown, LLaMA-3.3-70B (3-shot) achieves the highest F1 score and accuracy, demonstrating strong few-shot generalization. Smaller models like Gemma-2-9B and Qwen3-32B show moderate to weak performance, particularly in implicit causal cases.

For Experiment 2: Cause-Effect-Connective Extraction (Word Similarity Match): Interestingly, while Qwen and LLaMA exhibit competitive performance on effect and connective spans in zero-shot setups, models like DeepSeek and Gemma show high variance in span accuracy, particularly with implicit causality. We also analyzed performance of each models across the different domains of the dataset. Figure 1 and 2 reports the respective F1 scores of each model.

Error analysis revealed that implicit causal sentences and sentences with complex or reversed clause structures posed the greatest challenges. Models frequently misidentified the directionality of the causal relation or failed to detect long-distance dependencies. These results highlight the need for incorporating syntactic, semantic, or discourse-level features in future modeling efforts. In particular, the inability to resolve anaphoric references and nested clause boundaries often led to incorrect causal inference. Moreover, models struggled with cases where causal cues were subtle or distributed across multiple clauses, indicating a limitation in capturing global sentence structure. Addressing these issues will require integrating structured linguistic representations.

5 Conclusion

We present the first comprehensive resource suite for causal relation extraction in Bangla, including an annotated corpus, connective taxonomy, and baseline models. These resources aim to spur further research in Bangla discourse-level understanding and bridge the resource gap in low-resource languages.

6 Limitations

While our resource suite marks a significant step toward causal relation extraction in

Bangla, it has certain limitations. The annotated corpus, though comprehensive, may not fully capture the linguistic variability across dialects and informal registers. Additionally, the baseline models are trained on limited data and may not generalize well to more complex or implicit causal structures.

References

Nabiha Asghar. 2016. Automatic extraction of causal relations from natural language texts: a comprehensive survey. *arXiv preprint arXiv:1605.07895*.

Eduardo Blanco, Nuria Castell, and Dan I Moldovan. 2008. Causal relation extraction. In *Lrec*, volume 66, page 74.

Quoc-Chinh Bui, Breanndán Ó Nualláin, Charles A Boucher, and Peter MA Sloot. 2010. Extracting causal relations on hiv drug resistance from literature. *BMC bioinformatics*, 11(1).

Ki Chan, Boon-Toh Low, Wai Lam, and Kai-Pui Lam. 2002. Extracting causation knowledge from natural language texts. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 555–560. Springer.

Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Mohammad Shakir. 2022. A joint model for detecting causal sentences and cause-effect relations from text. In *Towards a Knowledge-Aware AI*, pages 191–205. IOS Press.

Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. 2018. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 306–316.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics.

Levin B. Fleiss, J.L. and M.C. Paik. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2:212–236.

Daniela Garcia. 1997. Coatis, an nlp system to locate expressions of actions connected by causality links. *Knowledge acquisition, modeling and management*.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*. Association for Computational Linguistics.

Roxana Girju, Dan I Moldovan, and 1 others. 2002. Text mining for causal relations. In *FLAIRS Conference*, pages 360–364.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2009. Classification of semantic relations between nominals. *Language Resources and Evaluation*, 43(2):105–121.

Ralph Grishman. 1988. Domain modeling for language analysis. Technical report, DTIC Document.

Siyi Guo, Liuqi Jin, Jiaoyun Yang, Mengyao Jiang, Lin Han, and Ning An. 2020. Causal extraction from the literature of pressure injury and risk factors. In *International Conference on Knowledge Graph (ICKG)*, pages 581–585. IEEE.

Brad Hershowitz, Melinda Hodkiewicz, Tyler Bikaun, Michael Stewart, and Wei Liu. 2024. Causal knowledge extraction from long text maintenance documents. *Computers in Industry*, 161:104110.

Jerry R Hobbs. 2005. Toward a useful concept of causality for lexical semantics. *Journal of Semantics*, 22(2):181–209.

