Shahjalal University of Science and Technology
Department of Computer Science and Engineering

SUNER: A Named Entity Recognition Corpus for Bangla Text

MD. Tariful Islam

Reg. No.: 2018331042

4^{th} year, 2^{nd} Semester

Raisa Fairooz

Reg. No.: 2018331050

4^{th} year, 2^{nd} Semester

Department of Computer Science and Engineering

Supervisor

Md. Eamin Rahman

Assistant Professor

Department of Computer Science and Engineering

29th May, 2024

SUNER: A Named Entity Recognition Corpus for Bangla Text

A Thesis submitted to the Department of Computer Science and Engineering, Shahjalal University of Science and Technology, in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering.

By

MD. Tariful Islam
Reg. No.: 2018331042
4^{th} year, 2^{nd} Semester

Raisa Fairooz
Reg. No.: 2018331050
4^{th} year, 2^{nd} Semester

Department of Computer Science and Engineering

Supervisor

Md. Eamin Rahman
Assistant Professor
Department of Computer Science and Engineering

29th May, 2024

Recommendation Letter from Thesis Supervisor

The thesis entitled *SUNER: A Named Entity Recognition Corpus for Bangla Text* submitted by the students

1. MD. Tariful Islam
2. Raisa Fairooz

is under my supervision. I, hereby, agree that the thesis can be submitted for examination.

Signature of the Supervisor:
Name of the Supervisor: Md. Eamin Rahman
Date: 29th May, 2024

Certificate of Acceptance of the Thesis

The thesis/project entitled *SUNER: A Named Entity Recognition Corpus for Bangla Text* submitted by the students

1. MD. Tariful Islam
2. Raisa Fairooz

on 29th May, 2024, hereby, accepted as the partial fulfillment of the requirements for the award of their Bachelor Degrees.

Head of the Dept.	Md Masum	Supervisor
Md Masum	Professor & Head	Md. Eamin Rahman
Professor & Head	Department of Computer	Assistant Professor
Department of Computer	Science and Engineering	Department of Computer
Science and Engineering		Science and Engineering

Abstract

Named-Entity Recognition(NER) is an important task in the field of Natural Language Processing (NLP). It plays a vital role for further work in this field ranging from information extraction to text summarization and question answering. NER in the field of Bengali language is ever more important because of the recent growth in digital content in Bengali. This report presents a comprehensive study that underscores the critical importance of NER in Bengali. We have developed an annotated dataset of 52,200 sentences tagged using 10 different named-entity tags, subsequently refined through validation to approximately 31,000 sentences. Employing BanglaBERT as a baseline model, we achieved a macro average F1 score of 0.6707 on the validated data, indicating scope for further improvement. Additionally, through collaboration with the B-NER dataset, our combined dataset yielded an F1 score of 78%, emphasizing the significance of data fusion in advancing Bengali NER research.

Keywords: NER, BanglaBERT

Acknowledgements

We would like to begin our gratitude to Allah for granting us the strength and ability to complete our first part of the thesis. Without His divine assistance, we would not have been able to do so. Then we would like to express our sincere gratitude to our honorable supervisor, **Md. Eamin Rahman** for his guidance, support, and encouragement throughout this project. His expertise and knowledge were invaluable to us, and we are truly grateful for his help. We would also like to thank our faculty members for their teaching and mentoring. Their insights and advice helped us to develop our research skills and to think critically about our work. We are grateful to our classmates for their support and collaboration. Finally, we would like to thank the Computer Science department for their support of this project. We are grateful for the resources that were made available to us.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Bangla, a language spoken by millions and sixth most widely spoken language by total number of speakers [2] has experienced a remarkable surge in digital content creation. This surge not only reflects the rapid growth of Bangla's digital landscape but also underscores the expanding reach and significance of this language in the modern world. As technology continues to touch the lives of diverse communities, the role of Natural Language Processing (NLP) in Bangla becomes ever more vital. NLP in Bangla empowers inclusive and accessible technology, ensuring that advancements are not limited to a select few, but extend to a broader spectrum of users.

At the heart of NLP lies Named Entity Recognition (NER), a task of paramount importance. NER involves identifying and classifying named entities like names of individuals, locations, organizations, time, drug [3], disease [4], gene, facility, brand and so on. This process, while crucial for information extraction, extends its influence to different downstream tasks like topic modelling [5], domain specific chatbot building [6], conference [7], and anaphora resolution. It underpins information retrieval, aids in extracting pertinent data, enables context comprehension, generates personalized recommendations, supports Information Extraction via relation identification, contributes to Text Summarization [8–10], Question Answering [11, 12], Machine Translation [13] and even assists in topic detection. NER is a cornerstone in the development of NLP, unraveling new possibilities in language understanding and utilization.

There are several variants in the way Bengali words are formed, and the language has a vast

Problem	Sentence
Multiple Meaning	1. গোলাপ ফুল(flower) সুন্দর। 2. ফুল(the word "full" in transliterated text)সং ডাউনলোড করুন।
	3. আহাদ আলী সরকার(a person's name)। 4. আওয়ামী লীগ সরকার(government) সফলভাবে দায়িত্ব টি পালন করেছে।
Multiple Expression	5. উনিশ্য একাত্তর সালের ২৬ শে মার্চ এই বর্বর হামলা সংঘটিত হয়। 6. ১৯৭১ সালের ২৬ শে মার্চ এই বর্বর হামলা সংঘটিত হয়।

Table 1.1: Sample of challenging sequence

vocabulary [16]. Furthermore, Bengali phrases' varied word choices complicate their syntactic and semantic structures. Due to this kind of variation, Bengali NER tasks are very challenging. Table 1.1 provides some sequence examples and provides insight into the Bengali language and how it contributes to the difficulty of NER assignments. One of the biggest benefits of English NER is the capitalization of nouns. Bengali does not use the concept of capitalization like English does. Additionally, depending on the context of the word sequence in Bengali, a single word may have several meanings. The first two pairs of sentences in Table 1.1 represent the problem of multiple meanings. In sentence 1, the token ফুল refers to a flower's name, whereas in sentence 2 the same token is used to refer to an English word in Bangla form to download the full song. In sentence 3, the token সরকার is used as the person end name while in sentence 4 the same token indicates the end of the organization name. The use of idiom makes Bengali text incomprehensible for machines. In sentence 5, the tokens উনিশশো একাত্তর indicates a time expression whereas in sentence 6, the single token ১৯৭১ also indicates a time.

In the pursuit of enhancing NLP's capabilities for Bangla, our work embarks on a journey with two pivotal objectives. Our first goal is to construct a robust Named Entity Recognizer, one capable of comprehensively identifying and classifying entities within Bangla text. To support this endeavor, we are curating a rich dataset that encapsulates the linguistic nuances and intricacies of the Bangla language. The dataset consists of texts from varying domains making it a balanced and rich one. However, our ambition does not end with data collection. Our second goal involves refining this dataset, strengthening its annotations, and exploring model architectures that can leverage its potential to the fullest. Through meticulous efforts, we aspire to present a tool that not only

uplifts NLP in Bangla but also inspires further studies and innovations in this field.

Chapter 2

Literature Review

The inception of Bangla Named Entity Recognition (NER) saw limited progress prior to 2008. Asif Ekbal and his collaborators initiated significant advancements during the period from 2007 to 2009. Their pioneering research marked the commencement of NER exploration within the Bangla language. Utilizing newspaper data, their efforts led to the development of machine learning models, including Hidden Markov Models (HMM) [14], Support Vector Machines (SVM)[15], Conditional Random Fields (CRF) [16], and Maximum Entropy (ME) [17]. Notably, the focus of these initial models was on entities such as Person, Location, Organization, and Object. It is noteworthy that Asif Ekbal's approach employed a word or token-based strategies rather than sentence-based ones. Despite their contributions, limitations in terms of dataset accessibility and source diversity were apparent. Moreover, it's important to emphasize that the early stages of Bangla NER primarily embraced machine learning models, as deep learning approaches had not yet entered the scene.

