Bangla Natural Language Inference

by

Sheikh Ayatur Rahman
23141051

Atif Ronan
20201075

Syed Saleh Mohammad Sajid
22241161

MD Ajmain Mahtab
23141034

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2024

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Sheikh Ayatur Rahman	MD Ajmain Mahtab
23141051	23141034
Atif Ronan	Syed Saleh Mohammad Sajid
20201075	22241161

Approval

The thesis titled "Bangla Natural Language Inference" submitted by

1. Sheikh Ayatur Rahman (23141051)
2. Atif Ronan (20201075)
3. Syed Saleh Mohammad Sajid (22241161)
4. MD Ajmain Mahtab (23141034)

Of Spring, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 16, 2024.

Examining Committee:

Supervisor:
(Member)

Dr. Farig Yousuf Sadeque

Assistant Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi

Associate Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

Natural Language Inference (NLI) plays a vital role in our interpretation of textual data. Understanding texts is often difficult due to the logical and contextual motivations behind them. However, with the help of a text inference model, we can decode it. Our focus will be on Bengali Language Text inference, and we believe it will be useful in understanding the meaning of texts. In this thesis, we will introduce a high-quality Bangla Natural Language Inference dataset. We will also develop a benchmark model that will be able to effectively comprehend the complex semantic and logical relations among texts. The model will use complex deep-learning techniques to draw more meaningful conclusions from the texts. The research topic proposes many benefits, e.g., creating machines that will implement this model to create an effective question-answering system, an information retrieval system, sentiment analysis, and a decision maker.

Keywords: Natural Language Inference(NLI); Bangla NLI; Deep Learning; Machine Learning; Premise; Hypothesis; Entailment; Contradiction; Neutral

Acknowledgement

Firstly, we want to thank our parents for always supporting us during the course of our academic careers. It is their continuous backing that has enabled us to complete our P3 smoothly and successfully.

Secondly, we want to thank our supervisor, Dr. Farig Sadeque sir, who has guided us during the course of our thesis so far, and his valuable insights have helped us greatly during our research work.

Finally, we want to thank our friends and colleagues whose advice has helped us greatly over the last few months.

We are incredibly grateful for all of their support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Natural Language Inference(NLI) is a way of interpreting various textual data. The idea involves a textual statement to be interpreted by matching a particular text with a different yet similar text to understand the meaning behind it. Each textual data(i.e., hypothesis) given in the dataset will be paired with another textual data(i.e., premises) in the dataset. This establishes a pre-defined hypothesis-premise relationship, which is fed to a model in order to automate the process. For example, a model Y will take X_{prem} and X_{hyp} from the given NLI dataset and formulate a logical and arithmetical relationship to interpret any textual statement given to the model. If we formulate the example, Y(X_{hyp}, X_{prem}); this will produce and output Y = [entailment], Y = [contradiction], Y = [neutral]. In NLI tasks we are interested in working out the logic of how this can be achieved for any given text.

Although it may seem like a simple process it has a wide range of issues that are still being solved to this day. The challenging process is understanding the semantic features of the text and to define those our model needs background knowledge and common sense [15]. Each Language has its unique way of expressing what is being said in the text and without the knowledge of these, a model is unable to perform at its optimal level. Additionally, the textual arguments(i.e., hypothesis) can also be written in multiple ways which makes it even more difficult to draw the hypothesis-premise relationship. Even the state-of-the-art models fail at times to generalize an NLI task given.

An NLI model such as BERT when trained with an NLI dataset is able to identify the semantics of a textual argument given to it. But, as mentioned above it will still struggle to generalize the NLI task. In [15] it is mentioned that the state-of-the-art model breaks easily when given a different textual statement, despite having the same meaning. It is still too soon to say that Natural Language Processing(NLP) has become perfect as NLI models still struggle to handle simple NLI tasks. However, if we consider improving the dataset or perhaps creating one with fewer biases, the models will appear to perform well. As [15] mentioned that apart from fine-tuning a model the dataset needs to be more nuanced. A dataset creates the foundation of an

NLI model. It is very important that the dataset used in the experiment produces unbiased results. In [16] during the experiment their results revealed the exploitation of the datasets, which the models used to achieve high accuracy. Therefore, it is evident that the use of a proper dataset is crucial. In our study, we will create our own dataset before training the model since Bengali lacks non-translated NLI datasets.

Bengali is one of the most used languages with over 200 million users. Bengali is spoken by most people in Bangladesh so Bengali NLI applications show great promise. In order to work on our study the availability of the dataset is crucially needed. However, Bengali is a resource-scarce language. There is [13] but since it is a translated dataset, problems in the translation model will be carried over to any model that trains using it. Thus, we need a raw Bengali NLI dataset that can allow an NLI model to understand the semantics of textual statements. Hence, we developed a Bengali dataset that allowed us to evade these cons. The creation of a Bengali dataset will serve well in our research and also lay the foundation for future research and studies in the field of NLI.

Chapter 2

Problem Statement

2.1 Problem Statement

Natural Language Inference (NLI) is useful for many applications such as semantic search and questions answering [9]. However, research into the potential application of NLI in the Bangla language has been limited due to the lack of availability of high-quality datasets with an adequate amount of data for training and testing machine learning models. While attempts at creating such a dataset have been made such as in [13], these data are often automatically generated and labeled. For instance, [13] was created by translating the MultiNLI training data from English to Bangla. However, using translation models affects the quality of data produced as errors in the translation model can seep into the dataset. Furthermore, [10] states that translation models can introduce artifacts and have an effect on the performance of models. In NLI, translation can reduce the lexical overlap between the hypothesis and the premise, and thereby affect the performance of the models. Furthermore, [2] found that NLI datasets often include spurious patterns that are learned by the machine learning model instead of the logical relation between two sentences. In an experiment where they only input the hypothesis to the model, it gave an accuracy of 67%, showing that the model had learned patterns in the hypothesis and used them to produce the outputs as opposed to learning its relation with the premise. [2] further found that augmenting the training dataset with explanations and making the model explain its classification of the relations in addition to predicting the label vastly reduces the problem of spurious patterns. Finally, when constructing NLI datasets, entity, and event coreference are major problems as assumptions about whether the premise and hypothesis refer to the same event and entity can change the classification of the relationship between two sentences. [1] addresses this problem by using captions of photos from the Flickr30k corpus and thus preventing ambiguity in which entity or event is being referred to as the context is restricted to the photo itself.

Keeping the above facts in mind, our research problem statement is as follows, Introducing a Bangla Natural Language Inference dataset where the statements are

written by humans in a naturalistic context.

In addition to introducing a new dataset, we will train baseline natural language inference models on the dataset by fine-tuning popular large language models such as BanglaBertLarge [14] and XLM-RoBERTa and compare their performances. We will make these models freely available. We hope that these models will help the Bangla NLP community incorporate NLI into their NLP research and use them in their products.

2.2 Research Objectives

Bangla being a resource-scarce language lacks datasets for NLI. Thus, this paper aims to introduce a new Bengali dataset for NLI. Then various NLP models will be trained using the dataset and their performances will be evaluated. The models will then be tested on the XNLI_bn dataset (English to Bangla translated MultiNLI dataset) to check if the performance generalizes to other datasets.

To summarize:

- Introduce a new Bengali dataset for NLI
- Train different NLI models without the premises to check if the dataset has artifacts
- Fine-tune different models on this dataset and evaluate their performances

Chapter 3

Literature Review

There have been various Natural Language Inference datasets and models built over the years in English and other languages as well.

In this section, we will critically examine some of these approaches.

3.1 SNLI

[1] introduces the Stanford Natural Language Inference (SNLI) corpus, which consists of 570,000 pairs of sentences with each sentence labeled as "contradiction", "entailment", or "neutral". This dataset is notable for several reasons. Firstly, the dataset consists of 570,152 pairs of sentences, which was two orders of magnitude larger than other resources of its type during its time of release. Secondly, 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 is crucial as automatic generation and labeling can introduce spurious patterns into the dataset which may be picked up by machine learning models and prevent it from focusing exclusively on the logical connections between two sentences. 56,941 of the pairs were later annotated by four other annotators where 98% of the instances reached a three-annotator consensus while 58% of them reached a unanimous consensus. As for the construction of 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. There were two advantages to this approach. Firstly, 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 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 just using string editing methods. The drawback with the SNLI dataset is that it consists of short and simple sentences and thus training a robust NLI model using this dataset alone is very difficult. However, the SNLI sets a standard for NLI datasets and provides a blueprint with which future NLI datasets can be made.

3.2 NLI with Natural Language Explanations

[2] provides an extension to the SNLI dataset e-SNLI, which consists of the same sentence pairs in SNLI but is augmented with explanations for the label the pair of sentences was given. The advantages of such an approach were twofold. Firstly, by forcing a model to provide explanations for the labels they predicted for a pair of sentences, the chances of a model making predictions based on spurious results are drastically reduced. This problem is quite pervasive as [2] mentions that a model can predict the correct label 67% of the time despite having been provided only the hypothesis, indicating that the model is picking up spurious correlations. As such, by providing explanations, we can be more confident in a model's predictions. Secondly, providing explanations generates trust in the machine learning model. This is crucial, especially in areas such as law and healthcare, where people often expect explanations as to why a model made a particular decision. The explainability of such models also allows engineers to debug the model and understand what exactly it is learning. For the collection of data, annotators were provided sentence pairs from the SNLI dataset and asked annotators to first highlight the parts of the hypothesis and premise that they thought were important for the relation provided and then asked them to write out explanations for them using the words they highlighted. Annotators were encouraged to focus on the non-obvious parts of the statements that induced the relation between the sentences as well as give self-contained explanations so that understanding them does not require the premise and hypothesis to be read. Finally, [2] also provides an architecture for performing natural language inference with explanations. It consists of a bi-directional LSTM which encodes the hypothesis and premise. The feature vectors are then passed to an MLP classifier which outputs a distribution over the three labels, and thus giving the relation between the sentences. A one-layer LSTM decoder which takes the feature vector is used to generate explanations. The output relation is prepended to the start of the explanation so that the label has an effect on the final explanation.