Ashwin Ittoo and Gosse Bouma. 2011. Extracting explicit and implicit causal relations from sparse, domain-specific texts. In *International Conference on Application of Natural Language to Information Systems*, pages 52–63. Springer.

Xinxi Jiang, Xiang Li, Qifeng Zhou, and Qing Wang. 2024. Grace: Generating cause and effect of disaster sub-events from social media text. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 999–1002.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Christopher SG Khoo. 1995. *Automatic identification of causal relations in text and their use for improving precision in information retrieval*. Ph.D. thesis.

Christopher SG Khoo, Jaklin Kornfilt, Robert N Oddy, and Sung Hyon Myaeng. 1998. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, 13(4):177–186.

Christopher SG Khoo, Sung Hyon Myaeng, and Robert N Oddy. 2001. Using cause-effect relations in text to improve information retrieval precision. *Information processing & management*, 37(1):119–145.

Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. 2021. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *Neurocomputing*, 423:207–219.

Jintao Liu, Zequn Zhang, Kaiwen Wei, Zhi Guo, Xian Sun, Li Jin, and Xiaoyu Li. 2023. Event causality extraction via implicit cause-effect interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6792–6804.

Boon-Toh Low, Ki Chan, Lei-Lei Choi, Man-Yee Chin, and Sin-Ling Lay. 2001. Semantic expectation-based causation knowledge extraction: A study on hong kong stock movement analysis. In *Advances in Knowledge Discovery and Data Mining: 5th Pacific-Asia Conference, PAKDD 2001 Hong Kong, China, April 16–18, 2001 Proceedings 5*, pages 114–123. Springer.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st international conference on World Wide Web*, pages 909–918.

Solat J Sheikh, Sajjad Haider, and Alexander H Levis. 2023. On semi-automated extraction of causal networks from raw text. *Engineering Applications of Artificial Intelligence*, 123:106189.

Antonio Sorgente, Giuseppe Vettigli, and Francesco Mele. 2013. Automatic extraction of cause-effect relations in natural language text. *DART@ AI* IA*, 2013:37–48.

Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Nelleke Oostdijk, Tommaso Caselli, Tadashi Nomoto, Onur Uca, Farhana Ferdousi Liza, and See Kiong Ng. 2023. Recess: Resource for extracting cause, effect, and signal spans. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–82.

Victor Veitch, Dhanya Sridhar, and David M Blei. 2019. Using text embeddings for causal inference. *arXiv preprint arXiv:1905.12741*.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1785–1794.

Fang Xuelan and Graeme Kennedy. 1992. Expressing causation in written english. *RELC Journal*, 23(1):62–80.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2021. A survey on extraction of causal relations from natural language text. *arXiv preprint arXiv:2101.06426*.

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting causal language use in science findings. In *2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 4664–4674, Hong Kong, China.

Zhengqiao Zeng, Zhongyuan Han, Jingyan Ye, Yaozu Tan, Haojie Cao, Zengyao Li, and Runjin Huang. 2024. A conspiracy theory text detection method based on roberta and xlm-roberta models. *Working Notes of CLEF*.

Sendong Zhao, Ting Liu, Sicheng Zhao, Yiheng Chen, and Jian-Yun Nie. 2016. Event causality extraction based on connectives analysis. *Neurocomputing*, 173:1943–1950.