Chowdhury et al. (2018) [18] played a pivotal role in advancing Bangla Named Entity Recognition (NER). Their study marked a new phase by focusing on a dataset encompassing 2,137 sentences. In an effort to broaden entity recognition, they extended the scope to include 7 categories. Their innovative approach hinged on the integration of Part-of-Speech (POS) tagging to refine entity classification. However, challenges emerged: their strategy inadvertently tagged pronouns like 'who' and 'doctor' as 'person' entities, and 'organization' entities suffered from inconsistent tagging. Despite these hurdles, Chowdhury et al.'s work laid the groundwork for refining entity

recognition approaches in the future.

In 2019, Rifat et al. [19] contributed significantly to the field of Bangla Named Entity Recognition (NER). Their study introduced a sizable dataset encompassing 96,697 tokens, including punctuation, aimed at capturing a broader linguistic range. However, the study faced challenges due to mixed tagging methods, involving both manual and rule-based automatic tagging. This hybrid approach introduced potential errors, impacting the accuracy of entity recognition. Additionally, the study highlighted issues of overgeneralization, where generic terms like 'country' and 'time' were erroneously tagged as named entities. Rifat et al.'s work emphasized the complexities of expanding Bangla NER and underscored the importance of precise annotation methodologies.

Karim et al. [20] contributed significantly to Bangla Named Entity Recognition (NER), presenting a substantial dataset of 71,284 sentences. This dataset drew from diverse sources, including Wikipedia and online newspapers, ensuring a robust representation of linguistic contexts. Their approach extended recognized entity categories to four basic entities and introduced the "Adhikery" annotation management tool. Methodologically, they employed Deep Convolutional Networks (DCN) and Bidirectional Long Short-Term Memory (BiLSTM) models, showcasing the potential of deep learning. However, the study grappled with limitations due to automatic tagging, leading to inconsistencies and inaccuracies in labeling. Karim et al.'s work emphasized the importance of addressing tagging challenges for accurate entity recognition in Bangla.

The landscape of Bangla Named Entity Recognition (NER) has seen remarkable advancements, particularly in recent years. Ashrafi et al. (2020) [21] contributed significantly by harnessing the dataset introduced by Karim et al., employing a BERT-based deep neural network with a CRF layer. Their sophisticated approach underscored the synergy between deep learning and sequential modeling techniques.

In 2023, the B-NER project led by Haque et al. [22] took a comprehensive approach to dataset construction and validation. Their dataset, derived from newspapers, blogs, forums, and Wikipedia pages, was manually annotated and verified by linguistic experts. Covering a diverse range of 8

entities using the IOB tagging scheme, the B-NER project emphasized the importance of both meticulous data collection and rigorous validation.

These recent works exemplify the convergence of advanced methodologies and high-quality datasets, with Ashrafi et al. showcasing the potential of BERT-based models and B-NER emphasizing the necessity of rigorous annotation. As the field continues to evolve, the integration of state-of-the-art techniques and robust datasets will be pivotal in achieving precise entity recognition in the Bangla language.

Looking forward, the challenges of dataset size, imbalance, model transparency, and comprehensive entity coverage remain important avenues for further exploration. Future research endeavors will likely focus on refining these aspects to continue pushing the boundaries of Bangla NER.

The table 2.1 shows a summary of the aforementioned works, their approaches and their respective best performance metrics.

Study	**Model**	**Performance Metrics**
Ekbal et al	HMM CRF SVM ME	F-92.28% R-93.98% P-90.63%
Chowdhury et al.	LSTM CRF	R-0.67 P-0.78 F-0.72 (Best Model)
Rifat et al.	BGRU + CNN	R-72.27% P-73.32% F-72.66%
Karim et al.	DCN-BiLSTM	R-58.62% P-68.95% F-63.37%
Ashrafi et al.	BERT + BiLST+CRF+CW	Micro F-90.64% Macro F-65.96% MUC F- 72.04%
Haque et al.	BanglaBERT mBERT	Macro-F1 - 0.74 R-0.76

Table 2.1: Comparison between different approaches and their respective performance metrics in related studies

Chapter 3

Dataset

3.1 Data Source

The entire dataset we collected initially consisted of 78,595 Bengali sentences. After checking for duplicates, a portion of the data has been discarded, and our final set of sentences includes 73,581 sentences. This dataset comprises sentences from some well-known benchmarking datasets. Table 3.1 provides a thorough summary of the sources. The majority of the sentences were gathered from the 'SUPara: A Balanced English-Bengali Parallel Corpus' [23] - 6,5847 sentences from the training set, and another thousand sentences combined from the development and test sets. The other sources used here are the Tatoeba dataset and the Rising News dataset.[24]

Data Source	**Sentence Count**
SUPara Training set	65,847
SUPara Development set	500
SUPara Test set	500
Tatoeba Development set	2,637
Tatoeba Test set	2,500
Rising News Development set	597
Rising News Test set	1,000

Table 3.1: Data sources for our final dataset and the number of sentences collected from each source

The dataset we collected are diverse in nature due to the fact that these sentences have been gathered from various domains. SUPara, which is the major source of our data is a balanced dataset

with texts from five domains - literature, journalistic texts, instructive texts, administrative texts and texts treating external communication. In the field of NLP and particularly for NER, a balanced dataset is considered an important asset. It enables our dataset to have different named entities coming from a great variety of domains. The diversity of the datasets, stemming from various domains, enhances the versatility and generalizability of our named entity recognition (NER) models. This diversity equips our models to excel in recognizing named entities across a wide range of subject matters, ensuring their robustness and effectiveness in real-world applications.

The following table in the figure 3.1 shows the two-level typography of SUPara.

SUPERORDINATE	BASIC LEVEL
1.Literature	1.Novels
	2.Essayistic texts
	3.(Auto)biographies
	4.Expository non-fictional literature
2.Journalistic texts	1.News reporting articles
	2.Comment articles (background articles, columns, editorials)
3.Instructive texts	1.Manuals
	2.Internal Legal documents
	3.Procedure descriptions
4.Administrative texts	1.Legislation
	2.Proceedings of debates
	3.Minutes of meetings
	4.Yearly reports
	5.Correspondence
	6.Official speeches
5.External Communication	1. (Self-)presentations of organizations, projects, events
	2.Informative documents of a general nature
	3.Promotion and advertising material
	4.Press releases and newsletters
	5.Scientific texts

Figure 3.1: SUPara's two-level typography, adapted from Mumin et Al.(2012)

Our dataset selection process was guided by another critical factor that underscore the quality and applicability of our research. That is, the datasets we curated are characterized by their minimal noise content, contributing to the precision and reliability of the named entity annotations. This quality control measure ensures that the data is highly trustworthy, setting a strong foundation for accurate NER model development.

3.2 Data Annotation

3.2.1 System Development

As a part of the dataset creation process, we have designed and developed a system for sentence-level human annotation, namely the SUST CSE NER Annotation Tool. The tool serves as a robust platform for annotators with distinct profiles to contribute collaboratively on the manual annotation process. A brief description of different aspects of the tool is given below:

1. **Tools and Frameworks**

   Table 3.2 shows the tools and frameworks used for the development of the SUST CSE NER Annotation Tool.

Technology	**Tools/Frameworks**
Frontend	Next.js
Backend	Next.js API Routes
Design Library	Mantine.dev
Database	MongoDB
Deployment	Vercel

