BanglaNLI: A Benchmark Dataset for Bangla Natural Language Inference

MD Ajmain Mahtab AJMAIN1234@GMAIL.COM
Atif Ronan ATIF.RONAN@G.BRACU.AC.BD
Sheikh Ayatur Rahman SHEIKH.AYATUR.RAHMAN@G.BRACU.AC.BD
Saleh Mohammad Sajid SALEH.MOHAMMAD.SAJID@G.BRACU.AC.BD
Sanjida Tasnim SANJIDA.TASNIM@G.BRACU.AC.BD
Farig Sadeque FARIG.SADEQUE@BRACU.AC.BD
BRAC University, Dhaka, Bangladesh

Editor: Sakinat Folorunso, Roseline Ogundokun, and Francisca Oladipo

Abstract

Natural Language Inference (NLI) is one of the basic tasks of Natural Language Processing (NLP) that measures the logical implication of a premise and a hypothesis. English and other numerous languages possess highly standardized, high-quality resources such as SNLI and MultiNLI. Bangla, with more than 200 million native speakers, is an under-resourced language, and existing NLI corpora mostly depend on machine translation, such as that used in BanglaBERT. These translated datasets are riddled with errors, biases, and semantic inconsistencies and are not as useful in training models. To address this gap, we present BanglaNLI, a high-quality Bangla Natural Language Inference dataset with expert annotations. The dataset was constructed from scratch based on the SNLI corpus methodology. We initiated this process with 4,200 image captions from the BanglaLekha-ImageCaption dataset written by native Bangla speakers. We took every caption as a premise and generated three hypotheses for every premise to address the three inference categories: entailment, contradiction, and neutral. This created 12,600 sentence pairs carefully annotated. In order to minimize annotation artifacts, annotators were asked to avoid simple heuristics such as negation-only contradictions or word-for-word overlap. Counterexamples were included intentionally to ensure robustness in the dataset. All annotators were native Bangla speakers, and the data was validated using Cohen's Kappa score, wherein high inter-annotator agreement ($\geq$0.88) was found.

Keywords: Low-resource Language; Semantic Relations; Sentence pair Classification; Entailment Detection; Contradiction Analysis

1. Introduction

Natural Language Inference (NLI) is one of the central Natural Language Processing (NLP) tasks in which a hypothesis is created using a premise. Then the hypothesis is classified as entailment, neutral, or contradiction. Although it may seem like a simple process it has a wide range of issues that are still being solved to this day.

Bangla is an Indo-Aryan language with over 200 million speakers, making it the sixth most spoken language globally, yet it remains under-resourced in the NLP landscape. Its complex script and morphological richness present unique challenges for NLI, similar to those faced by many African languages. By introducing a native-authored benchmark, we

© 2026 M.A. Mahtab, A. Ronan, S.A. Rahman, S.M. Sajid, S. Tasnim & F. Sadeque.

provide a template for other low-resource communities to move beyond machine translation and build culturally and linguistically grounded datasets.

The challenging process of inference classification is understanding the semantic features of the text and defining them for our model, which requires background knowledge and common sense Schuster et al. (2022). Large-scale NLI corpora such as SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) have driven the majority of research in English and other high-resource languages. However, Bangla is spoken by over 200 million people worldwide but lacks high-quality, expert-annotated NLI corpora. There are existing Bangla resources that were created using machine translation from English corpora (Hasan et al., 2020), but they have semantic inconsistencies, mistranslations, and annotation artifacts. Translation models are prone to adding artifacts into the dataset (Artetxe et al., 2020).

Automated transformation from question-answering pairs Demszky et al. (2018) is another common method for scaling NLI resources. However, they also introduce artifacts or semantic inconsistencies. Our work, by contrast, focuses on a manually curated, native approach to ensure higher data quality for the Bangla language.

