IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 February 2023, accepted 29 March 2023, date of publication 17 April 2023, date of current version 12 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3267746

APPLIED RESEARCH

B-NER: A Novel Bangla Named Entity Recognition Dataset With Largest Entities and Its Baseline Evaluation

MD. ZAHIDUL HAQUE[1], SAKIB ZAMAN[1], JILLUR RAHMAN SAURAV[2], SUMMIT HAQUE[3], MD. SAIFUL ISLAM[4], AND MOHAMMAD RUHUL AMIN[5], (Member, IEEE)

[1]Department of Computer Science and Engineering, Sylhet Engineering College, Tilagarh, Sylhet 3100, Bangladesh
[2]Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX 76019, USA
[3]Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh
[4]Department of Computing Science, University of Alberta, Edmonton, AB T6G 2R3, Canada
[5]Department of Computer and Information Science, Fordham University, Bronx, NY 10458, USA

Corresponding authors: Md. Saiful Islam (mdsaifu1@ualberta.ca) and Mohammad Ruhul Amin (mamin17@fordham.edu)

ABSTRACT Within the Natural Language Processing (NLP) framework, Named Entity Recognition (NER) is regarded as the basis for extracting key information to understand texts in any language. As Bangla is a highly inflectional, morphologically rich, and resource-scarce language, building a balanced NER corpus with large and diverse entities is a demanding task. However, previously developed Bangla NER systems are limited to recognizing only three familiar entities: person, location, and organization. To address this significant limitation, we introduce a novel Bangla NER dataset B-NER, which was created using 22,144 manually annotated Bangla sentences collected from Bangla newspapers and Bangla Wikipedia. This dataset includes a total of 9,895 unique words which were manually categorized into eight different entity types, such as a person, organization, event, artifact, time indicator, natural phenomenon, geopolitical entity, and geographical location. Inter-annotator agreement experiments were conducted to validate the quality of annotations performed by three annotators, resulting in a Kappa score of 0.82. In this paper, we provide an outline of the annotation guideline illustrated with examples, discuss the B-NER dataset properties, and present benchmark evaluations of the dataset. To establish that B-NER is more comprehensive and balanced in comparison to other publicly accessible datasets, we conducted cross-dataset modeling and validation, i.e. trained NER model on one dataset while tested on another, and found that the model trained on B-NER performed the best in that settings. Furthermore, we performed exhaustive benchmark evaluations based on Bidirectional LSTM with fastText embeddings and sentence transformer models. Among these models, fine-tuned NR/IndicbnBERT achieved noticeable results with a Macro-F1 of 86%. This dataset and baseline results will be publicly available under a CC-BY 4.0 license in the CoNLL-2002 format to facilitate further research on Bangla NER.

INDEX TERMS Named entity recognition (NER), natural language processing, bangla NER dataset, information extraction, B-NER.

I. INTRODUCTION

Named Entity Recognition (NER) plays a vital role in extracting essential information from text content for executing key information-based Natural Language Processing (NLP) operations down the line, including but not limited to understanding the context, information retrieval, and generating recommendations, etc. NER seeks to recognize and classify *proper nouns* [1], [2], [3] from texts to extract key content elements such as a person, organization, location, etc. After the introduction of NER at the MUC-6 conference [4], it has become an integral part of almost all types of information extraction tasks, such as relation extractions, text summarization, question answering, topic detection, etc.

The associate editor coordinating the review of this manuscript and approving it for publication was Sabah Mohammed.

Initial approaches for solving NER were mainly focused on rule-based methods as described in [5], [6], and [7] where name entities were extracted from texts for certain domains by following a set of handcrafted rules. However, those rule-based methods were designed for specific languages, such as English, Spanish, French, etc., and they also required more relevancy regarding the grammatical structure of sentences. Therefore, researchers shifted their efforts to Machine Learning models as these methods can produce better inference in recognizing name entities on a wide range of domains [8]. As more NER-labeled data became available for other languages, researchers also developed multilingual Machine Learning models for a language family sharing similar linguistic properties [9].

A wide variety of Machine Learning algorithms were proposed for NER, such as Hidden Markov Models (HMMs) [10], [11], Conditional Random Fields (CRFs) [12], [13], Maximum Entropy (ME) [14], Maximum Entropy Markov Models (MEMMs) [15], Support Vector Machines (SVMs) [16], [17], and Ensemble Modeling techniques [18]. However, it has been observed that these models cannot recognize all probable name entities despite utilizing a large annotated corpus. Thus, a combination of Machine Learning methods and a set of handcrafted features, referred to as hybrid NER models, were proposed to achieve higher accuracy [19]. In recent days, with the advancement of deep learning methods and large-scale language models like BERT, it has become easier to develop a robust NER system [20], [21], [22]. Still, a language-specific large annotated corpus of name entities is required to achieve high performance and extend those NER language models in practice.

Despite having much recent success in resource-rich languages, like English, German and French, no such prominent work has been done in NER for most low-resource languages. In this paper, we focus on building the Bangla NER system, which still needs a well-annotated corpus of name entities. A few labeled datasets exist for NER in Bangla, but unfortunately, none of them follow the established standard. The principal reason for such a shortcoming is the unavailability of qualified labelers who are knowledgeable about the name entities in Bangla. Another reason is that the annotation process is very costly as it is time-consuming and tedious. In particular, NER labelers need to go through each word of a sentence and understand the context in which the words were used before labeling them with appropriate tags, followed by another pass to inspect the correctness of tags.

Previously published Bangla NER datasets suffer from several demerits: (1) while other languages have well-established studies to identify name entities like geopolitical, natural phenomena, artifacts, and events [23], none of them attempted to recognize those aforementioned entities in Bangla; and (2) previously produced datasets suffer from weak supervision as the annotation was done automatically by using automated tools developed for other popular online languages. For example, Karim et al. [24], and Rifat et al. [25] both prepared Bangla NER datasets that contain only four entity classes

Entity name	Definition	Example
Person (Per)	Defines the name of a person. Entities could be individual or with a rank or designation. Sometimes honorific addressing terms before a person's name also acts as person entities.	সাকিব আল হাসান (Shakib Al Hasan), লিওনেল মেসি (Lionel Messi), প্রধানমন্ত্রী শেখ হাসিনা (Prime Minister Sheikh Hasina), মেয়র আনিসুল হক (Mayor Anisul Haque), মেজর হুমায়ুন কবির (Major Humayun Kabir), মি হক (Mr Haque), ড প্রফুল্ল চন্দ্র (Dr Profullo Chandra)
Geographical Entity (Geo)	Indicates geographical locations like cities, towns, villages, continents, countries, seas, streets, roads.	ঢাকা (Dhaka), লন্ডন (London), সুবর্নপুর (Subornopur), এশিয়া (Asia), ইউরোপ (Europe), ব্রাজিল (Brazil), প্রশান্ত মহাসাগর (Pacific Ocean), বিমানবন্দর সড়ক (Airport Road)
Organization (Org)	Defined by companies, enterprises, banks, societies, associations schools, academies, universities, newspapers	গুগল (Google), বাংলাদেশ ব্যাংক (Bangladesh Bank), ফিফা (FIFA), মোহাম্মদপুর মডেল স্কুল (Mohammadpur Model School), বিকেএসপি (BKSP), বাংলা একাডেমি (Bangla Academy), বুয়েট (BUET)
Geopolitical Entity (Gpe)	Geopolitical Entities are entities associated with some sort of political structure	বাংলাদেশি (Bangladeshi), ভারতীয় (Indian), রোহিঙ্গা (Rohingya), ব্রিটিশ (British), আমেরিকান (American)
Time Indicator (Tim)	Entities indicating time	সকাল (Morning), ১২ই জুলাই (12th July), আগামীকাল (Tomorrow), মঙ্গলবার (Tuesday), জানুয়ারি (January), পহেলা বৈশাখ (Pahela Baishakh)
Artifact (Art)	Entities defining artistic works or efforts made by human	নরওয়েজিয়ান উড (Norwegian Wood), বিগ লিটল লাইস (Big Little Lies), টম এন্ড জেরি (Tom and Jerry)
Natural Phenomenon (Nat)	Mentions that define entities for natural incidents and natural objects	হারিকেন মারিয়া (Hurricane Maria), পূর্ণিমা (Full Moon), সাইক্লোন সিডর (Cyclone Sidr)
Event (Eve)	Defined for entities about several events	ঈদুল আযহা (Eid-ul-Ajha), দূর্গাপূজা (Durgapuja), নববর্ষ (Noboborsho), বিশ্বকাপ (World Cup)