Table 3.2: System Development Environment for SUST CSE NER Annotation tool

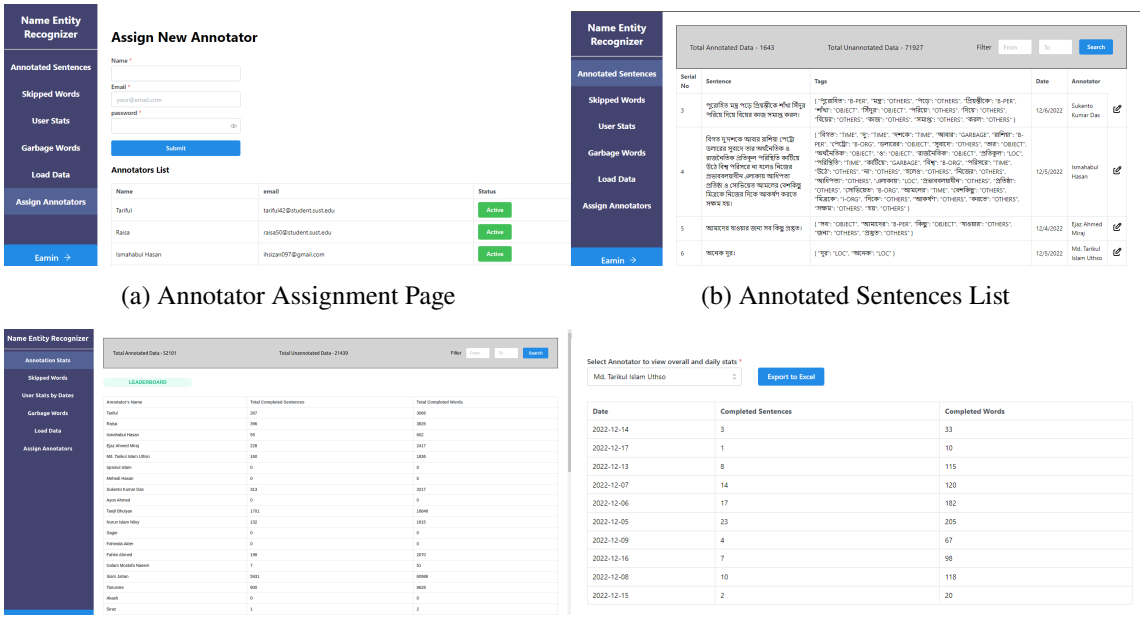
2. **Roles**

   - **Admin**

     The Admin can assign annotators, load data and overlook the annotation process. Admin has access to all the annotated data to edit (fig 3.2b) , and also has both a high level view and a detailed view of all the annotator's respective annotation statistics (fig 3.2c and 3.2d). The admin can manually enable and disable a particular annotator (fig 3.2a) if some issue arises regarding the quality of annotations.

   - **Annotator**

     The annotators have access to the sentences to tag, and their respective annotation statistics.



(a) Annotator Assignment Page

(b) Annotated Sentences List

(c) Annotator Leaderboard

(d) Detailed statistics of a user

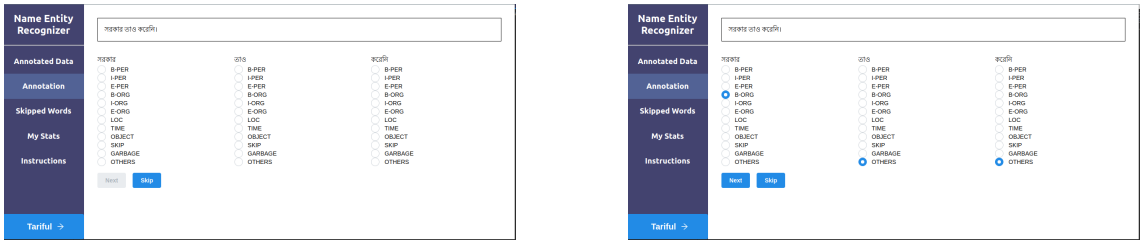
Figure 3.2: Screenshots of Different System Pages for Admin

3. **Annotation Workflow**

Every annotator possesses individual profiles. The administrator has the ability to manually activate or deactivate specific annotators. Once an annotator logs into the system, they are presented with a sentence for annotation. The sentence is presented at the top of the interface, followed by the arrangement of tokens on the screen. Subsequently, the annotator labels the sentence in accordance with the guidelines outlined on the website. Upon completing the annotation of a sentence, the annotator proceeds to annotate the subsequent one. The annotation is demonstrated in the figure 3.3

3.2.2 Tag Descriptions

In the landscape of Named Entity Recognition tasks, several tagging schemes have gained widespread usage. Notably, the IOB (Inside, Outside, Beginning) and IOBES (Inside, Outside, Beginning, End, Single) schemes are well-known frameworks employed to classify tokens into specific named entity categories, such as ”person,” ”location,” and ”organization.”



(a) Annotator Interface - Sentence Annotation

(b) Annotator Interface - After Tagging

Figure 3.3: Screenshots of Annotator Interface

In the context of our study, we have the tagging scheme referred to as the IOBE (Inside, Outside, Beginning, End) scheme. This is similar to the IOBES tagging scheme with the alteration that the single entities are marked as the beginning entities. Table 3.3 represents a complete list of tags used in our work with their respective implications and examples in sentences.

Entity Name	**Description**	**Example**
B-PER	The word is the name of a person. If there are multiple words that represent a single person, the first word will be tagged as ”B-PER”. If the name consists of only one word, it will be tagged as ”B-PER”	জাহিদ (B-PER) হোসেনও একজন বৈমানিক। আমাদের প্রামাণ্যচিত্রে টমাস (B-PER) ডিক্সটার বলেছেন,আমরা সবাই বিল(B-PER)গেটস হতে পারি না।
I-PER	All the words in a person’s name, except the first and the last one.	পার্টির চেয়ারম্যান গোলাম মোহাম্মদ(I-PER) কাদের
E-PER	The last word in a person’s name.	তার নাম জাহানারা বেগম(E-PER)
B-ORG	The first word in the name of an organization or institution.	শাহজালাল (B-ORG) বিজ্ঞান ও প্রযুক্তি বিশ্ববিদ্যালয় এবছর দ্বিতীয়বারের মতো অনুষ্ঠানটি আয়োজন করতে যাচ্ছে।
I-ORG	All words except the first and the last one in the name of an organization or institution.	বাংলাদেশ প্রকৌশল(I-ORG)বিশ্ববিদ্যালয়ের স্থাপত্য-বিদ্যার স্নাতক পর্যায়ের শিক্ষার্থীদের প্রাকৃতিক ভূ-দৃশ্য স্থাপত্যবিদ্যার দুটি কোর্স পড়ানো হয়।
E-ORG	The last word in the name of an organization or institution.	ইউনিয়ন পরিষদের(E-ORG)কার্যকাল ৫ বছর। গার্ল গাইড(E-ORG)একটি অরাজনৈতিক এবং সমাজসেবামূলক যুব আন্দোলন।
LOC	The word represents a location or a physical place.	বাংলাদেশের (LOC)অভ্যন্তরে ভারতের (LOC) ১১১টি ও ভারতের (LOC)অভ্যন্তরে বাংলাদেশের (LOC)৫১টি ছিটমহল রয়েছে।
TIME	The word specifies a time or event.	গত বসন্তে (TIME)তার সাথে আমার শেষ দেখা হয়েছিলো। সাতচল্লিশের (TIME)দেশভাগের পরেই শুরু ছিটমহল আখ্যানের।
OBJECT	The name of an item or any object.	ডিনার সেট (OBJECT)কেনায় চামেলী কোন বিষয়টি বিবেচনা করেছেন?
OTHERS	The word that does not belong to any Name Entities.	সরকার তাও (OTHERS)করেনি (OTHERS)

Table 3.3: Entity Tagging Examples

3.3 Data Validation

Data validation, especially the validation of annotations, is of paramount importance in ensuring the reliability and credibility of our curated dataset. Every sentence within our current dataset has been meticulously tagged by a single annotator. However, to guarantee the accuracy and consistency of the annotations, a rigorous validation process is essential.

To accomplish this, we are undertaking a comprehensive validation process by re-tagging all sentences within the dataset. This time, we are employing well-established metrics like Cohen's Kappa to gauge the inter-annotator agreement of the annotations. This approach involves having multiple annotators independently re-tag the sentences. The resulting inter-annotator agreement score provides valuable insights into the degree of consistency among annotators, thus serving as a measure of annotation quality.

By engaging in this validation process, we aim to identify and rectify any discrepancies or inconsistencies that may have arisen during the initial annotation phase. The validation effort not only ensures that our dataset maintains a high standard of quality but also enhances the reliability of the annotations for downstream applications.

3.4 Data Description

3.4.1 Annotated Data

In this subsection, we represent a comprehensive analysis of our annotated dataset shedding light on key patterns, trends, and observations. Following is an analysis of our annotated dataset which contains 52,124 sentences in total.

1. **Sentence Length Distribution:**

   In this analysis, we delve into the distribution of sentence lengths within our annotated dataset. By examining the variation in sentence lengths, we gain insights into the linguistic characteristics of the text and uncover potential patterns that might influence named entity annotation. Our statistics reveal that the minimum word count for sentences was 1, and the word count even ranges up to 125 words for the sentences. Almost half of the sentences have the word count of below 30. The figure 3.4 further demonstrates the sentence length

distribution of our dataset.