To address this gap, we compiled the BanglaNLI dataset by adapting the methodology used in SNLI but applying it directly to Bangla text. Premises were derived from the BanglaLekha-ImageCaption dataset (Mansoor et al., 2019), and hypotheses were manually written by native Bangla speakers to cover all inference categories. Inter-annotator agreement was validated using Cohen's Kappa scores, ensuring reliability of the annotations. This dataset provides a structured, high-quality dataset, enabling reproducibility of experiments and supporting further exploration of Bangla-specific language inference challenges. Through this research:

- We introduce BanglaNLI, a 100% human-authored Bangla Natural Language Inference dataset constructed directly in Bangla, addressing the lack of high-quality native NLI resources for the language.
- We adapt the SNLI-style data collection methodology to Bangla by generating premise–hypothesis pairs from the BanglaLekha-ImageCaption dataset, covering the entailment, contradiction, and neutral inference categories.
- We validate the dataset through inter-annotator agreement using Cohen's Kappa, achieving a high agreement score (≥ 0.88), showing the reliability of the annotations.

2. Related Work

2.1. SNLI

The Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) consists of little over 570,000 sentence pairs labeled as "entailment", "neutral", or "contradiction". This dataset is notable for several reasons. The dataset consists of 570,152 sentence pairs, which was two orders of magnitude larger than other resources of its type during its time of release. Additionally, the sentences in this dataset were written and labeled by human annotators as opposed to being algorithmically generated and automatically or semi-automatically labeled.

This inspired us to create our own Bengali NLI dataset. To construct the dataset, annotators were provided with image captions from the Flickr30k corpus without the images themselves. For each caption, annotators were asked to write an alternate caption that is definitely true for the original caption, another alternate caption that might be true for the picture, and also a definitely false caption of the photo.

This approach had two advantages. First, it solved the indeterminacy problem caused by event and entity coreference problems. These problems are caused by the decision to label the logical relationship of two sentences, which is greatly affected by our assumption as to whether the sentences are referring to the same sentence or entity. Both of these problems are solved as photo captions restrict the context of the photo to a specific event and entity, and thereby prevent any such ambiguity.

Secondly, this method creates a richer set of sentences as opposed to algorithmically generated ones, as the sentences are written by humans in a natural context, which is far superior to creating entailment and contradiction statements by using string editing methods. While datasets like SNLI provide a foundation for generic inference, other research has explored end-task oriented entailment. For instance,the SciTail dataset (Khot et al., 2018), which focuses on inter-sentence interactions within the context of scientific question answering, demonstrating that domain-specific NLI requires deeper exploration of word-to-word relationships between the premise and hypothesis.

2.2. AfriXNLI

(Adelani et al., 2025) introduced AfriXNLI, a natural language inference dataset that includes 15 diverse and low-resource African languages. They created this dataset by having professional translators, fluent in the target languages, translate a subset of the English XNLI dataset. The authors saw a need for more NLI resources for African languages. Most multilingual benchmarks either ignore African languages or only include Swahili, which is better resourced and does not show the continent's full language diversity. Instead of using machine translation, which can lead to mistakes and cultural mismatches, the authors worked with human translators to make the data more accurate and culturally relevant, especially for languages that machine translation does not handle well.

2.3. XNLI_BN

The only other Bengali NLI dataset we found was introduced by BanglaBERT (Hasan et al., 2020). It is a large dataset consisting of 388,763 pairs of sentences. The dataset was made by translating the MultiNLI (Williams et al., 2018) training data from English to Bangla using the translation model introduced in BanglaBERT (Hasan et al., 2020). This has posed a problem, as this dataset was found to be riddled with translation errors and biases, which models have exploited. Some of the translation errors result in the meaning of the sentence being changed completely, which in turn has an adverse effect on the patterns learned by the NLI models. While this dataset can be a valuable source of data, we believe there is scope for improvement in the quality of training data, as these data do not come from a naturalized context.

2.4. Artifacts

An NLI dataset has many problems, and one notable problem is the presence of some heuristic bias, otherwise known as artifacts. Annotators use simple heuristics to annotate faster and more efficiently, which introduces artifacts that the model could exploit (Gururangan et al., 2018). For example, in case of contradiction, the annotator could simply negate the statement, so the model will look for any negation word to determine if it's a contradiction or not. If a particular word is always present in the data under the entailment category, then the model can simply look at that word and predict the sentence as entailment without reviewing the other relevant information. These heuristic patterns are picked up by the models as they are trained and then use them to predict and produce false results.

