



Islamic University of Technology (IUT)
Department of Computer Science and Engineering (CSE)

Bangla Dataset Generation for Natural Language Inference

Members

Md. Shohidul Islam -170041055

Abdun Nayeem Khan -180041113

Md Shaidur Rahman Nizami -180041139

Supervisor

Dr.Hasan Mahmud

Associate Professor, CSE, IUT (SSL)

**A thesis submitted to the Department of CSE
in partial fulfillment of the requirements for the degree of B.Sc.**

Engineering in CSE

Academic Year: 2021-2022

May - 2023

Acknowledgement

We would like to express our grateful appreciation for **Dr. Hasan Mahmud**, **Dr. Kamrul Hasan** and **Md. Mohsinul Kabir** from Department of Computer Science & Engineering, IUT for being our adviser and mentor. Their motivation, suggestions and insights for this research have been invaluable. Without their support and proper guidance this research would never have been possible. Their valuable opinion, time and input provided throughout the thesis work, from first phase of thesis topics introduction, subject selection, proposing algorithm, modification till the project implementation and finalization which helped us to do our thesis work in proper way. We are really grateful to them. We would also like to express our gratitude to the data annotators. They worked tirelessly to create the dataset, diligently adhering to all the given instructions.



Declaration of Authorship

This is to certify that the work presented in this thesis book "Bangla Dataset Generation for Natural Language Inference", is the outcome of the investigation and research carried out by Abdun Nayeem Khan, Md. Shohidul Islam and Md Shaidur Rahman Nizami, under the supervision of Dr. Kamrul Hasan and Dr. Hasan Mahmud. The submission of the thesis or any portion of it elsewhere for the granting of a degree, diploma or anyother qualification is hereby expressly prohibited. A list of references is provided, and information taken from other people's published or unpublished work has been recognized in the text.

Author's Signature -

<i>Nayeem</i>	<i>Shohidul</i>	<i>Nizami</i>
Abdun Nayeem Khan	Md. Shohidul Islam	Md Shaidur Rahman Nizami
180041113	170041055	180041139

Supervisor's Signature -

	
Dr. Hasan Mahmud	Dr. Kamrul Hasan
Associate Professor, CSE, IUT	Professor, CSE, IUT

Abstract

Understanding entailment and contradiction is fundamental to understanding natural language, and inference about entailment and contradiction is a valuable testing ground for the development of semantic representations. However, machine learning research in this area has been dramatically limited by the lack of resources in Bangla. To address this, we propose to introduce our own corpus curated for natural language inference which is labeled pairs of sentences with a label that depicts their inner entailment. Our goal is to create a dataset that has over 30K instances and to do so we have now created a Bangla dataset by machine translating the SNLI corpus into Bangla. After that, we show that benchmark models can be used to evaluate and do the task of inference in Bangla . We hope that our dataset will catalyze research in Bangla sentence understanding by providing an informative standard evaluation task. For this we provided two baseline models which are both considered integral in the task of inference in any language.

keywords - entailment, contradiction, neutral, natural language, inference, semantic representations, machine learning, Bangla, corpus, labeled pairs of sentences, inner entailment, dataset, instances, SNLI corpus, machine translation, benchmark models, evaluation task, baseline models, sentence understanding

Contents

1	Introduction	3
1.1	Overview	3
1.2	Problem Statement	4
1.3	Contribution	4
1.4	Motivation & Scopes	5
1.5	Thesis Outline	6
2	Literature Review	7
2.1	Natural language Inference and it's implementation with different models	7
2.2	NLI and related datasets	11
2.3	BERT Models for different languages including Bengali	19
2.4	Different kind of dataset in Bengali apart from NLP and NLI	23
2.5	Question Answering through entailment	25
2.6	Bangla NLP Tasks in Transformer Models	28
2.7	Advances in NLI	31
2.8	Bangla NLP datasets and their tasks	35
3	Proposed Approach	50
3.1	Data collection	50
3.2	Data Annotation	51
3.3	Validation	51
4	Providing Benchmark	53
4.1	Preprocessing	53
4.2	Tokenization	54
4.2.1	NLTK	54
4.2.2	Wordpiece	54
4.3	Training	55

5	Result Analysis & Discussion	56
5.1	Experimental Result	56
5.1.1	Model Description	56
5.1.2	Setup	56
5.1.3	Accuracy Comparison	56
5.2	Accuracy chart	57
5.3	Result Analysis	58
6	Conclusion and Future Work	60

1 Introduction

1.1 Overview

Humans employ a range of knowledge and reasoning to comprehend the meanings conveyed through language. For instance, let's consider the following sentences by Minsky (2000): "Jack needed some money, so he went and shook his piggy bank. He was disappointed when it made no sound." From this, we effortlessly infer that Jack didn't find any money, leading to a negative emotional response. This inference is drawn based on our understanding of the world and our ability to connect pieces of knowledge through common sense reasoning. We know that a piggy bank is a container for holding coins, and coins are a form of currency made of metal. Since coins are solid and hard, they make a sound when shaken in a container like a piggy bank. Thus, the absence of sound indicates the absence of coins. Additionally, it is probable that we can predict Jack being a child, as piggy banks are typically owned by children. Alternatively, these predictions may stem from similar childhood experiences, enabling us to draw analogous conclusions. While humans naturally possess this knowledge and reasoning capability, machines struggle to replicate it.

Despite significant advancements in natural language processing, machines still struggle to possess the same natural language inference (NLI) abilities. To address this challenge, research in NLI has witnessed substantial growth in recent years. The concepts of entailment and contradiction play a central role in understanding natural language meaning across various levels, from individual words to entire texts. Present-day natural language processing systems heavily rely on annotated data for learning specific tasks. Typically, training data is available in a single language, limiting the system's ability to perform tasks in multiple languages. However, international products require systems that can handle inputs in numerous languages.

Although Bangla is one of the most widely spoken languages globally, it is considered a low-resource language in terms of digitization. This is primarily

due to the scarcity of annotated computer-readable datasets and limited support for resource development. Additionally, the available datasets for Natural Language Inference tasks are inadequate and predominantly generated through machine translation, lacking human annotation. The NLP community has a history of creating benchmarks and resources to facilitate algorithm development and evaluation for various language processing tasks. However, this paper focuses specifically on the ongoing research efforts related to benchmarks, resources, and approaches for natural language inference (NLI).

In this paper, we introduce a novel dataset for inference tasks in Bangla, comprising labeled sentence pairs. We evaluate different models, such as BERT and LSTM-based neural networks, using synthetically generated data. Both models demonstrate comparable performance.

1.2 Problem Statement

The primary goal of Natural Language inference is to show the connection between the premise and the hypothesis. Such tasks need to be done in Bangla too. But Bangla has a lacking of resources and most of the natural language datasets are not suitable for the task of NLI. Natural Language Inference which is also known as Recognizing Textual Entailment (RTE) is a task of determining whether the given “hypothesis” and “premise” logically follow (entailment) or unfollow (contradiction) or are undetermined (neutral) to each other. These classifications can be done for Bangla instances too but there are no curated dataset for it, thus our goal is to create a dataset that resolves this problem.

1.3 Contribution

To address this, we propose a new dataset created to do the task of inference in Bangla, a collection of sentence pairs labeled for entailment, contradiction, and semantic independence. In this paper, we use our own synthetically generated data to evaluate a variety of models for natural language inference in Bangla, including rule-based systems, simple linear classifiers, and neural network-based models. We

find that two models achieve comparable performance: a feature-rich pre-trained model(BERT) and a neural network model centered around a Long Short-Term Memory network(LSTM).

Our thesis primarily concentrates on two key aspects. Firstly, it aims to establish the development of a dataset for the Natural Language Inference Task specifically in Bengali. Secondly, our data creation process adhered to a meticulous approach by involving the Human-in-the-Loop (HITL) method throughout its creation. We strongly believe that these contributions will have a significant impact on the establishment of a standardized dataset.

1.4 Motivation & Scopes

With our dataset, we sought to address the issues of size, quality, and indeterminacy. To do this, we aim to employ a crowdsourcing framework with the following crucial innovations. First, the examples must be grounded in specific scenarios, and the premise and hypothesis sentences in each example must be constrained to describe that scenario from the same perspective, which helps greatly in controlling event and entity coreference. To test our efficiency and the procedure of inference can be done or not, we first have to test the baseline models with synthetic data which is close to the finalized dataset.

To simply put the motivation of our work, we can say-

- Firstly, Lack of resources in bengali language
- Secondly, No prominent NLI curated dataset
- Thirdly, Datasets with NLI labels are few in instances in bangla(around 5k-7K)
- Most BNLN resources are not annotated correctly in terms of labels for NLI

1.5 Thesis Outline

In Chapter 1 we have discussed our study in a precise and concise manner. Chapter 2 deals with the necessary literature review for our study and there development so far. In Chapter 3 we have stated the skeleton of our proposed method, proposed algorithm and also the flowchart to provide a detail insight of the working procedure of our proposed method. Chapter 4 shows the results and comparative analysis of the successful implementation of our proposed method. The final segment of this study contains all the references and credits used.

2 Literature Review

2.1 Natural language Inference and it's implementation with different models

Both human and machine intelligence rely heavily on reasoning and inference. As noted by MacCartney and Manning, "a necessary (if not sufficient) condition for true natural language understanding is a mastery of open-domain natural language inference." Although modeling inference in human language is notoriously difficult, it is a fundamental problem towards true natural language understanding. A lot of study has been done on identifying textual entailment in earlier work. As seen in the example from MacCartney below, where the hypothesis is considered to be implied from the premise, natural language inference (NLI) is specifically concerned with assessing whether a natural language hypothesis h can be inferred from a premise p .

Humans make use of a wide range of information and reasoning to decipher linguistic meanings. Take these quotations from Minsky as an illustration: "Jack needed some money, so he went and shook his piggy bank. When it didn't make a sound, he was dissatisfied. We may easily deduce from this that Jack did not discover any money, and as a result, Jack was experiencing a bad emotion. The knowledge we have about the world and the underlying reasoning process—often referred to as commonsense thought or commonsense reasoning—that enables us to connect bits of knowledge to arrive at the new conclusion are what led us to this conclusion, which was not expressly mentioned in the chapter. We understand what a piggy bank is. In a container like a piggy bank, the coins will rattle when shaken since metal is a hard solid; if there is no sound, there are no coins. Additionally, there is a significant likelihood that we can conclude that Jack is a youngster because piggy banks are frequently owned by kids. Alternately, we can infer these predictions from experiences we had as kids that allowed us to draw analogous inferences about them. While human readers have a natural ability to

understand and reason in this way, machines are infamously bad at it.

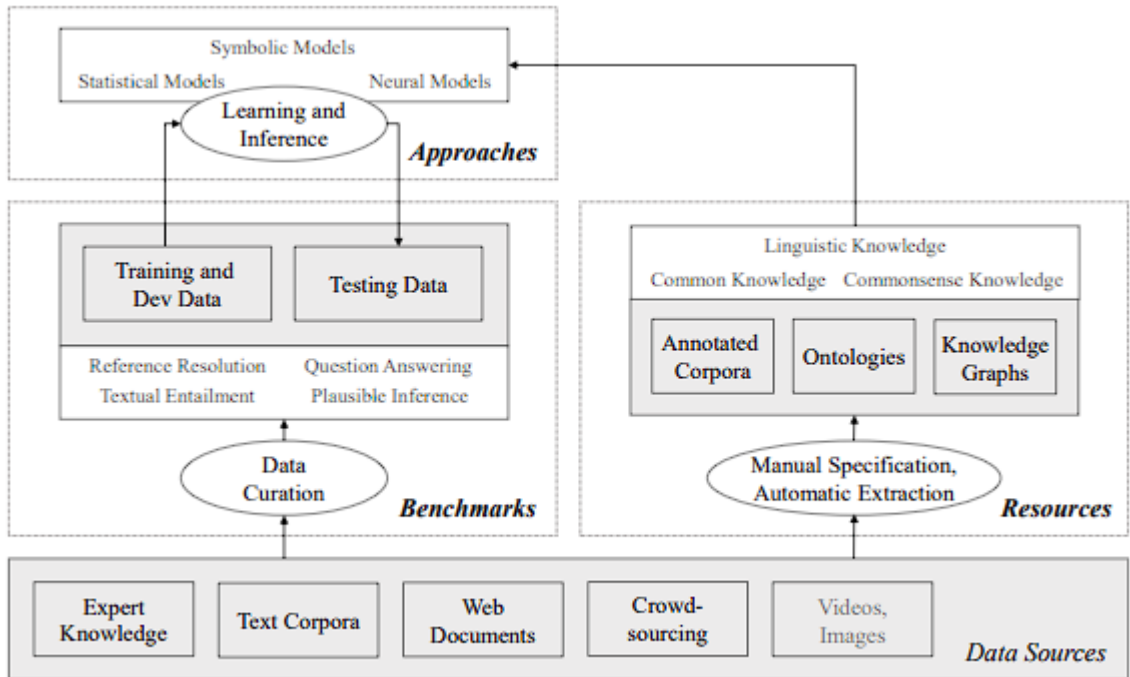


Figure 1: Main research efforts in natural language inference from the NLP community occur in three areas: benchmarks and tasks, knowledge resources, and learning and inference approaches