FIGURE 1. Snapshot of proposed B-NER eight Name Entities with definitions and few examples.

(Person, Location, Organization, Object / Time). And both datasets were produced using various tagging tools without maintaining linguistically defined standard tags for Bangla. Examples of such popular tools include contextual POS tagging by Benajiba and Rosso [26], FiNER[1] and SARPA[2] by Makela et al. [27], Stanford NER[3] tools [28], Spacy dependency annotation and PoS tagging web service by Colic and Rinaldi [29]. Attempts to use such tools for Bangla NER resulted in erroneous systems, as the Bangla language neither belongs to any of the resource-rich language families nor the complex inflectional linguistic properties of Bangla were considered by any of those tools.

There is a crying need for a well-annotated Bangla NER dataset that follows standard human annotation guidelines. In this work, we propose a novel Bangla NER dataset, named as B-NER, which is constructed from various Wikipedia data,

[1] https://github.com/flammie/omorfi
[2] http://demo.seco.tkk.fi/sarpa/
[3] http://nlp.stanford.edu/software/CRF-NER.shtml

online newspapers (e.g., Prothom Alo, Kaler Kantha, and Anandabazar, etc.), and annotated manually into eight entities (Person, Geographical Entity, Organization, Geopolitical, Time Indicator, Artifact, Natural Phenomenon, and Event). Figure 1 shows the snapshot of the B-NER NEs with definitions. Benchmark experiments were done using state-of-the-art (SOTA) neural network models to evaluate the manually annotated 22,144 sentences corpus.

A plethora of linguistic challenges appear when solving NER in Bangla due to its complex morphological structures. Traditional approaches are not proficient enough to overcome these challenges. According to recent studies, pre-trained language models can be more effective for NER, as they provide additional semantic information to the network. In addition, the pre-trained language model handles a large portion of the training, as it only requires fine-tuning, resulting in less run-time for optimizing the NER objective functions. In this work, we employed Bidirectional Encoder Representations from Transformers (BERT) [30], and it's variants as a language model. BERT is a pre-trained model with a vast amount of unlabeled data. We fine-tuned the recently developed multilingual BERT pre-trained model (mBERT). Through comprehensive evaluations of the B-NER dataset using different BERT models, we achieved a macro accuracy of 86% using the fine-tuned NR/IndicbnBERT. Moreover, to ensure the high quality of the corpus, zero-shot analysis was done on other publicly available datasets.

This study aims to extend the reach of the Bangla NER system to make Information Extraction tasks easier. The following are the key contributions of this work:

- We propose a novel dataset B-NER, a fine-grained largest Bangla NER dataset using the BIO tagging scheme.
- We present the dataset properties, the statistics of the datasets, an overview of the annotation guidelines, illustrated with examples, and an analysis of the inter-annotator agreement.
- We performed the benchmark evaluation of the B-NER dataset using both the sequential dependency models and sentence transformer models to induct state-of-the-art results.
- We employed the cross-dataset model validation experiments (trained NER model on one dataset while tested on another) to examine the impact of our dataset in comparison to other publicly accessible datasets.
- A public release of the dataset and codebase will be made available under a CC-BY 4.0 license in the CoNLL-2002 format to facilitate future work on Bangla NER.

II. RELATED WORK

With the expansion social networks and popularity of e-commerce platforms, NER has become an essential tool for information extraction. For decades, NER has taken the central place of various NLP applications [31]. Over the past few years, numerous research articles on NER have been published for high-resource languages like English, Russian, German, Spanish, French, etc. With the availability of large annotated NER datasets in those languages, and the advent of deep neural networks and pre-trained language model, NER detection methods have reached ∼95% F1 score.[4]

Some factors, like the unavailability of labeled gold standard dataset, restrict the development of the Bangla NER system. A few research works are known for their contributions to developing annotated NER Bangla datasets. Ekbal et al. made a handful contributions in the early stage of Bangla NER research [32], [33], [34], [35], [36], [37], [38], [39]. In [34], Ekbal et al. proposed a Bangla NER dataset, including its benchmark for NER detection. In [32] and [40], authors utilized POS tags as the features for recognizing 4 major entity categories, such as person, location, organization and object. Data was collected from limited domains such as national, state, and sports sections. But unfortunately, they did not make the dataset publicly available.

We present widely cited Bangla NER research works in the Table 1, including their corpus details, modeling approaches, and evaluation details. Chowdhury et al. created a dataset named B-CAB, containing 2,137 sentences with seven different entities [40]. As they relied heavily on POS tags to label the NER entities, their protocols resulted in wrong annotations for *common nouns, collective nouns, pronouns*, etc. For example, those works generated wrong annotations by tagging *'who', 'people', 'doctor'* as person (PER) entities. Similarly, they tagged many nameless words, like *'road', 'building'*, etc., as facility (FAC) entities. There were some inconsistencies in tagging organizations (ORG) as well. It is worthy to note here that among all different types of *nouns*, a *proper noun* is considered to be a named entity according to the conventional NER studies [1], [2], [3], [4].

Similar discrepancies, which are noted above, are also observed in another publicly accessible Bangla NER dataset created by Rifat et al. [25]. The authors labeled generic terms as the named entities, like *'country,' 'time,'* etc., which resulted in wrong annotations. Besides, only a small fraction of the full dataset was manually labeled, while the rest were labeled using rule-based tagging tools. Those tagging rules generated a lot of errors, in particular, as the tool used a substring matching for tagging all occurrences of a word as per the predefined entity. This dataset contains 96,697 tokens in total, including punctuation, which they tagged as well. The largest Bangla NER Dataset, which consists of 71,284 sentences, was created by Karim et al. [24]. They gathered information from Wikipedia and online Bangla news sources such as The Daily Ittefaq, Bangladesh Pratidin, and Kaler Kantho. They labeled the dataset with 4 basic entity types using the BIO tagging method using their in-house annotation management tool called **Adhikary.** Inconsistent labeling of entities was discovered, primarily due to using automated

[4]http://nlpprogress.com/english/named_entity_recognition.html

TABLE 1. An overview of the widely cited research articles that worked on the bangla NER datasets, including their corresponding benchmark models and assessments.