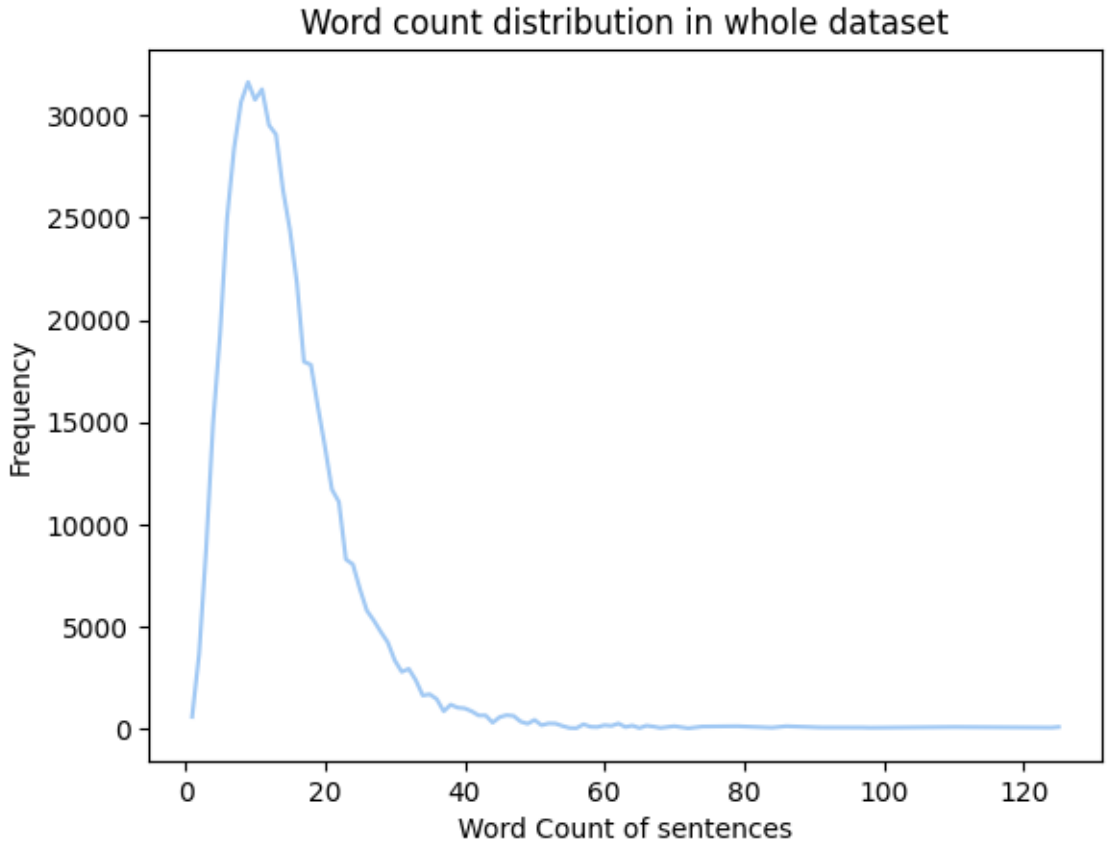


Figure 3.4: Sentence length distribution of the whole dataset.

2. **Quantifying Named Entities:**

   One of the key aspects of our data analysis involves quantifying the occurrences of named entities and comparing them to non-named entities. By numerically assessing the prevalence of different entity types, we can better understand the composition of our dataset and its relevance for various natural language processing tasks. The figure 3.5 shows that the occurrence of named entities are significantly lower than non named entity tokens.

3. **Distribution of Named Entity Categories:**

   In this analysis, we focus on the distribution of different named entity categories present in the dataset. By categorizing entities into classes such as person, organization, location, and

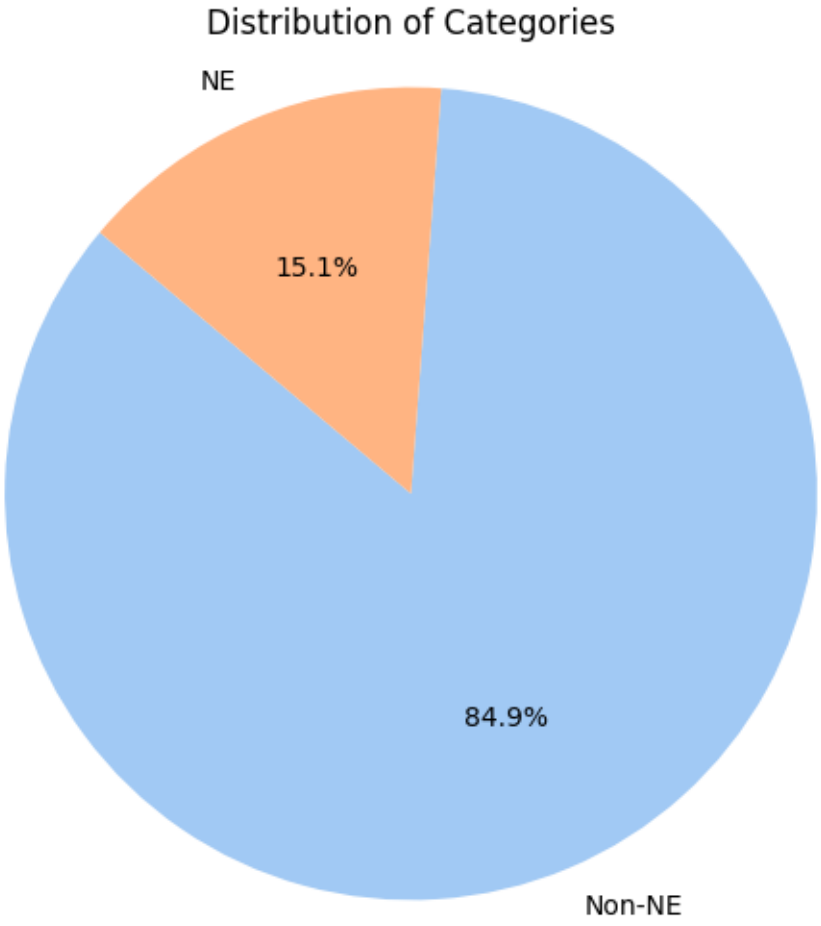


Figure 3.5: Ratio of Named Entities vs Non-NE tokens in the dataset.

others, we obtain insights into the prominence of each category and its potential impact on downstream applications. The figure 3.6 demonstrates the relative occurrences of each tags in the text.

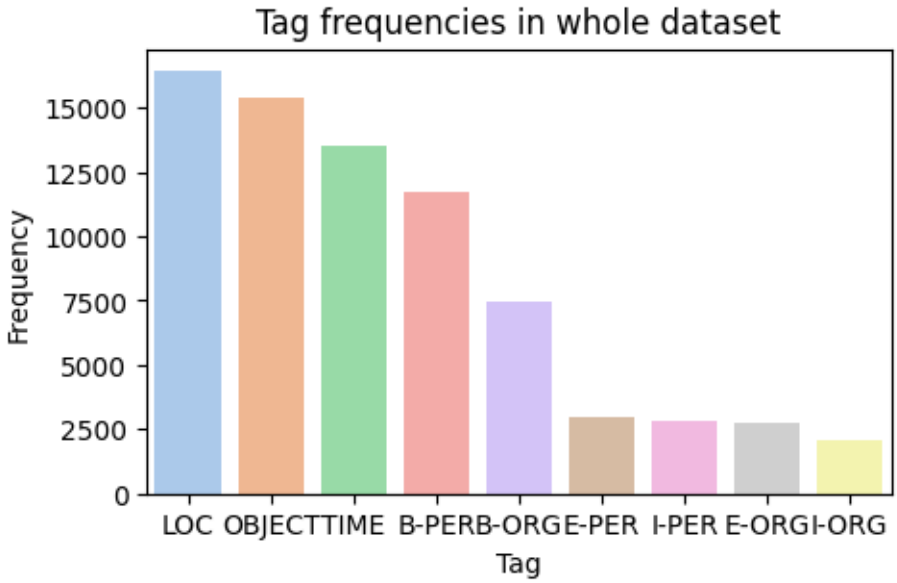


Figure 3.6: Distribution of each type of tokens in the dataset.