To test whether a dataset contains said artifacts, the dataset can be tested with models trained only using the hypothesis (Poliak et al., 2018). Using a hypothesis-only model, Poliak et al. (2018) tested 10 datasets and found 6 of the datasets had this problem, including SNLI. To combat this problem, datasets have to have a lot of examples that break these heuristics, so HANS dataset (McCoy et al., 2019) was created. McCoy et al. (2019) found that most NLI models fail to pass the HANS dataset. However, they found NLI models to do well when the datasets they were trained on were augmented with HANS-like examples. Hence, any model that crowdsources must be aware of such artifacts and find ways to make examples that do not follow these heuristics.

3. Methodology

3.1. Dataset Creation

The dataset was manually authored (without the use of generative AI tools) by a team of four native Bengali researchers with academic expertise in Natural Language Processing. Unlike traditional crowdsourced datasets that rely on non-expert workers, our annotators were specifically trained on the formal logical requirements of NLI. This expert-led approach allowed for the deliberate mitigation of common artifacts, such as negation-bias and sentence-length heuristics, ensuring a higher level of semantic rigor. We generated over 4200 captions from the BanglaLekha-ImageCaption dataset (Mansoor et al., 2019); these served as our Premise. For each premise, we wrote three alternative sentences that served as our Hypotheses, as shown in 1.

In total, we typed over 12,600 sentences, and each sentence is labeled under the following categories such as entailment, neutral, and contradiction.

- **Entailment:** if the premise is true, then the hypothesis is also true
- **Neutral:** if the premise is true, then the hypothesis could be true or false
- **Contradiction:** if the premise is true, then the hypothesis is false

Some samples for each class are presented in Table 1, along with their translations.

Premise	Hypothesis	Gold label
তিন জন মেয়ে মানুষ আছে। এক জন দাঁড়িয়ে আছে আর দুইজন বসে আছে Translation: There are three girls. One is standing and two are sitting.	কিছু মানুষ দাঁড়িয়ে আছে আর কিছু মানুষ বসে আছে। Translation: Some people are standing and some people are sitting.	Entailment
আচারের দোকানের সামনে দাঁড়িয়ে আছে একটি ছোট ছেলে Translation: A little boy is standing in front of the pickle shop	আচারের দোকানের বাইরে একটি ছেলে দাঁড়িয়ে ছবি তুলছে Translation: A boy is standing outside the pickle shop taking a photo.	Neutral
একজন মানুষ আরেকজন মানুষের ছবি তুলছে মোবাইল দিয়ে। পাশে দুইজন বসে আছে। Translation: One person is taking a picture of another person with a mobile phone.	একটি ছেলে ক্যামেরা দিয়ে ছবি তুলছে Translation: A boy is taking a photo with a camera	Contradiction
একটি সাদা শার্ট পরা ছেলে বাদিকে চেয়ে তাকিয়ে আছে যার চুল কালো ছোট। Translation: A boy wearing a white shirt is looking to the left.	সাদা শার্ট পড়া একজন ছেলে আছে Translation: There is a boy wearing a white shirt.	Entailment
পুরুষটির চোখ চিত্রকর্মের উপর Translation: The man's eyes are on the painting	একটি পুরুষ গ্রামের চিত্রকর্ম দেখছে Translation: A man is looking at a painting of a village.	Neutral

Table 1: Examples of Premise, Hypothesis and its Classification from the dataset

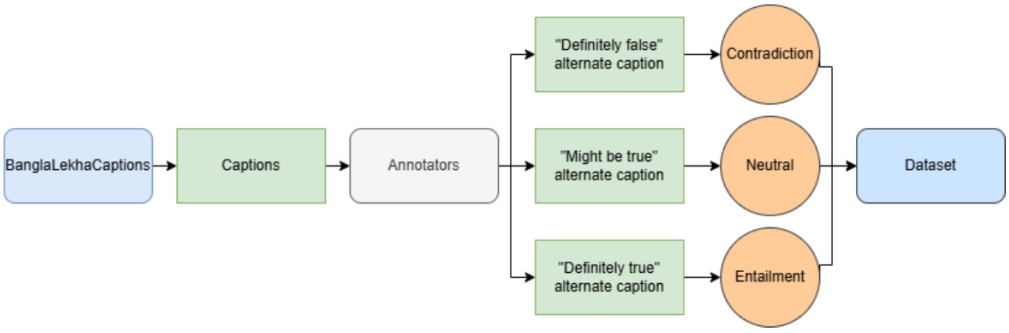

Figure 1: Methodological flowchart for the NLI Dataset Creation. The dataset is sourced from BanglaLekhaCaptions, annotated, and classified into three inference classes.