Study	Corpus Details	Model	Evaluation Metrics
Ekbal et al. (2007-2009)	Created using Bangla newspaper articles containing 34 Mill words. Training size - 150k words (POS annotated, not NE annotated) Test size/ Methods - 10 fold Cross-Validation Tagset - 17 as input and 4 as output	HMM CRF SVM ME	R-93.98% P-90.63% F-92.28% (only Bengali language)
Hasanuzzaman et al. (2009)	IJCNLP-08 NER Shared Task on South Asian Languages. Training size - 122,467 tokens Test size/ Methods - 10 fold cross-validation Tagset - 4 tags	ME (Maximum Entropy)	R-88.01% P-82.63% F-85.22%
Shamima Parvez (2017)	1 sentence with 21 tagged words Test size/ Methods - 2 sentences	HMM	R-1 P-.7 F-.82 (chunked person name)
Chowdhury et al. (2018)	Bangla Content Annotation Bank (B-CAB) Features used: token + POS + gazetteers Tagset - 7 tags	LSTM CRF	R-.67 P-.78 F-.72 (Best Model)
Rifat et al. (2019)	Bangla newspaper articles Training size - 67,554 words annotated with NE tags Test size/ Methods - 29,143 words Tagset - 7 including 4 major tags	BGRU + CNN	R-72.27% P-73.32% F-72.66%
Karim et al. (2019)	Online news sources and Wikipedia Training size - 71,000 sentences Tagset - 8 including 4 major tags	DCN-BiLSTM	R-58.62% P-68.95% F-63.37%
Ashrafi et al. (2020)	Dataset from Karim et al. Tagset - 8 tags	BERT + BiLST + CRF + CW	Micro F-90.64% Macro F-65.96% MUC F- 72.04%

Dataset	Sample Sentence	Their Annotation	Annotation By Experts
Rifat et al.	দেশে দারিদ্র্যের হার প্রত্যাশার বেশি কমেছে (Poverty rate in the country has decreased more than expected)	B-LOC, O, O, O, O, O	O, O, O, O, O, O
	এ সময় পুলিশ তাঁদের মিছিলে বাধা দেয় (At that time the police obstructed their procession)	O, B-Tim, B-ORG, O, O, O, O	O, O, B-ORG, O, O, O, O
Karim et al.	অন্য কিছু শিক্ষক তাঁর পাশে এসে দাঁড়িয়েছিলেন (Some other teachers stood beside him)	O, O, B-PER, B-PER, O, O, O	O, O, O, O, O, O, O
	ধারণামতে ৬৩০ খ্রিস্টাব্দের কোন এক সময়ে তিনি ভারতবর্ষে প্রবেশ করেছিলেন (It is believed that he entered India at some point in 630 AD)	O, O, O, O, O, O, B-PER, B-LOC, O, O	O, B-TIM, O, O, O, O, O, B-LOC, O, O

FIGURE 2. By comparing the annotation samples from Rifat et al. and Karim et al. with linguistic expert annotation, we show the NER annotation flaws in the available Bangla NER datasets.

tagging systems. In the Figure 2, we show multiple examples with their wrong labeling for both datasets discussed above.

Researchers in the field of Bangla NLP employed different machine-learning models performing NER. Ekbal et al. utilized various machine learning techniques, such as Conditional Random Fields (CRF), Support Vector Machines (SVM), Hidden Markov Model (HMM), including their different combination which ultimately achieved the best performance for their tasks [37], [38]. Ekbal et al. further reported that achieving better performance by the hybrid models actually required a large labeled dataset, which they thought to be the major bottleneck. Chowdhuri et al. proposed a three-stage approach for NER task: 1. rule-based; 2. dictionary-based; and 3. n-gram-based models. Hasan et al. used a share task dataset from the IJCNLP-08 NER task [41] and showed that jointly adding a POS tagger with NE recognizer and affix induction can significantly improve the performance. Margin Infused Relaxed Algorithm (MIRA) by Crammer and Singer [42] was first used by Banerjee et al. [43] to detect named entities in Bangla. This model reportedly outperformed other contemporary systems based on CRF, SVM, or HMM. Banik and Rahman [44] used a Recurrent Neural Network (RNN) based approach for Bangla NER modeling. Later, Rifat et al. [25] proposed a similar system with a combination of Bidirectional Gated Recurrent Unit (BGRU) and Convolutional Neural Network (CNN). Among recent studies, Karim et al. used Densely Connected Network (DCN) in combination with Bidirectional Long Short Term Memory (BiLSTM). Ashrafi et al. [45] implemented BERT-based deep neural architecture that uses the contextual embeddings from BERT [30] as input for multi-label classification. Table 1 illustrates a comprehensive overview of recent works on Bangla NER. To avoid confusion and repetitiveness, the datasets of Karim et al. [24] and Rifat et al. [25] will be referred to as Dataset-K and Dataset-R in the later sections.

Entity	Root Word	Affix	Root Word + Affix
Geopolitical	রোহিঙ্গা	দের	রোহিঙ্গাদের
Geographical	কক্সবাজার	এর	কক্সবাজারের
Geographical	ভারত	কে	ভারতকে
Person	হোসেন	কে	হোসেনকে
Person	কবিরুল	এর	কবিরুলের
Organization	জাতিসংঘ	এর	জাতিসংঘের
Artifact	টুইটার	এ	টুইটারে
Geographical	বাংলাদেশ	এ	বাংলাদেশে
Geographical	বরিশাল	এর	বরিশালের
Geographical	সিলেট	এর	সিলেটের
Geographical	শ্রীনগর	এ	শ্রীনগরে
Time	একসপ্তাহ	এর	একসপ্তাহের

FIGURE 3. Here we are including some examples of inflected words, along with the base and affix of the respective words.

As discussed above, all previously published research works focused mainly on improving NER modeling accuracy despite the lacking of a sizeable NER-labeled dataset. In this paper, we try to address that major lacking by creating a large and quality dataset as the performance of models also relies on the it. According to Nobata et al., [46], annotating NER data is a work that is intrinsically complicated since it requires a lot of time, extensive human labor, and specialized expertise. The existing datasets were mostly labeled by using automated tools and in a few cases with the help of non-experts due to the constraints in terms of time and money, and unavailability of linguistic specialists. To overcome these shortcomings of Bangla NER datasets, we developed the B-NER dataset, which contains 22,144 data, that were manually annotated for 9 different named entities. To lower the human annotator's bias, we developed and adhered to a strict annotation standard. The annotators were instructed to follow the standard which was later verified by linguistic experts.

III. CHALLENGES OF BANGLA NER

Due to the complex language structure and resource scarcity, Bangla NER appeared to be a difficult task compared to other languages [38], [45]. The common challenges for solving NER in Bangla are discussed below.

A. INCLUSION OF AFFIXES

As Bangla is a morphologically rich language, we need to consider the morphological structure into account to ensure more human-readable output. Affixes play a vital role in developing the contextual meaning of a sentence. When an affix is combined with the base form of a word, it produces a derived or inflected form of that word, which substantially influences the context. Unlike English, affixes are often included with entity words in Bangla without changing the entity class of the word. Since, a Bangla entity with its affix may result in about 150 different forms, it is difficult for NER models to recognize those. Here, the most frequently used affixes are shown in Figure 3.