3.3 Translated datasets

3.3.1 xnli_bn dataset

The only Bangla dataset for NLI we found was introduced by [13]. It is a large dataset consisting of 388,763 pairs of sentences. The dataset was made by translating the MultiNLI [6] training data from English to Bangla using the translation model introduced in [11]. 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 and this issue is compounded by errors of the translation model itself. Furthermore, the translation may introduce spurious correlations in the dataset.

3.4 Artifacts

[4] talks about the prevalence of artifacts in NLI training data. The paper states that one of the most common methods for data collection is to crowdsource. However, [4] finds the annotators use simple heuristics to annotate faster and more efficiently which introduces artifacts that the model could exploit. 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. To determine the degree to which this problem is present in the dataset, they trained fastText, a bag of words and bigrams text classifier, without the premise. This model correctly classified a large number of the hypotheses, suggesting that the dataset is filled with these artifacts. Another paper [5], tested 10 datasets using a hypothesis-only model 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 [5] created such a dataset, HANS. They found most NLI models fail to pass the HANS dataset matching the findings with [4]. 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.

3.5 Breaking NLI systems with sentences that require simple lexical inferences

[3] claims that the current NLI models are lacking in their generalization ability and fail to grasp many basic inferences. First of all, their aim is to introduce a new NLI test set that exposes the limitations of current state-of-the-art models in capturing inferences that require lexical and world knowledge. Secondly, they propose a new model that incorporates external knowledge sources to improve performance on this test set. They constructed a test set with the goal of evaluating the ability of state-of-the-art NLI models to make inferences that require simple lexical knowledge by taking the premise from the SNLI training set, and for each premise, multiple hypotheses were created by simply replacing a word within the premise with a different word. The labels were generated through crowdsourcing. For evaluation without external knowledge, they used RESIDUAL-STACKED-ENCODER, which is a biLSTM-based single-sentence-encoding model; ESIM (Enhanced Sequential Inference Model), which is an RNN-based attention model; and decomposable attention, which is an attention model without RNN. This model was tested with the SNLI dataset and on both the SNLI test set and the new test set. For evaluation with external knowledge, they propose KIM (Knowledge-based Inference Model), which incorporates external knowledge from WordNet. All the models have high accuracy in the SNLI test set, whereas the accuracy on the new test models is significantly low. On the other hand, the difference between the accuracy of the SNLI test set and the new test set by KIM is very small compared to the models without exter-

nal knowledge. This shows that neural NLI models are lacking in making lexical inferences without lexical and world knowledge.

3.6 Testing the generalization power of neural network models across NLI benchmarks

[8] is a comprehensive study of how NN models help in accessing NLI tasks for any datasets given to them. However, they discovered that even the state-of-the-art models are not well equipped to generalize any NLI tasks; rather some NLI models that have shown excellent accuracy and precision were seen to fail. When the model was tested with a different test dataset, the accuracy of these models was seen to drop significantly when testing with a test dataset from a different corpus. The compared metrics between the models on different datasets are vastly different hence, raising the issue of transfer learning.

NLI models are shown to have significant results when trained on SNLI and MultiNLI datasets. Therefore, they decided to try out the models on different kinds of datasets; however, they found metrics of lower accuracy. The authors of the paper decided to improve the outcomes by implementing various techniques in order for the models to work effectively on any given dataset. In this experiment six state-of-the-art models each of which uses the techniques involved to produce a better output metric. The models would be trained and tested on different combinations of test and train data from SNLI, MultiNLI, and SICK. The models included BiLSTM-max and HBMP, which utilize sentence encoding; ESM and KIM, which utilize cross-sentence attention and finally ESIM and BERT-base, which utilize cross-sentence and pre-trained language models.

After the experiment was completed the new metrics of these six models were shown to be significantly better out of which the BERT-base model outperformed all of them. Another dataset, SICK was used as well however, the six models performed poorly in that with the BERT-base model being significantly better in them. The BERT-base achieved an accuracy in the range of 59-92 percent for the given combination. It can be interpreted from this outcome that pre-trained models along with cross-sentence attention technique-based models can produce the highest accuracy with any combination of datasets. This also lays the foundation that NLI tasks are better performed if such techniques are placed.

However, even if the results were good for some datasets it was not good when it came to SICK. Therefore, this suggests that the dataset itself has issues of its own. We are familiar with the common issues faced in NLI tasks such as the lack of background knowledge and common sense. The models are not yet dynamic enough to handle certain texts. As models are trained on these datasets the datasets also need to be fine-tuned in order for them to be effective to the models. NLI datasets such as SNLI and MultiNLI were seen to show good results as their architecture of the dataset is almost similar which is in contrast to SICK. The author highlights that in order to produce effective NLI models we need to design more nuanced

datasets. Moreover, he also suggests that we put more research into the underlying mechanisms of NLI.

3.7 Non-English Datasets

There have also been attempts to create datasets in languages other than English. [12] introduced the first-ever Chinese original dataset for natural language inference, which is made up of 56,000 annotated sentence pairs. Instead of using automatic translation of large-scale annotations such as SNLI, the annotations used here are produced by native speakers who excel in linguistics because, when it comes to automatic translation, the quality of the data is a question, and the translation of the dataset also carries the cultural context of the primary language. For the OCNLI dataset, state-of-the-art models such as the RoBERTa model have been used, which showed high accuracy on XNLI dataset among other transformers. For the dataset, the standard format of NLI has been followed, which consists of a sentence pair of premise and hypothesis, annotated with one of the three labels, which are entailment, contradiction, and neutral. For the premise, it is selected from five genres, which are government documents, news, literature, TV talk shows, and telephone conversion. To make the data more realistic and challenging, the strategy Multi was introduced, where instead of introducing 1 sentence for each label, which was done in MNLI, 3 sentences were provided for each label, summing up to 9 sentences per hypothesis. Two more strategies were also introduced, as the strategy multi brings more hypothesis-only bias. The first method is known as multiple encouragement, where the annotator is encouraged to create high-quality hypotheses based on some criteria. The second method is known as MULTICONSTRAINT, where hypotheses are generated following some constraint. Among all the models, RoBERTa shows the highest performance (78.2%), which is 12 points behind human performance (90.4%). The result of the XNLI dataset, which is a translated version of the MNLI dataset, produced 70.4% accuracy, and a combined dataset of OCNLI and XNLI showed 75.6% accuracy. This shows OCNLI outperforms both datasets, considering XNLI is the largest multi-genre Chinese dataset.

[16] shows us the FarsTail dataset which was designed by the authors in order to create the Farsi NLI(Natural Language Inference) dataset. A major problem in the field of NLI is the availability of NLI datasets for different languages. We already know the major well-known datasets like SNLI and MultiNLI which are used widely for NLI tasks. Besides, significant contributions are being made to the field of NLI and the development of advanced Deep Learning Models began to play a role in the understanding of natural languages generated by humans. Similar to the Stanford NLI dataset(NLI), which caused a massive development of the NLI models; the Farsi corpus was generated to aid in understanding the Farsi(Persian) language.

Farsi language apparently was one of the most influential languages and it can be seen in languages like Turkic, Georgian, Armenian, and many Indo-Aryan languages. The author decided to model the FarsTail dataset similar to SciTail where the sentences are made from multiple-choice questions(MCQ). These questions are based on real-world natural sentences known to exist in the world. The dataset also contains

MCQs which are based on realistic examples that occur daily. Its design process is fairly simple. Each person will generate three outputs(i.e., entailment, contradiction, neutral) from the MCQ. The first step involves simply answering some MCQ questions where the participants pick out the answers from the website. The participants are asked to search for the correct(entailment), incorrect(contradiction), and unsure(neutral). These are inserted into the questions and used to generate the outputs. The second step involves data cleaning, where four annotators are asked to relabel the samples. The sample with an agreement of at least 80% was kept while the others were rejected, however, the rejected sample was given another chance by giving them to the original annotator, the first step participant, to be relabeled again. Finally, they were checked for spelling, grammatical mistakes, and repetition of data.

The author also wanted to test the dataset on common models such as ELMo, BERT and LASER, ESIM, HBMP, etc. The approach was taken to ensure a baseline for future research using FarsTail. The study used different embedding methods such as traditional TF-IDF, word2vec, fastText, ELMo, and BERT. For the BERT method, two pre-trained models from the Hugging Face Transformers library, ParsBERT and BERT-base-multilingual-cased (mBERT), were fine-tuned. As for the modeling approaches, the study exploited different methods including Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). After the model was trained on the mentioned datasets the metrics of each model were categorized as the test and training dataset. The BERT model was shown to obtain an accuracy of 82 percent and 83 percent in the training and test dataset respectively. The results signified the room for more improvement in the current methods used thus encouraging future research in the matter and in NLP applications. The metrics were later compared to metrics of the SNLI and other datasets to see how well it performed. The results between SNLI and FarsTail on the same models were significant, but regardless showed promises considering the challenges of natural language inference in Persian language and the relatively small size of the FarsTail dataset.