3.4.2 Validated Data

Our validated dataset consists of 31,124 sentences. Following are the observations on different aspects of the validated data:

1. **Quantifying Named Entities in the Validated Dataset:**

   In evaluating the impact of re-annotation on our dataset, a notable shift is observed in the distribution of named entity tags. The validated dataset reveals a decrease in the ratio of named entity tags from 15.1% in the annotated dataset to 12.2%. This change is attributed to a meticulous re-annotation process where a focused approach on proper nouns for named entity tagging was implemented. The observed drop underscores our commitment to precision, ensuring a more accurate representation of named entities in the dataset.

   Figure 3.7 shows the occurrence of named entities after validating the dataset.

2. **Distribution of Named Entity Categories:**

   Figure 3.8 illustrates the relative occurrences of each tag in the annotated data. Upon comparison with the validated data, the overall ratio remains consistent, with a noteworthy observation: the LOC (Location) tags have been slightly surpassed by the OBJ (Object) tags.

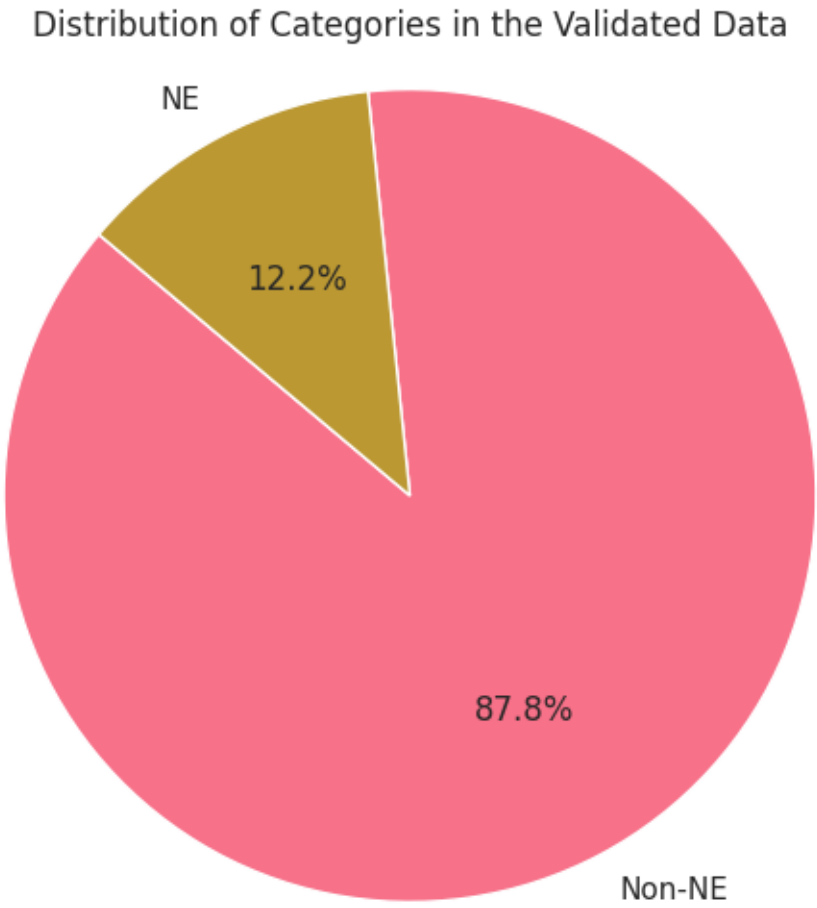


Figure 3.7: Ratio of Named Entities vs Non-NE tokens in the dataset.

3.4.3 Existing Data

3.4.3.1 The B-NER Dataset

For our research, we leveraged the B-NER dataset, a comprehensive collection of sentences sourced from diverse sources such as blogs, news articles, forums, and Wikipedia. The dataset creation process involved web crawling, targeting newspapers, blogs, and forums. Initially, around 100,000 sentences were collected, and after filtering out code-mixed texts and sentences with spelling errors, the dataset was refined to 22,144 sentences containing 297,409 tokens.

The B-NER dataset encompasses 8 named entity (NE) categories, divided into 17 tags including 'B-geo,' 'O,' 'B-gpe,' 'B-per,' 'I-per,' 'B-tim,' 'B-org,' 'I-org,' 'B-art,' 'I-art,' 'I-tim,' 'B-eve,'

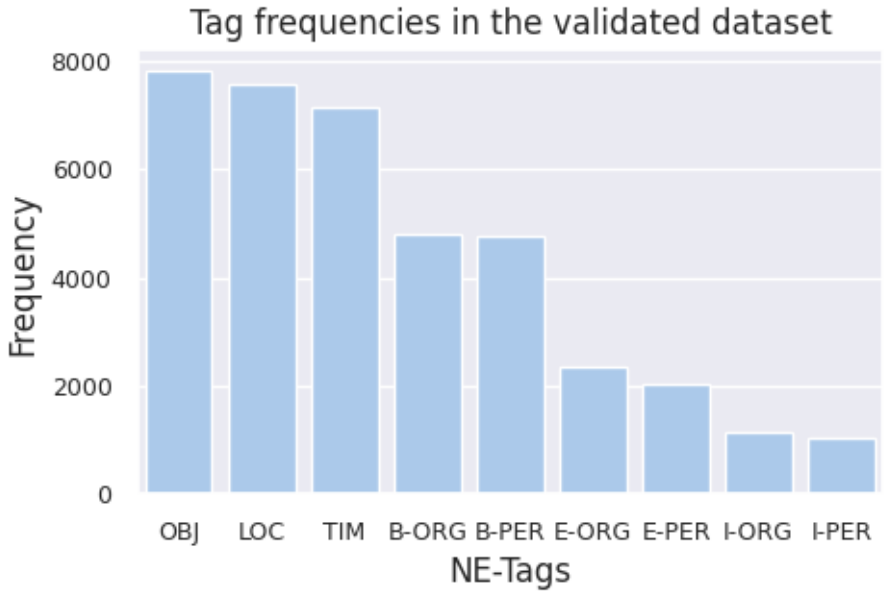


Figure 3.8: Distribution of each type of token in the dataset.

'I-eve,' 'I-geo,' 'I-gpe,' 'B-nat,' and 'I-nat.' The dataset adheres to the BIO tagging scheme, inspired by the Groningen Meaning Bank (GMB), where the beginning token of an entity is tagged as 'B-entity name,' and the subsequent tokens are tagged as 'I-entity name.' Tokens not associated with any of the eight entity types are labeled as 'O.' The dataset's design ensures a structured representation for effective named entity recognition (NER).

A visual representation of the tag frequencies in the B-NER dataset is presented in Figure 3.9, highlighting the distribution of entity types within the dataset.

NE Tags	B Tag	I Tag	Total
PER	9,201	9,024	18,225
GEO	10,421	695	11,116
ORG	5,465	4,322	9,787
GPE	2,453	9	2,462
NAT	31	17	48
TIM	4,736	2,449	7,185
ART	422	316	738
EVE	515	398	913
MISC	-	-	2,46,943

Figure 3.9: Distribution of each type of token in the B-NER dataset, adapted from Haque et al(2023)

3.4.4 Combined Dataset

In pursuit of a more expansive and diverse dataset for our experimentation, we embarked on a strategic integration of the B-NER dataset with our recently validated dataset. This collaborative effort aimed to leverage the strengths of both datasets while addressing specific nuances in their tagging schemes.

Notably, the B-NER dataset introduced unique tags such as 'B-art' and 'I-art' to denote artifact names. To reconcile this divergence, we seamlessly mapped these tags to our 'OBJ' category, which encapsulated a similar semantic space of object-related terms. This harmonization not only enhanced the richness of our dataset but also provided a unified framework for handling diverse entity types.

Another significant alignment involved the representation of events ('B-eve,' 'I-eve') and natural phenomena ('B-nat,' 'I-nat') in the B-NER dataset. Recognizing the overlap with our 'TIM' (time) category, encompassing events and natural occurrences, we strategically converged these entities under a cohesive umbrella. This deliberate mapping aimed to create a more cohesive and interoperable dataset, laying the groundwork for nuanced named entity recognition tasks.