A Example Appendix

Type	Sentence	Connective	Cause	Effect	Explanation
Explicit + Marked	বৃষ্টির কারণে খেলা বাতিল করা হয়েছে।	কারণে	বৃষ্টির কারণে	খেলা বাতিল করা হয়েছে	Connective directly marks the causal clause
Explicit + Unmarked	সে এত ক্লান্ত ছিল যে মাঠে পৌঁছাতে পারেনি।	যে (used structurally, not a standard causal marker)	সে এত ক্লান্ত ছিল	মাঠে পৌঁছাতে পারেনি	Causal relation is clear but not flagged with a typical marker
Implicit + Marked	সে হঠাৎ থেমে গেল, সম্ভবত ক্লান্তি।	সম্ভবত (modal indicating inferred cause)	ক্লান্তি (inferred)	সে হঠাৎ থেমে গেল	Connective used, but full cause is not syntactically marked
Implicit + Unmarked	টেস্টে ফেল করেছে। এখন সারাদিন মন খারাপ।	NA	টেস্টে ফেল করেছে	এখন সারাদিন মন খারাপ	Causal relation inferred from discourse

Table 3: Bangla Causality Types

	Politics	Editorials	Sports	International	Entertainment	Finance	Science and Tech	Story	Miscelleneous
llama-3.3-70b-versatile (3-shot)	0.851851852	0.826688038	0.770606115	0.854166667	0.654166667	0.853535354	0.89349112	0.862962963	0.707352941
llama-3.3-70b-versatile (0-shot)	0.776167472	0.784688995	0.59403794	0.923076923	0.684615385	0.877622378	0.89764436	0.67025641	0.688888889
llama-3.1-8b-instant (3-shot)	0.829420373	0.576351753	0.366758242	0.929032	0.523809524	0.986	0.82007722	0.364800759	0.523269834
llama-3.1-8b-instant (0-shot)	0.868791869	0.708602151	0.486190476	0.93224	0.627692308	0.988	0.90909091	0.532142857	0.675757576
gemma2-9b-it (3-shot)	0.811295911	0.464516129	0.203976608	0.923076923	0.70173913	0.791208791	0.76190476	0.413308913	0.526748971
gemma2-9b-it (0-shot)	0.686868687	0.379986477	0.212579577	0.6	0.619365079	0.615384615	0.57279693	0.387421241	0.525482094
deepseek-r1-distill-llama-70b (3-shot)	0.666666667	0.52515015	0.583937198	0.933333333	0.53	0.768115942	0.74285714	0.575	0.537254902
deepseek-r1-distill-llama-70b (0-shot)	0.696672716	0.571663866	0.329545455	0.736842105	0.693466759	0.742424242	0.56506239	0.619047619	0.639246032
qwen3-32b-instant (3-shot)	0.669947226	0.480421885	0.391774892	0.857142857	0.365079365	0.516363636	0.65564738	0.695739348	0.513580247
qwen3-32b-instant (0-shot)	0.780952381	0.428991326	0.49161365	0.833333333	0.35942029	0.577777778	0.69230769	0.412307692	0.598430141

Figure 1: Heatmap for the F1 score of the classification

	Politics	Editorials	Sports	International	Entertainment	Finance	Science and Tech	Story	Miscelleneous
llama-3.3-70b-versatile (3-shot)	0.623333333	0.583333333	0.6	0.82	0.743333333	0.653333333	0.426666667	0.603333333	0.83
llama-3.3-70b-versatile (0-shot)	0.726666667	0.546666667	0.616666667	0.85	0.676666667	0.626666667	0.403333333	0.64	0.893333333
llama-3.1-8b-instant (3-shot)	0.503333333	0.51	0.39	0.603333333	0.666666667	0.47	0.39	0.706666667	0.576666667
llama-3.1-8b-instant (0-shot)	0.6	0.533333333	0.766666667	0.58	0.64	0.513333333	0.373333333	0.516666667	0.586666667
gemma2-9b-it (3-shot)	0.613333333	0.583333333	0.65	0.79	0.716666667	0.6	0.373333333	0.633333333	0.716666667
gemma2-9b-it (0-shot)	0.596666667	0.573333333	0.52	0.75	0.64	0.34	0.406666667	0.636666667	0.71
deepseek-r1-distill-llama-70b (3-shot)	0.473333333	0.453333333	0.496666667	0.61	0.603333333	0.546666667	0.742857143	0.44	0.473333333
deepseek-r1-distill-llama-70b (0-shot)	0.506666667	0.526666667	0.366666667	0.453333333	0.64	0.46	0.223333333	0.366666667	0.673333333
qwen3-32b-instant (3-shot)	0.716666667	0.54	0.516666667	0.643333333	0.623333333	0.516666667	0.41	0.496666667	0.48
qwen3-32b-instant (0-shot)	0.593333333	0.66	0.596666667	0.693333333	0.57	0.763333333	0.396666667	0.51	0.456666667