The author also mentions the biases in the FarsTail dataset which the model exploits to achieve high accuracy. Even though the authors tried to keep the annotation clues low by reducing the amount of task-specific human-generated texts, some biases emerged in FairTail hypotheses. Some words were used to confine the general point presented in the premise to make a contradicting hypothesis. To investigate these biases, the authors evaluated two biased models that classified instances based on incomplete input data. The classification accuracy of these models gives an estimate of the degree to which the superficial clues can be exploited by the learning algorithms. The results showed that the models' accuracy on the hard subset obtained by the overlap-based biased model is usually lower than that of the hypothesis-only biased model. This reveals that the models exploit more of the overlap information between premises and hypotheses than the biased patterns in the hypotheses. These results suggest that the models' success in recognizing textual entailment is partly due to their exploitation of available biases in the dataset. We can reduce the biases by reducing the amount of task-specific human-generated texts to keep the

annotation clues low; however, it is not an effective method to tackle it. The author suggests that in the future more research should be done to reduce the biases in a dataset.

Chapter 4

Preliminary

The methodology of the Standford Natural Language Inference dataset [1] has been followed, where annotators were provided with image captions from the corpus Flickr30k. For each caption, the annotators generated three different captions that are either completely true given the original caption is true, might be true given the original caption is true, or completely false given the original caption is true.

Following this methodology, we used 4200 image captions so far from the BanglaLekha-Dataset. From each caption, we generated 3 more sentences that either entail, are neutral, or contradict given the original caption is true. The first benefit of using this methodology is that it solves the issue of indeterminacy, which is caused by the coreference between an event and an entity. Secondly, this method generates sentences far better than using string editing methods, as the sentences generated are human-made which is more natural and free of translation errors.

	Premise	Entailment	Neutral	Contradiction
SNLI	A person on a horse jumps over a broken down airplane.	A person is outdoors, on a horse.	A person is training his horse for a competition.	A person is at a diner, ordering an omelette.
Bangla NLI	একটি শিশু ছবি আঁকছে।	একটি বাচ্চা ছবি আঁকছে	একটি বাচ্চা একটি গাড়ির ছবি আঁকছে	একটি বাচ্চা গান গাচ্ছে

Table 4.1: Bangla NLI

4.1 Other Datasets

4.1.1 Translated Datasets

xnli_bn

While there have been previous attempts at making a Bangla NLI dataset, their effectiveness has been held back by several factors. For instance, the xnli_bn dataset [14] was made by traslating the MultiNLI [6] dataset from English to Bangla. This dataset has several translation errors which reduce its effectiveness as a Bangla NLI dataset.

Some of the translation errors result in the meaning of the sentence changed completely, which in turn has an adverse effect on the patterns learnt by the NLI models. Some instances of these errors in the xnli_bn dataset and their explanations are listed in 4.2.

Despite that, this data can be potentially used as auxiliary training data.

Issue of gendered pronouns

Aside from translation errors, another issue translated datasets may face is that Bangla has no gendered pronouns while English has gendered pronouns. This can result in changes in the meaning of sentences, which in turn can render classification labels incorrect.

For instance, consider the following premise-hypothesis pair,

He is playing the guitar.

She is playing the guitar.

Naturally, annotators would label this pair as a contradiction since the two sentences refer to two different individuals.

However, is we were to translate both sentences, we would get the following premise-hypothesis pair,

সে গিটার বাজাচ্ছে

সে গিটার বাজাচ্ছে

The ”contradiction” label for the above sentences would be rendered incorrect since the lack of gendered pronouns in the Bangla language causes both sentences to become the same sentence after translation from English. The correct label in this case would thus have been ”entailment”.

While the above example is rudimentary, how the presence or absence of gendered pronouns is crucial to the meaning behind a sentence in a language, and how assumptions regarding them can affect the way people express meaning. As such, it is an important consideration when building an NLI dataset in Bangla.

Idiomatic expressions

Translation models often fail to properly translate sentences involving idiomatic expressions. This can result in the intended meaning behind a sentence being warped completely. This happens because most translation models translate sentences literally, as opposed to considering the intended meaning behind a sentence. This is exacerbated by the fact that many idiomatic expressions make sense in only a handful of languages at a time. To illustrates this issue, we have listed some English idiomatic expressions and their Bangla translations, using the BanglaT5 english-to-bangla translation model [17] in 4.3 along with explanations for the issues behind them.

These examples illustrate that idiomatic expression in one language can hinder translation models when translating to Bangla.

The BNLI dataset we created was annotated by native Bengali people. We believed this approach would provide us with significantly better results than the translated corpus. As the dataset was annotated with daily Bangla speakers, we were able to attain a native view of how Bangla can be said or spoken by an individual. The dataset contains 4200 rows, where each row showcases everyday examples. This allowed our model to understand the semantic and contextual meanings of a text much better.

English sentence	Bangla sentence	Explanation
Product and geography are what make cream skimming work.	পণ্য এবং ভূগোল হচ্ছে ক্রিমের স্কিমিং কাজ।	While the English sentence implies that "product" and "geography" are factors that make cream skimming work, the Bangla sentence implies that "product" and "geography" are the "skimming" tasks of "creaming", which is clearly a translation error.
The tennis shoes have a range of prices.	টেনিসের জুতার দাম অনেক।	The English sentence indicates that the tennis shoes have a range prices. However, the Bangla translation is ambiguous as it may either mean that the tennis shows are expensive or they may have multiple prices.
Read for Slate's take on Jackson's findings.	জ্যাকসনের আবিষ্কারগুলো স্লেট এর জন্য পড়ুন।	While the English statement implies the reader can read Slate's opinion on Jackson's finding, the Bangla sentence implies the reader should read Jackson's findings for Slate, which once again is a translation error.
The analysis proves that there is no link between PM and bronchitis.	বিশ্লেষণটি প্রমাণ করে যে প্রধানমন্ত্রী এবং ব্রংকাইটিসের মধ্যে কোন সংযোগ নেই।	Here, the translation model mistook the abbreviation PM for Prime Minister.
well it's been very interesting	ভাল, এটা খুবই আকর্ষণীয় ছিল	The word "well" in this case is being misconstrued as "good"

Table 4.2: Examples of translation errors within the first 25 rows and their explanations from xnli_bn dataset

English sentence	**Bangla sentence**	**Explanation**
To break the ice at the party, he told a funny joke.	পার্টিতে বরফ ভাঙ্গানোর জন্য সে একটা মজার কৌতুক বলেছিল।	The Bangla sentence implies that the person told a funny joke in order to literally break ice at a party.
You'll have to bite the bullet and start your own business if you want to be your own boss.	তুমি যদি নিজের বস হতে চাও, তাহলে তোমাকে গুলি খেয়ে নিজের ব্যবসা শুরু করতে হবে।	The Bangla sentence the person literally has to consume a bullet.
He kicked the bucket after a long battle with cancer.	ক্যান্সারের সাথে দীর্ঘ লড়াইয়ের পর তিনি বালতিতে লাথি মেরেছিলেন।	The Bangla sentence implied the person kicked a bucket as opposed to passing away.
He let the cat out of the bag about the surprise party.	তিনি ব্যাগ থেকে বিড়াল বের করে সারপ্রাইজ পার্টি সম্পর্কে বললেন।	The Bangla sentence indicates that a cat was literally let out of the bag instead of the person letting out a secret.
She visits her hometown once in a blue moon.	তিনি একবার একটি নীল চাঁদে তার নিজ শহর পরিদর্শন করেন।	The Bangla indicates that the person visited her hometown on the occasion of a blue moon instead implying that her visiting her hometown was a rare occurence.

Table 4.3: Examples of translation errors for English idiomatic expressions when translating to Bangla and their explanations.

Chapter 5

Data

5.1 Dataset Creation

This dataset was constructed by all four members of our study. Each of us are native Bangla speakers and all of us familiar with natural language inference and the concepts of entailment, neutral, and contradiction. This allowed us to construct an expert-constructed dataset, which minimizes mistakes during construction stemming from a lack of understanding of the concepts of entailment, neutral, and contradiction.

5.1.1 Dataset description

Our dataset currently consists of 12,600 pairs of sentences. Each pair is labeled as one of "entailment", "neutral", and "contradiction". Some examples of datapoints from our dataset are given in 5.1.

Premise	**Hypothesis**	**Classification**
তিন জন মেয়ে মানুষ আছে। এক জন দাড়িয়ে আছে আর দুই জন বসে আছে।	কিছু মানুষ দাঁড়িয়ে আছে আর কিছু মানুষ বসে আছে।	Entailment
আঁচারের দোকানের সামনে দারিয়ে আছে একটি ছোট ছেলে	আচারের দোকানের বাইরে একটি ছেলে দাঁড়িয়ে ছবি তুলছে	Neutral
ছেলেটির কাছে একটি ক্যামেরা আছে	ছেলেটি খালি হাতে বসে আছে	Contradiction
একটি মহিলা কলসিতে পানি ভরছে	একটি নারী চাপকল থেকে কালো পানি কলসিতে নিচ্ছে	Entailment
পুরুষটির চোখ চিত্রকর্মের উপর	একটি পুরুষ গ্রামের চিত্রকর্ম দেখ-ছে	Neutral

Table 5.1: Examples from our dataset

Each row of data consists of a "premise", "hypothesis", and "contradiction". The "premise" is the initial sentence providing some context. The "hypothesis" is a

statement of which we have to find its logical relationship with the ”premise”. The classification is the relationship between the premise and hypothesis. The classification can be one of the following,

Entailment: If the hypothesis logically follows from the premise.

Neutral: If the hypothesis is logically irrelevant to the premise.

Contradiction: If the hypothesis does not logically follow from the premise.

5.2 Dataset Construction

To construct the dataset, we used the BanglaLekhaImageCaptions dataset [7], which contains 9,154 images and annotations written in Bengali by two native Bengali speakers. From that dataset, we picked 4200 annotations, and for each caption, we wrote three alternative statements to make sure all the classes are balanced.

- An alternate caption that is **definitely true** given the caption is true. This serves as **entailment**.
- An alternate caption that **might be true** given the caption is true. This serves as **neutral**.
- An alternate caption that is **definitely false** given the caption is true. This serves as **contradiction**.