Challenges	Example
Free order sentence structure	সেদিন রাতে বাড়ি ফিরে **নজরুল** That night **Nazrul** returned home
	নজরুল সেদিন রাতে বাড়ি ফিরে **Nazrul** returned home that night
	সেদিন রাতে **নজরুল** বাড়ি ফিরে **Nazrul** returned home that night
Capitalization	**মনির** ঢাকায় থাকে **Monir** lives in **Dhaka**
Word-sense disambiguation	**পূজা** টিভি দেখছে **Puja** is watching TV
	তিনি **পূজায়** গ্রামে আসবেন He will come to the village on **Puja**
Phrases and idioms	বড় লোকের **খয়ের খাঁর** অভাব নেই Rich people have no shortage of **flatterers**

FIGURE 4. Here we present a list of challenges of Bangla NER including example sentences.

B. FREE ORDER SENTENCE STRUCTURE

Bangla is a highly *inflected* language from the Indo-Aryan language family. As words within a Bangla sentence can be positioned *free-order* without loosing its context, the same entity word may appear in different positions in a sentence. For example, in Figure 4, the person name *'Nazrul'* can reside at several positions within the sentence preserving the correct meaning. This is why, it is challenging to identify named entities in Bangla sentences.

C. ABSENCE OF CAPITALIZATION FEATURE

In many languages, especially in English, proper noun generally refer to an entity and always starts with a capital letter. As Bangla language does not require the word capitalization criteria for *proper nouns*, recognizing those named entities has additional difficulty [1], [2], [3]. In some widely used NER models, the capitalization feature plays a vital role in identifying named entities. For example, in Figure 4, *'Dhaka'*, the name of a city starts with a capital letter in English. Thus it refers to a named entity and is easily identifiable. But Bangla language does not have this benefit.

D. WORD-SENSE DISAMBIGUATION

In Bangla language, several words have different meanings despite having similar spellings [47]. It creates additional difficulty for the model to understand which sense of the word is used in a context. For example, the token *'Paris'* stands for the name of a city, while the token *'Paris'* in the phrase *'Paris Saint Germain'* expresses the beginning token of an organization (i.e., a football club). Similarly, a person's name in Bangla often expresses multiple meanings like in other South Asian languages. For example, in Figure 4, *'Puja'* is a female name, which also refers to the name of a religious

Entity Words	Tag
শ্রীপুর (Sreepur)	B-geo
মো (Md) আসাদুজ্জামান (Asaduzzaman)	B-per, I-per
বিনিয়োগ (Investment)	O
তাছলিমা (Taslima) আক্তার (Akter)	B-per, I-per
সোমবার (Monday) দুপুরে (Noon)	B-tim, I-tim
উত্তর (North) আমেরিকা (America)	B-geo, B-geo
যাতায়াত (Travel)	O
ঢাকা (Dhaka) বিশ্ববিদ্যালয় (University)	B-org, I-org
ভারতীয় (Indian)	B-gpe
অভিযান (Adventure)	O
আগস্ট (August)	B-tim

FIGURE 5. A snapshot of the B-NER annotation process employing the BIO tagging technique. If the word is not a *proper noun*, an "O" or outside tag is used. Otherwise, the named entity will consist of two or more words or tokens, with the first word tagged with "B" for beginning and the remaining words being "I" for inside tag.

event. Similarly, the word *'Kartik'* refers to both a person's name and a month's name in the Bangla calendar. Since, Bangla is highly inflected and free-order language, and it does not require the use of *preposition* as much like in English, the word-sense disambiguation is quite difficult in Bangla; hence making it more challenging to recognize named entities.

E. PHRASES AND IDIOMS

Phrases and idioms in Bangla often contain words that do not refer to their actual meaning. For example, in Figure 4, 'the word '*Khoyer Kha*' generally means a person's name in Bangla. Nevertheless, when it is used as an idiom, it refers to a flatterer. *'Alaler Ghorer Dulal'* is another example of a different meaning. Although *'Alal'* and *'Dulal'* are person names, it refers to a spoilt kid of a rich dad. In general, NER systems face difficulty in identifying the named entities within phrase and idioms, but the word-sense disambiguation issues as discussed above makes it much more difficult for Bangla.

IV. DATASET

Since blogs, news, forums, and Wikipedia sentences tend to include a large number of entities, the B-NER dataset is sourced from those. We used python built-in library *beautifulsoup* for crawling newspapers, blogs and forums websites. An overview of the complete workflow of dataset preparation is illustrated in Figure 6. Initially, we collected nearly 100,000 sentences from sources mentioned above. We filtered all code-mix texts including any sentences that include words with spelling errors. Finally, We created B-NER using 22,144 sentences that comprises 297,409 tokens. Our dataset incorporates eight NEs as shown in Figure 1 with respective examples and entity definitions. Miscellaneous words that do not belong to any of the previously stated eight categories are tagged with the letter ''O''. Table 2 displays the distribution of the annotations per entity class of the B-NER dataset.

TABLE 2. Count of "B" tags and "I" tags in the entire B-NER dataset. Here, "MISC" refers to miscellaneous words that has the highest number with tag "O" (for outside tag entities).

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

A. ANNOTATION SCHEME

Even though there is no mandated list of NE types for creating an NER corpus in Bangla, we developed an extensive annotation guideline following the definition of Groningen Meaning Bank (GMB) [23], and Sekine's extended named entities [48] for Bangla Language. The annotation assignment was led by a Bangla linguistic expert, who trained two native senior undergraduate students for the task. Annotation was performed in two phases. In the first phase, two annotators had performed annotations, which in case of disagreement, were revised in the second phase by the linguistic expert. For the annotations which resulted in agreement from both the annotators, were tested in two ways: first, using a random sample of 1,000 annotated NEs; and second, using a selection of 1,000 annotated NEs drawn from Wikipedia related to locations and organizations. The linguistic expert reviewed both samples, resulting in a 1.1% and 2.3% correction rate, respectively. The annotation was done using an open-source toolkit licensed under the Apache License v2.0 named WebAnno-based Toolkit.[5] It supports with several annotation layers including an easy to use interactive web interface.

For tagging named entities, researchers proposed and used various tagging schemes, such as BILUO, BIO, BIO2, IO, BIL2, BIOES, IOE, IOE2, etc. [49]. Each of the annotation strategies has its own benefits and flaws. Most schemes deliver distinct benefits to high-resource language. For post-positional languages like Bengali, Hindi, and Japanese, the BIO tagging scheme is widely adopted by researchers [50]. Our proposed B-NER dataset was designed following Groningen Meaning Bank (GMB) [23] and used the BIO tagging scheme, which was first introduced by Ramshaw and Marcus [51]. According to the BIO tagging scheme, if a chunk is found belonging to any entities, then the beginning token is tagged as the ''B-entity name'' format, and the remaining tokens are tagged as the ''I-entity name'' format. All other tokens of the dataset belonging to none of the eight types of entities are tagged as ''O-entity name''.