To navigate the challenge of different IOB (Inside, Outside, Beginning) tagging schemes—B-NER using IOB and our dataset utilizing IOBE — we executed a meticulous transformation. In the merged dataset, we systematically converted our 'E-entity' tags to 'I-entity' tags, aligning with the B-NER dataset's tagging convention. This careful maneuver ensured a seamless integration, fostering a dataset with consistent tagging patterns crucial for cross-domain NER tasks.

To elucidate the harmonization process, we present the detailed tables 3.4 and 3.5 showcasing the conversion of tags between the two datasets:

B-NER Tags	**Our Tags**
B-art, I-art	OBJ
B-eve, I-eve, B-nat, I-nat	TIM

Table 3.4: Adjustments made to the B-NER dataset for merging with our dataset

Our Tags	Adjusted Tags
E-ORG	I-ORG
E-PER	I-PER

Table 3.5: Adjustments made to our dataset tagging scheme for the merging

This meticulous integration not only addresses disparities but also enhances the overall coherence of the combined dataset. We obtain a combined dataset of 53,763 sentences for further experimentation.

3.4.5 Synthetic Data

To augment our annotated dataset and enrich the diversity of our training data, we employed a rule-based approach for synthetic data creation. Leveraging a collection of 500,000 people's names, we systematically crafted sentences and annotated them using rule-based tagging methods. This approach resulted in the addition of approximately 700,000 sentences, each containing named entities, to our synthetic dataset.

Sample from Synthetic Dataset:

As part of our synthetic data generation process, we present a brief excerpt to showcase the diversity and structure of the artificially created sentences. This sample includes a subset of sentences along with their corresponding named entity tags:

1. Sentence 1:
    - Text: "মোছলেহ উদ্দিন বগুড়া জেলায় বসবাস করেন।"
    - Named Entity Tags: [B-PER, E-PER, LOC, O, O, O]
2. Sentence 2:
    - Text: "রঞ্জন চৌধুরী গত বুধবার ঢাকা এসেছেন।"
    - Named Entity Tags: [B-PER, E-PER, TIM, TIM, LOC, O]

3.5 Word-Based Data

Our dataset comprises a total of 315,168 words, including duplicates. Upon removing duplicates, we are left with 38,363 distinct words. In the development of our word-based dataset, we opted for a straightforward approach—associating each word with its most frequent tag. While this method simplifies the dataset, it introduces a loss of context and potential ambiguity in tag assignment. The word-based dataset offers potential applications for experimenting with classical machine learning models, though we haven't conducted such experiments in this study.

Additionally, we can provide the corresponding word count for each of the tags, offering a more detailed perspective on the distribution of tags within the word-based dataset in table 3.6.

Tag	Number of Unique Words
B-ORG	1551
B-PER	1788
E-ORG	684
E-PER	882
I-ORG	567
I-PER	461
LOC	2640
OBJ	3594
TIM	1477
O	31452

Table 3.6: Number of Unique Words in Each Tag

Chapter 4

Methodology

The methodology section outlines the following steps: Dataset preparation, Model Training, Model Evaluation, Experiment and Analysis in Named Entity Recognition (NER) model. After developing NER dataset, the aim of using the BanglaBERT [1] and Indic-Bert [25] which are transfer learning based model.

4.1 BanglaBERT on Validated Data and Combined Data

BanglaBERT, a Natural Language Understanding (NLU) model built upon the foundation of BERT [26] architecture which stands for Bidirectional Encoder Representations from Transformers, is a state-of-the-art language model designed to capture contextual information from text. BanglaBERT also employs a transformer architecture, characterized by multiple layers of self-attention mechanisms that allow the model to effectively understand and represent contextual relationships between words in a sentence.

To pretrain BanglaBERT, a comprehensive dataset known as 'Bangla2B+' was compiled. This dataset is generated by crawling a diverse collection of 110 popular Bengali websites, resulting in a rich and varied source of Bengali text.

The illustration on figure 4.1 sheds light into the detailed architecture of the BanglaBERT model.

4.2 Data Preprocessing

The process of preparing and formatting the labeled Named Entity Recognition (NER) dataset is a crucial step in the journey of model development. This section explains the careful steps we took to convert raw data into a structured format suitable for training and evaluating the BanglaBERT-based NER model.

4.2.1 Getting the Data Ready

We started by keeping our labeled NER dataset in a MongoDB database. This acted like a safe storage place where we could put and get back our dataset easily.

4.2.2 Asking for the Data Using API

To get the data out of the storage, we used API requests. Think of it like asking the database nicely for the data we needed. The data came back in JSON format, which is a simple way to organize information.

4.2.3 Turning JSON into Tables

We wanted to make things even neater, so we turned the JSON data into something like tables using CSV format. This made the data easier to work with and helped us remove unnecessary parts.

4.2.4 Cleaning Up: Removing the Garbage Data

We did not require all the words and tags in the dataset, so we cleaned them up by removing the redundant things. This step helped us keep only the important information.

4.2.5 Making it Fit for the Model

The BanglaBERT model takes a special kind of data format called JSONL. So, we changed our cleaned-up CSV data into JSONL format. This way, the data was ready for the BanglaBERT-based

NER model to use.

4.2.6 Dataset Split

The dataset was divided into three distinct subsets:

- **Training Set (80%)**: Used for training the model.
- **Validation Set (10%)**: Employed to fine-tune model parameters and prevent overfitting.
- **Test Set (10%)**: Used to evaluate the model's performance on unseen data.

Dataset	Train	Validation	Test	Total
SUNER	25638	2849	3166	31653
SUNER + B-NER	43547	4839	5377	53763

Table 4.1: Dataset split for BanglaBERT

The Table 4.1 shows the split count of our dataset and the combined form of our dataset and the B-NER dataset. it shows that our dataset SUNER is split into 25638 training samples, 2849 validation samples, and 3166 test samples, for a total of 31653 samples. The combined form of "SUNER + B-NER", shows that the dataset is split into 43547 training samples, 4839 validation samples, and 5377 test samples, for a total of 53763 samples

4.3 Model Training and Evaluation

In the first phase, the model underwent an initial training period spanning 10 epochs. During this period, the model was exposed to the labeled dataset, learning to identify and categorize named entities embedded within the text. Following the initial training, the model's efficacy was evaluated using the validation set. This evaluation centered on a crucial metric known as eval_loss, which indicated how well the model performed on the validation data. A rising trend in eval_loss beyond a certain epoch count hinted at overfitting, where the model might become too specialized for the training data. To address this, a second training phase was executed, lasting for 2 additional epochs. This fine-tuning phase aimed to further enhance the model's performance while preventing potential overfitting. This meticulous approach ensured that the model's generalization capabilities were optimized for a wide range of unseen data.

Following the training phases, the trained model was tested using the independent test dataset. This allowed us to assess its ability to generalize and accurately predict named entities in unseen text.

4.4 IndicBERT on Validated Data

The Indic BERT model is based on the ALBERT [27] model, a recent derivative of BERT. It is pre-trained on 12 Indian languages: Assamese, Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu. To get the performance of our dataset we use this multilingual model. We take the help of simpletransformer [28] system to load the model architecture from huggingface [29].

4.4.1 Dataset Preparation

The data format is similar to our csv format where each row contains a sentence and one token with its true label. To fit the model instead of sentence we have to convert it to sentence_id, tokens to words and labels.

The dataset was divided into three distinct subsets:

- **Training Set (80%)**: Used for training the model.
- **Validation Set (10%)**: Employed to fine-tune model parameters and prevent overfitting.
- **Test Set (10%)**: Used to evaluate the model's performance on unseen data.

Train	Validation	Test	Total
25638	2849	3166	31653

Table 4.2: Dataset split for Indic-bert

The table 4.2 shows that after splitting our dataset we got 25,638 train data, 2,849 validation data and 3,166 test data.

4.4.2 Model Training and Evaluation procedure

The training process was configured with a train batch size of 16, facilitating efficient utilization

of computational resources and batch-wise gradient descent optimization. Additionally, an evaluation strategy was employed, whereby the model's performance was evaluated periodically on the validation set to monitor training progress and prevent overfitting. The model underwent iterative training over a total of 20 epochs, with each epoch comprising a complete pass through the training dataset. During each epoch, the model was exposed to batches of training data, and the gradients of the loss function with respect to model parameters were computed and utilized to update the model weights via backpropagation.