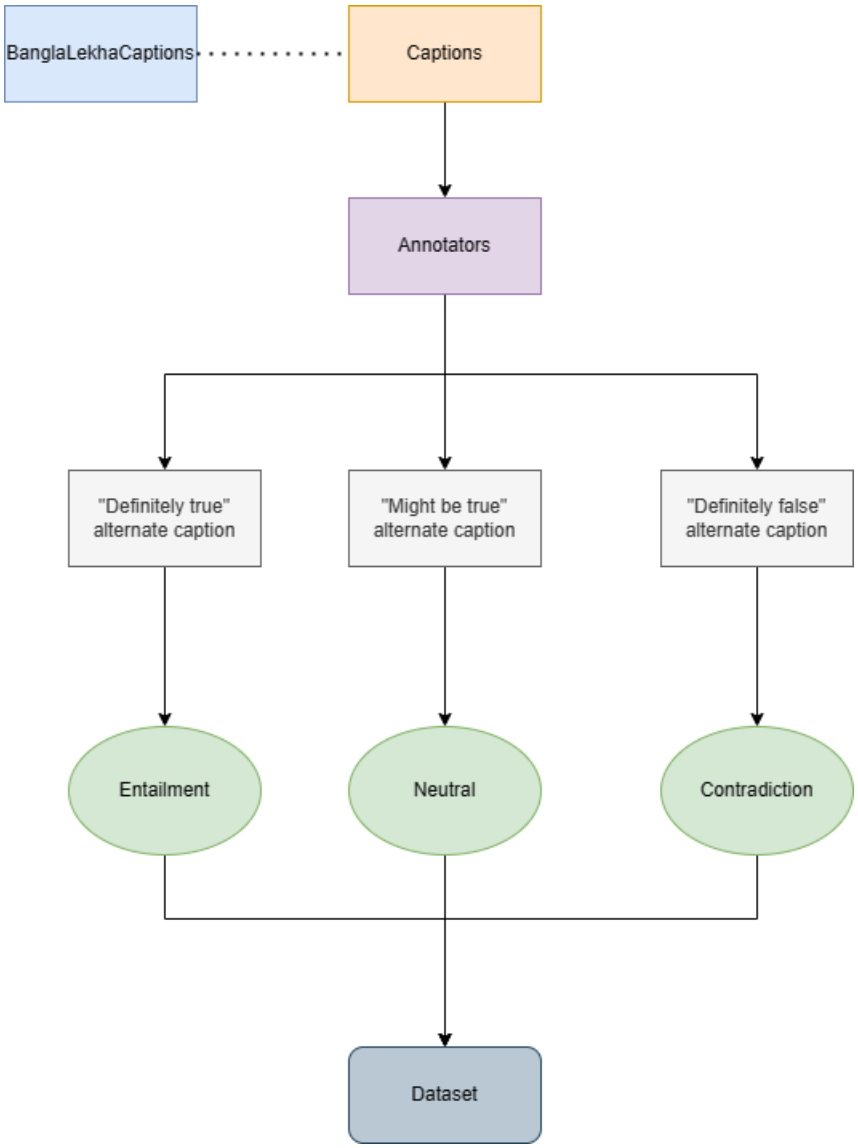

Figure 5.1: Overview of dataset creation process

5.3 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 to be not 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 to be not 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, we can 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 contradiction instead of neutral since the second sentence’s “I” is forced to be referring to “Aleef”.

To solve this, just like in [2], we made sure each hypothesis were about specific scenarios and hypothesis annotators would write would be 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 not seeing any of the images during annotation. This ensures that sentence pairs can be labeled using only the sentences.

5.4 Artifacts

When making datasets, consciously or unconsciously annotators use heuristics. These heuristics cause different patterns in the dataset that models can exploit. From Table 5.2 we can see some examples of heuristics used by annotators. For example, a common heuristic could be simply negating the premise to create contradiction. Then the model could learn to recognize negation words and simply label sentences with these kinds of words as contradictions. Another example is, to use all the words present in the premise in the hypothesis. The model can exploit these patterns. To counter this our annotators made counter-examples for common heuristics.

Premise	**Hypothesis**	**Gold_label**
অনেকগুলো বালিকা পাশাপাশি বসে আছে।	অনেক বালিকা বসে আছে।	Entailement
একটি শিশু বই দেখছে।	একটি শিশু বই দেখছে না।	Contradiction

Table 5.2: Conscious or unconscious use of heuristic by annotator

Table 5.3 shows examples to counter heuristic made by annotators. 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 very similar to the hypothesis since there are

Premise	Hypothesis	Gold_label
একটি শিশু রাস্তায় চাকা নিয়ে খেলছে।	রাস্তায় কোন গাড়ি নাই	Neutral
সামনের সাড়িতে কয়েকজন মানুষ বসে আছে। পিছনের সাড়িতে কয়েকজন মানুষ দাড়িয়ে আছে।	মানুষ ছবি তুলছে	Neutral
একটা অনুষ্ঠানে শিক্ষক, শিক্ষার্থী একসাথে দাঁড়িয়ে বসে ছবি তুলছে।	শিক্ষক ও শিক্ষার্থী সবাই দাঁড়িয়ে ছবি তুলছে	Contradiction

Table 5.3: Counter examples

examples of entailments being very similar to the premise. To further analyse artifacts in the dataset, a hypothesis-only model was used.

5.5 Data 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, we used the Cohen Kappa score. It is a number which ranges from -1 to 1. An explanation of what the score means is as follows,

- -1 indicates complete disagreement between the two annotators.
- 0 indicates there is no significant agreement or disagreement between the annotators beyond agreement or disagreement by chance.
- 1 indicates perfect agreement between the annotators.

The Cohen Kappa scores of our cross-validation is as follows:

Original Annotator	Annotator Relabelling	Cohen 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 5.4: Cohen Kappa scores for each pair of annotators

Since all Cohen Kappa scores are above 0.75, it indicates strong agreement between all annotators.

Chapter 6

Methodology

6.1 Methodology

Several preprocessing steps were applied on the pairs of sentences so that they could be fed into the pre-trained model which would then classify the pair as an entailment, neutral, or contradiction.

Preprocessing

- **Adding special tokens**: The model has to differentiate between the premise and hypothesis. As such, a [CLS] token is added in front of the premise and a [SEP] token is added at the end of the premise, which in turn is the start of the hypothesis. A [SEP] token is also added at the end of the hypothesis.

- **Tokenization**: Both sentences are broken down into units that can be fed into a pre-trained model using a tokenizer for that particular model from [18].

- **Converting tokens to ids**: Each tokenizer objects maintains a mapping of tokens to ids. We used this to convert the tokens to ids that are used to query the static embedding look-up table.

- **Adding token type ids**: In order to differentiate the two sentences, token type ids are used for each token. Each token in the premise is labeled as 0, and each token in the hypothesis is labelled as 1.

- **Adding attention masks**: Since we want the model to pay attention to all tokens in both sentences, we have a mask of all ones for each token in both sentences.

- **Batching and padding**: In order to stabilize the learning process, we batched the sentence pairs and padded them with a tokenizer specific pad token to ensure all sentences in a given batch have the same length.

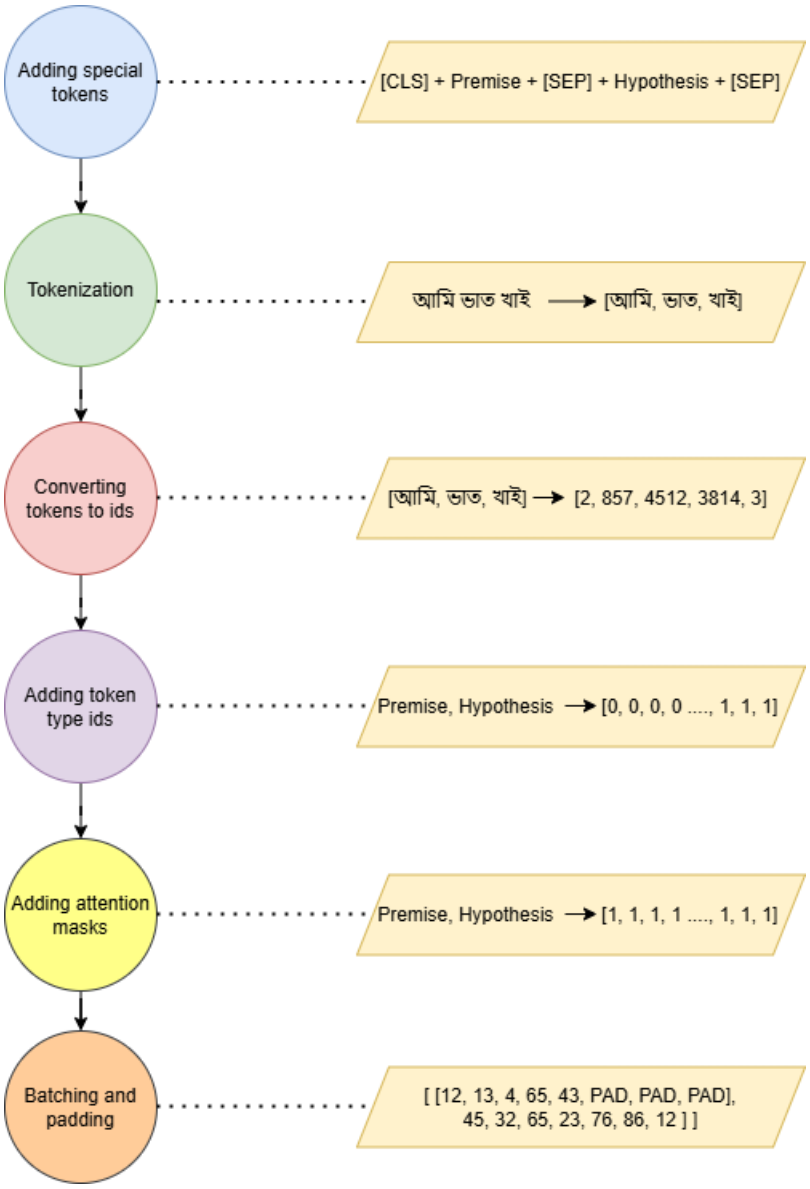

Figure 6.1: Preprocessing steps

6.2 Reforming data for training

Initially, our dataset contained 4 features, and they are premise, which contained the premise sentence; entailment, which contains a sentence that entails the premise; neutral, which contains a sentence that is neutral to the premise; and contradiction, which contains a sentence that contradicts the premise. We reformed our dataset into a new DataFrame where each row from the previous row is separated into 3 new rows, and each of those rows contains 3 features, gold_label, which contains the label; sentence1, which contains the premise; and sentence2, which contains the hypothesis.