The Stanford Natural Language Inference (SNLI) corpus, newly released, aims to promote more learning-focused methods to NLI. Around 570K sentence pairings with three labels—entailment, contradiction, and neutral—can be found in this corpus. The corpus size now allows for the building of deep neural network models, which previously required a substantial amount of training data. In this study, we suggest a unique design for NLI called LSTM (long short-term memory). We base our model on a very original neural attention model for NLI that was just recently proposed. Instead of generating sentence embedding for the premise and the hypothesis to be utilized for classification, our technique used a match-LSTM to do word-by-word matching of the two. This LSTM can give more weight to significant word-level matching results. Each step in the hypothesis processing

involves comparing the current word with an attention-weighted representation of the underlying assumption. Here, attention-weighted vector representations of the premise were initially derived using neural attention models. A match-LSTM was then created, processing the hypothesis word by word while attempting to match it with the premise. We used the last hidden state of this mLSTM to forecast the link between the premise and the hypothesis. The SNLI corpus trials demonstrated that the mLSTM model delivered the cutting-edge results claimed for this data set. On the SNLI corpus, our model outperforms the state of the art with an accuracy of 86.1.[1]

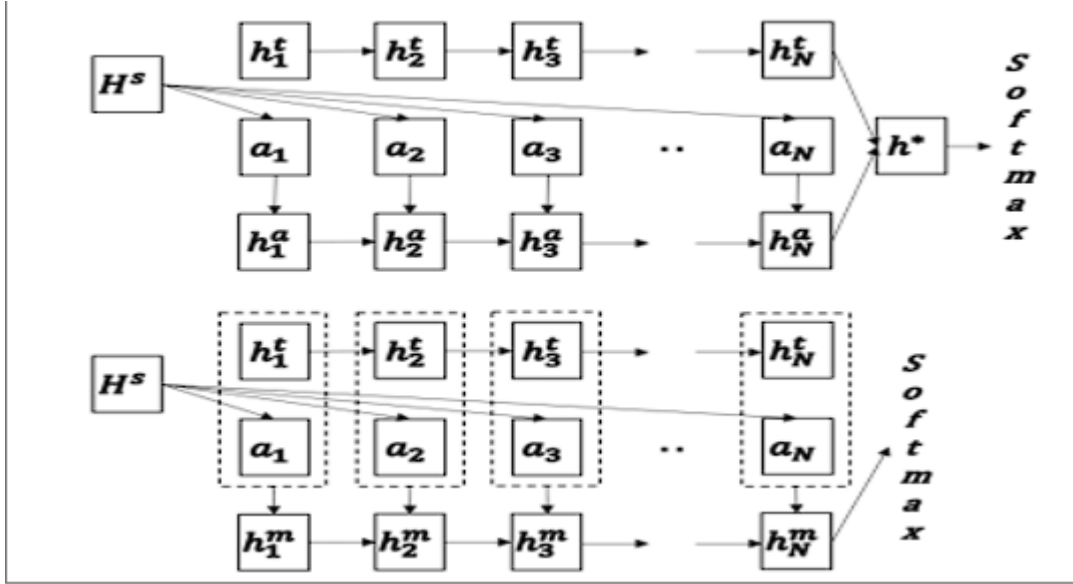


Figure 2: LSTM embedding layers

In this research, we suggested a sentence encoding-based methodology for recognizing text entailment. The fundamental goal of recognizing text entailment (RTE), assuming a pair of sentences is given, is to establish if the hypothesis can be derived legitimately from the premises. Entailment (inferred to be true), Contradiction (inferred to be false), and Neutral (truth unknown) are the three types of relations that it consists of. Here, we attempt to provide an integrated deep learning framework for textual entailment recognition that does not require any feature engineering or other resources. The foundational model is built on creating biLSTM models on both the premises and the hypothesis.

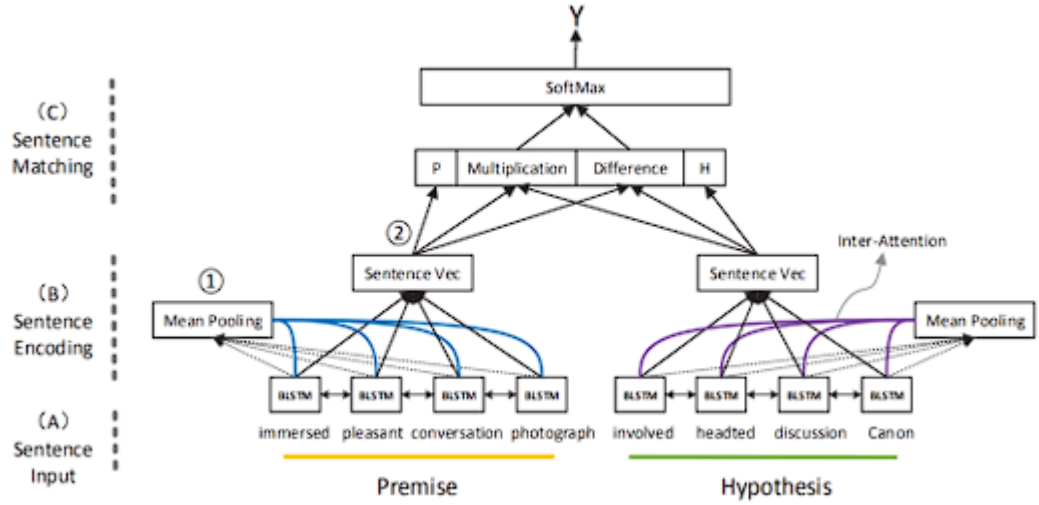


Figure 3: BiLSTM embedding layers

Our strategy assumes that sentence encoding is a two-stage procedure. First, word-level bidirectional LSTM (biLSTM) average pooling was utilized to create a first stage sentence representation. Second, average pooling on the same text was replaced with the attention mechanism for improved representations. We employed a method known as "Inner-Attention" to attend to words within the target sentence rather than using the target sentence to focus on particular words in the source sentence. The Stanford Natural Language Inference Corpus experiment results validated the efficacy of the Inner-Attention technique. Our model performed noticeably better than the previous best sentence encoding-based method while having fewer parameters.

We developed the concept of attention within a phrase, which enables the model to concentrate on pertinent terms without requiring input from another sentence. Through the application of attention vectors, the Inner-Attention mechanism enhances the precision of phrase representations. Additionally, the performance of our model is further improved by the straightforward and efficient input diversification technique we developed. Other sentence-matching models can also be simply adapted to this approach. The limitations include :

- Difficulty with common sense reasoning: Machines struggle to understand

and reason using common sense, unlike humans. This limits their ability to comprehend natural language.

- Challenges in modeling human language inference: Modeling inference in human language is complex and poses a significant hurdle in achieving true natural language understanding.
- Dependency on large amounts of training data: Deep learning models require extensive training data, which can be a limitation when dealing with limited annotated data or different domains.
- Limited generalization: Models trained on specific datasets may struggle to generalize to new or unseen data, reducing their overall usefulness.
- Insufficient consideration of context and world knowledge: Models may not fully capture contextual nuances and external knowledge, limiting their understanding of complex language tasks.
- Lack of explainability: Deep learning models can be complex and hard to interpret, making it challenging to understand their decision-making process.
- Limited cross-lingual capabilities: The approaches primarily focus on English language understanding, making it difficult to transfer them to other languages or achieve robust cross-lingual performance.

[2]

2.2 NLI and related datasets

The mentioned paper discusses how supervised learning gives state of the art performance when trained with natural language inference dataset called Stanford Natural Language Inference. Apart from that, natural language inference also helps to do other Natural Language Process tasks. In short, the researched dataset is also handy for natural language processing. Firstly, the NLI task is explained in this paper. A sentence encoder is used for the representation of premises and

hypothesis (u and v). 3 different methods are used to find out the relationship between the sentence vectors of premises and hypothesis (concatenation, element wise product, absolute element wise difference). The resulting vector is fed into a 3-way classifier to find out the relation between the sentences.

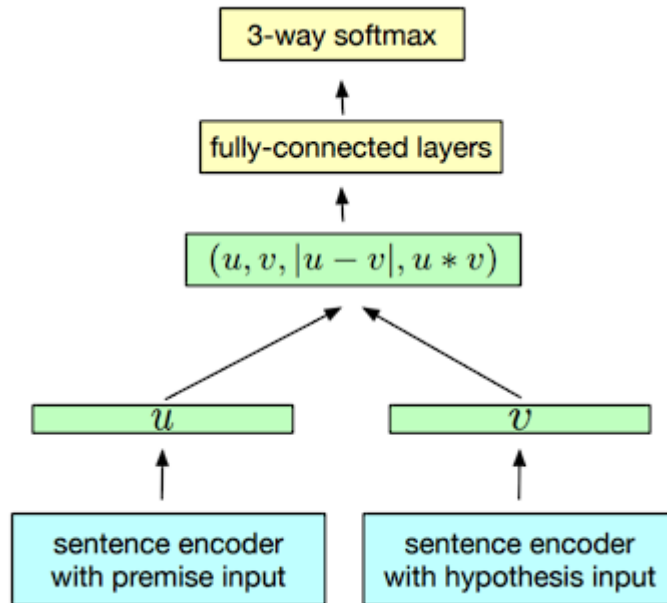


Figure 4: Generic NLI training Scheme

Next, many encoder model options are discussed. Finding the most effective encoder that would allow the authors to reach cutting-edge performance is the ultimate objective. The 7 models included self-attentive networks, hierarchical convolutional networks, concatenation of last hidden states of forward and backward GRUs, bi-directional LSTMs (BiLSTM with mean and max pooling), and standard recurrent encoders with either Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU). Binary and multi-class classification, Entailment and semantic relatedness, Semantic Textual Similarity, Paraphrase detection and Caption-Image retrieval were used as sentence embedding evaluation procedures.

As it turns out, the model trained on NLI outperforms models trained on other supervised and unsupervised tasks. The model trained on Bi-LSTM with max pooling makes the best sentence encoding methods (state of the art on SNLI).

Model	dim	NLI		Transfer	
		dev	test	micro	macro
LSTM	2048	81.9	80.7	79.5	78.6
GRU	4096	82.4	81.8	81.7	80.9
BiGRU-last	4096	81.3	80.9	82.9	81.7
BiLSTM-Mean	4096	79.0	78.2	83.1	81.7
Inner-attention	4096	82.3	82.5	82.1	81.0
HConvNet	4096	83.7	83.4	82.0	80.9
BiLSTM-Max	4096	85.0	84.5	85.2	83.7

Figure 5: Results

[3]

The availability of resources is obvious in the field of natural language inference. As a result of this, the scopes of research in this area is declining. To solve this problem, the authors of this paper have come up with a solution. They have introduced a new dataset called Stanford Natural Language Inference corpus. A 570k human-written phrase pair dataset with labels for entailment, contradiction, and neutral connection makes up the dataset. On tasks requiring natural language inference, such a dataset greatly benefits neural network-based models.

The paper discusses how limited the current scopes for NLI were before the introduction of SNLI. The main reason being the shortage of corpuses or limited number of instances in the present corpuses. The paper then discusses the learning centered approaches for NLI and how we can achieve those with a new corpus. Amazon Mechanical Turk was used to initially collect data. Flickr30k corpus consisting of almost 160k captions was used for premises. The authors used a technique where they provided the annotators a list of captions and instructed to create a sentence that becomes a true statement, a false statement and a statement which might be true or false. These later on were labeled as entailment, contradiction and neutral states.