3.2. Indeterminacy

Disagreement about event and entity coreference can cause problems with labeling data. Coreference is the relationship between two words or phrases in which both refer to the same person or thing, and one stands as a linguistic antecedent of the other, as the two pronouns in "She taught herself" but not in "She taught her".

Take the example pair: “A man is in Dhaka” and “A man is in Chittagong”. If it is assumed that the pairs refer to different entities, then the pair can be labeled as neutral. However, if it is assumed that they refer to the same entity, then it is a contradiction. Hence, one option must be chosen.

However, if events and entities are assumed not to be coreferent, then most claims will be neutral. For example, take the pair: “A man lives in Manhattan” and “A man lives in Dhaka”. Since we are assuming them not to be coreferent, they both could be true at the same time or not true. Thus, like this, most pairs will be neutral. Moreover, only broad universal generalizations can contradict. Take, for example, the pair: ”All boats sink in the Pacific Ocean” and ”No boats sink in the Pacific Ocean”. Only in sentences like this can we find contradictions.

On the other hand, taking the opposite assumption also has problems. For example, the pair “Aleef is bartending” and “I am walking” will be labeled as a contradiction instead of neutral, since the second sentence’s “I” is forced to refer to “Aleef”.

To solve this, we followed the same approach as in SNLI, ensuring that each hypothesis was about a specific scenario and that hypothesis annotators wrote about the same scenario. This was done by using image captions that restricted the scenario to the specific image. Thus, the premise simply describes the image. Entailment is an alternate caption describing the image. Neutral might be a true description of the image. Lastly, contradiction is a false description of the image.

Furthermore, only the captions were used by the annotators (they did not see any of the images during annotation). This ensures that sentence pairs can be labeled using only the sentences.

3.3. Artifacts

When making datasets, consciously or unconsciously, annotators use heuristics (Gururangan et al., 2018)(Glockner et al., 2018). These heuristics cause different patterns in the dataset that models can exploit.

Premise	Hypothesis	Gold label
অনেকগুলো বালিকা পাশাপাশি বসে আছে। Translation: Many girls are sitting side by side.	অনেক বালিকা বসে আছে। Translation: Many girls are sitting.	Entailment
একটি শিশু বই দেখছে। Translation: A child is looking at a book.	একটি শিশু বই দেখছে না। Translation: A child is not looking at a book.	Contradiction

Table 2: Conscious or unconscious use of heuristic by annotators

Table 2 shows examples of Heuristics used by annotators. In the First example, the annotator used all the words present in the premise. The models can see this as a pattern that they can exploit if they find a hypothesis containing words that are all present in the premise and predict it as entailment. In the Second row, the hypothesis is a simple negation

of the premise, a very common heuristic that machine learning models exploit. Thus, whenever the model finds a hypothesis that contains negation words, it will automatically predict it as a contradiction.

Table 3 shows counterexamples for mitigating artifacts. For example, in the first row, a neutral sentence is created using negation words. In the second row, a short sentence for neutral was made since neutral generally tends to have longer sentences as information, in general, is added to it. For the third example, a contradiction sentence is made similar to the hypothesis since there are examples of entailments that are very similar to the premise.

Premise	**Hypothesis**	**Gold label**
একটি শিশু রাস্তায় চাকা নিয়ে খেলছে। Translation: A child is playing with a wheel on the road.	রাস্তায় কোন গাড়ি নাই Translation: There are no cars on the road	Neutral
সামনের সারিতে কয়েকজন মানুষ বসে আছে। Translation: Several people are sitting in the front row.	মানুষ ছবি তুলছে Translation: A person is taking a picture	Neutral
একটি অনুষ্ঠান শিক্ষক, শিক্ষার্থী একসাথে দাঁড়িয়ে বসে ছবি তুলছে। Translation: At an event, teachers and students are taking pictures together, standing and sitting.	শিক্ষক ও শিক্ষার্থী সবাই দাঁড়িয়ে ছবি তুলছে Translation: Both the teacher and the students are standing and taking a photo.	Contradiction

Table 3: Counter examples created to ensure that the model does not memorize heuristics.