Premise	Entailment	Neutral	Contradiction
নববর্ষের উৎসবে কয়েকজন মিলে হাতে ঢোল গলায় ঝুলিয়...	আজ নববর্ষের আয়োজন হচ্ছে	চার পাশে অনেক মেলা হচ্ছে	কোন শব্দ হচ্ছে না

Table 6.1: Before Data Reformation

gold_label	sentence1	sentence2
Entailment	নববর্ষের উৎসবে কয়েকজন মিলে হাতে ঢোল গলায় ঝুলিয়...	আজ নববর্ষের আয়োজন হচ্ছে
Neutral	নববর্ষের উৎসবে কয়েকজন মিলে হাতে ঢোল গলায় ঝুলিয়...	চার পাশে অনেক মেলা হচ্ছে
Contradiction	নববর্ষের উৎসবে কয়েকজন মিলে হাতে ঢোল গলায় ঝুলিয়...	কোন শব্দ হচ্ছে না

Table 6.2: After Data Reformation

6.2.1 Data Splitting

After the dataset is done being reformed, it is split into an 80% training dataset and the remaining 20% as test dataset.

6.2.2 Tokenization

For tokenization, firstly, we have 3 Bert model which are banglabert_large, banglabert, and bangla-bert-base and for each model, we load the corresponding tokenizer. Then we retrieve CLS,SEP,PAD, and UNK tokens and their ids. After that, for each model, we make 2 new sentences, which are prepped_sent_1_ which contains sentence1 with CLS token at the beginning of the sentence and SEP token at the end and prepped_sent_2_ which contains sentence2 with only SEP token in the end. Then we tokenize the prepped sentences to convert them into token lists, and then convert the tokens into their corresponding ids with the help of the tokenizer. Lastly, we generate token type ids, which differentiates between the first sentence and the second sentence. The table below shows the tokenization for banglabert_large.

6.2.3 Filtering, Encoding, and Resetting Indexes

After tokenization, we filter our dataset and only keep rows that contain the labels Entailment, Neutral, or Contradiction. Then we encode the labels where Entailment is 0, Neutral is 1, and Contradiction is 2. Lastly, on both the train and test DataFrames, we drop the old indexes and add new indexes to ensure consistency.

gold_label	sentence1	sentence2	prepped_sent_1_csebuetnlp/banglabert_large	prepped_sent_2_csebuetnlp/banglabert_large	tokenized_sent_1_csebuetnlp/banglabert_large	tokenized_sent_2_csebuetnlp/banglabert_large	sent1_token_type_csebuetnlp/banglabert_large	sent2_token_type_csebuetnlp/banglabert_large
Entailment	অনেক মেয়ে মানুষ বসে আছে।.	সব মেয়েদের মধ্যে কেউ দাঁড়িয়ে নেই।	2 অনেক মেয়ে মানুষ বসে আছে। 3	সব মেয়েদের মধ্যে কেউ দাঁড়িয়ে নেই। 3	[22, 1011, 1, 1019, 1346, 972, 205, 23]	[889, 1, 1021, 1206, 1, 1052, 205, 23]	[0, 0, 0, 0, 0, 0, 0, 0]	[1, 1, 1, 1, 1, 1, 1, 1]

Table 6.3: Tokenization for banglabert_large

6.2.4 Input Sequence, Attention Mask, and token_type

Then, for each model, for both train and test DataFrame, we prepared input sequence that contains tokenized sentence 1 concatenated with tokenized sentence 2. After that, attention masks was added to the input sequence using the the get_sent2_token_type function. Lastly, we set token_type_MODEL_NAME to sent1_token_type_MODEL_NAME concatenated with sent2_token_type_MODEL_NAME and kept the columns ”gold_label”, ”input_sequence_csebuetnlp/banglabert_large”, ”attention_mask_csebuetnlp/banglabert_large”, ”token_type_csebuetnlp/banglabert_large”, ”input_sequence_csebuetnlp/banglabert”, ”attention_mask_csebuetnlp/banglabert”, ”token_type_csebuetnlp/banglabert”, ”input_sequence_sagorsarker/bangla-bert-base”, ”attention_mask_sagorsarker/bangla-bert-base”, and ”token_type_sagorsarker/bangla-bert-base” in our Final DataFrame.

gold_label	input_sequence_csebuetnlp/banglabert_large	attention_mask_csebuetnlp/banglabert_large	token_type_csebuetnlp/banglabert_large
0	[22, 1011, 1, 1019, 1346, 972, 205, 23, 889, 1...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]	[0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

Table 6.4: After preparing input_sequence, adding attention mask and setting token_type

6.2.5 Custom Collate

Lastly, we used custom collate on our dataset instead of the default collate because:-

1. In order to form a batch, all sequences need to be of the same length. Using custom collate, we mitigate this issue by adding pad sequences, so all sequences are of equal length.

2. As we have multiple models of Bert that we are ensembling, using custom collate, we can ensure that the data from each model is correctly padded and batched.

3. Apply One-hot Encoding to golden_labels, as our dataset is a classification dataset.

Chapter 7

Models

7.1 Abstract model architecture

- In order to compare the performances of various pre-trained models, we designed a simple classification model where a multi-class classification head is added on top of the pre-trained model.

- The input to the model are the premise-hypothesis pair, attention mask, and token type ids while the output is one of entailment, neutral, and contradiction.

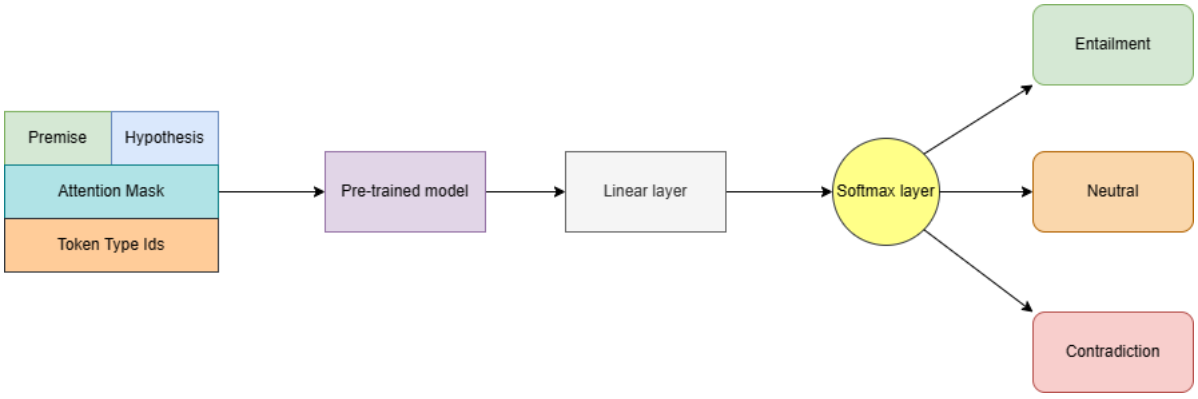

Figure 7.1: Abstract Model Architecture

7.1.1 BanglaBert

Using the strong BERT architecture, the BanglaBERT model is a transformer-based language model created especially for the Bengali language. It can interpret and produce native Bengali language with ease since it has been pre-trained on a large corpus of Bengali texts. BanglaBERT is used in the field of Natural Language Inference (NLI) to determine the connections between pairs of phrases in Bengali, determining whether one implies, is neutral towards, or

contradicts another.

7.1.2 ELECTRA-based BanglaBERT base

The BanglaBert base is a slightly upgraded version of the BanglaBert. It is a pre-trained discriminator model pre-trained using the replaced token detection objective, making it suitable for scenarios requiring faster inference and lower computational resources. Additionally, it is capable of comprehending and producing native Bengali language thanks to pre-training on a sizable corpus of text using the BERT architecture. Both models perform very well in NLI tasks, but the base model strikes a balance between computational effectiveness and language comprehension depth, making it a useful substitute in resource-constrained settings without appreciably compromising accuracy.

7.1.3 ELECTRA-based BanglaBERT Large

With more characteristics than its base version, the ELECTRA-based BanglaBERT Large model is a sophisticated transformer-based language model designed only for the Bengali language. It can comprehend and produce Bengali with great accuracy since it has been pre-trained on a sizable corpus of text in the language. When it comes to complex NLI tasks like advanced content analysis and automated reasoning with subtleties, this model performs very well. Because of its bigger size and increased ability to capture complex language patterns, the ELECTRA-based BanglaBERT Large model performs better than the ELECTRA-based BanglaBERT and BanglaBERT Base models. ELECTRA-based BanglaBERT Large prioritizes accuracy and depth of knowledge, making it perfect for situations where precision is crucial.