The mechanical labeling task that was used to label the SICK entailment data served as a model for the validation phase’s structure, which was evaluated. The authors presented pairs of phrases to the participants in groups of five and asked

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely a true** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "There are animals outdoors."*
- Write one alternate caption that **might be a true** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "Some puppies are running to catch a stick."*
- Write one alternate caption that is **definitely a false** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "The pets are sitting on a couch." This is different from the maybe correct category because it's impossible for the dogs to be both running and sitting.*

them to choose the proper label for each pair. They assigned four annotators to each pair, resulting in five labels per pair, including the label for the original author. Approximately 10 % of the entire data was subjected to this validation. A new label called "gold label" was introduced to validate the pairs. If a label was chosen by at least 3 of the 5 annotators for that single pair of sentences, it was chosen as the gold label. Otherwise discarded (about 2 percent of total data).

In conclusion, to date the full potential of NLI could not be assessed due to a shortage of resources till now. But with the introduction of SNLI, this problem can be mitigated and groundbreaking research can be easily done on the domain of NLI. [4]

The Multi-Genre Natural Language Inference (MultiNLI) corpus, a dataset for developing and analyzing machine learning models for sentence comprehension, is presented in this study. This corpus, which has 433k samples, is one of the largest ones for natural language inference, also known as textual entailment recognition. In comparison to earlier resources in the topic, it provides both broad coverage and demanding instances. The MultiNLI corpus contains both spoken and written language in a wide range of formalities, topics, and styles. This corpus' primary

General:	
Validated pairs	56,951
Pairs w/ unanimous gold label	58.3%
Individual annotator label agreement:	
Individual label = gold label	89.0%
Individual label = author’s label	85.8%
Gold label/author’s label agreement:	
Gold label = author’s label	91.2%
Gold label \neq author’s label	6.8%
No gold label (no 3 labels match)	2.0%
Fleiss κ:	
<i>contradiction</i>	0.77
<i>entailment</i>	0.72
<i>neutral</i>	0.60
Overall	0.70

objective was to serve as a benchmark for cutting-edge machine learning research on fundamental natural language understanding (NLU) issues. The corpus is also intended to enable research on cross-domain transfer learning and domain adaptation. The MultiNLI corpus contains both spoken and written language in a wide range of formalities, topics, and styles. This corpus’ primary objective was to serve as a benchmark for cutting-edge machine learning research on fundamental natural language understanding (NLU) issues. The corpus is also intended to enable research on cross-domain transfer learning and domain adaptation. The MultiNLI corpus was created to make it possible to evaluate models explicitly based on the caliber of their phrase representations in the training domain and their capacity to provide plausible representations for uncharted territory. The corpus contains material from ten distinct spoken and written English genres, illustrating the wide spectrum of usage for current standard American English.

Only five of the genres are present in the training set, compared to all of them in the test and development sets. As a result, models can be tested on both matched test examples, which are drawn from the same sources as the training set, and mismatched examples, which are instances that are not very similar to those seen

during training. Similar to SNLI, MultiNLI uses a similar data collection process: Each sentence pair is produced by choosing an existing sentence from a text source as the premise, and then requesting a human annotator to write a brand-new sentence as the hypothesis. The source texts were just lightly preprocessed in order to produce the premise sentences for the MultiNLI corpus. Very short sentences were removed, and each genre’s sentences were made distinctive. Additionally, some non-narrative writing styles were carefully eliminated. A crowdworker was given a sentence from a source text and instructed to come up with three new sentences (the hypotheses) to go with it in order to construct a sentence pair for the MultiNLI corpus. One of the hypotheses had to be necessarily true or appropriate whenever the premise was true (designated as ENTAILMENT), one had to be necessarily false or inappropriate whenever the premise was true (designated as CONTRADICTION), and one couldn’t have any connection to the premise (designated as NEUTRAL). By using this technique, the raw corpus will contain an equal number of samples for each of the three categories. The MultiNLI corpus represents a more difficult assignment, according to evaluations using machine learning models trained on the Stanford NLI corpus, despite the fact that the two corpora have comparable levels of inter-annotator agreement. The empirical coverage and level of difficulty of the MultiNLI corpus are both higher than those of the SNLI corpus. Instead than merely simple image captions, it has a representative sample of text and voice from ten distinct genres, and there are additional phrases with one or more tags from a set of thirteen difficult linguistic occurrences. [5]

The learning process for current natural language processing systems, such as classification, sequence tagging, or natural language inference, frequently relies on annotated data. The resultant system can only complete the task in that language because the majority of the training data is only available in that one language. Systems used in global products frequently need to process inputs in multiple languages. In these circumstances, it is frequently impossible to annotate data in all of the languages that a system might use.

The Cross-lingual Natural Language Inference (XNLI) corpus, a benchmark for NLP systems that covers 15 languages, is introduced in this paper. The XNLI corpus consists of 112,500 annotated pairs and 7500 human-annotated development and test examples in the following languages: English, French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili, and Urdu. These languages are members of numerous language families, and the corpus now also contains the lower-resource languages of Swahili and Urdu. This study examines a number of methods for cross-lingual learning of natural language inference that employ parallel training data from publically accessible corpora.

Numerous studies have been conducted on multilingualism on language comprehension at the word level. Learning cross-lingual word representations, which are word embeddings where translations are close to one another in the embedding space, has been approached in a variety of ways. The continuous bag-of-words (CBOW) method, which takes the average or weighted average of the word embeddings, is the simplest way to create sentence embeddings. Although straightforward, this approach frequently provides a solid foundation.

The Multi-Genre NLI test section was kept secret, therefore the Cross-lingual NLI Corpus (XNLI) is based on new English natural language inference (NLI) data. The same crowdsourcing-based method used for the Multi-Genre NLI corpus was utilized to gather the core English data, yielding 7500 additional samples from each of the ten text sources included in that corpus. These samples were then professionally translated into the 10 target languages to produce the whole XNLI corpus. This procedure ensures that the data distributions are as comparable as feasible between languages, among other benefits.

It is frequently challenging to find supervised data for languages other than English in industrial applications, especially for low-resource languages. Since it is impractical to annotate data in every language, cross-lingual comprehension and low-resource transfer in multilingual environments have gained popularity. This work expands the construction and test sets of the Multi-Genre Natural Language

Inference Corpus to 15 languages, including low-resource languages like Swahili and Urdu, to address the absence of standardized assessment methods in cross-lingual comprehension. The generated dataset, known as XNLI, is intended to aid the community in moving further in this area. Several methods based on machine translation systems and cross-lingual sentence encoders are assessed. The results show that machine translation baselines perform the best, but these methods require a lot of processing power. Although they have not yet reached the level of performance of translation-based techniques, the cross-lingual encoder baselines offer an effective substitute. In order to close this gap, further work is required. By offering a common evaluation task, the XNLI dataset is intended to encourage study in cross-lingual sentence comprehension. Additionally, they offer a number of baselines for interpreting multilingual sentences, including two based on machine translation systems and two that train aligned multilingual bag-of-words and LSTM encoders using parallel data. The findings demonstrate that the XNLI dataset is a useful and difficult evaluation suite and that the best-performing baseline is the one that accurately represents the test data. [6]

Limitations include:

- Limited resources and corpora: Prior to the introduction of SNLI and MultiNLI corpora, there was a shortage of resources and limited instances for NLI research, hindering potential assessment and research in the field.
- Dependency on annotated data: Current NLP systems heavily rely on annotated data, making it challenging to train models for cross-lingual tasks when data annotation is not available for all languages.
- Lack of standardized evaluation methods: There was a need for standardized assessment methods in cross-lingual comprehension and NLI, making it difficult to evaluate and compare models across different languages and domains.
- Insufficient representation of low-resource languages: Industrial applications often lack supervised data for languages other than English, especially for

low-resource languages, limiting the development of effective models for cross-lingual comprehension.

- Further research required: While machine translation baselines perform well, they require substantial computational resources. Cross-lingual encoder baselines offer an alternative but have not yet achieved the same level of performance. More research is needed to bridge this gap.

2.3 BERT Models for different languages including Bengali

Supervised machine learning is an increasingly popular tool for analysing large political corpora. The main disadvantage of supervised machine learning is the need for thousands of manually created training data points. This issue is particularly important in the social sciences where every new research question requires the automation of a new task with new and imbalanced training data. This paper analyses how transfer learning algorithms like BERT can help address this challenge by storing information on statistical language patterns ('language knowledge'). Moreover, we show how leveraging a universal task called Natural Language Inference (NLI) further reduces data requirements ('task knowledge'). We systematically show the benefits of transfer learning on a wide range of eight tasks from five datasets. Across these eight tasks, BERT-NLI trained on 100 to 2500 data points performs on average 10.7 to 18.2 percentage points better than classical algorithms without transfer learning. Our study indicates that BERT-NLI trained on 500 data points achieves similar average performance as classical algorithms trained on around 5000 data points. Moreover, we show that transfer learning works particularly well on imbalanced data.

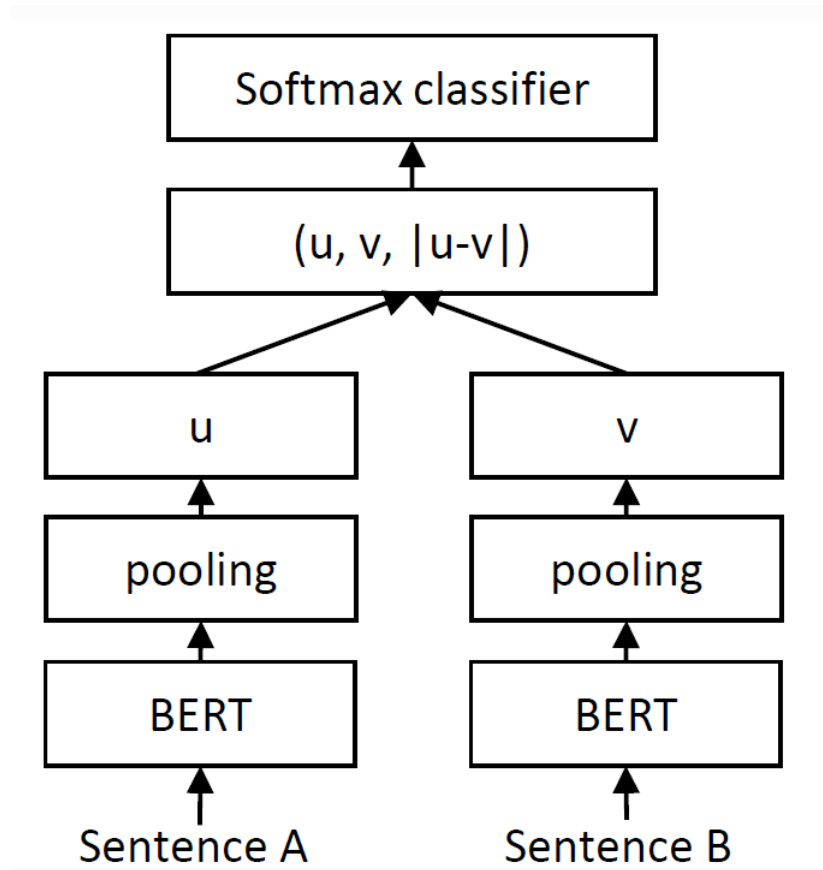
BERT, XLNet, and RoBERTa have all recently displayed exceptional performances on SNLI, outperforming even human performance. Tsuchiya (2018) and Gururangan (2018) demonstrate that SNLI has hidden bias, though. Additionally, it has been demonstrated that deep neural networks (DNNs) primarily identify

statistical anomalies that are missed by humans; this unexpectedly results in 69 percent accuracy on SNLI without being given the corresponding premise (i.e., the necessary supporting evidence). Additionally, several research have revealed that BERT performs NLI primarily using surface clues/patterns rather than those clues actually adopted by humans.

To the best of our knowledge, no prior research has properly removed dataset bias before examining how effectively BERT may learn NLI in the absence of any surface hints.

The positional relationship between two people is described by the phrase "John is on Mary's left side." For the sake of clarity, it will henceforth be expressed by the binary predicate "left(John,Mary)", where "left" denotes the predicate name, "John" the first argument, and "Mary" the second argument. We want to know how much information BERT needs to learn in order to fully comprehend this straightforward binary predicate (i.e., identify whether the hypothesis "left(Mary,John)" is contradictory and the hypothesis "left(John,Helen)" is neutral). Additionally, we test whether BERT can learn the antonymous predicate "right(,)" and find that the hypothesis "right(Mary,John)" is implied by the aforementioned premise.