3.4. Validation

The four annotators were split into two pairs. Annotators in each pair shared 10% of their data without the label. Each annotator then relabelled them. To measure the degree of agreement between the two annotators, Cohen's Kappa score was used. Since all Cohen's Kappa scores presented in Table 4 are above 0.88, this indicates strong agreement between all annotators.

Original Annotator	**Annotator Relabelling**	**Cohen's Kappa Score**
Annotator 1	Annotator 2	0.88
Annotator 2	Annotator 1	0.96
Annotator 3	Annotator 4	0.95
Annotator 4	Annotator 3	0.91

Table 4: Cohen's Kappa scores for each pair of annotators

4. Pre-trained Models

Model	Accuracy	Macro F1	Macro Precision	Macro Recall
Bangla BERT Base	0.36	0.28	0.40	0.36

Table 5: Bangla BERT Base performance without training on Bangla NLI dataset.

To evaluate the dataset's difficulty, BanglaBERT Base was tested without any fine-tuning, achieving a near-random accuracy of 0.36 (Macro F1: 0.28), as shown in Table 5. The model failed almost entirely on the Neutral class (Recall = 0.00), defaulting instead toward Contradiction predictions. This near-random performance, compared to 0.77 accuracy after fine-tuning, confirms that BanglaNLI cannot be solved without task-specific training.

We tested several different pre-trained models on our dataset: Bangla Bert Base, ELECTRA-based Bangla Bert Base (Bhattacharjee et al., 2022), mBERT uncased (Devlin et al., 2019), ELECTRA-based Bangla Bert Large, and XLM RoBERTa base (Conneau et al., 2020). We used an 80%-20% train-test split and fine-tuned the hyperparameters with grid search using combinations of values of [16, 32, 64] for batch size and [1e−3, 1e−5, 1e−7] for learning rates for all models except XLM-RoBERTa-base. To prevent overfitting, we used early stopping during training for all models. In Table 7, the hyperparameters that gave the best performance in terms of macro F1 score for each model are given. These models, trained on their corresponding set of hyperparameters listed in Table 7, were used for all subsequent tests in this paper.

Model	Accuracy	Macro F1	Macro Precision	Macro Recall
mBERT uncased	0.79	0.77	0.79	0.80
XLM-RoBERTa-base	0.81	0.80	0.79	0.80
ELECTRA-based Bangla BERT Base	0.86	0.84	0.86	0.86
Bangla BERT Base	0.77	0.74	0.76	0.76
ELECTRA-based Bangla BERT Large	0.86	0.85	0.86	0.87
Model ensemble	0.87	0.85	0.86	0.87

Table 6: Validation metrics for each model using the hyperparameters listed in Table 7

Model	Batch size	Optimizer	Epsilon	Learning Rate	Epochs
mBERT uncased	32	Adam	$1e^{-6}$	$1e^{-5}$	16
XLM-RoBERTa-base	8	Adam	$1e^{-6}$	$1e^{-5}$	12
ELECTRA-based Bangla BERT Base	32	Adam	$1e^{-6}$	$1e^{-5}$	15
Bangla BERT Base	64	Adam	$1e^{-6}$	$1e^{-5}$	14
ELECTRA-based Bangla BERT Large	32	Adam	$1e^{-6}$	$1e^{-5}$	10
Model ensemble	32	Adam	$1e^{-6}$	$1e^{-3}$	10

Table 7: Hyperparameters that gave the best macro-F1 score for each model

From Table 6, we can see that ELECTRA-based Bangla BERT base, ELECTRA-based Bangla BERT large are the best performing models. The ensemble model concatenated the results of ELECTRA-based Bangla BERT base, ELECTRA-based Bangla BERT large, and Bangla BERT base, and a linear layer afterwards to predict the labels.

4.1. Cross dataset Validation

Model	Accuracy	Macro F1	Macro Precision	Macro Recall
ELECTRA-based Bangla BERT Large	0.69	0.69	0.69	0.70

Table 8: ELECTRA-based Bangla BERT Large performance on XNLI_BN

We have trained ELECTRA-based Bangla BERT on our dataset and tested it on XNLI_BN to check its robustness.