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

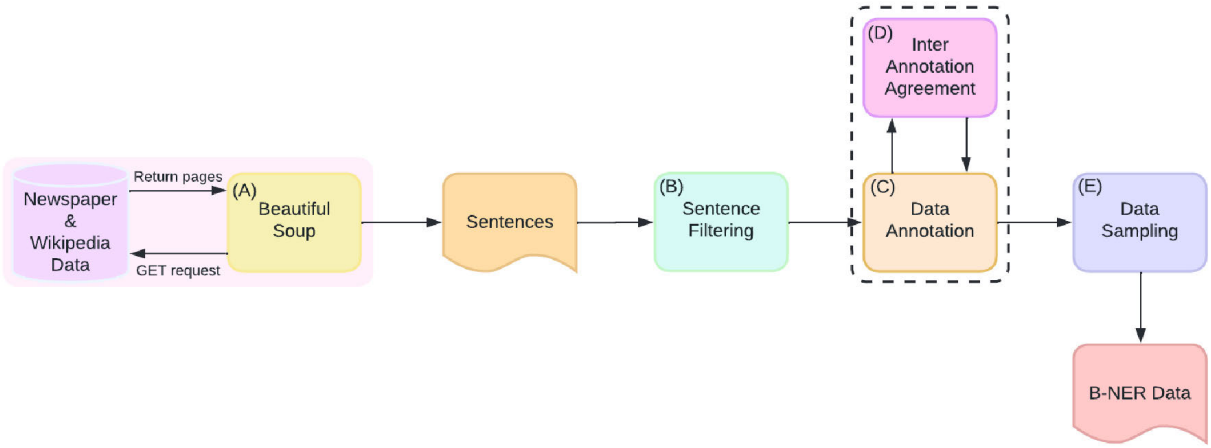

FIGURE 6. An overview of the steps involved in creating the B-NER dataset using newspaper and Wikipedia text.

TABLE 3. Statistics of our proposed B-NER dataset comparing with publicly accessible Bangla NER dataset Dataset-K (Karim et al. [24]) and dataset-R (Rifat et al. [25]).

Dataset Properties	B-NER	Dataset-K	Dataset-R
Sentences	22,144	71,284	6,561
Tokens	2,97,418	9,83,663	96,705
Unique Tokens	34,237	96,155	15,985
Sentence Length	[1-233]	[5-30]	[2-182]
Avg. Sentence Length	13.43	13.80	14.12
Entities	8	4	4
Tagging Scheme	BIO	BIO	BIO
Number of Tags	17	9	8

Some examples of annotation methods of the tokens are displayed in Figure 1.

B. DATASET STATISTICS

Trained human annotators manually annotated the 2,97,409 tokens, then validated by another linguistic expert. The dataset contains a total of 83% or 2,46,943 miscellaneous terms that do not correspond to any entity. These non-entity tokens provide valuable contextual information that helps to determine NEs from a sentence. Among the remaining tokens, the distribution of each entity is shown in Table 2. Person, geographical location, and organization entity cover most of the entity words of the dataset, while natural events cover the lowest number of entity words. Figure 7 demonstrates most frequently occurring tokens using word cloud for each entity class in our proposed dataset. To compute the inter-annotator agreement (IAA) we used Cohen's kappa score that was found to be 82.7% for the B-NER dataset. In addition, the comparison between the proposed B-NER with the publicly available Dataset-K and Dataset-R are shown in Table 3.

V. METHODOLOGY

To perform the benchmarking of our dataset, we experimented with sequence to sequence learning models using

FIGURE 7. Entity-wise Most Frequent B-NER Words.

both the recurrent neural network (RNN) based models and the pre-trained sentence transformer model (BERT) to recognize named entities from B-NER. We discuss both the methods including different settings are discussed below.

A. RECURRENT NEURAL NETWORKS

We used the Bidirectional Long Short Term Memory (BiLSTM) [52] model to conduct sequence to sequence modeling. Later, we tried a hybrid approach which we referred as BiLSMT-CRF by combining BiLSTM with Conditional Random Field (CRF) methods. In addition, we also attempted

modeling with fastText, the pre-trained embedding for Bangla, with both the BiLSTM and BiLSMT-CRF methods individually as shown in Table 4.

B. BIDIRECTIONAL ENCODER REPRESENTATION FROM TRANSFORMER (BERT)

BERT is a pre-trained language model which produces deep bidirectional representations of text based on contextual associations of its tokens. We analyzed the performance by fine-tuning BERT on B-NER. We only fine-tuned the output layer with our training data due to computing resource limitations. We chose one multilingual BERT: bert-base-multilingual-cased **(mBERT)** [30], and four mono-lingual pre-trained Bangla BERT: indic-transformers-bn-bert (**NR/IndicbnBERT**),[6], bangla-electra (**BanglaElectra**),[7], **SagorBERT** [53], and **K/Bangla-BERT** [54] using multiple Bangla task-specific dataset related to NER rather than using them directly. SagorBERT is a pre-trained Bangla language model created using BERT-based mask language modeling over 17 GB of data. In contrast, BanglaBERT was trained using 27.5 GB of data gathered from crawling 110 popular Bangla websites. The Latest K/Bangla-BERT was trained on 40 GB of structured data to overcome the limitations of low-resource Bangla Language. It has about 102,000 vocabularies, nearly three times the size of the original BERT.

We observed that NR/IndicbnBERT's performance was better than other well-known monolingual models, and it was trained on 3GB of Bengali data collected from OSCAR.[8] The huggingface transformers library was used to fine-tune the NR/IndicbnBERT model for B-NER. The Figure 8 shows the NR/IndcicbnBERT discriminator for NER modeling over the B-NER dataset.

VI. RESULTS ANALYSIS

A. MODEL SELECTION

Recent studies show that Bidirectional Long Short Term Memory (BiLSTM) [52], Conditional Random Field (CRF) [55], and pre-trained language models with deep transformers [30] have become a strong encoder for NER [56]. We thus follow the empirical settings, including BERT as the backbone of our experiments. We evaluated our methods using standard performance metrics, such as precision, recall, and macro averaged F1. To address the unbalanced data distribution of NEs, we used average macro F1 as the baseline evaluation metric.

In our study, we performed a total of 11 experiments on the B-NER dataset as cataloged in the Table 4. For all experiments, we employed exactly same data samples that were randomly selected using an 80-20 train-test data split. During the tuning process for the BiLSTM models, we adjusted parameters like learning rate, batch size, dropout rate, number of LSTM cells, and number of layers. For the fine-tuning of BERT models, we gave special consideration to the learning rate and batch size. Our benchmark results, displayed in Table 4, are presented in two separate groups: RNN-based models and BERT-based models. The fine-tuned NR/IndicbnBERT model achieved the highest F1-score of 0.86 among all models; thus setting the current state-of-the-art for NER modeling on the B-NER dataset.

[6]https://huggingface.co/neuralspace-reverie/indic-transformers-bn-bert
[7]https://huggingface.co/monsoon-nlp/bangla-electra
[8]https://oscar-project.org/

TABLE 4. Benchmark results on our eight-entity B-NER dataset utilizing precision, recall, and macro averaged F1. A total of ten methods are split into two groups, such as RNN and pre-trained BERT fine-tuning. The best model for RNN methods was found to be BiLSTM with fastText embeddings, while NR/IndicbnBERT produced the best results among all models.