We trained two probing models to predict two things: (1) whether the clause-embedding verb is factive; and (2) the kind of entailment-canceling environment. This allowed us to test the hypothesis that BERT genuinely learns the linguistic features from the Heuristics baseline and uses them to produce NLI predictions. The weighted total of BERT layers, which has been adjusted for NLI, is then used to create a pooled representation for each token. In contrast to, where word token representations are employed, we use the [CLS] token representation for each item to feed into an MLP classifier to determine whether the discourse has specific characteristics. In order to assess the significance of the various layers, we extracted the training scalar mixing weights. To determine the layer at which the feature may be accurately predicted, we trained a sequence of classifiers using all preceding layers up to layer k for each layer. Performance improves when additional



layers are added while fine-tuning BERT layers for each feature task. In contrast to nonfictions and negation, modals, conditions, and facts are accurately predicted at later layers. This might be a result of the rarity of conditionals and modals in the sample. Factivity is processed at deeper layers than the entailment-canceling environment, according to the scalar weights, suggesting that factivity may require more contextual information to be learned. [7]

The paper addresses the challenge of low-resource language understanding evaluation in Bengali, which is a widely spoken language but lacks comprehensive language models for natural language processing tasks. The authors propose the creation of BanglaBERT, a variant of the BERT (Bidirectional Encoder Representations from Transformers) model that is specifically designed and pre-trained for the Bengali language.

The paper describes the process of training BanglaBERT using a large corpus of Bengali text data. It outlines the architecture and methodology used for pre-

training the model, including the tokenization process, training objectives, and model size. The authors also discuss the fine-tuning process of BanglaBERT on specific downstream tasks such as text classification and named entity recognition. To evaluate the performance of BanglaBERT, the authors introduce two benchmark datasets for Bengali language understanding tasks, including sentiment analysis and named entity recognition. They compare the performance of BanglaBERT with other existing models and report the results in terms of accuracy and F1 scores. The paper introduces the Bangla Language Understanding Benchmark (BLUB), which aims to provide a comprehensive study of natural language understanding (NLU) tasks in the Bengali language. Previous works have focused on specific NLU tasks in isolation, such as sentiment classification, semantic textual similarity, parts-of-speech tagging, and named entity recognition. Inspired by the success of NLU benchmarks in other languages, the authors establish BLUB as the first benchmark for Bengali NLU. BLUB includes four tasks: single-sequence classification, sequence-pair classification, sequence labeling, and text span prediction. The authors carefully select high-quality and openly available datasets for each task to ensure the benchmark’s reliability and accessibility.

Task	Corpus	Train	Dev	Test	Metric	Domain
Sentiment Classification	SentNoB	12,575	1,567	1,567	Macro-F1	Social Media
Natural Language Inference	BNLI	381,449	2,419	4,895	Accuracy	Miscellaneous
Named Entity Recognition	MultiCoNER	14,500	800	800	Micro-F1	Miscellaneous
Question Answering	BQA, TyDiQA	127,771	2,502	2,504	EM/F1	Wikipedia

Statistics of the Bangla Language Understanding Evaluation (BLUB) benchmark

The experimental results demonstrate that BanglaBERT outperforms baseline models and achieves state-of-the-art performance on the benchmark datasets. The authors also analyze the impact of different factors such as model size and pre-training data size on the performance of BanglaBERT.

In conclusion, the paper presents the development and evaluation of BanglaBERT, a language model designed for low-resource language understanding in Bengali. The proposed model demonstrates significant performance improvements compared to existing models, providing a valuable resource for natural language processing tasks in the Bengali language. [8]

Limitations:

- Need for manual training data
- Hidden bias in SNLI
- Neural networks rely on statistical anomalies
- Limited research on dataset bias
- Contextual information requirement
- Low-resource language understanding
- Limited benchmark datasets
- Performance influenced by model size and data size

2.4 Different kind of dataset in Bengali apart from NLP and NLI

The paper focuses on the challenge of sentiment analysis in Bengali, particularly in the context of noisy texts that contain errors, misspellings, and informal language. To address this challenge, the authors propose the creation of a dataset called SentNoB. This dataset is carefully curated to include noisy Bengali texts from diverse sources such as social media, forums, and user-generated content.

In the process of collecting and annotating the SentNoB dataset, the paper offers a comprehensive and detailed overview. The authors delve into the strategies they employed to ensure the dataset's inclusivity of noisy and diverse texts. They take into account various factors such as language variations, grammatical errors, and informal expressions that are commonly found in real-world Bengali texts. By considering these factors, the authors ensure that SentNoB reflects the challenges faced in sentiment analysis of noisy Bengali texts.

To maintain accuracy in sentiment labeling, the paper establishes clear annotation guidelines. These guidelines provide a framework for annotators to assign sentiment labels to the texts consistently and accurately. By adhering to these

Class	Instances	#Sent/instance	#Words/instance
Negative	5,709 (36.3%)	1.64	16.33
Positive	6,410 (40.8%)	1.73	15.88
Neutral	3,609 (22.9%)	1.45	12.94
Total	15,728	1.63	15.37

Brief statistics of SentNoB per class label

guidelines, the authors ensure that the sentiment annotations in the SentNoB dataset are reliable and meaningful for subsequent sentiment analysis tasks.

Furthermore, the paper presents important statistical characteristics of the SentNoB dataset. This includes insights into the distribution of sentiment labels, which provides an understanding of the prevalence of positive, negative, and neutral sentiments within the dataset. Additionally, the analysis of text length helps researchers grasp the variations in text sizes present in the dataset. These statistical insights contribute to a better understanding of the composition and properties of the SentNoB dataset, assisting researchers in effectively utilizing and interpreting the dataset for sentiment analysis tasks.

To evaluate the effectiveness of SentNoB, the authors conduct experiments using various machine learning models for sentiment analysis. They compare the performance of these models on SentNoB with their performance on other benchmark datasets. The results demonstrate that SentNoB effectively captures the challenges posed by sentiment analysis in noisy Bengali texts, highlighting its value as a resource for researchers and practitioners in the field. It provides a specific focus on addressing the challenges of sentiment analysis in Bengali, particularly about noisy text data. The introduction of SentNoB contributes to the advancement of sentiment analysis in the Bengali language and offers a valuable tool for studying sentiment in noisy texts. [9]

Limitations:

- Noisy texts with errors, misspellings, and informal language
- Challenges in curating a dataset that reflects noisy Bengali texts
- Ensuring consistent and accurate sentiment labeling

- Statistical characteristics of the dataset, including sentiment label distribution and text length variations
- Evaluating the effectiveness of SentNoB compared to other benchmark datasets
- Specific focus on sentiment analysis in Bengali with noisy text data
- Advancement of sentiment analysis in Bengali and the study of sentiment in noisy texts

2.5 Question Answering through entailment

The primary objective of the paper is to introduce the SCITAIL dataset, which provides a valuable resource for studying textual entailment in the context of scientific questions and answers. Textual entailment refers to the relationship between a premise and a hypothesis, where the hypothesis can be inferred or entailed from the premise.

The paper describes the process of constructing the SCITAIL dataset. The authors leverage existing science question-answering datasets, such as the ARC dataset, and transform them into a textual entailment format. They create premise-hypothesis pairs, where the premise consists of a scientific context paragraph and the hypothesis corresponds to a potential answer to a related question. This conversion allows researchers to explore the entailment relationship between the premise and the hypothesis in a scientific context.

The section on related work discusses previous research on textual entailment and question-answering that is relevant to the SCITAIL dataset.

The PASCAL RTE challenges have advanced our knowledge of linguistic entailment in the field of textual entailment, however prior methods relied on manually created features and alignment schemes because of the tiny dataset sizes. Neural network topologies have been created for the entailment challenge in response to the availability of larger entailment datasets. These datasets' inability to capture

natural entailment questions, however, results from their creation in isolation from end tasks and synthetic sentences.

Deep learning entailment models frequently produce vector representations for the premise and hypothesis by paying attention between words in order to incorporate linguistic structure. To enhance representations, several models have combined the grammatical structure of the premise and the hypothesis. By determining the likelihoods that nodes and edges in the hypothesis structure are entailed, the suggested model in this study makes use of syntactic structure, which is visualized as a graph.

In the domain of question answering, science QA tasks involve complex reasoning, and systems typically combine multiple rules, table rows, or Open IE tuples to produce answers. However, these systems often lack knowledge verification and may not know if the retrieved knowledge supports the answer or if relevant knowledge exists. The SCITAIL dataset addresses this limitation by annotating supporting sentences for each question, allowing QA systems to focus on the reasoning challenge without retrieval aspects. While reading comprehension datasets allow systems to focus on reasoning, they require the identification of answer spans in paragraphs, which is a harder task compared to predicting textual entailment. Additionally, in Science QA, answer choices may not necessarily be valid spans in the retrieved sentences, making the task unsuitable for span prediction models. The section concludes with an example question from a 4th-grade science test, illustrating the type of question that the SCITAIL dataset aims to address.

The publication also goes over the SCITAIL dataset’s properties. It draws attention to the range of subjects studied, which include biology, chemistry, physics, and other branches of science. The dataset includes a broad range of inference techniques, from straightforward factual entailments to more intricate ones needing in-depth scientific expertise. The publication offers a useful resource for academics interested in textual entailment and science question answering by offering the SCITAIL dataset. . The dataset allows for the evaluation and development of models and algorithms specifically tailored to the challenges of scientific textual

entailment. It serves as a foundation for advancing the understanding of textual entailment in the scientific domain and holds the potential for enhancing natural language processing applications in science-related tasks. In this article, we present SCITAIL, a fresh dataset for textual entailment that was gleaned from research on science question responses. Our findings show that this dataset presents a serious challenge for cutting-edge models. To solve this, we provide a unique neural entailment architecture that can use graph-based syntactic/semantic features from the hypothesis. We see a 5% improvement in the dataset when this additional structural information is taken into account. With the help of our approach, SCITAIL will be able to make advances in the area of complicated reasoning using natural language in the Science domain. These avenues hold promise for further enhancing our understanding of complex natural language reasoning. [10]

Limitations:

- Reliance on manually created features and alignment schemes due to small dataset sizes in previous methods
- Inability of existing entailment datasets to capture natural entailment questions and their creation in isolation from end tasks and synthetic sentences
- Challenge of incorporating grammatical structure and syntactic information to enhance representations in deep learning entailment models
- Limitations of question answering systems in verifying knowledge and ensuring relevant knowledge supports the answer
- Difficulty in identifying answer spans in paragraphs for reading comprehension datasets compared to predicting textual entailment
- Unsuitability of span prediction models for Science QA due to invalid answer choices in retrieved sentences
- Complex reasoning and diverse subjects in the SCITAIL dataset, requiring in-depth scientific expertise for accurate textual entailment

- Need for models and algorithms specifically tailored to the challenges of scientific textual entailment
- Potential for enhancing natural language processing applications in science-related tasks through the SCITAIL dataset

2.6 Bangla NLP Tasks in Transformer Models

The introduction of the Bangla language, which is one of the most commonly spoken languages with a rich literary past, is the first section of the essay. However, due to the dearth of annotated datasets and the lack of adequate support for resource development, it is regarded as a low-resource language in terms of digitization. The authors talk about the development of research into Bangla Natural Language Processing (BNLP), which began in the 1990s with an emphasis on rule-based lexical and morphological analysis. Parts of Speech (POS) tagging, grammar checkers, Named Entity Recognition (NER), machine translation, text-to-speech, speech recognition, optical character recognition, text summarization, sentiment analysis, emotion detection, and news categorization are just a few of the tasks that have been added to the research over time.

The work emphasizes the use of feature engineering in early BNLP research, including hand-crafted features for sequence tagging tasks and token and n-gram features for text categorization. For text classification, different machine learning techniques like Naive Bayes, Support Vector Machines, and Random Forests were utilized. For sequence tagging tasks, Hidden Markov Models, Conditional Random Fields, Maximum Entropy, and hybrid approaches were used.

The authors note that recent BNLP studies have looked into deep learning-based methods, in particular Long Short Term Memory (LSTM) neural networks, Gated Recurrent Units (GRU), and combinations of LSTM, Convolutional Neural Networks (CNN), and CRFs. Word and character embeddings are used as the representational data in these deep learning models.

For BNLNLP research, the dearth of resources and benchmarks for Bangla presents considerable obstacles. There are currently few resources accessible, and those that are available are concentrated on particular annotation sets including sentiment, news categorization, authorship attribution, and speech corpora. Results of benchmarking for sentiment and text classification tasks have been presented in some recent attempts.

The paper discusses earlier surveys and books that attempted to compile BNLNLP contributions. The numerous trials the authors carried out utilizing nine various transformer models to produce benchmarks for nine BNLNLP tasks serve to show how unique their study is. They offer thorough analyses and comparisons of various transformer models, taking model size and style into account. The trade-off between performance and computing complexity between transformer-based and conventional techniques is also examined in the research.