4.2. Hypothesis-only Model

Following the methodology in Poliak et al. (2018), we evaluated our models using the hyperparameters in Table 7 on a hypothesis-only dataset to determine the prevalence of artifacts in the dataset. To do this, we only input the hypotheses into the model instead of premise-hypothesis pairing. The expected accuracy if there were no artifacts is 33.3% since the model should not be able to come to any conclusions without the premises, it has to guess randomly from three choices.

As shown in Table 9, accuracy scores range from 48.7% (ELECTRA-based Bangla BERT large) to 60.2% (XLM-RoBERTa). While these results indicate that some spurious patterns,

common to crowdsourced NLI data, have been introduced, they represent a significant improvement over standard English benchmarks. For comparison, the hypothesis-only baseline for the SNLI dataset is reported at approximately 69% Poliak et al. (2018).

Crucially, there is a substantial Inference Gap of 31% for our ensemble model (87% vs 55.8%), and an even larger gap of 37.3% for our native ELECTRA-based large model (86% vs 48.7%). This suggests that while models can exploit some linguistic cues in the hypothesis, the premise remains essential for accurate logical inference. Notably, our native-trained models (ELECTRA-based) demonstrated higher resistance to artifacts than the multilingual models, which further validates the choice of using native Bangla architectures for this task.

Model	**Accuracy**	**Macro F1**	**Macro F1**
mBERT uncased	0.597	0.57	0.60
XLM-Roberta	0.602	0.59	0.60
ELECTRA-based Bangla BERT base	0.537	0.51	0.54
ELECTRA-based Bangla BERT large	0.487	0.46	0.49
Bangla BERT base	0.536	0.51	0.54
Model ensemble	0.558	0.53	0.56

Table 9: Model Metrics when only hypotheses without premises were given

4.3. Error Analysis

Simple statistical tools were used to analyze possible reasons for models failing to label properly. The ensemble model performed the best; thus, its test cases were separated into two groups: correctly labeled and incorrectly labeled. Then the two groups were compared to find possible reasons for mislabeling, as shown in Table 10.

Category	**Correct Labels**	**Incorrect Labels**	**%Difference**
Number of cases	2176	342	-
Mean Premise Length	53.69	56.38	4.89%
Mean Hypothesis Length	38.14	39.33	3.07%
Mean number of Unknown words in the Premise	1.40	1.72	20.51%
Mean number of Unknown words in the Hypothesis	0.69	0.83	18.42%

Table 10: Comparison of Correctly and Incorrectly Labeled Cases.

There is not much difference in the premise and hypothesis length between the two groups. However, there is a significant difference in the average number of unknown words in the Premise and the Hypothesis between the groups. Hence, a possible reason the model struggles is because of the words that the model could not tokenize properly.

Moreover, as can be seen from Figure 2, the model struggles most with Contradiction. The most frequent error is Contradictions being labeled as Neutral. To further determine whether there is a significant difference in length and number of unknown tokens in premises and hypotheses between correctly and incorrectly labeled samples, we performed independent two-tailed t-tests.

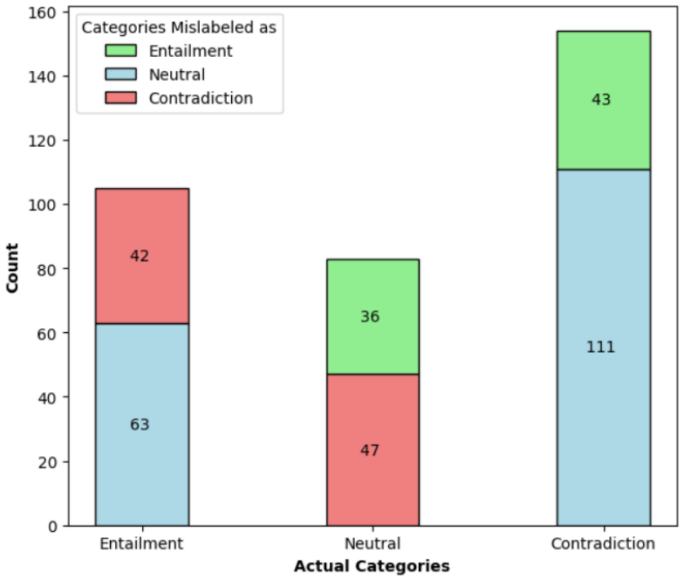

Figure 2: Count of labels given by our ensemble models for misclassified data points