7.1.4 Roberta-XLM

As a strong multilingual transformer-based model RoBERTa-XLM is intended for a range of multilingual natural language understanding problems. It combines the advantages of XML (Cross-lingual Language Model) with RoBERTa (a robustly optimized BERT technique), making it especially suitable for jobs requiring complex text interpretation across languages, such as Natural Language Inference (NLI). Using its cross-lingual skills and strong contextual comprehension, RoBERTa-XLM excels in NLI by identifying links between phrases, such as entailment, contradiction, or neutrality. This model is a useful tool for cross-lingual semantic comprehension and inference tasks since its application in NLI guarantees excellent accuracy and consistency in multilingual environments.

7.1.5 Ensemble

In machine learning, an ensemble model integrates predictions from several models to improve overall performance, maximizing the benefits of each individual model while reducing its drawbacks. Ensemble models are very useful in Natural Language Inference (NLI) because they combine several viewpoints on sentence connections to provide predictions that are more reliable and accurate. Ensemble approaches can perform better in identifying entailment, contradiction, or

neutrality in phrase pairs by combining outputs from many models, such as transformers, recurrent neural networks, or even separate fine-tuned instances of the same architecture. In NLI applications, this method is useful for guaranteeing high recall and precision, mitigating the influence of biases or mistakes in any one model, and enhancing generalization across a variety of language circumstances.

We ensembled three models, Bangla Bert Base, ELECTRA-based Bangla Bert Base, and ELECTRA-based Bangla Bert Large. We first pass pass the input through each model and obtain each of their embeddings. Bangla Bert Base, ELECTRA-based Bangla Bert Base, and ELECTRA-based Bangla Bert Large have embeddings of size 768, 768, 1024 respectively. We then concatenate the three embeddings to obtain an embedding of size 2560. We then pass that embedding a 2560-dimensional dense layer and then finally through a 3-dimensional dense layer and softmax layer for final classification.

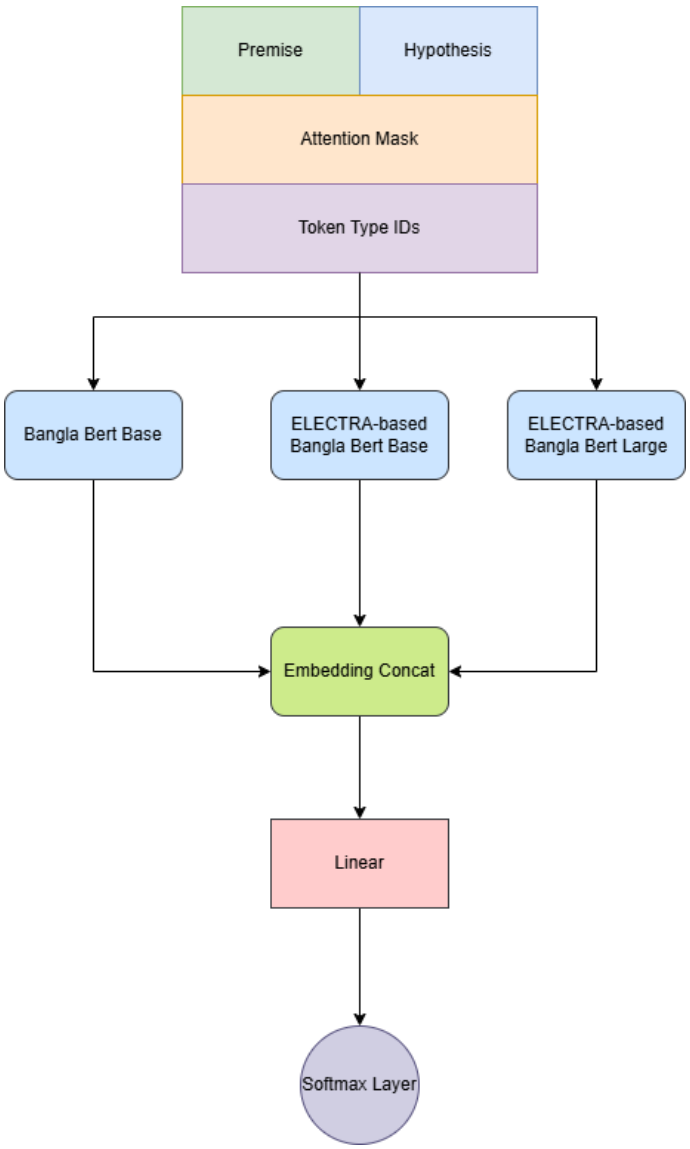

Figure 7.2: Overview of ensemble model

7.2 Results

We tested several different pre-trained models on our dataset: Bangla Bert Base, ELECTRA-based Bangla Bert Base, mBERT uncased, ELECTRA-based Bangla Bert Large, and XLM RoBERTa base. 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. Due to resource constraints, we used batch sizes of [2, 4, 8] and the same learning rates as before. To prevent overfitting, we used early stopping during training for all models. In 7.1, the hyperparameters that gave the best performance in terms of macro F1 score for each model is given. These models, trained on their corresponding set of hyperparameters listed in 7.1, were used for all subsequent tests in this paper.

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

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 7.2: Validation metrics for each model using the hyperparameters listed in 7.1

From 7.2 as well as 7.3, 7.4, 7.5, 7.6, we can see that ELECTRA-based Bangla BERT base, ELECTRA-based Bangla BERT large, are the best performing models, with our model ensemble of ELECTRA-based Bangla BERT base, ELECTRA-based Bangla BERT large, and Bangla BERT base giving a similar performance.

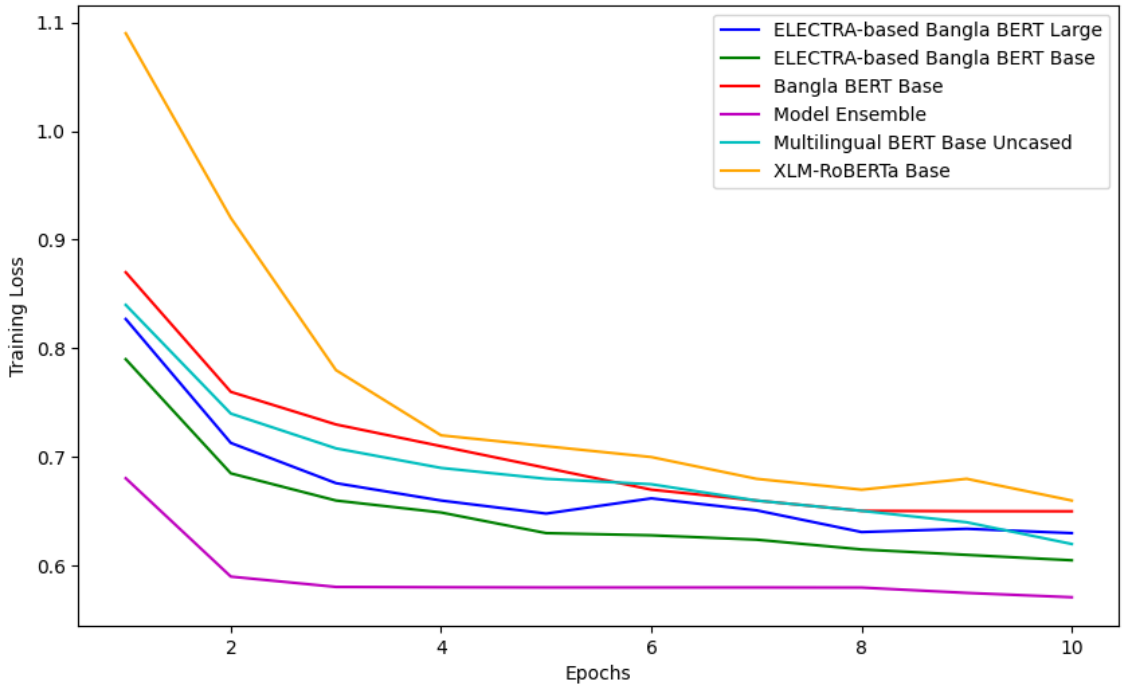

Figure 7.3: Training losses by epoch for the models used using the hyperparameters listed in 7.1

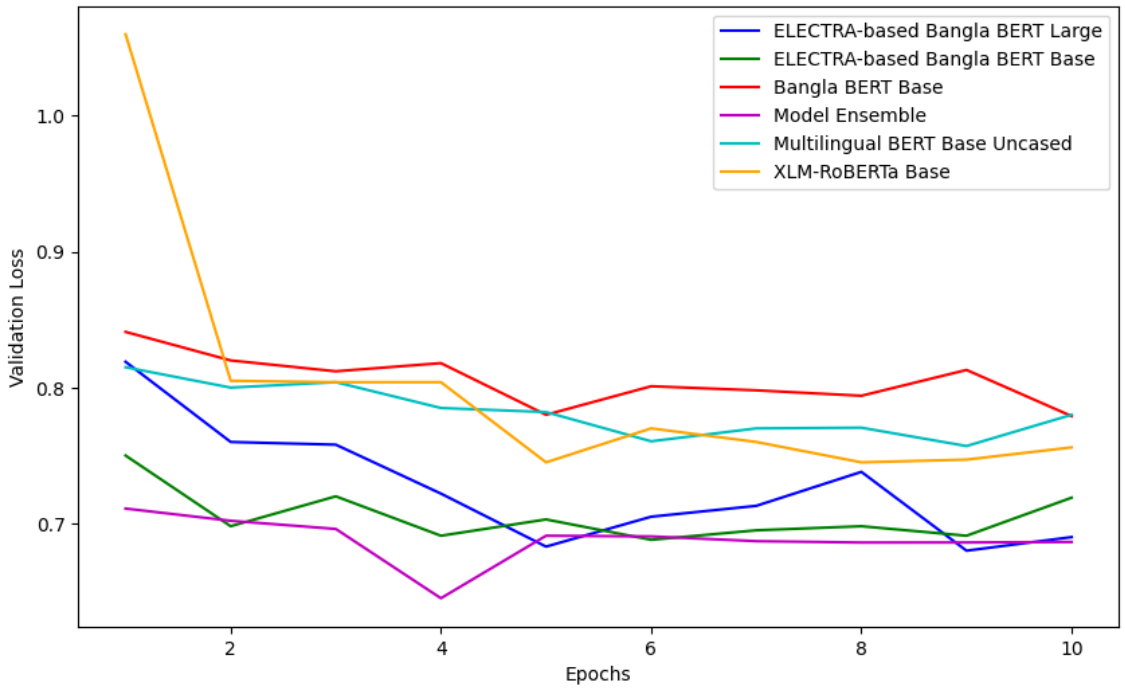

Figure 7.4: Validation losses by epoch for the models used using the hyperparameters listed in 7.1