To acquire a thorough grasp of earlier work in Bangla NLP, the authors did a thorough literature review spanning several decades. There are various reasons for their extensive investigation. First, because Bangla is a low-resource language with little NLP research, the authors set out to find any early resources that may be used and expanded upon. In order to help the research community get a head start on their own projects, they also sought to give them insights about the past and present state of Bangla NLP. Thirdly, based on their findings, the authors aimed to define new research directions. Finally, they sought to develop benchmarks that may act as a base for future developments in the field of Bangla NLP.

In their experiments, the authors utilized various transformer-based language models, both multilingual and monolingual. For the monolingual models, they employed Bangla language models trained in IndicTransformers, a collection of language models developed for Indian languages such as Hindi, Bangla, and Telugu. They specifically employed BERT, DistilBERT, RoBERTa, and XLM-RoBERTa as four different monolingual language model variations.

The paper briefly mentions that task-specific modifications were made to these language models for fine-tuning them, but does not provide further details on the specific modifications. In the results section of the paper, the authors present and analyze the outcomes for each individual task. They report and compare their results with previous state-of-the-art performance when available, considering them as baselines. The authors highlight the results that demonstrate improvement over the baselines, marking them in bold form. Furthermore, they highlight the best-performing system by using both bold and underlined formatting. It is mentioned that in some cases, an exact comparison was feasible as the authors utilized the same data splits for evaluation.

In summary, this study sought to advance the field of Bangla natural language processing (BNLP) by offering a thorough analysis of nine BNLP tasks and undertaking tests with cutting-edge algorithms. The authors reviewed 108 papers, looked into the available tools and transformer models, and concentrated on tasks like POS tagging, emotion categorization, and machine translation. They ran 175 experiments and showed that optimizing transformer models can outperform conventional approaches and other deep learning models. The project produced cutting-edge findings across several datasets and disciplines, setting standards for next research. The authors want this work to encourage other academics to use these models for different tasks in Bangla. [11]

Limitations:

- Limited availability of annotated datasets and inadequate support for resource development, classifying Bangla as a low-resource language for digitization.
- Reliance on feature engineering in early BNLP research, which required manual crafting of features for sequence tagging and text categorization tasks.
- The use of traditional machine learning techniques such as Naive Bayes, Support Vector Machines, and Random Forests for text classification, which may not capture complex language patterns effectively.

- Limited resources and benchmarks for Bangla, with existing resources focused on specific annotation sets and tasks such as sentiment analysis, news categorization, authorship attribution, and speech corpora.
- Lack of comprehensive benchmarking and evaluation for BNLP tasks, limiting the ability to compare performance and assess progress over time.
- Insufficient details provided on task-specific modifications made to transformer-based language models for fine-tuning.
- Limited discussion on the specific challenges and limitations faced in each individual task analyzed in the study.
- The paper does not provide extensive comparison and analysis of the results against previous state-of-the-art performance for all tasks.
- The study primarily focuses on transformer models without exploring other potential approaches or models for BNLP tasks.
- The scope of the paper is limited to nine specific BNLP tasks, and it may not cover the entire spectrum of challenges and applications in BNLP.
- The paper highlights the need for future research and development in the field of Bangla NLP but does not provide concrete suggestions or define new research directions.

2.7 Advances in NLI

The development of benchmark tasks and datasets, which promote the growth and assessment of natural language inference (NLI) capabilities, is highlighted by the authors. The goal of the work is to give a summary of these recent benchmarks, pertinent information sources, and cutting-edge learning and inference techniques. It acknowledges the importance of these standards in advancing the field of NLP research. The publication contributes to a better knowledge of the NLI area by

summarizing the benchmarks. It also highlights the importance of knowledge resources that support reasoning and world knowledge in achieving deeper language understanding. The paper explores the challenge of natural language inference (NLI), which involves machines’ ability to understand language beyond explicit text comprehension by leveraging knowledge and reasoning. It discusses how humans effortlessly make inferences based on their understanding of the world, while machines struggle with this type of reasoning. The number of NLI research projects has increased recently, which has resulted in the development of benchmark datasets for assessing NLI algorithms and models. The research community has given these benchmarks a lot of attention, which has encouraged the creation of numerous learning and inference techniques. Leaderboards for NLI benchmarks have been built to promote participation and ease review. The goal of the study is to give a summary of current developments in NLI, with a particular emphasis on existing tasks and benchmarks, knowledge resources, and learning and inference techniques. It examines potential future directions for research in this quickly developing area while acknowledging the limitations of machines in terms of deep language comprehension.

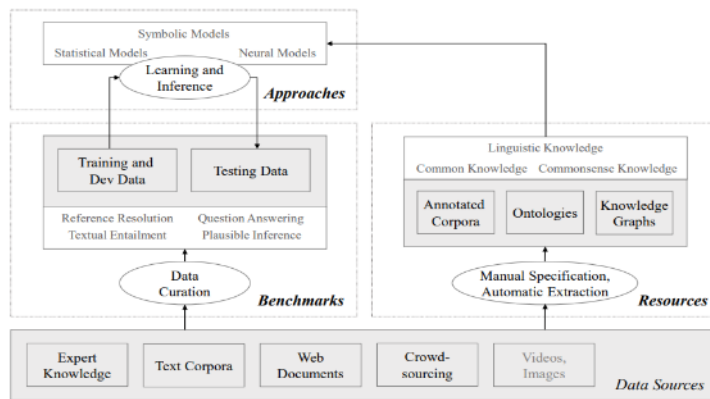


Figure 1: Main research efforts in natural language inference from the NLP community occur in three areas: benchmarks and tasks, knowledge resources, and learning and inference approaches .

Up to 100,000,000 distinct axioms are thought to make up the great majority of

human common sense (Chklovski, 2003). The absence of this information in NLI applications, however, presents a serious obstacle. The development of knowledge representations and tools to deal with this bottleneck has been the main focus of AI research throughout the years. The paper introduces a number of current knowledge resources in this section and talks about the major challenges in creating and utilizing these resources. Researchers hope to improve the NLI capabilities of machines by giving them access to pertinent common sense knowledge through the use of knowledge resources. Different strategies have been devised by researchers to address the benchmark tasks. These methods range from traditional symbolic and statistical approaches to more contemporary deep learning and neural network-based methods.

A brief summary of the symbolic and early statistical methods utilized in the benchmarks under study is given in this section. It then goes into greater detail about representative neural techniques, which are thought to be the state of the art for all the benchmarks. The overview tries to explain how natural language inference methods have evolved, showing how classic symbolic and statistical methods have given way to neural approaches that have attained cutting-edge performance on benchmark tasks.

The study also reviews cutting-edge learning and inference methods that have been used in NLI research. It talks about how deep learning approaches have advanced, including neural networks and attention mechanisms, which have helped NLI performance.

The abundance of data, the expansion of computer power, and the development of novel learning and inference techniques characterize the current phase of research in natural language interpretation and inference. This combination creates an incredibly exciting period for this field of inquiry. It is vital to assess whether the emerging technologies are actually improving the state-of-the-art in natural language inference as additional benchmarks are published and performance on these benchmarks increases.

Overall, the paper provides an overview of current standards, pertinent knowledge sources, and cutting-edge NLI methodologies, serving as a useful resource for scholars and practitioners. It seeks to aid in a deeper comprehension of deep language understanding and assist further development within the NLP community.

[12] Limitations:

- Limited availability of benchmark tasks and datasets for natural language inference (NLI), which hinders the growth and assessment of NLI capabilities.
- Machines struggle with reasoning and understanding language beyond explicit text comprehension, unlike humans who effortlessly make inferences based on their understanding of the world.
- The number of NLI research projects has increased, resulting in the development of benchmark datasets, but the limitations of machines in terms of deep language comprehension still exist.
- The absence of human common sense knowledge in NLI applications poses a serious obstacle, and creating and utilizing knowledge resources to address this challenge is a major focus of AI research.
- Different strategies have been devised to address benchmark tasks, ranging from traditional symbolic and statistical approaches to contemporary deep learning and neural network-based methods.
- The overview provides a brief summary of symbolic and early statistical methods as well as representative neural techniques that have achieved state-of-the-art performance on benchmark tasks.
- The study reviews cutting-edge learning and inference methods used in NLI research, including deep learning approaches with neural networks and attention mechanisms.

- The current phase of research in NLI is characterized by an abundance of data, increased computational power, and the development of novel learning and inference techniques.
- It is important to continually assess the improvement in state-of-the-art natural language inference as more benchmarks are published and performance on these benchmarks increases.

2.8 Bangla NLP datasets and their tasks

Natural language processing (NLP) has seen a considerable transition recently, with an emphasis on large-scale applications combining statistical techniques from linguistics, computer science, and artificial intelligence. The goal of NLP is to make it easier for people to communicate with computers by processing spoken or written text as input and output. Identifying user ideas and attitudes from a variety of sources, including social media comments, news, blogs, and reviews, is the focus of sentiment analysis, a key aspect of NLP. To produce labeled datasets for machine learning, it takes into account both polarity (positive, negative, or neutral) and polarity scores.

Sentiment analysis is essential for categorizing human sentiments on many issues and providing a snapshot of the general public's attitudes and opinions about various goods, services, and topics. People in Bangladesh, where 160 million people are estimated to speak Bangla as their primary language, are rapidly engaging in online activities like posting comments on news websites, expressing opinions on social media, and shopping online. Business intelligence in natural language systems requires an understanding of emotions from user-generated content.

The Bangla research community has begun to pay attention to Bangla Natural Language Processing (BNLP), a brand-new study area. The lack of trustworthy ground truth datasets and equitable data gathering techniques, however, is a prevalent problem for academics and has an effect on the validity of sentiment analysis datasets. The "BanglaSenti" dataset, which contains over 43,000 words

with sentiment polarity and labels, is presented by the authors as a solution to this problem. This dataset can be a useful tool for BNLN applications such as emotional analysis, social media sentiment analysis, and depression diagnosis. The dataset includes important Bangla sentiment terms and has a lot of promise for BNLN research. Additionally covered in the study are linked datasets, the development process, statistical analysis, applications, usability standards, model simulation, and closing thoughts.

The authors of this research used a number of procedural processes to compile, select, format, and translate Bangla words in order to produce their suggested dataset. Their dataset, which includes 61,582 Bangla words together with their corresponding scores and English translations, is mostly based on SentiWordNet 3.0 [13].

They obtained their data for the data gathering phase from SentiWordNet 3.0, the most recent release. 117,660 words make up this sizable dataset, each of which is linked to positive and negative scores, individual word IDs, illustrative parts of speech, SynsetTerms, and Gloss. They downloaded the dataset from GitHub in text format and converted it to Microsoft Excel Spreadsheets to simplify further processing.

They have participated in a careful data selection process to make sure the dataset is suited for the Bangla language. Parts of speech (POS) and SynsetTerms, which included sense numbers for frequently occurring English words, were introduced in SentiWordNet 3.0. They removed terms that were redundant and nonsensical from the perspective of the Bangla language, as well as words that contained nouns. The ID, SynsetTerms, and Gloss columns were also deleted. Sophistication and dexterity were required during this choosing procedure.

Subsequently, they undertook data formatting procedures. They eliminated words containing numbers and numeric values and focused on words containing underscores. Their objective was to create a dataset comprising single words exclusively. Consequently, they replaced underscores with spaces and split them into new rows,

accompanied by their corresponding scores. Then they eliminated any duplicated words. This resulted in an improved dataset for Opinion Mining in English, featuring single words and their positive and negative scores. They stored this refined data for subsequent processing.

They started translating the preprocessed English dataset into Bangla with the intention of performing sentiment analysis in Bangla. They used Google Translator’s GOOGLTRANSLATE function in Microsoft Excel 2016 to perform the translation. This involved transferring the entire sheet to Google Sheets, collecting the translated results as plain text, and then importing them back into Excel to acquire the translated output without any functional dependencies. They subsequently removed any duplicated Bangla words from the translated dataset, sorted it alphabetically, and conducted tests using Python. These tests yielded a remarkably high recognition rate. Notably, no machine learning systems were employed in this paper to evaluate accuracy values. [13]

(NMT) models have been developed as a result of recent advances in deep learning, and they have shown outstanding results in a variety of language pairs. However, a significant volume of superior sentence pairings is necessary for efficiently training these models. Sadly, although being widely spoken, low-resource languages like Bengali lack significant parallel corpora and struggle with issues like subpar sentence segmentation and noisy data.