Method	Precision	Recall	Macro-F1
BiLSTM	0.44	0.41	0.42
BiLSTM with Character Embeddings	0.45	0.43	0.44
BiLSTM (fastText)	**0.76**	**0.75**	**0.76**
BiLSTM-CRF	0.64	0.55	0.58
BiLSTM-CRF (fastText)	0.73	0.73	0.72
mBERT	0.71	0.68	0.69
NR/IndicbnBERT	**0.87**	**0.85**	**0.86**
BanglaElectra	0.68	0.48	0.52
SagorBERT	0.78	0.77	0.77
K/Bangla-BERT	0.62	0.60	0.61

TABLE 5. The outcome of training NER model on B-NER dataset and then testing on two other datasets: Dataset-K (Karim et al. [24]) and Dataset-R (Rifat et al. [25]).

Dataset	Precision	Recall	Macro-F1
Dataset-K	0.53	0.49	0.49
Dataset-R	0.62	0.66	0.64

B. CROSS-DATASET MODELING AND VALIDATION

To assess both the quality and efficacy of our dataset, we conducted a cross-dataset evaluation. We selected the best-performing model from our previous discussion, trained it fully on our dataset, and then evaluated its performance on two additional publicly available datasets. To ensure conformity, we selected three entities present in all three datasets: Person, Organization, and Location. The results of the experiment are presented in Table 5. We observed satisfactory accuracy on Dataset-R, but the model was lacking by greater margin in each performance metric on Dataset-K. To gain further insights into this discrepancy, we conducted a deeper analysis of the annotations and found a significant dissimilarity between the annotations of Dataset-K and B-NER.

To verify the annotation tags of Dataset-K, we decided to re-annotate the named-entities with standard tags. We selected 2,000 sentences randomly from Dataset-K and re-annotated them with the help of a linguistic expert annotator. We observed a significant improvement in the Macro-F1 score, increasing from 0.50 to 0.71 when B-NER model was validated on the re-annotated samples (see Table 6).

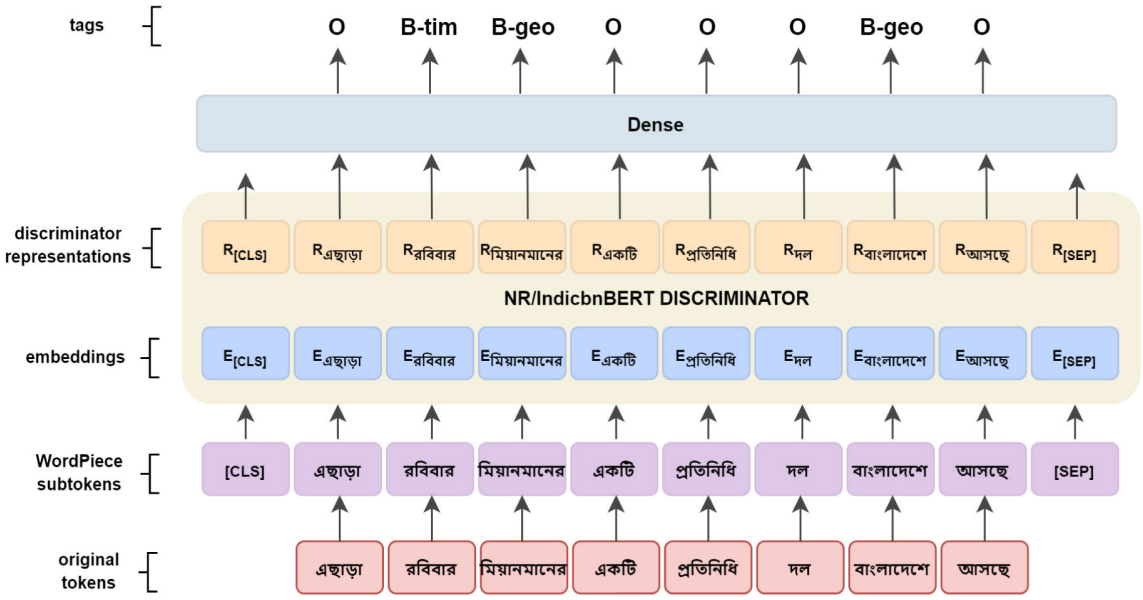

FIGURE 8. Architecture for NER modeling using NR/IndicbnBERT.

TABLE 6. Performance of B-NER trained model on re-annotated 2,000 samples of Dataset-K. Exactly same 2,000 samples were used for both the experiments before and after re-annotations. We observe the cross-dataset validation performance improved with B-NER model after re-annotations.

Comparison	Precision	Recall	Macro-F1
1) B-NER trained IndicBERT tested on the original Dataset-K tags for 2,000 samples	0.54	0.49	0.50
2) Re-annotated 2,000 samples of Dataset-K was tested with B-NER trained IndicBERT	**0.70**	**0.73**	**0.71**

TABLE 7. Statistics of the hand-annotated gold standard test dataset.

Dataset Features	Count
Sentences	2,300
Tokens	31,478
Unique Tokens	10,234
Sentence Length	[1-75]
Avg. Sentence Length	13.69
Entities	3
Tagging Scheme	BIO
Number of Tags	7

TABLE 8. List of datasets in different combinations with three entities and seven tags with their cross-dataset model validation performance when applied on gold standard dataset.

Dataset for Model Training	Precision	Recall	Macro-F1
B-NER	**0.73**	**0.76**	**0.74**
Dataset-K	0.55	0.64	0.57
Dataset-K + B-NER	0.60	0.67	0.62
Dataset-R	0.55	0.65	0.58
Dataset-R + B-NER	0.73	0.71	0.72
Dataset-K + Dataset-R + B-NER	0.62	0.68	0.64

C. EVALUATION ON GOLD STANDARD TEST DATASET

1) DATASET CREATION

To evaluate the quality of our dataset annotations relative to the other datasets, we created a separate gold standard test set for cross-dataset model validation. We trained the same model on the three datasets (B-NER, Dataset-K, Dataset-R) individually. The sources used to create the dataset was developed by Hossain et al. [57] that includes news from 22 newspapers, comprising a total of 1,055,220 sentences in 12 different categories. A random selection of 2,300 sentences was made and annotated by a team of 11 senior undergraduate students from diverse academic backgrounds, including Journalism (2), Bangla Literature (3), Political Science (2), Social Science (2), and Computer Science (2). This approach was taken to ensure that the annotations reflected knowledge from various domains. We only used three entities: Person, Organization, and Location, to create a uniform testbed for all datasets. At least three annotators verified each sample under the supervision of a linguistic expert and achieved 100 percent inter-annotator agreement. The statistics of this newly created dataset are presented in Table 7.

2) CROSS-DATASET MODEL VALIDATION ON GOLD STANDARD DATASET

We performed cross-dataset model validation in different combinations of three datasets: B-NER, Dataset-K, and Dataset-R. As presented in the Table 8, a total of 7 different NR/IndicbnBERT models were created and then applied on the hand-held gold standard dataset of 2,300 samples. We see that the best performing NER model was trained solely on the proposed B-NER dataset achieving the Micro-F1 score

of 0.74. These findings emphasize the essence of a cleanly annotated dataset, as the model's accuracy decreased when the datasets were mixed. These results also support our previous conclusion that Dataset-K has significant discrepancies. It demonstrated the poorest performance when trained in isolation with Dataset-K and even when combined with our B-NER dataset. The results obtained on these datasets detailed in Table 8.