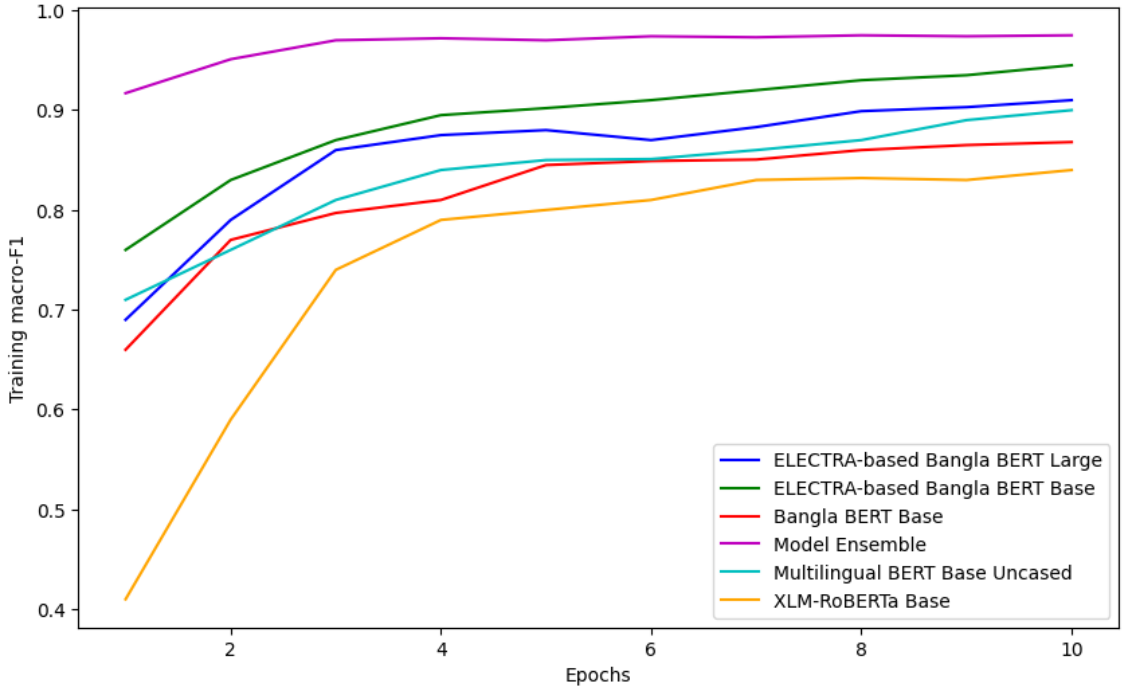

Figure 7.5: Training macro-F1s by epoch for the models used using the hyperparameters listed in 7.1

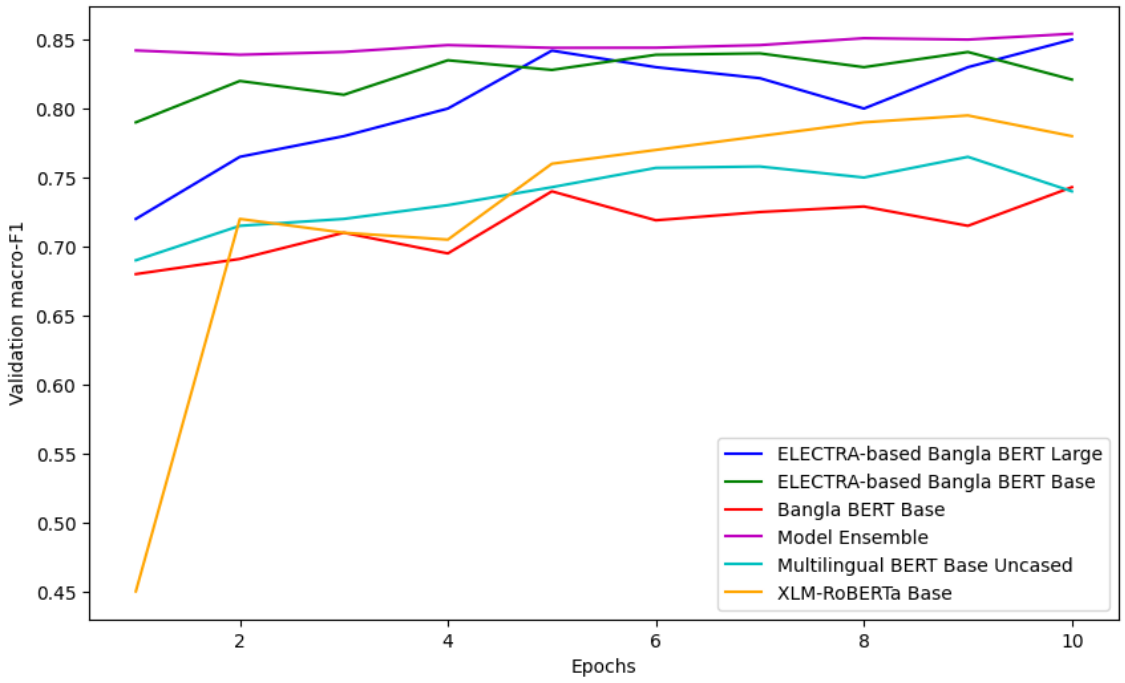

Figure 7.6: Validation macro-F1s by epoch for the models used using the hyperparameters listed in 7.1

Chapter 8

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 test cases were separated into two groups: correctly labeled and incorrectly labeled. Then the two groups were compared to find possible reasons for mislabeling.

Category	**Correctly Labeled**	**Incorrectly Labeled**	**Percentage 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 Premise	1.40	1.72	20.51%
Mean number of Unknown words in Hypothesis	0.69	0.83	18.42%

Table 8.1: 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 average number of unknown words in Premise and Hypothesis between the groups. Hence, a possible reason the model struggle is because of the words that the model could not tokenize properly.

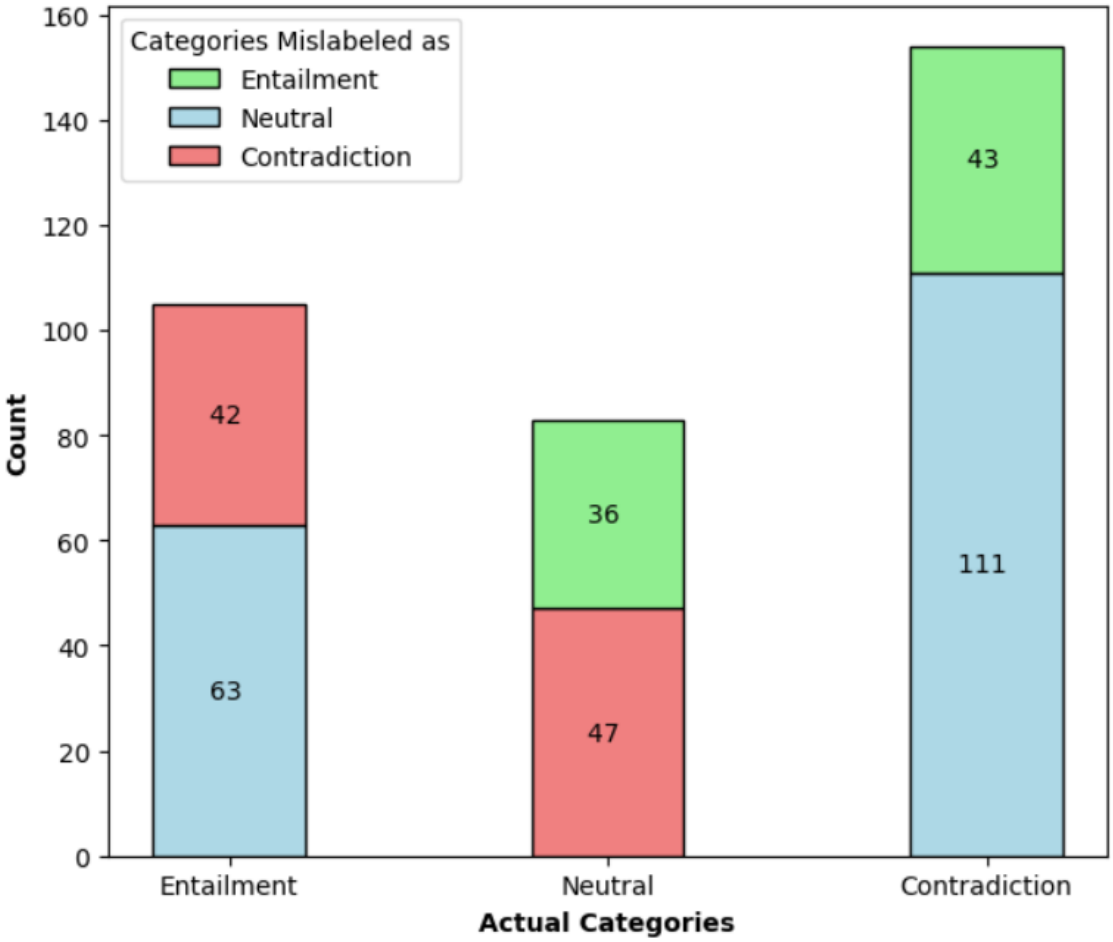

Figure 8.1: Count of labels given by our ensemble models for data points that were misclassified

Moreover, as can be seen from 8.1 the model struggles most with Contradiction. With the most frequent error being Contradictions being labeled as Neutral.