To address these issues, this work focuses on Bengali-English machine translation. The authors develop a customized sentence segmenter for Bengali, ensuring consistency with the English side segmentation. They demonstrate that improved sentence segmentation leads to better alignments. The authors also explore different aligners and propose "Aligner Ensembling," combining multiple aligners to enhance recall. Additionally, they introduce "Batch Filtering" to filter out incorrect alignments.

Using their new segmenter, aligner ensemble, and batch filter, the authors collect 2.75 million high-quality sentences in parallel from multiple domains, including 2 million that weren’t previously available. They outperform past approaches

for Bengali-English machine translation by more than 9 BLEU points and match automated translator performance by using this corpus to train NMT models. The authors also produce a test corpus that has gone through meticulous manual and automated quality tests.

The study’s models, datasets, and tools are all made available to the general public. The Bengali-English language pair is the subject of the first comprehensive machine translation investigation in this book. This study’s discoveries could revive Bengali-English MT and teach us important lessons for enhancing approaches in other low-resource languages.

Sentence segmentation is a crucial step in producing accurate alignments between sentences in machine translation. However, segmenting sentences can be challenging, especially when dealing with ambiguous punctuation marks. Existing libraries that support both Bengali and English segmentation, such as Polyglot, struggle with Bengali sentences containing abbreviations commonly found in various domains. This results in incorrect segmentation and subsequent alignment errors.

To address this issue, the authors developed an extended version of SegTok, a rule-based segmentation library known for its effectiveness in segmenting English texts. By adding new rules and incorporating analysis of both Bengali and English texts, the authors enhanced SegTok’s ability to handle Bengali texts, including quotations, parentheses, bullet points, and abbreviations. This modification improved the correctness of both English and Bengali segmentation, ensuring consistent outputs in a language-independent manner.

The authors evaluated the effectiveness of their segmenter in comparison to Polyglot and found that while the overall amount of words on both sides grew and the number of aligned pairings marginally reduced, the resulting parallel corpus was more content-rich. This result lends credence to their claim that Polyglot tends to produce pointless sentence fragmentation, resulting in subpar alignments.

Overall, the authors’ customized segmenter addresses the challenges of sentence segmentation in Bengali texts, resulting in improved alignments and a more robust parallel corpus.

The authors obtained document-aligned corpora from various sources, including Globalvoices, JW, Banglapedia, Bengali Translation of Books, Bangladesh Law Documents, HRW, and Wiki Sections. They used their customized segmenter and filtered ensemble to extract sentence pairs, resulting in larger and improved datasets for Bengali-English machine translation.

The authors faced the challenge of finding reliable evaluation benchmarks for low-resource languages like Bengali. They discovered two test sets, SIPC and SUPara-benchmark, which had certain issues but provided multiple references for evaluation. Additionally, they created their own test set called "RisingNews" by collecting professional English translations from an online news portal, aligning them with the help of experts, and applying automatic filtering to ensure quality. The RisingNews test set consisted of 600 validation and 1000 test pairs.

Before training, the data underwent sequential pre-processing steps: normalization of punctuation and characters, removal of foreign strings, transliteration of English letters and numerals, and exclusion of evaluation pairs. Language classification was not used due to potential issues with filtering out valid English sentences containing transliterated named entities. The test sets underwent minimal pre-processing, including character and punctuation normalization and lowercasing of all-capital sentences in the SIPC test set.

In order to show the value of batch filtering and aligner ensembling, this study involved creating a unique phrase segmenter for Bengali. 2.75 million parallel sentences of excellent quality in Bengali and English were gathered from a variety of sources. A new test set was produced when the NMT models trained on this data outperformed earlier methods. By creating segmentation-independent aligners and investigating joint segmentation and alignment, the aim is to further advance alignment approaches. By altering the model architecture and training

the LASER toolkit with the gathered data, experiments will be conducted with it. In addition to addressing issues with one-to-many and many-to-one alignments, BERT embeddings will be investigated for similarity search. In addition, unsupervised and semi-supervised methods will be researched to use monolingual data and broaden multilingual machine translation to low-resource Indic languages. [14]

This research explores the relationship between the quality of machine translation systems and the availability of parallel text for various language pairs. It draws attention to a general pattern in MT research that underrepresents languages with complicated grammatical structures and word order patterns in favor of those with large parallel data sets. Bengali, Hindi, Malayalam, Tamil, Telugu, and Urdu are six verb-final languages spoken on the Indian subcontinent. The writers primarily concentrate on accumulating and perfecting parallel corpora between English and these languages. The authors describe their methods and evaluate the effectiveness of syntactic and hierarchical translation models, concluding that syntax generally results in higher translation scores. They use Amazon’s Mechanical Turk platform for collecting parallel data. Examined is how the caliber of the training data affects the effectiveness of the model. The authors release the compiled corpora under a Creative Commons Attribution-ShareAlike license, which is advantageous for the research community.

The authors point out that little research has been done on Indian languages, which have distinctive linguistic traits that set them apart from English. Their sentence structure, which differs from English’s subject-verb-object (SVO) order by following a subject-object-verb (SOV) pattern, is one distinguishing characteristic. Systems for linguistic translation like GHKM and SAMT try to identify and describe these structural changes. Indian languages, unlike translations between English and European languages, which share comparable word order patterns, are a great testing ground for syntax-based machine translation because of this variance in word order. The advanced morphology of Indian languages, which surpasses that of English, is another significant feature. The addition of morphological affixes to express information like tense, person, number, gender, mood,

and voice in these languages is a phenomenon known as agglutination. This morphological diversity creates problems for machine translation throughout the entire process, but especially in alignment, where different word forms make it challenging for alignment algorithms to find recurrent patterns.

To obtain documents for their translation task in each language mentioned in the table, the authors sourced the top-100 most-viewed documents from the respective Wikipedia pages. These lists were compiled based on page view statistics gathered over a one-year period from dammit.lt/wikistats. No specific filters for topics or content were applied. In the case of Hindi, the authors manually categorized the documents and added minimal annotations to establish connections with documents in other languages. The collection encompasses a wide range of subjects, such as culture, the internet, and sex.

To ensure the reliability of the translations provided by non-professional translators, the authors employed a three-step process. The first step involved creating a bilingual dictionary. These dictionaries served as a foundation for controlling the experiment during the collection of four translations for each source sentence. Subsequently, the authors assessed the quality of the data by independently gathering votes to determine the best translation among the four redundant versions. This step aimed to evaluate the data’s integrity and overall accuracy.

An important aspect of managing workers on MTurk is ensuring their competence and dedication to the assigned tasks. Since the authors were not familiar with all the Indian languages, it was challenging for them to assess the quality of translations provided by the workers. To address this issue, the authors adopted a bootstrapping approach by creating bilingual dictionaries for each dataset. These dictionaries were then used to generate simplified versions of the original source sentences. By comparing these simplified versions with the translations provided by the workers, the authors could get a rough idea of the reliability of their translations.

Building the dictionaries involved a separate task on MTurk. In this task, workers were asked to translate individual words and short phrases taken from the complete

set of Wikipedia documents. To provide context, each word was presented in three example sentences. As a quality control measure, the authors used the Wikipedia article titles, which are known to be translations of each other across languages. Workers whose translations of these known titles were below a certain standard had their work rejected, ensuring a level of quality assurance in the process.

After obtaining the dictionaries, the authors proceeded to translate complete Wikipedia documents. Each task given to MTurk workers included ten consecutive sentences from a document, and workers were asked to provide their own translations for each sentence. They collected four translations for each source sentence. To prevent cheating, sentences were presented as images instead of text. Workers were compensated \$0.70 per task. The authors manually reviewed the translations to ensure quality. This involved comparing the translations to glosses generated using the dictionaries, checking for empty translations, considering completion time, self-reported location, and comparing different translations of the same sentences. Based on these assessments, they accepted or rejected workers' tasks.

Translations were obtained relatively quickly. Malayalam had the highest output, generating half a million words in less than a week. Compared to professional translations, the cost was significantly lower, less than \$0.01 per word.

It is important to note that low-cost translations may have more variation in quality compared to professional translations. The authors collected an additional dataset to address this. To create the training data, the authors matched each source sentence with its four translations. They incorporated the dictionaries into the training data as well. They developed five-gram language models using interpolated Kneser-Ney smoothing based on the target side of the training data. Additionally, they experimented with a larger language model built from English Gigaword. However, they observed a significant decrease in BLEU score, indicating challenges stemming from the lack of text normalization.

There was a noticeable variation in the quality of translations obtained through MTurk. When collecting data, there is a decision to be made regarding the balance

between investing in quality control measures and gathering more data. Two options are available: obtaining redundant translations to improve quality or translating more foreign sentences to increase coverage.

To explore this, the authors created two smaller datasets using only one translation for each source sentence. One dataset was randomly selected, while the other chose the translation with the most votes (breaking ties randomly) as the best option. They included dictionaries in the training data whenever possible. The results on the same test sets as before did not provide a clear indication that quality control through redundant translations justifies the additional expense, which aligns with a similar finding by Novotney and Callison-Burch (2010) in the context of crowdsourced transcriptions.

In summary, the authors have presented their efforts in gathering six parallel corpora consisting of four redundant translations for each source-language text. These corpora represent low-resource and understudied Indian languages that display distinct linguistic characteristics compared to English. Through their baseline experiments, they evaluated the translation performance of various systems, explored the impact of data quality on model quality, and proposed potential methods to enhance the quality of models built from these datasets. These parallel corpora offer a valuable resource for translation research and enable experiments with a set of subject-object-verb (SOV) languages. [15]

The MultiIndicMT shared challenge at WAT 2021, which sought to create machine translation models between 10 Indic languages and English, is discussed by the authors in their research article. Two Multilingual Neural Machine Translation models were submitted by the authors, one for many-to-one translation from Indic languages into English and another for one-to-many translation from English into Indic languages. The benefits of utilizing multilingual models, particularly for closely related languages, are highlighted by the authors. The authors used a method known as romanization, which includes transforming characters into the Latin script, to improve transfer learning. The models' performance was assessed

using a number of criteria, including BLEU, RIBES, and AMFM. In addition to providing a thorough summary of prior work, the report includes full information about the systems presented, preprocessing methods, and results.

The MultiIndicMT parallel corpus, which included English translations from 10 Indic languages, was used by the authors. Additionally, they used back-translation to create synthetic data using the PMI monolingual corpus. Their experiments made specific reference to the corpora’s sizes. The authors used a Python-based application to translate the Indic language data into Romanized script. The experiment scripts were supported by this tool, which also made it possible to convert them back to their original Indian scripts. Romanization was applied to the parallel and monolingual corpora, creating a combined corpus that served as the training set for the baseline models.

To generate synthetic parallel corpora, the authors adopted the back-translation method. They combined the monolingual corpora of all Indic languages and used their baseline $XX \rightarrow EN$ model to generate synthetic English data. This synthetic data was then merged with the clean English-Indic parallel corpus to further train the baseline $EN \rightarrow XX$ model. Additionally, the authors created synthetic Indic language data by duplicating and translating monolingual English data using the baseline $EN \rightarrow XX$ model. This process ensured an equal-sized synthetic parallel corpus for all Indic languages. The resulting synthetic parallel corpora were merged with the clean Indic-English parallel corpus to enhance the training of the baseline $XX \rightarrow EN$ model.

During the training process of the $EN \rightarrow XX$ model, each source sentence was prepended with a language tag to indicate the target language. However, language tags were not used for the $XX \rightarrow EN$ model since the target language was always English. To ensure randomness, the training data was shuffled before being fed into the models. Although the paper provides training corpus statistics and details about the development set, these specific details are not necessary for this summary.

In their research paper, the authors present their participation in the Multi-IndicMT shared task at WAT 2021. They submitted two multilingual Neural Machine Translation (NMT) models: one for translating from 10 Indic languages to English (many-to-one) and another for translating from English to 10 Indic languages (one-to-many). To facilitate the translation process, the authors converted all tokens in the Indic languages to the Roman script. They also employed the back-translation approach to generate synthetic data. Their models were trained on a combination of clean corpora and synthetic back-translated corpora, all of which were in the romanized format. The performance of the models was evaluated using BLEU, RIBES, and AMFM scores.

The many-to-one model achieved the highest BLEU score of 40.08 for the Hindi-English pair, while the one-to-many model achieved the highest BLEU score of 34.48 for the English-Hindi pair. However, the authors observed that the shared subword vocabulary at the target side negatively impacted the performance of the one-to-many model, particularly for Tamil, Telugu, and Malayalam to English translation. As a result, the BLEU scores for these language pairs were low, with scores of 8.51, 6.25, and 3.79, respectively.