Factor	T-statistic	Reject null-hypothesis
Hypothesis Length	0.077	No
Premise Length	0.069	No
Unknown token count in hypothesis	3.055	Yes
Unknown token count in premise	2.653	Yes

Table 11: Independent two-tailed t-test analysis result.

The results demonstrate that unknown tokens appear significantly more in the premises and hypotheses of incorrectly labeled samples than in correctly labeled samples. However, there is no statistically significant difference between the lengths of premises and hypotheses between correctly and incorrectly labeled samples, as shown in Table 11. This suggests that unknown tokens are a statistically significant reason behind mislabeling.

We also tested to check if the errors in the contradictions class are due to implicit contradiction as shown in 12. We checked the mislabelled contradictions without [UNK] tokens (54 sentences), which turned out to be 100% implicit. Then we took 50 random

correctly labelled contradictions and checked if they are implicit or not. 90% of them are implicit contradictions, implying the error is not due to artefact removal techniques.

Factor	Mislabelled (as Neutral)	Labelled correctly
Sentences without [UNK]	54	50
Implicit Contradictions	54	44
Percentage	100	90

Table 12: Implicit contradiction test

5. Conclusion

NLI is a crucial subfield in the field of NLP, through which we can determine whether a sentence is related to another sentence. And with the introduction of rich NLI datasets such as Stanford Natural Language Inference (SNLI) and advancements in models such as Transformer, BERT, RoBERTa, etc., this subfield has achieved tremendous progress. However, because of a scarcity of high-quality Bangla datasets, major discoveries of possible NLI applications in Bangla have yet to be made. Hence, through this research, we introduced a new Bengali dataset for NLI that has been made by inputting the texts using the everyday Bengali words and sentences everyone speaks. Our dataset's validity is also something we needed to consider as this is a new dataset it needs to be thoroughly checked for issues. Our main issue was the presence of many unknown words in our dataset. The pre-trained models are unable to grasp the semantics because of this and as such are not able to generalize properly thereby losing performance while training. The other issue we came across was the lack of diversity in the data. We also believe that adding more diverse data to the model can enhance its generalizing performance and mitigate any biases present. Above all, the dataset's creation has contributed a significant step towards Bangla NLP. Additionally, this has also laid a lot of groundwork for improvement in the dataset which we believe will bring out the best in this dataset.

6. Dataset Availability

The BanglaNLI dataset is publicly available on Kaggle at https://www.kaggle.com/datasets/ajmainmahtab/bangla-natural-language-inference-dataset/data (Mahtab et al., 2026). The dataset is released under [Attribution 4.0 International (CC BY 4.0)] for research purposes.

References

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, Tombekai Vangoni Sherman, and Pontus Stene-

torp. Irokobench: A new benchmark for african languages in the age of large language models, 2025. URL https://arxiv.org/abs/2406.03368.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.618. URL https://aclanthology.org/2020.emnlp-main.618.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.98. URL https://aclanthology.org/2022.findings-naacl.98.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

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

Dorottya Demszky, Kelvin Guu, and Percy Liang. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking nli systems with sentences that require simple lexical inferences, 2018.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Pro-*

ceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2612–2623, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.207. URL https://aclanthology.org/2020.emnlp-main.207.

Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5189–5197, 2018.

MD Ajmain Mahtab, Atif Ronan, Sheikh Ayatur Rahman, Syed Saleh Mohammad Sajid, Sanjida Tasnim, and Dr. Farig Sadeque. Bangla natural language inference dataset, 2026. URL https://www.kaggle.com/dsv/16106539.

Nafees Mansoor, Abrar Hasin Kamal, Nabeel Mohammed, Sifat Momen, and Md Matiur Rahman. Banglalekhaimagecaptions. Mendeley Data, 2019.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*, 2018.

Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. Stretching sentence-pair nli models to reason over long documents and clusters, 2022.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.