VII. CONCLUSION

In this paper, we presented the B-NER dataset, which contains 22,144 Bangla sentences with eight entities, the most comprehensive dataset for Bangla NER. We collected and filtered raw corpus data from various sources, including blogs, newspaper articles, and Wikipedia. We conducted experimental analysis using different methods, including RNN-based and pre-trained BERT-based models, and found that fine-tuned NR/IndicbnBERT produced the best results for B-NER (Macro-F1 86%). We also performed cross-dataset modeling and validation and observed a Micro-F1 of 74%, demonstrating the potential impact of B-NER on a gold standard dataset. Our experimental outcomes demonstrated that hand-annotated corpus could help achieving generalization for NER modeling. Despite the strengths of our suggested dataset, such as its use of real-world data, it suffers from dataset imbalance, and its size could be expanded further. Our future plans include addressing these issues and expanding the B-NER dataset. We aim to enhance our B-NER system and achieve generalization of information extraction for any Bangla data sources.

ACKNOWLEDGMENT

The authors would like to express our gratitude to Abu Toha Faisal, an Ex-Student of the Sylhet Engineering College, Sylhet, Bangladesh, for providing the B-NER raw data from a diverse set of domains, and filtering the data to avoid data repetition.

REFERENCES

[1] N. Chinchor and P. Robinson, "Appendix E: MUC-7 named entity task definition (version 3.5)," in *Proc. 7th Message Understand. Conf. (MUC-7), Conf. Held Fairfax, Virginia*, Apr. 1998, pp. 1–21.

[2] C. Le Meur, S. Galliano, and E. Geoffrois, "Conventions d'annotations en entités nommées-ESTER," Rapport Technique de la Campagne Ester, 2004.

[3] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticæ Investigationes*, vol. 30, no. 1, pp. 3–26, Jan. 2007.

[4] R. Grishman and B. M. Sundheim, "Message understanding conference—6: A brief history," in *Proc. 16th Int. Conf. Comput. Linguistics*, vol. 1, 1996, pp. 1–6.

[5] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Unsupervised named-entity extraction from the web: An experimental study," *Artif. Intell.*, vol. 165, no. 1, pp. 91–134, 2005.

[6] S. Sekine and C. Nobata, "Definition, dictionaries and tagger for extended named entity hierarchy," in *Proc. LREC*, Lisbon, Portugal, 2004, pp. 1977–1980.

[7] S. Zhang and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," *J. Biomed. Inform.*, vol. 46, no. 6, pp. 1088–1098, Dec. 2013.

[8] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12 pp. 2493–2537, Aug. 2011.

[9] G. Aguilar, F. AlGhamdi, V. Soto, M. Diab, J. Hirschberg, and T. Solorio, "Named entity recognition on code-switched data: Overview of the CALCS 2018 shared task," in *Proc. 3rd Workshop Comput. Approaches Linguistic Code-Switching*. Melbourne, VIC, Australia: Association for Computational Linguistics, Jul. 2018, pp. 138–147. [Online]. Available: https://aclanthology.org/W18-3219

[10] T. Brants, "TnT: A statistical part-of-speech tagger," in *Proc. 6th Conf. Appl. Natural Lang. Process.*, 2000, pp. 224–231, doi: 10.3115/974147.974178.

[11] G. Zhou and J. Su, "Named entity recognition using an HMM-based chunk tagger," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 473–480.

[12] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[13] W. Li and A. McCallum, "Rapid development of Hindi named entity recognition using conditional random fields and feature induction," *ACM Trans. Asian Lang. Inf. Process. (TALIP)*, vol. 2, no. 3, pp. 290–294, 2003.

[14] A. Ratnaparkhi, "A maximum entropy model for part-of-speech tagging," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 1996, pp. 1–10.

[15] A. McCallum, D. Freitag, and F. C. N. Pereira, "Maximum entropy Markov models for information extraction and segmentation," in *Proc. ICML*, vol. 17, 2000, pp. 591–598.

[16] T. Kudo and Y. Matsumoto, "Chunking with support vector machines," in *Proc. 2nd Meeting North Amer. Chapter Assoc. Comput. Linguistics*, 2001, pp. 1–8.

[17] H. Yamada and Y. Matsumoto, "Statistical dependency analysis with support vector machines," in *Proc. 8th Int. Conf. Parsing Technol.*, 2003, pp. 195–206.

[18] F. Alam, "Named entity recognition on transcription using cascaded classifiers," Work. Notes EVALITA, 2011.

[19] S. Abdallah, K. Shaalan, and M. Shoaib, "Integrating rule-based system with classification for Arabic named entity recognition," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*. Cham, Switzerland: Springer, 2012, pp. 311–322.

[20] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," *Trans. Assoc. Comput. Linguistics*, vol. 4, pp. 357–370, Dec. 2016.

[21] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1064–1074. [Online]. Available: https://aclanthology.org/P16-1101

[22] Z. Jie and W. Lu, "Dependency-guided LSTM-CRF for named entity recognition," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 3862–3872. [Online]. Available: https://aclanthology.org/D19-1399

[23] J. Bos, V. Basile, K. Evang, N. J. Venhuizen, and J. Bjerva, "The Groningen meaning bank," in *Handbook of Linguistic Annotation*. Springer, 2017, pp. 463–496.

[24] R. Karim, M. A. Islam, S. R. Simanto, S. A. Chowdhury, K. Roy, A. Al Neon, M. Hasan, A. Firoze, and R. M. Rahman, "A step towards information extraction: Named entity recognition in Bangla using deep learning," *J. Intell. Fuzzy Syst.*, vol. 37, no. 6, pp. 7401–7413, 2019.

[25] M. J. R. Rifat, S. Abujar, S. R. H. Noori, and S. A. Hossain, "Bengali named entity recognition: A survey with deep learning benchmark," in *Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, 2019, pp. 1–5.

[26] Y. Benajiba and P. Rosso, "ANERsys 2.0: Conquering the NER task for the Arabic language by combining the maximum entropy with pos-tag information," in *Proc. IICAI*, 2007, pp. 1814–1823.

[27] E. Mäkelä, "Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text," in *Proc. Eur. Semantic Web Conf.* Cham, Switzerland: Springer, 2014, pp. 424–428.

[28] R. Eiselen, M. Puttkammer, J. Hocking, and A. Kruger, "CTexTools 2," 2018.