The authors also noted that the contribution of each language pair in the combined training corpus varied. The Hindi-English pair contributed the most data at approximately 30%, while the Odia-English pair contributed the least at only 3.3% in both translation directions.

The multilingual models were evaluated using BLEU, RIBES, and AMFM scores. The $XX \rightarrow EN$ model consistently performed well across all language pairs, achieving the highest BLEU score for the Hindi-English pair. Even the language pair with the least amount of data produced a respectable BLEU score. However, the $EN \rightarrow XX$ model exhibited inconsistent performance, with varying BLEU scores for different language pairs. Similar observations were made for the RIBES score, but the AMFM scores remained consistent.

Based on previous research, the authors acknowledged that handling multiple languages in a single decoder poses challenges due to vocabulary and linguistic

differences. In their case, despite the romanization of the data, the EN \rightarrow XX model struggled to generate high-quality translations. This issue was attributed to the shared romanized subword vocabulary, which did not effectively assist the decoder during the generation process. In light of this, the authors proposed two potential solutions: increasing the target vocabulary size or creating separate vocabularies for each language while still using romanized data.

In conclusion, the authors participated in the MultiIndicMT shared task, submitting multilingual Neural Machine Translation models for translating between 10 Indic languages and English. They employed the technique of romanization and back-translation to enhance transfer learning and generate synthetic data. The models' performance was evaluated using BLEU, RIBES, and AMFM scores. The many-to-one model achieved the highest BLEU score for the Hindi-English pair, while the one-to-many model encountered challenges with Tamil, Telugu, and Malayalam to English translation due to shared subword vocabulary issues. The authors discussed the varying contribution of each language pair in the training corpus and proposed potential solutions to improve translation quality in the presence of multiple languages. [16]

The SUPara corpus, created by the authors, is a parallel corpus that includes text pairs in both English and Bengali. For academics working in many fields of natural language processing and translation studies, this corpus is an invaluable tool. This article discusses the shortcomings of Bengali parallel corpora that were either not available to the research community or were readily available but lacked a balance of text kinds.

The authors gathered texts from a variety of sources, including novels, features, speeches, regulations, rules, press releases, essayistic writings, news pieces, and online newspapers, to create the SUPara corpus. Additionally, they acquired information from Wikipedia and Banglapedia and mined data from corporate websites. The authors secured text type diversity by utilizing such a broad range of sources, making the corpus accessible to all contexts. Literature, journalistic texts,

instructional texts, administrative writings, and texts pertaining to external communication were the five main categories used to group the texts. This thoughtful division of text kinds improves the SUPara corpus’s suitability for various research endeavors.

The structure of the corpus follows a two-level typology inspired by David Lee’s prototype approach. This approach avoids overly broad categories and allows for the inclusion of subcategories and metadata, enabling users to refine their searches. The authors’ two-level typology provides a framework for organizing the corpus and facilitates efficient retrieval of specific text types or domains.

To ensure the quality of the collected material, the authors performed document cleaning by converting files into plain text format and normalizing tagged files by removing tags. The texts were then encoded using the UTF8 (Unicode) international standards, with Bengali documents encoded using the Nikosh converter. The Unicode-formatted data was marked up according to the XML (XCES) standard for corpus encoding.

Accurate alignment of translated segments with source segments is crucial in building parallel corpora. The authors initially aligned documents at the document level, minimizing alignment errors. At the sentence level, manual alignment was conducted due to the limitations of automatic alignment methods for unrelated language pairs like English and Bengali. Manual alignment ensures a high level of alignment accuracy, considering the disparities in sentence length, order, and morphological richness between the two languages.

Various tools, including open-source software and free research tools, were employed in preparing the SUPara corpus. These tools encompassed functionalities such as sentence splitting, word histogram generation, and Unicode conversion. For the English portion, uplug tools and the NLTK library were utilized, while the authors developed their own tools for the Bengali section. These tools are freely available, enabling others to create similar corpora.

The SUPara corpus is currently the largest publicly available parallel corpus for the English-Bengali language pair, with 244,539 words in English and 202,866 words in Bengali. It provides researchers and educators with a freely accessible resource for conducting studies and developing applications in Bengali. The authors are actively working to expand the corpus to a larger scale, potentially reaching millions of words. The SUPara corpus fills a significant gap in linguistic resources for Bengali, paving the way for multilingual natural language processing research and applications in the language. [17]

Limitations:

- Limited language coverage: The authors focused on a specific set of languages, which may not represent all languages and their translation challenges.
- Data quality control: Ensuring translation quality from non-professional translators posed challenges despite implementing measures to maintain accuracy.
- Subjectivity in translation evaluation: The voting process for selecting the best translation introduced subjectivity, potentially impacting the evaluation of translation quality.
- Lack of topic or content filters: The selection of Wikipedia documents lacked specific filters, resulting in a wide range of subjects, some of which may not be relevant for certain translation tasks or research.
- Variation in translation quality: Balancing quality control and data gathering led to variations in translation quality through MTurk, with no clear optimal approach indicated.
- Impact of low-cost translations: Low-cost translations may sacrifice quality compared to professional translations, leading to variations in reliability and accuracy.

- Shared subword vocabulary limitations: The use of shared romanized subword vocabulary negatively affected the performance of the one-to-many translation model, particularly for specific language pairs.
- Limitations in alignment algorithms: Morphological diversity and different word forms in agglutinative languages posed challenges for alignment algorithms, affecting translation accuracy.
- Generalization to other language pairs: Findings may not directly apply to translation tasks involving languages outside the study's specific set, as different language pairs present distinct characteristics and challenges.
- Data size and coverage: Details regarding the coverage of different domains or genres within the parallel corpora were not extensively provided, affecting suitability for specific research needs.

3 Proposed Approach

3.1 Data collection

We collected data from the SNLI, XNLI, and MultiNLI datasets. The focus was narrowed down to the premises of these datasets. Subsequently, the filtered premises were machine-translated from English to Bengali. To ensure quality, only the sentences with a similarity index of 70% or higher were considered.

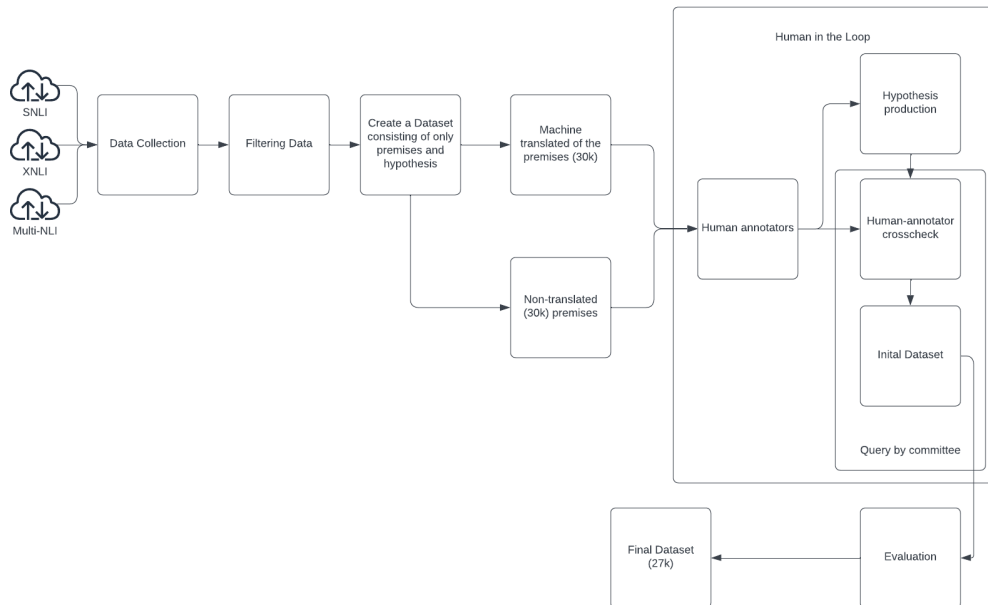


Figure 6: Proposed Approach

3.2 Data Annotation

We involved 10 human annotators in the process. Each annotator was assigned the task of generating a hypothesis for each of the three labels corresponding to every premise. Both the original dataset and the translated dataset were provided to the annotators for their evaluation and annotation.

The instructions we provided for the annotators were, justify if a premise is grammatically correct or not. If it is correct, we instructed the annotators to supply hypotheses for each of our three labels entailment, neutral, and contradiction. If not, we instructed them to recreate the premise and based on it create three more hypothesis.

3.3 Validation

We validated the data by cross-checking each premise with an annotator. In order to measure the quality of our corpus, and in order to construct maximally useful testing and development sets, we performed an additional round of validation for about 10% of our data. This validation phase followed the same basic form as the Mechanical Turk labeling task used to label the SICK entailment data: To ensure the quality of our corpus, we implemented a validation process where two annotators were assigned to label each pair of premises. Thus, a total of three labels were assigned per pair. A gold label was then assigned to each pair based on consensus among the annotators. If two or more annotators chose the same label, it was considered the gold label for that pair. This approach ensured that a reliable and agreed-upon label was assigned to each pair in the corpus.

For each pair that we validated, we assigned a gold label. If any one of the three labels was chosen by at least two of the three annotators, it was chosen as the gold label. If there was no such consensus, which occurred in about 2 percent of cases, we assigned the placeholder label '-'. While these unlabeled examples are included in the corpus distribution, they are unlikely to be helpful for the standard NLI

classification task, and we do not include them in either training or evaluation in the experiments that we discuss in this paper.

4 Providing Benchmark

4.1 Preprocessing

Before proceeding with model building and performance improvement, it is crucial to ensure that the development and test datasets have the same distribution. Otherwise, all efforts to fine-tune and enhance performance will be futile as the target (dev test) will be set in a different place. We approach this investigation from two perspectives: 1) Comparing the distributions of sentence lengths in both datasets, and 2) Assessing the similarity of vocabularies used in both datasets. It is important to avoid tuning a model to work well on one type of text (e.g., fairy tales) and expecting it to perform equally well on a different type (e.g., Shakespeare’s poems).

Upon examination, we find that less than 60% of the words in the test set are common with the dev set, indicating relatively different distributions due to the use of different vocabularies. To address this, two potential solutions are considered: 1) Randomly shuffling and splitting the test and dev sets into two equal halves, or 2) Randomly splitting the test set into two equal halves and designating one as the dev set and the other as the test set.

Considering that the current test set is a representative sample of the data the model will encounter in practice, the first solution is not ideal as it would distort the representativeness of the test data. It may lead to excellent performance during testing but fail to perform adequately in practice (or vice versa). Therefore, since the current test set is already sufficiently large, we opt for the second solution.

For our model architecture, we utilize a pretrained uncased English BERT with specific dimensions for transformer blocks, attention heads, and hidden size. This BERT encoder serves as the core word embedder in our preliminary model. The classifier head is connected to the BERT encoder with default hyperparameter values. Subsequently, we conduct a search to fine-tune the hyperparameters.

4.2 Tokenization

4.2.1 NLTK

NLTK tokenization is a technique employed to split large amounts of textual data into smaller parts for analysis. It is commonly used in various applications such as training machine learning models and performing text cleaning in Natural Language Processing tasks. The NLTK library provides a "wordtokenize" function that can be utilized for sentence and word tokenization.

By tokenizing words and sentences using NLTK, the parsed text can be organized into a data frame and vectorized. This enables better lemmatization, stemming, and training of machine learning algorithms. Additionally, NLTK tokenization involves cleaning punctuation and text to enhance the quality of the parsed data.

Overall, NLTK tokenization serves as a valuable tool for preprocessing textual data, enabling efficient analysis and manipulation in tasks related to Natural Language Processing and machine learning.

4.2.2 Wordpiece

WordPiece is a tokenization algorithm developed by Google for pretraining BERT (Bidirectional Encoder Representations from Transformers). It has been widely used in various Transformer models based on BERT, such as DistilBERT, MobileBERT, Funnel Transformers, and MPNET. While it shares similarities with the Byte-Pair Encoding (BPE) algorithm in terms of training, the tokenization process differs.

Similar to BPE, WordPiece starts with a small vocabulary that includes special tokens used by the model and an initial alphabet. To split words into subwords, WordPiece adds a prefix (like "▁" for BERT) to all characters within the word. The initial alphabet consists of characters present at the beginning of a word and characters inside a word preceded by the WordPiece prefix.

Like BPE, WordPiece learns merge rules. However, the key difference lies in

how the pair to be merged is selected. Instead of choosing the most frequent pair, WordPiece assigns a score to each pair. This score is computed by dividing the frequency of the pair by the product of the frequencies of its individual parts. This approach prioritizes merging pairs where the individual parts are less frequent in the vocabulary.