Figure 2: Heatmap of the average of the matching score of cause, effect, and implicit cause-effect

Sentence	Token	BIO Tag
সে অসুস্থ ছিল বলে স্কুলে যায়নি।	সে	B-Cause
	অসুস্থ	I-Cause
	ছিল	I-Cause
	বলে	B-Connective
	স্কুলে	B-Effect
	যায়নি।	I-Effect
বৃষ্টি হচ্ছে, ছাতা নিয়ে নাও।	বৃষ্টি	B-Cause
	হচ্ছে,	I-Cause
	ছাতা	B-Effect
	নিয়ে	I-Effect
	নাও।	I-Effect
পরীক্ষায় খারাপ ফল করেছে, তাই সে খুব মন খারাপ।	পরীক্ষায়	B-Cause
	খারাপ	I-Cause
	ফল	I-Cause
	করেছে,	I-Cause
	তাই	B-Connective
	সে	B-Effect
	খুব	I-Effect
	মন	I-Effect
	খারাপ।	I-Effect
ঘুম থেকে উঠেই সে কফ খেতে শুরু করল।	ঘুম	B-Cause
	থেকে	I-Cause
	উঠেই	I-Cause
	সে	B-Effect
	কফ	I-Effect
	খেতে	I-Effect
	শুরু	I-Effect
	করল।	I-Effect

Table 4: Bengali causal sentences annotated with BIO tags for Cause, Effect, and Connective