[29] N. Colic and F. Rinaldi, "Improving spaCy dependency annotation and PoS tagging web service using independent NER services," *Genomics Informat.*, vol. 17, no. 2, p. e21, 2019.
[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423
[31] L. Derczynski, D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak, and K. Bontcheva, "Analysis of named entity recognition and linking for tweets," *Inf. Process. Manage.*, vol. 51, no. 2, pp. 32–49, Mar. 2015.
[32] A. Ekbal and S. Bandyopadhyay, "A hidden Markov model based named entity recognition system: Bengali and Hindi as case studies," in *Proc. Int. Conf. Pattern Recognit. Mach. Intell.* Cham, Switzerland: Springer, 2007, pp. 545–552.
[33] A. Ekbal and S. Bandyopadhyay, "Bengali named entity recognition using support vector machine," in *Proc. IJCNLP Workshop Named Entity Recognit. South South East Asian Lang.*, 2008, pp. 1–8.
[34] A. Ekbal and S. Bandyopadhyay, "Development of Bengali named entity tagged corpus and its use in NER systems," in *Proc. 6th Workshop Asian Lang. Resour.*, 2008, pp. 1–8.
[35] A. Ekbal, R. Haque, and S. Bandyopadhyay, "Named entity recognition in Bengali: A conditional random field approach," in *Proc. 3rd Int. Joint Conf. Natural Lang. Process.*, vol. 2, 2008, pp. 1–6.
[36] A. Ekbal and S. Bandyopadhyay, "A web-based Bengali news corpus for named entity recognition," *Lang. Resour. Eval.*, vol. 42, no. 2, pp. 173–182, 2008.
[37] A. Ekbal and S. Bandyopadhyay, "Bengali named entity recognition using classifier combination," in *Proc. 7th Int. Conf. Adv. Pattern Recognit.*, 2009, pp. 259–262.
[38] A. Ekbal and S. Bandyopadhyay, "Named entity recognition in Bengali," *Northern Eur. J. Lang. Technol.*, vol. 1, pp. 26–58, Mar. 2009.
[39] M. Hasanuzzaman, A. Ekbal, and S. Bandyopadhyay, "Maximum entropy approach for named entity recognition in Bengali and Hindi," *Int. J. Recent Trends Eng.*, vol. 1, no. 1, p. 408, 2009.
[40] S. A. Chowdhury, F. Alam, and N. Khan, "Towards Bangla named entity recognition," in *Proc. 21st Int. Conf. Comput. Inf. Technol. (ICCIT)*, 2018, pp. 1–7.
[41] K. S. Hasan, M. A. Ur Rahman, and V. Ng, "Learning-based named entity recognition for morphologically-rich, resource-scarce languages," in *Proc. 12th Conf. Eur. Chapter ACL (EACL)*, 2009, pp. 354–362.
[42] K. Crammer and Y. Singer, "Ultraconservative online algorithms for multiclass problems," in *Proc. Int. Conf. Comput. Learn. Theory*. Cham, Switzerland: Springer, 2001, pp. 99–115.
[43] S. Banerjee, S. K. Naskar, and S. Bandyopadhyay, "Bengali named entity recognition using margin infused relaxed algorithm," in *Proc. Int. Conf. Text, Speech, Dialogue*. Cham, Switzerland: Springer, 2014, pp. 125–132.
[44] N. Banik and M. H. H. Rahman, "GRU based named entity recognition system for Bangla online newspapers," in *Proc. Int. Conf. Innov. Eng. Technol. (ICIET)*, 2018, pp. 1–6.
[45] I. Ashrafi, M. Mohammad, A. S. Mauree, G. M. A. Nijhum, R. Karim, N. Mohammed, and S. Momen, "Banner: A cost-sensitive contextualized model for Bangla named entity recognition," *IEEE Access*, vol. 8, pp. 58206–58226, 2020.
[46] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 145–153.
[47] S. Nazah, M. M. Hoque, and M. R. Hossain, "Word sense disambiguation of Bangla sentences using statistical approach," in *Proc. 3rd Int. Conf. Electr. Inf. Commun. Technol. (EICT)*, 2017, pp. 1–6.
[48] S. Sekine, "Extended named entity ontology with attribute information," in *Proc. 6th Int. Conf. Lang. Resour. Eval. (LREC)*, 2008, pp. 1–6.
[49] N. Alshammari and S. Alanazi, "The impact of using different annotation schemes on named entity recognition," *Egyptian Informat. J.*, vol. 22, no. 3, pp. 295–302, 2021.
[50] M. K. Malik and S. M. Sarwar, "Named entity recognition system for postpositional languages: Urdu as a case study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 10, pp. 141–147, 2016.
[51] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Natural Language Processing Using Very Large Corpora*. Cham, Switzerland: Springer, 1999, pp. 157–176.
[52] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[53] S. Sarker. (2020). BanglaBERT: Bengali mask language model for Bengali language understading," [Online]. Available: https://github.com/sagorbrur/bangla-bert
[54] M. Kowsher, A. A. Sami, N. J. Prottasha, M. S. Arefin, P. K. Dhar, and T. Koshiba, "Bangla-BERT: Transformer-based efficient model for transfer learning and language understanding," *IEEE Access*, vol. 10, pp. 91855–91870, 2022.
[55] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1529–1537.
[56] X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, and J. Li, "A unified MRC framework for named entity recognition," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5849–5859.
[57] M. Z. Hossain, M. A. Rahman, M. S. Islam, and S. Kar, "BanFakeNews: A dataset for detecting fake news in Bangla," in *Proc. 12th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, May 2020, pp. 2862–2871. [Online]. Available: https://aclanthology.org/2020.lrec-1.349

MD. ZAHIDUL HAQUE received the B.Sc. (Eng.) degree in computer science and engineering from the Sylhet Engineering College, Sylhet, Bangladesh. He is currently with Graaho Technologies. He is also a former Research Engineer with Mayalogy Ltd. His current research interests include natural language processing, romanized bangla for medical patient text, entity linking, and computer vision.

SAKIB ZAMAN was born in Kishoreganj, Bangladesh. He received the B.Sc. (Eng.) degree in computer science and engineering from the Department of Computer Science and Engineering, Sylhet Engineering College, Sylhet, Bangladesh. His current research interests include machine learning, deep learning, and natural language processing.

JILLUR RAHMAN SAURAV received the B.Sc. (Eng.) degree in computer science and engineering from the Shahjalal University of Science and Technology, Sylhet, Bangladesh. He is a former Software Engineer with Pipilika. He is currently a Graduate Teaching Assistant with The University of Texas at Arlington, USA. His research interests include machine learning, deep learning, and natural language processing.

SUMMIT HAQUE received the B.Sc. (Eng.) degree in computer science and engineering from the Shahjalal University of Science and Technology, Sylhet, Bangladesh. In 2018, he was a Faculty Member with the Shahjalal University of Science and Technology, where he is currently an Assistant Professor. His research interests include data science, machine learning, and natural language processing.

MD. SAIFUL ISLAM is a Graduate Research Fellow of computing science with the University of Alberta, where he works on human-in-the-loop, natural language processing (NLP), and the intersection of human and AI. Prior to joining the University of Alberta, he worked as an Assistant Professor with the Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet, Bangladesh, where he has been a Faculty Member, since 2013. During this time, he also led a research group that focused on developing new NLP applications for social good and education. He has continued working on cutting-edge NLP and human-in-the-loop research, collaborating with diverse researchers and industry partners. He has published extensively in top-tier computer science and NLP conferences. He regularly serves as a reviewer for leading NLP conferences.

MOHAMMAD RUHUL AMIN (Member, IEEE) received the Ph.D. degree from the Computer Science Department, Stony Brook University (SBU). He is an Assistant Professor with the Computer and Information Science Department, Fordham University. His team develops computational ways to solve cutting-edge problems in public health, bioinformatics, natural language processing, computational social science, and accessibility. Recently, he has worked on a multi-wheel input device for non-visual interaction with computing platforms for people with visual impairments and stereotypical gender associations in language and their change over time. He is currently collaborating with research groups at the University of Toronto, University of British Columbia, University of Maryland, Pennsylvania State University, New York University, Columbia University, and Stony Brook University, to solve several data science research problems. His research interests include the intersection of big data, statistical methods, and artificial intelligence.