8.1 Independent two-tailed t-test analysis

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 for correctly labeled and incorrectly labeled samples. There are four factors under consideration here,

- Premise Length
- Hypothesis Length
- Number of unknown tokens in premise
- Number of unknown tokens in hypothesis

Thus, our null (H_0) and alternative hypotheses (H_A) are as follows:

H_0: There is no statistically significant difference between the means of the factor under consideration (premise length, hypothesis length, number of unknown tokens in premise, number of unknown tokens in hypothesis) between correctly labelled and incorrectly labelled samples.

H_A: There is a statistically significant difference between the means of the factor under consideration (premise length, hypothesis length, number of unknown tokens in premise, number of unknown tokens in hypothesis) between correctly labeled and incorrectly labeled samples.

Since the number of correct labels and incorrect labels are 2176 and 342 respectively, we assumed both samples have equal variance since both have over 30 samples.

We calculated the t-statistic using the formula,

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{S_1}{n_1} + \frac{S_2}{n_2}}} \tag{8.1}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means, S_1 and S_2 are the variances, and n_1 and n_2 are the sample sizes.

Degrees of freedom was calculated using,

$$\nu = n_1 + n_2 - 2 \tag{8.2}$$

where n_1 and n_2 are the sample sizes.

We used a significance level, α, of 0.05 along with our sample sizes of $n1$ and $n2$ of 2176 and 342 respectively.

We thus obtained a critical value of 1.96.

We then calculated the t-statistic for each factor and the results are summarized in 8.2

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

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 8.2: Independent two-taield t-test analysis results

8.2 Hypothesis-only Model

Like in [5], we trained our models using the hyperparameters in 7.1 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.

From 8.3, we can see that all the models are able to identify patterns in the dataset despite there not being any premises, indicating some spurious patterns have been introduced into the dataset. Regardless, comparing the results in 8.3 to the metrics obtained in 7.2, we see that premises are still needed to obtain good results on this dataset.

Model	Accuracy	Macro F1	Micro 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 8.3: Model Metrics when only hypotheses without premises were given

Chapter 9

Limitations and Future Work

9.1 Limitations

Although results have shown some promises, our dataset has limitations that can be addressed and improved upon in future works.

1. **Mislabelled Data:** Our error analysis has indicated that Contradiction has been mislabelled the most as a Neutral label. According to the metrics, 45.03% of the time contradiction statements were mislabeled, and 32.46% of the time that contradiction statement was mislabeled as a neutral statement. It would indicate that there are similarities between these texts and the model is unable to distinguish the semantics behind them.

2. **Unknown Words:** The presence of unknown words within the dataset is also something we discovered. Given that the dataset is new; pre-trained models, such as BanglaBERT base and BanglaBERT and others are unable to find these words in their corpus, hence causing a downfall in the dataset's efficiency. Some Bengali words carry important semantic information which may go unnoticed, thus making inaccurate predictions.

3. **Artifacts:** Artifacts were another concern that we discovered while annotating our data. We have tested out our models with the premises removed from our dataset. After reviewing the results we found that the hypothesis only approaches the best performance was an accuracy of 0.558 against an accuracy of 0.87 in the previous model. This suggests there are some artifacts, but they are not significant. The artifacts pose a problem in making a nuanced dataset and if not handled properly it could make the dataset unusable as model's will not be able to learn properly.

4. **Lack of Diversity:** Our dataset did not contain enough diversity of information. Even with 12,600 rows of data, diversity is missing. A NLI dataset or rather a language dataset should contain a vast reservoir of information from different backgrounds. For example, from newspaper articles, journals, and academic papers. Without them the model's ability to generalize across different contexts becomes limited.

5. **Insufficient Data:** In our study, we only managed to create 12,600 rows of data. And our model is missing a variety of crucial data which it can use to generalize further. Additionally, having more data will allow the model to achieve higher results, thus increasing the dataset's validity further. We only used four annotators however, if we use more annotators we can attain more data and as well as a diverse set of information.

9.2 Future Work

Our results and outcome show room for major improvements. As we have created and annotated the dataset from scratch using image captions; we believe this dataset has more potential than any other Bangla NLI dataset available. Our annotations were generated with the help of Bangladeshi people who are well-versed in speaking their native language. However, the dataset can be further enhanced for more usability. Here are some examples.

1. **Data Diversity:** As an NLI dataset, it should contain a more diverse and informative set of texts, each relevant to certain topics. Our sentences cover everyday Bangla sentences spoken by the native. However, we believe that the model's generalizing ability is limited as it does have access to different kinds of texts. With only four annotators it is difficult to achieve this task. Hence, we believe that adding more data, and as well as a diverse set of topics will improve the dataset's capability in training models.

2. **Annotation Consistency:** With only four annotators annotation may not be consistent. Even with our high Kappa Score some annotations can still be mislabeled or inconsistent. Hence, the dataset's annotation can be better with the help of more annotators and especially people from Bangladesh who are well-versed in Bangla. We can use crowdsourcing as an effective way to attain more data with more consistency.

3. **Labels with Explanations:** Having labels with explanations can provide more semantic information that the model can utilize. We found that our models have mislabelled data quite a few times, especially between neutral and contradiction statements. With these explanations, models can understand more about why a statement is an entailment, neutral, or contradiction statement. Moreover, this could also be used to mitigate the effects that artifacts have on the dataset.

4. **Annotation Methods:** Our annotators filled up by typing sentences. This method is prone to creating or generating unknown words which may not be available or recognized by the NLI models. Rather than typing text we can change the data annotation method by using multiple choice questions where the annotators simply pick out the sentences and label them as such. Moreover, this could ensure more consistency and remove dataset biases as well.

Chapter 10

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 many groundwork for improvement in the dataset which we believe will bring out the best in this dataset.

Bibliography

[1] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075. [Online]. Available: https://aclanthology.org/D15-1075.

[2] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom, *E-snli: Natural language inference with natural language explanations*, 2018. arXiv: 1812.01193 [cs.CL].

[3] M. Glockner, V. Shwartz, and Y. Goldberg, *Breaking nli systems with sentences that require simple lexical inferences*, 2018. arXiv: 1805.02266 [cs.CL].

[4] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith, "Annotation artifacts in natural language inference data," *arXiv preprint arXiv:1803.02324*, 2018.

[5] A. Poliak, J. Naradowsky, A. Haldar, R. Rudinger, and B. Van Durme, "Hypothesis only baselines in natural language inference," *arXiv preprint arXiv:1805.01042*, 2018.

[6] A. Williams, N. Nangia, and S. 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)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1112–1122. DOI: 10.18653/v1/N18-1101. [Online]. Available: https://aclanthology.org/N18-1101.

[7] N. Mansoor, A. H. Kamal, N. Mohammed, S. Momen, and M. M. Rahman, *Banglalekhaimagecaptions*, Mendeley Data, version 2, 2019. DOI: 10.17632/rxxch9vw59.2.

[8] A. Talman and S. Chatzikyriakidis, *Testing the generalization power of neural network models across nli benchmarks*, 2019. arXiv: 1810.09774 [cs.CL].

[9] X. Wang, P. Kapanipathi, R. Musa, *et al.*, "Improving natural language inference using external knowledge in the science questions domain," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 7208–7215, Jul. 2019. DOI: 10.1609/aaai.v33i01.33017208. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/4705.

[10] M. Artetxe, G. Labaka, and E. Agirre, "Translation artifacts in cross-lingual transfer learning," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 7674–7684. DOI: 10.18653/v1/2020.emnlp-main.618. [Online]. Available: https://aclanthology.org/2020.emnlp-main.618.

[11] T. Hasan, A. Bhattacharjee, K. Samin, *et al.*, "Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, Nov. 2020, pp. 2612–2623. DOI: 10.18653/v1/2020.emnlp-main.207. [Online]. Available: https://aclanthology.org/2020.emnlp-main.207.

[12] H. Hu, K. Richardson, L. Xu, L. Li, S. Kuebler, and L. S. Moss, *Ocnli: Original chinese natural language inference*, 2020. arXiv: 2010.05444 [cs.CL].

[13] A. Bhattacharjee, T. Hasan, K. Samin, *et al.*, *Banglabert: Combating embedding barrier in multilingual models for low-resource language understanding*, 2021. arXiv: 2101.00204 [cs.CL].

[14] A. Bhattacharjee, T. Hasan, W. Ahmad, *et al.*, "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla," in *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1318–1327. DOI: 10.18653/v1/2022.findings-naacl.98. [Online]. Available: https://aclanthology.org/2022.findings-naacl.98.

[15] T. Schuster, S. Chen, S. Buthpitiya, A. Fabrikant, and D. Metzler, "Stretching sentence-pair nli models to reason over long documents and clusters," *arXiv preprint arXiv:2204.07447*, 2022.

[16] H. Amirkhani, M. AzariJafari, S. Faridan-Jahromi, Z. Kouhkan, Z. Pourjafari, and A. Amirak, "FarsTail: A persian natural language inference dataset," *Soft Computing*, Jul. 2023. DOI: 10.1007/s00500-023-08959-3. [Online]. Available: https://doi.org/10.1007%2Fs00500-023-08959-3.

[17] A. Bhattacharjee, T. Hasan, W. U. Ahmad, and R. Shahriyar, *Banglanlg and banglat5: Benchmarks and resources for evaluating low-resource natural language generation in bangla*, 2023. arXiv: 2205.11081 [cs.CL].

[18] *Hugging face*. [Online]. Available: https://huggingface.co/.