connectives	Sentence	English
অতএব	আদালতে প্রমাণ উপস্থাপন করা হয়নি, অতএব মামলাটি খারিজ করা হয়েছে।	No evidence was presented in court, therefore the case was dismissed. (formal/logical)
উক্ত কারণে	উক্ত কারণে, অভিযুক্ত ব্যক্তিকে দোষী ঘোষণা করা হলো।	For the said reason, the accused was declared guilty. (formal/legal)
এই কারণে	আমি অসুস্থ ছিলাম, এই কারণে পরীক্ষা দিতে পারিনি।	I was sick, for this reason I couldnt take the exam.
এই জন্যে	আমি বাইরে যাচ্ছি না, এই জন্যে দেরি করছি।	Im not going out, thats why Im getting late.
এজন্য	আমি ব্যস্ত ছিলাম, এজন্য ফোন ধরতে পারিনি।	I was busy, thats why I couldnt answer the call.
এর কারণ	ট্রেন দেরি হয়েছে, এর কারণ আবহাওয়া খারাপ ছিল।	The train was late, the reason being bad weather.
এর ফলে	দেরিতে বের হয়েছিলাম, এর ফলে বাস মিস করেছি।	I left late, consequently, I missed the bus.
কাজেই	সে নিয়ম মানেনি, কাজেই তাকে শাস্তি পেতে হলো।	He didnt follow the rules, hence he was punished.
কারণ	সে আসেনি কারণ তার জ্বর ছিল।	He didnt come because he had a fever.
কারণে	বৃষ্টির কারণে খেলা বন্ধ হয়ে গেছে।	The game was stopped due to rain.
কারণেই	এই কারণেই আমি ওকে বিশ্বাস করি না।	Thats the very reason I dont trust him.
কেননা	কেননা সে অসুস্থ, সে স্কুলে যায়নি।	Because he is sick, he didnt go to school.
তদুপরি	সে নিয়মিত পড়ে এবং তদুপরি সে স্মার্টও।	He studies regularly and moreover, hes smart too.
তাই	ওর পরীক্ষা আছে, তাই সে এখন পড়ছে।	She has an exam, so she is studying now.
নইলে	তুমি পড়াশোনা করো, নইলে তুমি ফেল করবে।	Study, otherwise youll fail.
ফলে	সে নিয়মিত পড়াশোনা করেছে, ফলে সে ভালো রেজাল্ট করেছে।	He studied regularly, as a result, he scored well.
ফলে দেখা যায়	মেঘ করেছে, ফলে দেখা যায় বৃষ্টি আসবে।	Its cloudy, so it seems itll rain.
যদি...তবে	যদি তুমি মনোযোগ দাও, তবে তুমি ভালো ফল পাবে।	If you pay attention, then youll get good results.
যাতে	সে দ্রুত কাজ শেষ করলো যাতে সবাই খুশি হয়।	He finished the work quickly so that everyone is happy.
যাতে না	সে চুপ করে থাকে যাতে কেউ বিরক্ত না হয়।	He stays quiet lest anyone gets annoyed.
যার কারণে	সে সময়মতো কাজ শেষ করেনি, যার কারণে বস রেগে গেছে।	He didnt finish the work on time, because of which the boss got angry.
যার ফলে	প্রচণ্ড গরম পড়েছে, যার ফলে অনেকেই অসুস্থ হয়ে পড়েছে।	Theres a heatwave, as a result of which many people have fallen ill.
যেন	আমি চুপ ছিলাম যেন ঝগড়া না হয়।	I kept quiet so that there wouldnt be a fight.
যেহেতু	যেহেতু বৃষ্টি হচ্ছে, আমরা বাইরে যাচ্ছি না।	Since its raining, were not going outside.
সেই কারণে	রাস্তায় কাজ চলছিল, সেই কারণে যানজট হয়েছে।	There was construction on the road, thats why there was traffic.
সেজন্য	সে অসুস্থ, সেজন্য সে স্কুলে যায়নি।	He is sick, so he didnt go to school.

Table 5: sample list of Bangla causal connectives.

Initial Connective	Extended Connectives
কারণ	কারণে, কারণ, এ কারণে, কারণেই, এর কারণে, এ কারণেই, যে কারণে, কারণ হিসেবে, সে কারণে, যার কারণে, এই কারণে, কারণ হলো, এটা একটা কারণ, এর কারণ হিসেবে, একই কারণে, কারণ দেখিয়ে, সে কারণেই, যেসব কারণে, তার কারণ, আর এ কারণেই, আর সে কারণেই, এ কারণে যে, আর এ কারণে, সম্ভবত এ কারণেই, সেই কারণে, সুস্পষ্ট কারণে, কারণ হচ্ছে, এসব কারণে, এবং সে কারণে, সম্ভবত সে কারণেই, এই ব্যতিক্রমধর্মী অপরাধের কারণ, দুটি কারণে, আনার কারণে, এর বড় কারণ, এর কারণ, এবং এ কারণে, এর প্রধান কারণ, অন্যতম কারণ, এর আরেকটা কারণ, কারণটি, এর কারণ হচ্ছে, কারণ হিসেবে উল্লেখ করা হয়, কোন কারণে, ওই কারণেই, হয়তো এ কারণে যে, এর মূল কারণ, তার অন্যতম কারণ, তার কারণ হচ্ছে, এর পেছনের কারণ হচ্ছে, এবং সে কারণেই, প্রধান কারণ, মূল কারণ, এর প্রধান কারণ ছিল, এবং সেই কারণেই, এবং এর কারণে, এর পেছনের কারণ, এর অন্যতম কারণ, মুখ্য কারণ, এর কারণ হলো, কারণ হিসেবে বলেন, সেকারণেই, এর চেয়েও বড় কারণ, ওই কারণে, কারণও,

Table 6: Illustration of initial causal connective and the generated connectives