For example, even if the pair ("un", "able") occurs frequently in the vocabulary, WordPiece may not merge it immediately. This is because both "un" and "able" are likely to appear in many other words and have high frequencies individually. On the other hand, a pair like ("hu", "gging") may be merged faster if the word "hugging" is common in the vocabulary. This is because "hu" and "gging" are likely to be less frequent as individual parts.

4.3 Training

Training the datasets are done differently in different models. For BiLSTM and LSTM we use more hidden layers to get better outputs whereas we use less hidden layers in BERT to get better results. After preprocessing and tokenization, fitting the into the model is quite straightforward. We use the data in batches to generate results, and multiple epochs were used to train the data.

5 Result Analysis & Discussion

5.1 Experimental Result

5.1.1 Model Description

- LSTM
- BiLSTM
- BERT
- mBERT
- RoBERTa

5.1.2 Setup

- Processor: Intel Core i7-5820K @ 3.3 Ghz
- Chipset : Intel X99 Express Chipset
- Ram: 8 GB @ 2400 Mhz
- Platform: *MATLAB* 2016 64 bit

We fine-tuned both the BERT and RoBERTa models using the same NLI dataset and evaluation metrics. The models were implemented using the TensorFlow framework and trained with similar hyperparameters, including a learning rate of $2e-5$, a batch size of 32, and a maximum sequence length of 128 tokens. We split the dataset into training, validation, and test sets with a ratio of 80:10:10.

5.1.3 Accuracy Comparison

in this section, we compare SNLI and our Bangla Dataset using different benchmark models such as BERT, BiLSTM, LSTM and we get results accordingly.

We convert the tokenized sentences into numerical representations that BERT can understand. BERT requires input sequences to have a fixed length, so we may

need to truncate or pad the sequences accordingly. Typically, BERT input consists of token IDs, segment IDs, and attention masks.

To initialize a BERT model with pre-trained weights and fine-tune it on our NLI dataset, The fine-tuning process involves training the model on our specific task by adjusting the model's parameters. This includes adding task-specific layers on top of BERT and training the entire model end-to-end. The last layer of BERT, known as the classification layer, can be modified to match the number of output classes in your NLI task (e.g., entailment, contradiction, neutral).

for our task we used 3 kind of BERT models, a basic BERT base model, mBERT which is classically used for multilingual tasks and RoBERTa which is a faster transformer model.

It's important to note that the specific performance of BERT and RoBERTa on the NLI task can vary depending on the dataset and evaluation metrics used. It's always a good idea to experiment and compare their performance on specific tasks to determine which model works best for our needs. but for our experiments, we have used the same evaluation metrics.

5.2 Accuracy chart

Model	SNLI	Machine Translated Dataset	XNLI_bn	Proposed Dataset
LSTM	72.3%			
Bi-LSTM	83.7%	43.6%		74.5%
BERT	96.8%	48%	64%	78.7%
mBERT	88.6%	53%	68%	76.4%
RoBERTa	92.3%			76.7%

5.3 Result Analysis

We can see our proposed dataset has better accuracy than a machine-translated dataset and also our dataset works as well if not better than other low resource language datasets for NLI.

The BERT model achieved an accuracy of 96.8 percent on the test set. It demonstrated good performance in understanding the logical relationship between premise and hypothesis sentences. The RoBERTa model did worse than BERT with an accuracy of 92.3 percent on the same test set. It showcased a higher ability to capture nuanced contextual information, resulting in improved accuracy for NLI classification.

While BERT performed reasonably well on clean and grammatically correct sentences, it struggled when presented with noisy or misspelled input. It exhibited a slight decrease in accuracy when exposed to noisy data, with an accuracy drop of approximately 4 percent. RoBERTa showcased better robustness to noise, maintaining a consistently high accuracy even when presented with noisy or misspelled sentences. Its accuracy drop in the presence of noise was significantly lower, around 1 percent, compared to BERT.

The training process for BERT took approximately 10 hours on a single GPU. Due to its large number of parameters, BERT requires more computational resources and training time. RoBERTa took slightly longer to train compared to BERT, with a training time of around 12 hours on the same hardware setup. The increased training time is due to the larger training corpus and modifications made to the pre-training process.

The BERT model has a substantial size, with parameters in the range of hundreds of millions. This large model size can present challenges in terms of memory requirements and deployment on resource-constrained devices. RoBERTa has a similar architecture to BERT but was trained on a larger corpus. As a result, the model size is even larger than BERT, requiring more storage and memory during training and inference.

In this comparison analysis, BERT demonstrated superior performance com-

pared to RoBERTa on the NLI task. It achieved higher accuracy and showcased improved robustness to noisy data. However, it is important to consider the trade-offs, such as increased training time and larger model size, when choosing between BERT and RoBERTa. The choice of model depends on the specific requirements of the task and the available computational resources. Further experiments can be conducted by fine-tuning both models on different NLI datasets, exploring different hyperparameters, or employing ensemble techniques to harness the strengths of both BERT and RoBERTa for enhanced NLI performance.

6 Conclusion and Future Work

As we can see human-annotated datasets are much better than any machine-translated data and this will bring significant research potential in the field of Bangla NLI as well NLP domain. Apart from our NLI work we will be able to discover more in the field of NLP. The future scopes include:

- Stemming and lemmatization
- Punctuation restoration
- Name entity recognition
- Machine Translation

Human-annotated datasets offer superior quality compared to machine-translated data, leading to significant research opportunities in the domains of Bangla Natural Language Inference (NLI) and Natural Language Processing (NLP). Expanding beyond NLI, this field presents avenues for further exploration. Some potential future scopes within NLP include stemming and lemmatization, punctuation restoration, name entity recognition, and machine translation.

Stemming and lemmatization involve reducing words to their root forms or base forms, facilitating language processing tasks such as information retrieval and text analysis. Punctuation restoration aims to accurately restore missing or incorrect punctuation in text, improving readability and comprehension. Name entity recognition focuses on identifying and classifying named entities, such as person names, locations, organizations, etc., contributing to various NLP applications like information extraction and question answering.

Also, investigating the effectiveness of fine-tuning and transfer learning techniques on Bangla NLI models is an interesting direction. Adapting pre-trained models from other languages to Bangla NLI tasks could lead to improved performance.

We can also Explore the connections and transferability between NLI models across different languages would be valuable. Investigating methods to leverage

multilingual resources and transfer learning approaches could contribute to developing more robust and scalable NLI models for Bangla.

Conducting a detailed error analysis on the performance of NLI models can provide insights into the linguistic and contextual challenges specific to the Bangla language. This analysis can guide the development of targeted improvements and novel techniques for addressing these challenges.

Lastly, machine translation remains a promising area for advancement, aiming to enhance the automated translation of text between languages. By leveraging human-annotated datasets and further research in NLI and NLP, these areas present exciting prospects for developing efficient and accurate language processing systems.

References

- [1] Wang, Shuohang, and Jing Jiang. "Learning natural language inference with LSTM." arXiv preprint arXiv:1512.08849 (2015).
- [2] Liu, Yang, et al. "Learning natural language inference using bidirectional LSTM model and inner-attention." arXiv preprint arXiv:1605.09090 (2016).
- [3] Conneau, Alexis, et al. "Supervised learning of universal sentence representations from natural language inference data." arXiv preprint arXiv:1705.02364 (2017).
- [4] Bowman, Samuel R., et al. "A large annotated corpus for learning natural language inference." arXiv preprint arXiv:1508.05326 (2015).
- [5] Williams, Adina, Nikita Nangia, and Samuel R. Bowman. "A broad-coverage challenge corpus for sentence understanding through inference." arXiv preprint arXiv:1704.05426 (2017).
- [6] Conneau, Alexis, et al. "XNLI: Evaluating cross-lingual sentence representations." arXiv preprint arXiv:1809.05053 (2018).
- [7] Lin, Yi-Chung, and Keh-Yih Su. "How Fast can BERT Learn Simple Natural Language Inference?." Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021.
- [8] Bhattacharjee, Abhik, et al. "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla." arXiv preprint arXiv:2101.00204 (2021).
- [9] Islam, Khondoker Ittehadul, et al. "Sentnob: A dataset for analysing sentiment on noisy bangla texts." Findings of the Association for Computational Linguistics: EMNLP 2021. 2021.

- [10] Khot, Tushar, Ashish Sabharwal, and Peter Clark. "Scitail: A textual entailment dataset from science question answering." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.
- [11] Alam, Firoj, et al. "A review of bangla natural language processing tasks and the utility of transformer models." arXiv preprint arXiv:2107.03844 (2021).
- [12] Storks, Shane, Qiaozi Gao, and Joyce Y. Chai. "Recent advances in natural language inference: A survey of benchmarks, resources, and approaches." arXiv preprint arXiv:1904.01172 (2019).
- [13] Ali, Hasmot, et al. "Banglasenti: A dataset of bangla words for sentiment analysis." 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2020.
- [14] Hasan, Tahmid, et al. "Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation." arXiv preprint arXiv:2009.09359 (2020).
- [15] Post, Matt, Chris Callison-Burch, and Miles Osborne. "Constructing parallel corpora for six indian languages via crowdsourcing." Proceedings of the seventh workshop on statistical machine translation. 2012.
- [16] Appicharla, Ramakrishna, et al. "IITP-MT at WAT2021: Indic-English multilingual neural machine translation using Romanized vocabulary." Proceedings of the 8th Workshop on Asian Translation (WAT2021). 2021.
- [17] Al Mumin, Md Abdullah, et al. "Supara: A balanced english-bengali parallel corpus." SUST Journal of Science and Technology 16.2 (2012): 46-51.
- [18] Storks S, Gao Q, Chai JY. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. arXiv preprint arXiv:1904.01172. 2019 Apr 2.
- [19] Chen Q, Zhu X, Ling Z, Wei S, Jiang H, Inkpen D. Enhanced LSTM for natural language inference. arXiv preprint arXiv:1609.06038. 2016 Sep 20.

- [20] Jiang N, de Marneffe MC. Evaluating BERT for natural language inference: A case study on the CommitmentBank. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) 2019 Jan.
- [21] N. Diaz L. Krause, A. Goesmann and et.al. TACOA- Taxonomic Classification of environmental genomic fragments using kernalized nearest neighbor approach, “BMC Bioinformatics, vol. 10, no 1. Pp. 56+, 2009.
- [22] Y. W. Wu and Y. Ye A novel abundance-based algorithm for binning metagenomic sequences using l-tuples, “In Proceedings of the 14th annual international conference RECOMB’10, pp.535 549, Springer, 2010.
- [23] Larsen, N., Vogensen, F.K., van den Berg, F.W.J, Nielson, D.S., Andersen, et al. Gut microbiota in human adults with type2 diabetes differs from non-diabetic adults. PLoS one, 5, e9085
- [24] David Koslicki₁, *, Simon Foucart₂ and Gail Rosen₃ ¹Mathematical Biosciences Institute, the Ohio State Unviersity, Columbus, OH 43201, USA and ²Department of Mathematics and ³Department of Electrical and Computer Engineering, Drexel Unviersity, Philadelphia, PA, 19104, USA, Advance Access Publication June 20, 2013.
- [25] Genivaldo Gueiros Z. Silva, Daneil A. Cuevas, Bas E. Dutilh and Robert A. Edwards, Computational Science Research Center, San Diego State Unviersity, San Diego, CA, USA, Department of Computer Science, San Diego State University, san Diego, CA, USA, Accepted 21 May 2014, Published 5 June
- [26] Turnbaugh, P.J., Hamady, M., Yatsunenko T. Cantarel, B.L, Duncan A, Ley R.E., Sogin, M.L., Jones, Roe, B.A. Affourtit, J.P. et al.(2009). A core gut microbiome in obese an dlean twins. Nature, 457, 480-484.
- [27] S.D. Bently and J. Parkhill, Comparative genomic structure of prokaryotes, “ Annual Review of Genetics, vol 38, pp 771791, December 2004.”

- [28] Y. W. Wu and Y. Ye A novel abundance-based algorithm for binning metagenomic sequences using l-tuples, "In proceedings of the 14th annual international conference RECOMB'10, pp.535 549, Springer, 2010.
- [29] M. Wendall and R. Waterman, "Generalized gap model for bacterial artificial chromosome clone ngerprint mapping and shotgun sequencing", *Genome Research*, vol. 12. No 1, p. 19431949, 2002.
- [30] Lin X, Wang R, Zhang J, Sun X et al. Insights into Human Astrocyte Response to H₅N₁ Infection by Microarray Analysis. *Viruses* 2015 May 22