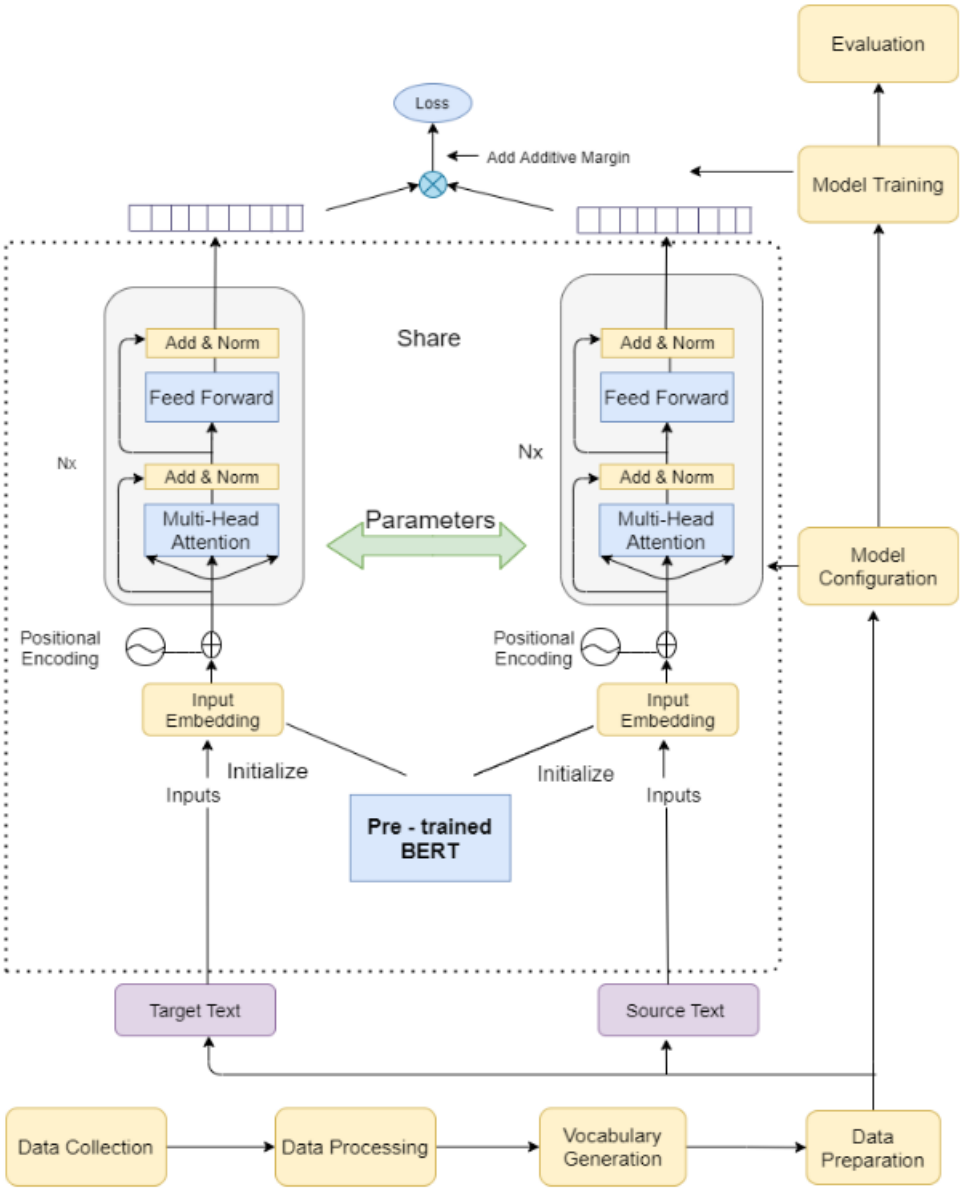


Figure 4.1: Encoder's Structure of Bangla-BERT and weight sharing mechanism. The right encoder is to develop the Bangla-BERT pre-training model using BanglaLM unsupervised dataset. The left encoder accepts the trained parameters from the pre-trained model(right encoder) and is used as fine-tuning for downstream tasks adapted from Kowsher et al. [1]

Chapter 5

Results and Discussion

5.1 Results on BanglaBERT on our Validated Dataset

After running the baseline BanglaBERT model in our dataset, we found the following performance metrics, illustrated on the table 5.1.

Metric	Value
Predict Weighted Avg F1 Score	0.6707
Predict weighted Avg Precision	0.677
Predict weighted Avg Recall	0.6736

Table 5.1: Performance Metrics for Training BanglaBERT

The obtained metrics shown in Table 5.1 give insights into the model's overall performance on our validated dataset. The predicted weighted Avg F1 Score of 0.6707 reflects the balance between precision and recall, indicating a reasonable level of accuracy. The model was evaluated on a dataset of 4151 instances, and the predicted weighted average Precision of 0.6777 highlights its ability to correctly identify positive cases. The predicted weighted average recall of 0.6736 signifies the model's capacity to capture a significant proportion of actual positive cases. These scores collectively contribute to understanding the model's effectiveness in BanglaBERT training.

Table 5.2 presents the average F1 scores achieved by the trained model on the test dataset across different named entity categories.

The average F1 scores are reported for the following named entity categories:

Tags	Avg F1
PER	0.736
ORG	0.619
LOC	0.715
TIME	0.604

Table 5.2: Performance Metrics for each label in BanglaBERT

- **PER (Person)**: The model attained an average F1 score of 0.736 for identifying and classifying person names within the text.
- **ORG (Organization)**: For recognizing organization names, the model achieved an average F1 score of 0.619, indicating its capability in discerning organizational entities.
- **LOC (Location)**: In terms of identifying location names, the model demonstrated proficiency with an average F1 score of 0.715, signifying accurate localization of geographic entities.
- **TIME (Time)**: The model exhibited competence in recognizing temporal entities, achieving an average F1 score of 0.604 for time-related expressions.

5.2 Results on BanglaBERT on our Merged Dataset

We further experimented with the BanglaBERT model leveraging the combined dataset obtained after processing and merging our validated dataset and the B-NER dataset.

Metric	Value
Predict Weighted Avg F1 Score	0.7843
Predict weighted Avg Precision	0.7826
Predict weighted Avg Recall	0.7865

Table 5.3: Performance Metrics on combined dataset in BanglaBERT

The metrics found in the experiment are shown in table 5.3. By merging our dataset with the B-NER dataset, we observed a significant improvement in the performance of our model. This finding suggests that enriching the training data with additional labeled examples led to a more robust and generalizable model. This can be attributed to the increased exposure to diverse linguistic patterns

and named entity variations, enabling the model to better capture the intricacies of named entity recognition.

Tags	Avg F1
PER	0.859
ORG	0.730
LOC	0.825
TIME	0.683

Table 5.4: Performance Metrics for each label on combined dataset in BanglaBERT

Table 5.4 presents the average F1 scores achieved by the trained model on the test dataset across different named entity categories.

The average F1 scores are reported for the following named entity categories:

- **PER (Person)**: The model attained an average F1 score of 0.859 for identifying and classifying person names within the text.
- **ORG (Organization)**: For recognizing organization names, the model achieved an average F1 score of 0.730, indicating its capability in discerning organizational entities.
- **LOC (Location)**: In terms of identifying location names, the model demonstrated proficiency with an average F1 score of 0.825, signifying accurate localization of geographic entities.
- **TIME (Time)**: The model exhibited competence in recognizing temporal entities, achieving an average F1 score of 0.683 for time-related expressions.

it also proves that combining the dataset also improves the performance of predicting each labels correctly.

5.2.1 Results on indicBERT on Validated Dataset

Metric	Value
F1 score	0.65

Table 5.5: Performance of Indic-bert on Combined Dataset

The table 5.5 shows that the evaluation of the Indic-Bert model for our dataset yielded an average F1 score of 0.65. While the model demonstrated competency in identifying named entities within Bengali text, achieving a respectable F1 score, it is noteworthy that this performance does not surpass the results obtained by state-of-the-art models reported in the literature.

Chapter 6

Conclusion

In this study, we embarked on the journey of building a comprehensive dataset for named entity recognition tasks. With painstaking efforts, we collected a dataset containing a substantial corpus of 73,000 sentences. Among these, we meticulously annotated 52,200 sentences and further validated 31,000 sentences marking the entities that characterize the text.

Our initial analysis involved applying the BanglaBERT and indicBERT models as a baseline for named entity recognition. The results have illuminated the significant room for improvement that lies ahead.

As we proceed, there are numerous avenues we intend to explore in order to enhance our results.

To further improve the performance, we envisage several strategies. We aim to enrich our dataset further, incorporating more diverse and domain-specific sentences. Tuning hyperparameters and experimenting with different models, including ensemble methods, will be instrumental in refining our approach.

Our contributions thus far have been substantial. Our dataset has the potential to develop into an effective tool for testing and refining named entity identification models. Additionally, our ongoing efforts in annotator validation signal our commitment to elevating the quality and reliability of our annotations.

While we know that our journey has only just begun, we are determined to improve our performance and impact. We are convinced that by pursuing these approaches attentively, we will get improved results, ultimately contributing to the improvement of named entity recognition and its applications in natural language processing.

Chapter 7

Future Work

As we conclude this phase of our research, our journey forward holds a plethora of exciting opportunities for advancement in the field of named entity recognition. The avenues we intend to explore include:

1. **Exploring Cutting-Edge Models**: In our pursuit of enhancing accuracy, we will delve into the application of state-of-the-art machine learning models, particularly transformer-based architectures. The integration of pre-trained language models promises to imbue our entity tagging process with heightened robustness and context sensitivity. Our literature review indicate that using different BERT base models like sagorBERT, and mBERT models can be used for NER tasks for Bangla language although these are more appropriate for multilingual natural language processing.

2. **Completing Validation on Data Annotation**: Our ongoing validation of dataset annotations is a critical step toward refining the quality of our dataset. By rigorously validating each annotation, we aim to rectify errors, ensuring the accuracy and reliability of the final dataset.

3. **Exploring Ensemble Techniques**: Ensembling, a powerful approach, beckons us to explore its potential in our context. By combining predictions from multiple models, we aspire to attain heightened performance metrics, potentially unlocking new levels of precision in named entity recognition.

4. **Merge NER Datasets**: Our vision extends to amalgamating diverse named entity datasets

into a unified resource. This larger dataset will be an invaluable asset to the natural language processing community, facilitating training and evaluation across a wide spectrum of applications.

5. **Harnessing Data Augmentation**: In our quest to fortify our dataset, we're excited to implement data augmentation techniques. These techniques hold the promise of enriching the dataset by generating variations, thus contributing to model resilience and adaptability. We plan to refine our process of generating the synthetic dataset, so it can contribute to the model's performance.

The journey ahead is marked by innovation, exploration, and relentless dedication to advancing the boundaries of our work. By embracing these future directions, we hope to contribute significantly to the evolution of named entity recognition and its potential to transform the landscape of natural language processing.

References

[1] M. Kowsher, A. Sami, N. Prottasha, M. Arefin, P. Dhar, and T. Koshiba, "Bangla-bert: Transformer-based efficient model for transfer learning and language understanding," *IEEE Access*, vol. 10, pp. 1–1, 01 2022.

[2] B. L. [Internet]., "Definitions," p. Bengali, 2021.

[3] K. I, P. D, D. AW, and A. S., "Boosting drug named entity recognition using an aggregate classifier," *05*, vol. 007, 2015.

[4] D. RI, L. R, and L. Z., "Ncbi disease corpus: A resource for disease name recognition and concept normalization," *J Biomed Inform [Internet]. 2014; 47:1–10. Available from: https://doi.org/http3A//dx.doi.org/10.1016/j.jbi.2013.12.006 PMID:*, p. 24393765, 2014.

[5] K. K and J. S., *Improving Topic Quality by Promoting Named Entities in Topic Modeling*. In: ACL 2018 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 2018.

[6] A. N., *Chatbot: A Conversational Agent employed with Named Entity Recognition Model using Artificial Neural Network*.

[7] K. S and R. C., *Using Type Information to Improve Entity Coreference Resolution*. In: Proceedings of the First Workshop on Computational Approaches to Discourse, 2020.

[8] J. S, S. S, and L. A., *Named Entity Recognition and Normalization in Tweets Towards Text Summarization*. In: 8th International Conference on Digital Information Management, ICDIM 2013, 2013.

[9] M. N and D. S., *Graph Based Text Summarization using NER and POS*. Int J Sci Res Dev, 2018.

[10] M. P, M. S˘, K. J, L. P, and P. J, *S˘ edivy´ J*. Text Summarization of Czech News Articles Using Named Entities, 2021.

[11] T. A, N. E, L. F, and M. R. I. Q. A. U. N. E. Recognition, *In: International Conference on Application of Natural Language to Information Systems*, 2005.

[12] M. D, van Zaanen M, and S. D., *Named Entity Recognition for Question Answering*. In: Proceedings of the Australasian Language Technology Workshop 2006, 2006.

[13] U. A, T. A, N. T, T. H, and O. M. N. M. T. I. N. Entity, *In: Proceedings of the 27th International Conference on Computational Linguistics*, 2018.

[14] A. Ekbal and S. Bandyopadhyay, "A hidden markov model based named entity recognition system: Bengali and hindi as case studies," in *Pattern Recognition and Machine Intelligence*, A. Ghosh, R. K. De, and S. K. Pal, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 545–552.

[15] ——, "Bengali named entity recognition using support vector machine," in *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008. [Online]. Available: https://aclanthology.org/I08-5008

[16] N. Patil, A. Patil, and B. Pawar, "Named entity recognition using conditional random fields," *Procedia Computer Science*, vol. 167, pp. 1181–1188, 01 2020.

[17] A. Ekbal and S. Saha, "Maximum entropy classifier ensembling using genetic algorithm for NER in Bengali," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), May 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/718$_{P}aper.pdf$

[18] S. A. Chowdhury, F. Alam, and N. Khan, "Towards bangla named entity recognition," in *2018 21st International Conference of Computer and Information Technology (ICCIT)*. IEEE, 2018, pp. 1–7.

[19] M. J. Rahman Rifat, S. Abujar, S. R. Haider Noori, and S. A. Hossain, "Bengali named entity recognition: A survey with deep learning benchmark," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2019, pp. 1–5.

[20] R. Karim, M. A. Islam, S. Simanto, S. Chowdhury, K. Roy, A. Neon, M. Hasan, A. Firoze, and M. Rahman, "A step towards information extraction: Named entity recognition in bangla using deep learning," *Journal of Intelligent Fuzzy Systems*, vol. 37, pp. 1–13, 07 2019.

[21] I. Ashrafi, M. Mohammad, A. Shawkat, G. Nijhum, R. Karim, N. Mohammed, and S. Momen, "Banner: A cost-sensitive contextualized model for bangla named entity recognition," *IEEE Access*, vol. PP, pp. 1–1, 03 2020.

[22] M. Z. Haque, S. Zaman, J. R. Saurav, S. Haque, M. S. Islam, and M. R. Amin, "B-ner: A novel bangla named entity recognition dataset with largest entities and its baseline evaluation," *IEEE Access*, vol. 11, pp. 45 194–45 205, 2023.

[23] M. A. A. Mumin, A. A. M. Shoeb, M. R. Selim, and M. Z. Iqbal, "Supara: A balanced english-bengali parallel corpus," *SUST Journal of Science and Technology*, vol. 16, no. 2, pp. 46–51, 2012, submitted: October 16, 2011; Accepted for Publication: May 12, 2012.

[24] T. Hasan, A. Bhattacharjee, K. Samin, M. Hasan, M. Basak, M. S. Rahman, and R. Shahriyar, "Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, November 2020.

[25] D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., A. Bhattacharyya, M. M. Khapra, and P. Kumar, "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages," in *Findings of EMNLP*, 2020.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[27] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” 2020.

[28] T. Rajapakse, “Simple transformers.” [Online]. Available: https://simpletransformers.ai/

[29] “Hugging face.” [Online]. Available: https://huggingface.co/ai4bharat/indic-bert